Preprints.org

Review

Not peer-reviewed version

Distinguishing Reality from AI: Approaches for Detecting Synthetic Content

David Ghiurău * and Daniela Elena Popescu

Posted Date: 18 November 2024

doi: 10.20944/preprints202411.1219.v1

Keywords: artificial intelligence; generative pre-trained transformers; security vulnerabilities; deep fake; social engineering; blockchain; watermarking; stylometric

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Review

Distinguishing Reality from AI: Approaches for Detecting Synthetic Content

David Ghiurău [1,*] and Daniela Elena Popescu [2]

[1] Department of Computers and Information Technology, Politehnica University of Timisoara, 2 V. Parvan Blvd, 300223 Timisoara, Romania
[2] Department of Computers and Information Technology, Faculty of Electrical Engineering and Information Technology, University of Oradea, 410087 Oradea, Romania; depopescu@uoradea.ro
* Correspondence: david.ghiurau@student.upt.ro

Abstract: The advent of sophisticated artificial intelligence technologies, including Generative Pre-trained Transformers for text and advanced generative models for image, audio, and video creation, have revolutionized content generation, presenting both innovative opportunities and significant challenges for information security. This paper introduces the characteristics, techniques, and challenges for detecting AI-generated content across text, image, audio, and video modalities, aiming to safeguard the integrity and authenticity of digital information.

Keywords: artificial intelligence; generative pre-trained transformers; security vulnerabilities; deep fake; social engineering; blockchain; watermarking; stylometric

1. The Evolution of Artificial Intelligence Technologies in Content Creation

Artificial intelligence is a concept that had early foundations and theoretical concepts since the 1950s, a technology focused on replicating human cognitive functions directly through programmed rules and logic [1]. Fast forward to the 1980s due to computational power increase, the concept saw a shift towards machine learning, where algorithms could learn from data rather than being explicitly programmed.

The growth of the internet and digital data created massive datasets that were necessary for training AI models, a period that saw various applications from search engines to personal assistants. The following years have seen groundbreaking advancements in deep learning, especially in the capability to generate complex content like text, images, videos, and audio autonomously, this is characterized by the rise of generative adversarial networks and transformer models, which have revolutionized content creation [2].

Artificial intelligence-generated content (AIGC) has become a significant force in today's digital landscape, driven by advancements in machine learning and artificial intelligence. This technology encompasses the creation of various types of content, including text, images, audio, and video, through sophisticated AI algorithms [3]. Notable advancements in natural language processing (NLP) and generative models like GPT-3 (Generative Pre-trained transformer), GPT-4, Electra, BERT (Bidirectional Encoder Representations from Transformers) and many more have greatly enhanced AIGC capabilities, leading to its adoption across multiple sectors such as journalism, marketing, entertainment, and education.

In the realm of media and journalism, AIGC has revolutionized content creation by enabling the production of news articles, summaries, and reports with remarkable efficiency. This not only boosts productivity in newsrooms but also facilitates the personalization of news content to suit individual preferences, thereby increasing reader engagement.

However, this personalization can also contribute to the formation of echo chambers and the spread of misinformation, presenting a significant ethical challenge. In marketing and advertising, AIGC tools are employed to generate personalized ads, social media posts, and email campaigns,

© 2024 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

leveraging consumer data to create highly targeted content. This has transformed how businesses interact with their audiences, making marketing efforts more effective and personalized.

The entertainment industry has also benefited immensely from AIGC, with AI assisting in scriptwriting, storytelling, and even music composition. Furthermore, AIGC plays a crucial role in generating realistic visual effects and animations, enhancing the overall quality and appeal of entertainment products. In education, AI-generated content is used to create interactive learning materials and customized educational resources, making learning more engaging and accessible. The technology also aids in providing educational content in various languages and formats, thereby addressing diverse learning needs and improving accessibility.

Despite these benefits, the rise of AIGC brings about several ethical and social concerns. The potential for AI-generated fake news and deepfake videos to spread misinformation is a significant threat that needs to be addressed [4]. Additionally, the automation of content creation could lead to job displacement in creative industries, raising concerns about the future of employment in these sectors. Intellectual property issues related to the ownership and copyright of AI-generated works are also emerging as critical legal challenges. Moreover, ensuring that AI-generated content does not perpetuate biases present in training data is essential to maintain fairness and prevent discrimination [5].

The interest of this area has seen an outstanding increase over the recent years, growth visible via the number of articles written about the generative models and artificial intelligence in general. Based on the numbers found in the most common research papers databases and Google Trends tool the following graph can be reproduced (Figure 1).

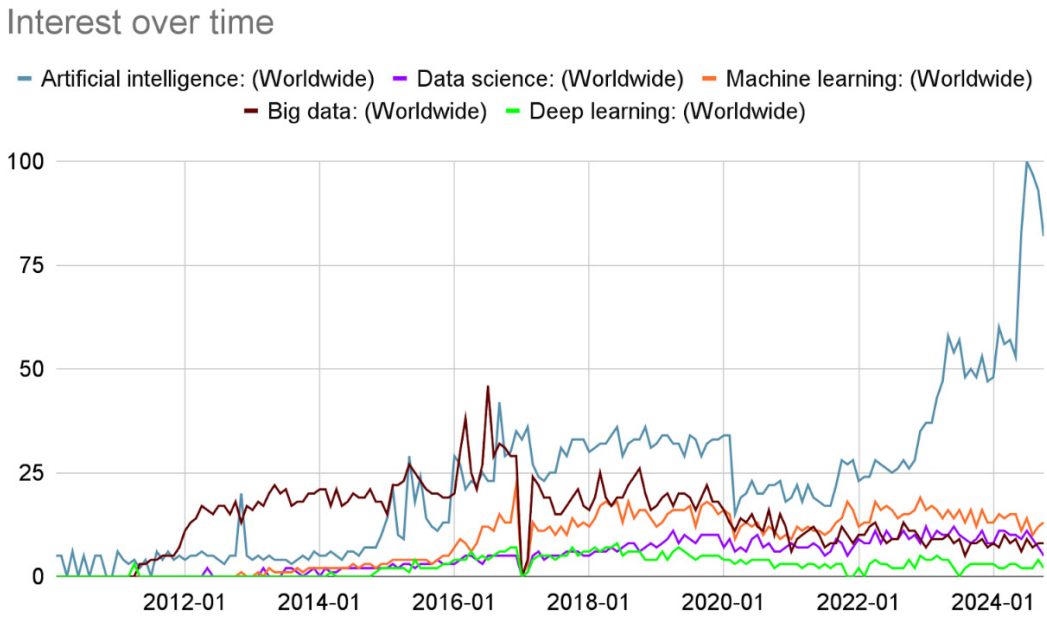


Figure 1. Interest over time for Artificial Intelligence [6].

The numbers represent search interest relative to the highest point on the chart for the given region and time. This results emphasize the need of research and development in this area.

AI now plays a pivotal role in automating and enhancing content creation across various fields, this digital transformation needs also safeguarding, particularly from synthetic content, and validating information authenticity. Recent developments have focused on creating synthetic data that not only alleviates bias and privacy concerns but also fulfills the requirements of being true to real data. Auditing frameworks for synthetic datasets are now being developed to ensure they meet critical socio-technical safeguards [2].

The purpose of this research paper is to give a comprehensive understanding of the current state of AI-generated content. This paper has at its core a thorough literature review, encompassing

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

existing research papers, articles, and reports on AIGC from academic databases, industry reports, and relevant publications. The review highlights recent advancements, key technologies like GPT-4 and GANs (Generative adversarial networks), and sector-specific applications, providing a solid foundation of knowledge on the topic.

The paper delves into technological advancements, practical applications, ethical concerns, and future trends in AIGC. By capturing the perspectives of those directly involved in the field, the research can provide a nuanced understanding of the current state of AIGC. Additionally, the paper includes case studies that analyze specific instances of AIGC implementation in various sectors. These case studies will help illustrate the practical benefits and challenges of AIGC, offering concrete examples of its impact. Both qualitative and quantitative data analysis methods are employed to process the information gathered from surveys, interviews, and case studies. Qualitative analysis will involve thematic analysis of interview transcripts and open-ended survey responses to identify common themes and insights. Quantitative analysis will involve statistical examination of survey data to quantify trends, usage patterns, and impacts.

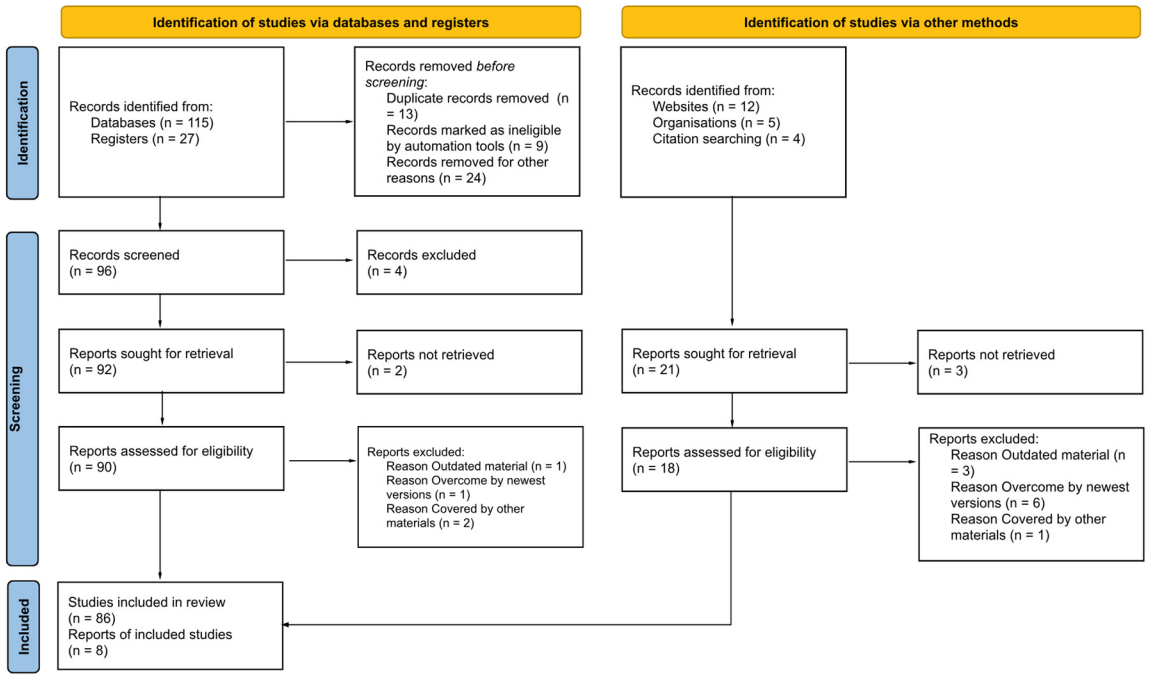
The data sources used are *Cochrane Library*, *Scopus*, *Web of Science*, *MDPI*, *IEEE Xplorer*, *Elsevier*, *Cornell University Arxiv*, and *EBSCO*. The inclusion and exclusion criteria are essential in defining the scope and quality of the research paper, the criteria (Table 2) used for this article are based on the following:

Table 2. Exclusion / Inclusion criteria of research papers.

Exclusion Criteria	Inclusion Criteria
Study type: exclusion of non-peer-reviewed studies	**Language**: articles published in English
Non-Published data: exclude grey literature such as reports or unpublished data	**Seniority**: articles published within the last seven years
Lack of outcome reporting: exclude studies that do not report on the primary outcomes of interest	**Applicability**: articles that remain relevant given the latest scientific advances and policy changes
Duplications: exclude duplicate studies or preliminary reports of already published research	**Outcome relevance**: studies that investigate or report on specific outcomes for the generative models
	Sample size: studies that ensure a robustness of tests and analysis

The findings from the literature review, surveys, interviews, and case studies are synthesized into a comprehensive report, the following Prisma flow chart (Figure 3) mentions the number of articles, case studies, registers, and databases analyzed.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1



From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71.

Figure 3. Prisma flow chart.

The paper is structured into six chapters as follows: Chapter 1: *The evolution of artificial intelligence technologies in content creation*, this chapter provides an overview of the evolution of artificial intelligence in the area of content creation and also mentions the methodology used for this research paper.

Chapter 2: *Characteristics of synthetic content* highlights the main types of artificial intelligence-generated content and how they can be analyzed and identified as synthetic content for each category: text, audio video, and image.

Chapter 3: *Challenges and difficulties* reviews existing literature relevant to the research topic. It identifies key theories, models, and empirical studies that have shaped the current state of the AIGC content detection tools.

Chapter 4: *Analysis of a current detection tools / algorithms* details the current frameworks and their advantages and difficulties in identifying the content authenticity and validity. The methods analyzed are stylometric, watermarking and digital fingerprints, adversarial and robust detection techniques, machine learning models and the integration of the blockchain in enhancing the trustworthiness of content.

Chapter 5: *Directions of innovation* discuss about the future implications of the AIGC technology. The chapter also highlights the current frameworks and offers recommendations for future research, suggesting areas where further investigation is needed.

Chapter 6: *Conclusion and Ethical concerns* summarizes the key findings of the research, drawing final conclusions based on the analysis, highlighting the current ethical concerns raised by the AIGC. It provides a concise summary of what has been learned from the study and its contributions to the field.

2. Characteristics of Synthetic Content

Identifying synthetic content across different media types, such as text, audio, image, and video, involves recognizing various specific characteristics. A prominent issue in synthetic media is the lack of coherence and continuity. In natural content, there is a seamless flow across frames in videos or across sentences and paragraphs in text.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

Recent researches [7,8] also suggests the use of multimodal approaches that integrate signals from various inputs (audio, visual, textual) to improve the detection of synthetic content, highlighting the need for comprehensive analysis tools that consider all aspects of multimedia content.

However, synthetic media often exhibits abrupt changes that disrupt this flow. For instance, in synthetic videos, there might be sudden shifts in the background, lighting, or character positions that are jarring to the viewer [9–11]. Similarly, in AI-generated text, the narrative may lack logical progression [12–14], with sentences or ideas that do not smoothly connect. These discontinuities can break the immersion and are indicative of the artificial origin of the media.

Another common characteristic of synthetic media is the presence of detail anomalies. Fine details, such as textures, shadows, and reflections, are often not as precise or realistic as in natural content [15,16]. For example, in synthetic images or videos, the textures of surfaces might appear uniform or lack the intricate variations seen in real-world materials. Shadows may fall incorrectly, and reflections might not accurately correspond to the surrounding environment. These anomalies can be subtle but significantly affect the overall realism of the media. The inability to replicate these fine details accurately highlights the limitations of current synthetic media generation technologies.

Understanding these characteristics helps in developing more effective detection systems and contributes to the broader efforts in digital media forensics, ensuring the authenticity and reliability of media content.

2.1. Text

Modern AI, like OpenAI's GPT series, can generate text that mimics human writing styles convincingly. This makes it challenging to detect AI-generated content using traditional text analysis tools. AI can subtly alter text in ways that can change meanings and spread misinformation without being overtly detectable.

One specific feature of synthetic text is the presence of unnatural syntactic patterns. Unlike human authors, who typically exhibit a rich and varied use of syntax, AI-generated text often demonstrates less variability. This can manifest in the repeated use of certain sentence structures or an over-reliance on specific grammatical forms [17]. Additionally, AI-generated content may incorporate uncommon phrase structures that seldom appear in human writing. These syntactic peculiarities can serve as indicators of text generation by artificial intelligence.

Another characteristic of AI-generated text is the occurrence of semantic anomalies. While human writers generally maintain consistency in their use of world knowledge and facts, AI-generated text can include subtle errors or factual inconsistencies [18]. These semantic anomalies might arise from the AI's limited understanding of context or its reliance on incomplete data. For instance, an AI might generate statements that, while grammatically correct, contain inaccuracies or logical inconsistencies that a human writer would likely avoid. These errors, although often subtle, can provide clues about the non-human origin of the text.

The GPTs have evolved very quickly in regards of semantic phrasing, but there are still errors of hints that can be detected. In Table 4. it is sumarised and analyzed a few phrases generated by a language model.

Table 4. Hints to identify synthetic text.

Phrases generated by a language model	Hints for detecting synthetic text
The intricate dance between science and nature provides an undeniable synergy, fostering innovation and breakthroughs	**Grandiose phrasing:** Uses broad, impressive-sounding words without real insight
In modern global economics, the balance of trade and currency fluctuations are vital components	**Surface-level complexity:** The text sounds technical but lacks specific, actionable insights

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

The historical development of art has traversed various eras, from the Renaissance to post-modernism, each reflecting societal values	**Inconsistent specificity:** Mentions detailed terms (e.g., *Renaissance*) but glosses over other key details
The relationship between quantum mechanics and classical physics has puzzled scientists for decades	**Generic technical jargon:** The language sounds academic but doesn't provide new or meaningful details
While many individuals believe in the importance of education, the future of technology seems to advance progressively towards artificial intelligence	**Overly generic or predictable patterns:** Complex, yet vague ideas lacking specificity or depth

2.2. Audio

One notable limitation of synthetic voices is their lack of naturalness. Unlike human speech, which is rich with subtle intonations, hesitations, and emotional expressions, AI-generated audio often falls short in these areas. Human speakers naturally vary their pitch, tone, and rhythm to convey meaning, emotion, and emphasis. In contrast, synthetic voices frequently exhibit a more monotone or mechanical quality. This absence of natural variation and emotional depth can make AI-generated speech sound less authentic and more artificial.

Another distinctive feature of synthetic audio is its consistency in tones. While human speech typically involves natural variations in pitch, speed, and volume, synthetic speech often maintains a uniform tone throughout. This can result in audio that sounds overly consistent or even robotic [19]. Human speakers adjust their vocal characteristics dynamically in response to context, emotion, and conversational flow, whereas AI-generated voices may lack this adaptability. The uniformity in tone and pace can serve as an indicator of artificial generation, highlighting the differences between human and synthetic speech [20].

As it can be visualised in Figure 5 the audio waveform can be interpreted and there are hints in synthetic content due to abrupt endings, instant cuts, monotone pauses or even arythmic steps besides the lack of intonation, emotion or human warmth. Nevertheless the generative models are evolving in this area, major development being supported by large corporations that want their digital assistant to sound as human as possible.

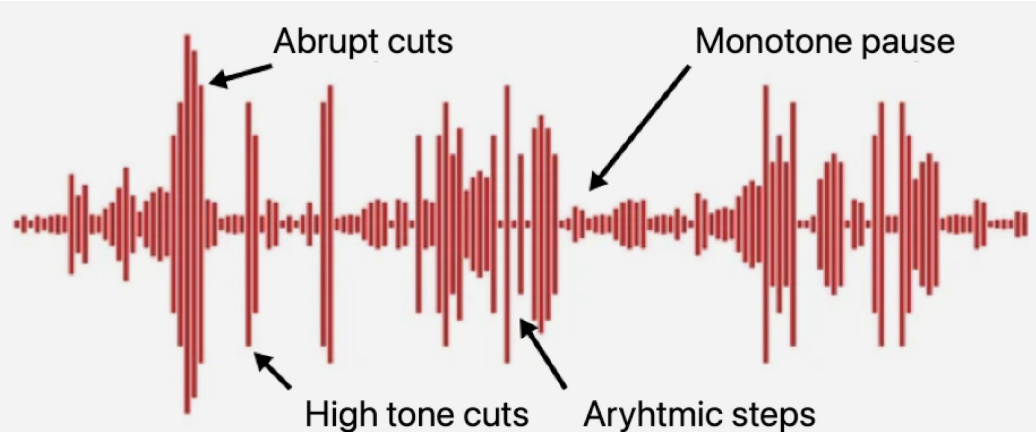


Figure 5. Audio waveform analysis.

2.3. Images

Generative Adversarial Networks can create highly realistic images that are difficult to differentiate from real photographs without detailed analysis. AI-generated images often carry metadata that mimics genuine content, misleading detection algorithms that rely on metadata analysis.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

One of the telltale signs of synthetic images is the presence of texture irregularities. In human-created or natural images, textures typically blend seamlessly, creating a coherent and consistent visual experience.

Synthetic images often struggle with maintaining this consistency. Certain areas within these images may exhibit textures that do not match the surrounding areas, leading to noticeable discrepancies. These irregularities can disrupt the visual harmony of the image and serve as indicators of artificial generation. Such texture inconsistencies highlight the limitations of current AI techniques in replicating the complexity and subtlety of natural textures [15].

Another common issue in synthetic images is the presence of unusual patterns in backgrounds or shadows. In real images, shadows and background elements adhere to the principles of physical lighting and environmental conditions, creating a natural and realistic appearance.

Synthetic images, however, often contain anomalies where shadows or background patterns do not align with expected physical properties. These discrepancies can manifest as incorrect shadow angles, inconsistent lighting, or patterns that defy the laws of physics. Such unnatural features can be indicative of an image's synthetic origin, revealing the challenges faced by AI in accurately simulating real-world environments [16,21].

The image from Figure 6 is a synthetic one created by the AIGC, as it can be seen the textures are not consistent throught the entire image, the apple has some texture defects on the right side and also the texture from the rain drops is flattened wrongly on the bottom part.

Introducing the AIGC image in an image processor to highlight the textures that are not common (Figure 7 in this case for an apple) it is visible that this image can be labeled as synthetic content due to high percentage of detections.

Figure 6. AIGC of an apple.

Figure 7. Analyzed AIGC image.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

Even if this one is easier to identify, the new models require more attention to details and more complex tools and frameworks to identify the synthetic content. The entanglement of multiple streams like audio and video used by the AIGC tools becomes even more difficult to identify.

2.4. Video

Tools such as *DeepFaceLab [22]* and *FaceSwap* have popularized the creation of hyper-realistic fake videos. Detecting these requires analyzing not only the visual elements but also checking for inconsistencies in audio and behavioral patterns. Videos combine multiple streams of data (visual, audio, and metadata), each of which can be independently manipulated, complicating the detection process.

A notable issue in synthetic videos, particularly those that have been dubbed or where the audio has been generated separately, is the occurrence of lip-sync errors. In natural human speech, there is a precise synchronization between the movement of the lips and the spoken words [20]. However, in synthetic videos, this synchronization is often imperfect. Viewers may notice mismatches where the lip movements do not align accurately with the audio, creating a disjointed and unnatural viewing experience. These lip-sync discrepancies are a clear indicator of video synthesis and highlight the challenges in achieving seamless audio-visual integration [23].

Another characteristic of synthetic videos is the presence of facial expression inconsistencies. In real human interactions, facial expressions are closely tied to the emotional tone and content of the speech, providing visual cues that enhance communication [11].

Synthetic videos may struggle to replicate this natural correspondence. As a result, the facial expressions displayed may not appropriately match the emotional tone or content of the spoken words. For instance, a synthetic character might smile while delivering sad news or maintain a neutral expression during an emotionally charged dialogue. These inconsistencies can detract from the video's realism and are indicative of the current limitations in AI-driven facial animation.

One common approach in detecting synthetic videos is split the entire video in frames (Figure 8) and analyze them as an array of images, therefore using the same methodology as in the case of a single image, this will allow a better understanding if the content is generated or genuine [24,25].

Figure 8. Frame analysis of a video.

Recognizing these characteristics is crucial for developing more accurate and efficient detection systems, which play an essential role in preserving media integrity. This effort supports the broader objectives of digital media forensics, where ensuring media authenticity and reliability is key to maintaining trust in information sources. As digital content continues to expand, such initiatives are vital not only for protecting individuals but also for defending institutions, industries, and public dialogue from misinformation and manipulation.

3. Challenges and Difficulties

To enhance the robustness of detection systems, it is crucial to propose the use of integrated systems that analyze text, image, and video data in tandem. These multimodal techniques leverage the strengths of each data type to identify inconsistencies or signs of manipulation more effectively. For instance, discrepancies between the audio and visual elements in a video, or between the text and the corresponding imagery, can serve as indicators of synthetic content. By cross-referencing multiple

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

data streams, these integrated systems can provide a more comprehensive and accurate assessment of authenticity, significantly improving the detection of manipulated media.

In the face of rapidly evolving manipulation techniques, it is essential to develop adaptive algorithms. These algorithms should be capable of learning from new data and adapting to emerging methods of content creation and manipulation. As AI-generated fakes become increasingly sophisticated, detection systems must continuously evolve to counter these advancements. Implementing machine learning models that can update and refine their detection capabilities based on the latest manipulation techniques will be critical in maintaining the effectiveness of these systems. The adaptability of these algorithms will ensure they remain relevant and capable of identifying even the most subtle and advanced fakes.

At the moment the rate of success in identifying artificial intelligence generated content is very fluctuant and human manipulation of the synthetic content makes it even more complex to differentiate the generated compared to genuine. In the studies [26,27] the results of the current tools are frightening:

Legend: 1. Human written (95%), 2. Human written in a non-English language and automatically translated into English (80%), 3. AI-generated (70%), 4. AI generated with multiple prompts (67%), 5. AI generated with human/manual edits (41%), 6. AI-generated with subsequent AI/machine paraphrasing (25%).

As visible in the Figure 9 the rate of accuracy drops from 70% to 25% after the text was paraphrased, therefore increasing the difficulty of any tool and framework to identify the AIGC [28–31].

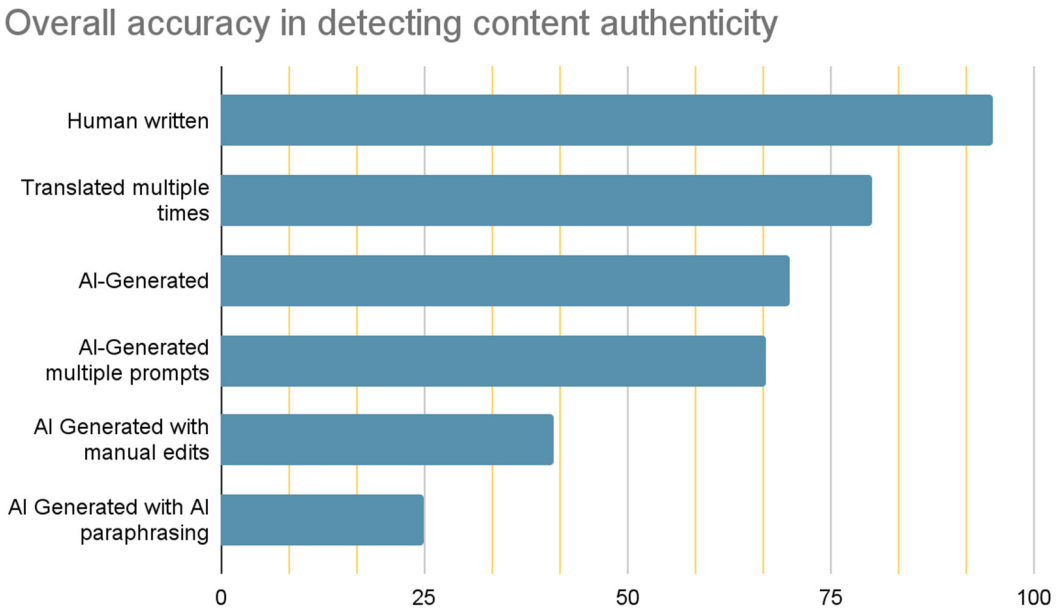


Figure 9. Accuracy of identifying generated content.

To detect manipulations in various types of media, specific machine learning models have proven highly effective. Convolutional Neural Networks (CNNs) are particularly well-suited for detecting image manipulations. CNNs can analyze pixel-level details to identify anomalies that are not visible to the human eye, such as subtle alterations in texture or lighting that indicate tampering. The study [8,32,33] highlights the efficacy of CNNs in this area comparing the application of neural networks in identifying deepfake images and videos. Similarly, Recurrent Neural Networks (RNNs) are adept at identifying text-based inconsistencies. RNNs can analyze sequential data and detect contextual anomalies within the text, such as improbable word usage or grammatical errors that signal synthetic generation [33,34]. By leveraging these neural network models, detection systems can achieve high accuracy in identifying manipulated media.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

Despite the advancements in AI-driven detection systems, there is a compelling argument for a hybrid approach that combines AI's computational power with human judgment. AI systems excel at processing large volumes of data and identifying patterns that may elude human detection [11]. However, human experts bring contextual understanding and critical thinking skills that can enhance the accuracy of these systems. By integrating human oversight into the detection process, we can improve detection rates and reduce false positives. This collaborative approach ensures that nuanced and complex cases, which may be challenging for AI alone, are evaluated with a level of discernment that only human judgment can provide. This synergy between human and AI capabilities can lead to more robust and reliable detection outcomes.

While AI-driven detection systems offer significant benefits, they also present notable challenges. One major limitation is the black-box nature of many AI systems. These systems often lack transparency in how they make decisions, making it difficult to understand and interpret their outputs. This opacity can be particularly problematic in legally sensitive contexts, where the basis for a decision must be clear and explainable. Additionally, the reliance on complex algorithms and large datasets can introduce biases and errors, further complicating their use. Addressing these challenges requires developing more transparent and interpretable AI models, as well as establishing rigorous standards for their deployment and evaluation. By acknowledging and addressing these limitations, we can work towards more reliable and ethically sound detection technologies.

Each area of the synthetic content generation faces multiple and difficult challenges, in the Table 10 there is a brief summary of the research papers that highlight these challenges.

Table 10. AIGC content detection challenges.

Area	Challenges	Reference
AI Text Detection	Detecting AI-generated text across domains is challenging due to **differences in context** and **writing style**	[7,12–14,17,18,35–37]
AI Video/Image Detection	Real-time detection of AI-generated images is **complex due to processing demands** and **frequent updates in image-generation models**	[10,15,16,20,21,25]
AI Audio Detection	Ensuring AI-generated music **maintains quality** and **originality** while **adhering to copyright regulations**	[2,19,20]

The development of detection technologies raises significant ethical and legal considerations that must be addressed. The use of surveillance to monitor and analyze media content can impinge on privacy rights, leading to ethical dilemmas about the extent and manner of data collection. It is imperative to handle these ethical implications, ensuring that the deployment of such technologies respects individual privacy and freedoms. Additionally, there is a need for robust legal frameworks to manage and regulate the use of detection technologies. These frameworks should establish clear guidelines for responsible use, ensuring that the technologies are employed in ways that are fair, transparent, and accountable.

4. Analysis of a Current Detection Tools/Algorithms

Traditional methods for detecting computer-generated content often fall short when faced with sophisticated AI-generated content. Innovations in detection technologies are crucial to address these shortcomings. One such innovation is the development of dual-stream networks that employ cross-attention mechanisms to better identify subtle anomalies distinguishing AIGC from genuine content. The research paper [38] demonstrates that these advanced detection systems can effectively capture fine-grained inconsistencies that traditional methods might miss. By integrating multiple streams of data and focusing attention mechanisms across them, these networks enhance the capability to detect nuanced irregularities in AI-generated media [8].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

As AI content generation techniques continue to improve, so does the ability to evade existing detection systems, including those reliant on watermarks. This escalation necessitates new approaches that can withstand adversarial manipulations while maintaining high visual quality. The study [39] suggests developing robust detection methods that are less susceptible to adversarial attacks, ensuring reliable identification of AIGC. These approaches could include adversarial training and the use of more resilient features that are less likely to be manipulated without detection. The detection of AI-generated content is increasingly moving beyond single modalities, such as text or images alone, to multimodal approaches. These approaches leverage combined cues from text, images, and possibly audio to enhance the accuracy and reliability of detection systems across various content types.

In academic contexts, distinguishing between human-written and AI-generated texts is critical for maintaining scientific integrity. As generative AI becomes more prevalent in research, there is a pressing need to develop frameworks that can reliably identify AI contributions in scientific publications. Implementing these frameworks will help preserve the credibility and trustworthiness of scholarly communication.

To keep pace with new generations of AI-generated content, incremental learning approaches are being explored. These approaches allow detection systems to adapt over time and incorporate new data, improving their effectiveness against the latest generative models. In the research paper [40] is highlighted that incremental learning can enhance the flexibility and responsiveness of detection systems, ensuring that they remain effective as AI technologies evolve. By continuously refining detection algorithms, it is possible to maintain robust defenses against increasingly sophisticated AI-generated content.

The following section presents approaches with their major characteristics from the literature review, highlighting the use cases, challenges, and benefits. For clarity, the summary of the findings is presented in Table 11.

Table 11. Current tools and algorithms.

Approach	Challenges	Benefits	References
Stylometric analysis	Detecting AI-generated text across different domains due to varied contexts	Uses **stylometric analysis** with **explainable AI**, improving transparency in AI text detection	[41–46]
Watermarking and digital fingerprints	Detecting realistic deepfakes, which constantly evolve in quality	Applies **watermarking** to help verify authentic versus synthetic content	[47–61]
Adversarial and robust detection techniques	Real-time detection is challenging due to processing speed and model updates	Develops **robust detection** for real-time image analysis and authentication	[8,10,24,40,62–68]
Machine learning models	GAN-generated images closely resemble real images, making detection challenging	Builds **machine learning models** for reliable detection of AI-generated images	[3,18,69–71]
Blockchain	Utilizing blockchain for AI trust while addressing scalability and data privacy concerns	**Blockchain** provides verified, tamper-proof information management	[53,57,72–76]

4.1. Stylometric Analysis

This approach analyzes the style of writing, including syntax, word usage, and grammar, to identify inconsistencies or anomalies that suggest content might be AI-generated. Stylometric tools

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

can be particularly effective in cases where a specific author's style is well-documented. Stylometric analysis refers to the study of the unique stylistic choices of an author, which can be used to attribute authorship to anonymous or disputed texts. It's a method often used in literary studies, forensic linguistics, and digital text analysis, leveraging the way language is used and arranged by different authors [41,45]. Stylometrics focuses on quantifying and analyzing aspects such as word frequencies, sentence length, grammatical patterns, and other linguistic features that tend to be consistent and individualistic within a writer's work.

In the context of detecting AI-generated content, stylometric analysis can be employed to distinguish between human-written and machine-generated texts. This is possible because AI models, even sophisticated ones like GPT, often produce text with certain detectable patterns and anomalies that may not align with the natural linguistic style of a human author [42,46]. For instance, AI-generated text might show repetitive syntactic structures, unusual word choice combinations, or consistency errors in style across a text. By analyzing these features, stylometric methods can help identify whether a piece of writing was generated by a human or a machine. Stylometric analysis in practice for language models would cover elements as presented in Table 12 [41,45].

Table 12. Stylometric analysis of generated text.

Phrases generated by a language model	Stylometric analysis
Humanity has reached a significant point of reflection, where both the moral and existential questions about artificial intelligence must be addressed to avoid potential consequences that could reshape our world irreversibly	**High word frequency in filler words**: Overuse of vague connectors (*where both*, *must be addressed*, *potential consequences*) and modal verbs (*must*, *could*). Stylometric analysis may show unnatural filler-word frequency or lack of diversity in key terms
With the dawn of big data, companies are now in possession of vast amounts of information, enabling better decision-making processes but raising serious concerns regarding privacy, ethics, and control over user data.	**Phrase repetition across contexts**: Frequent recycling of *vast amounts of information* and *serious concerns regarding privacy* without depth. Stylometric analysis could indicate a repetitive n-gram pattern, suggesting the model's tendency to repeat high-frequency phrases.
Emerging technologies bring with them not only innovation but also responsibility, as organizations strive to harness their potential while upholding the ethical principles that shape a just and equitable future for all	**Overuse of platitudes**: Terms like *upholding ethical principle* and *just and equitable future* sound polished but convey little actionable insight. Stylometric analysis might reveal an excessive use of long phrases that overgeneralize complex topics
While the digital transformation accelerates, institutions face unprecedented challenges in adapting to new paradigms, requiring a commitment to transparency, collaboration, and agility in navigating the unknown	**Unnatural lexical choices and phrasing**: Phrases like *navigating the unknown* combined with *commitment to transparency* show abrupt topic shifts. Stylometric analysis may highlight odd lexical choices or inconsistent lexical richness across sentences

Some tools focus on detecting paraphrased content or unusually high similarities to known AI-generated texts. This includes analyzing sentence structures, vocabulary, and thematic consistency across large datasets to identify potential matches with known AI outputs. Content similarity and paraphrasing detection involve identifying and analyzing texts to determine how closely they resemble each other in terms of meaning, even if their words and structures differ. These techniques are crucial in various fields such as education, where they help detect plagiarism, and in natural language processing applications, where they ensure the uniqueness and diversity of generated content.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

Another type of framework [12] is one that utilizes Explainable AI (xAI) techniques to analyze stylistic features and improve the interpretability of machine learning model predictions in distinguishing AI-generated texts, showing high accuracy and identifying key attributes for classification.

Content similarity detection focuses on measuring how similar two pieces of content are. This can be done at different levels. The first level would be lexical which measures the overlap in vocabulary between two texts. Techniques like cosine similarity, Jaccard index, and others that analyze term frequency are commonly used. Then on a semantic level, this goes beyond mere word overlap and assesses the meanings conveyed by the texts.

4.2. Watermarking and Digital Fingerprints

Emerging techniques involve embedding watermarks or digital fingerprints in AI-generated content at the time of creation. These markers are designed to be undetectable to readers but can be identified by specialized detection tools, providing a direct method for distinguishing AI-generated content [50]. Watermarking and digital fingerprints are two techniques used to protect and verify the authenticity of digital content. These methods are essential in the realms of copyright protection, content authentication, and tracking the distribution of digital media.

Digital watermarking involves embedding a secret or unique mark within a piece of digital media that can be detected or extracted later to confirm its authenticity or ownership. Watermarks can be designed to be either perceptible or imperceptible to the user, depending on the application.

One of the primary applications of digital watermarking is in the realm of copyright protection. By embedding a unique watermark into digital content, the copyright owner can unequivocally prove their ownership. This watermark acts as a digital signature that is difficult to remove without compromising the content's integrity. In the event of a copyright dispute, the presence of the watermark can serve as compelling evidence to support the copyright owner's claims. This technique is particularly valuable in protecting intellectual property in digital media such as images, videos, and audio files.

Digital watermarking also plays a crucial role in content authentication. By embedding a watermark within the content, it becomes possible to verify its integrity over time. If the watermark remains intact, it indicates that the content has not been tampered with or altered. This capability is essential for applications where maintaining the original state of the content is critical, such as in legal evidence, news reporting, and archival preservation. Watermark verification can quickly detect any unauthorized modifications, thereby ensuring the authenticity and reliability of the content [51].

In the media industry, digital watermarking is used for broadcast monitoring. Watermarks embedded in audio or video clips can be tracked to monitor where and how often the content is played . This information is invaluable for various purposes, including royalty calculations, audience measurement, and usage monitoring. For example, broadcasters can use watermark data to determine the reach and frequency of specific advertisements or programs, ensuring accurate and fair compensation for content creators. Additionally, this tracking capability helps in enforcing licensing agreements and preventing unauthorized broadcasts.

In the Figure 13 the watermarking process consists in adding whitenoise to an image in order to prove their authenticity and also include digital rights and royalty calculations for future use of this image.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

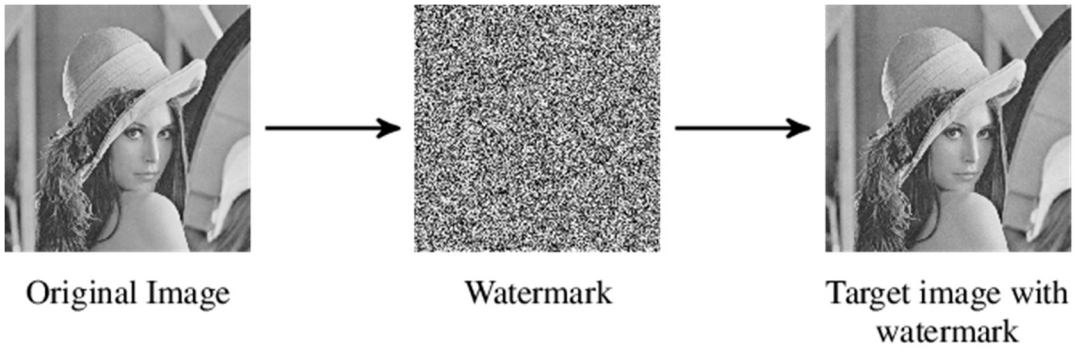


Figure 13. Watermarking an image [35].

Digital fingerprinting, on the other hand, involves creating a unique identifier or *fingerprint* for digital content based on its characteristics or features [59,60]. This fingerprint is unique to each piece of content, much like a human fingerprint.

One of the primary applications of digital fingerprinting is content tracking. By generating unique fingerprints for each piece of digital content, it is possible to monitor its distribution across various channels. This ensures that the content does not appear in unauthorized locations [52,58]. For example, digital fingerprints can help media companies track how their videos are shared across social media platforms, identifying any instances of unauthorized uploads. This capability is essential for protecting intellectual property and ensuring that content distribution adheres to licensing agreements.

Digital fingerprinting is also a valuable tool in forensic investigations, particularly in cases of illegal content distribution. By analyzing the fingerprints embedded in digital media, investigators can trace the source or distribution path of the content. This helps identify the origin of pirated material and the individuals or networks involved in its dissemination. Digital fingerprints provide a trail of evidence that can be used in legal proceedings to prosecute those responsible for copyright infringement or other illegal activities related to digital content.

Media companies leverage digital fingerprinting for efficient content management. With large libraries of digital content, it can be challenging to organize and retrieve specific media files. Digital fingerprints enable precise identification and categorization of content, making it easier to manage extensive media archives [59]. This technology supports the automation of content indexing and retrieval processes, improving operational efficiency and ensuring that media assets are easily accessible when needed.

On a comparison level there are implementational, security and other key differences between watermarking and digital fingerprinting. Table 14 summarises the dissimilarities of those two approaches highlighting the varianton of specific criteria.

Table 14. Differences between watermarking and digital fingerprints.

Criteria	Watermarking	Digital fingerprinting
Visibility	Can be **visible** or **invisible**, visible watermarks assert ownership directly on the content, while invisible ones protect without altering appearance	Always **invisible** and does not alter content in any perceptible way, maintaining the original user experience
Purpose	Primarily used to assert **ownership** and maintain content **integrity**	Used mainly for **tracking** and **identification** to monitor distribution and detect unauthorized use
Robustness to alterations	Designed to be **robust against manipulations** like compression,	**Sensitive to content changes**, significant alterations, like re-

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

	cropping, and minor edits, persists through moderate transformations	encoding or major edits, can produce a different fingerprint
Content protection	Helps in **copyright enforcement** by embedding proof of ownership, aiding in verification of authenticity and integrity	**Assists in monitoring distribution** and forensic analysis to track content's path and identify unauthorized distribution
Use in copyright enforcement	**Critical for asserting copyright** ownership and deterring misuse through visible or invisible marks	Valuable for **tracking** usage patterns and identifying sources in case of illegal distribution or infringement
Suitability for AIGC	Used in **AI-generated content detection,** watermarks can be embedded in media to detect authenticity in AI-generated visuals	**Less commonly used for AIGC** detection but can help track specific user activity when fingerprints are embedded in distributed content
Susceptibility to removal	May be **vulnerable to watermark removal attacks** if methods of embedding are publicly known or poorly implemented	**Difficult to remove** without changing the content significantly, as fingerprints are embedded uniquely and are integral to the content distribution
Implementation complexity	Generally **simpler to implement**, as it relies on embedding identifiable markers, visible marks are particularly straightforward	More complex, **requiring unique identifiers** per copy and often higher computational resources to embed and retrieve fingerprints
Example applications	**Common** in image copyright marking, video watermarking for brand logos, and preventing reuse of visual content in unauthorized contexts	**Widely used** in music streaming to monitor piracy, digital video distribution to track views, and document tracking for unauthorized sharing
Robustness research	Research [16] indicates some watermarking methods **may be vulnerable** to advanced removal techniques, particularly in **AIGC**	Less focus on robustness against tampering but can be **affected by significant modifications** that alter the original fingerprint

4.3. Adversarial and Robust Detection Techniques

As AI models become more sophisticated, detection methods also evolve to counteract evasion tactics. This includes developing models that can detect AI-generated content even when it has been altered or paraphrased to evade simpler detection algorithms.

Adversarial and robust detection techniques are critical components in the field of machine learning, particularly in the context of enhancing the security and reliability of models against adversarial attacks. These techniques aim to detect and mitigate the effects of inputs specifically designed to deceive or confuse models [63].

Adversarial detection focuses on identifying and responding to adversarial attacks, which are manipulations of input data intended to cause a machine learning model to make errors [24,39]. These attacks can be subtle, such as slight alterations to an image that cause it to be misclassified.

The adversarial detection follows a few key approaches, one of them is input reconstruction. This technique involves the use of autoencoders to reconstruct the input data and then compare the reconstructed input to the original. Autoencoders are neural networks designed to learn efficient codings of input data, typically for dimensionality reduction or noise removal. In the context of adversarial detection, large discrepancies between the original and reconstructed inputs may indicate the presence of adversarial manipulations [64]. By analyzing these differences, it is possible to detect and potentially mitigate the effects of adversarial attacks on the model.

Another crucial method for enhancing adversarial robustness is adversarial training. This approach involves incorporating adversarial examples directly into the training process. By exposing

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

the model to these challenging examples during training, the model learns to recognize and withstand adversarial inputs. Adversarial training effectively improves the model's robustness by making it more capable of handling variations and distortions that adversarial attacks introduce. This method has been shown to significantly enhance the resilience of machine learning models against various types of adversarial manipulations [68].

Statistical methods are also implemented to detect adversarial tampering by analyzing input data for statistical anomalies. These methods involve examining the input data's statistical properties and identifying deviations from typical data distributions. Statistical anomalies that diverge from expected patterns can indicate adversarial interference [71]. By leveraging statistical techniques, such as anomaly detection algorithms and distributional analysis, it is possible to identify inputs that do not conform to normal data behavior, thereby signaling potential adversarial attacks.

There are also techniques for enhancing model robustness, one of them is the regularization technique (Figure 15). This method, including dropout, L1/L2 regularization, and data augmentation, helps in preventing the model from overfitting to noise in the training data. Dropout involves randomly dropping units from the neural network during training to prevent co-adaptation of features. L1/L2 regularization adds penalty terms to the loss function to constrain the model's complexity. Data augmentation artificially expands the training dataset by applying transformations to the original data. These techniques collectively enhance the model's generalization capabilities, making it more resilient to noisy and adversarial inputs [62,67].

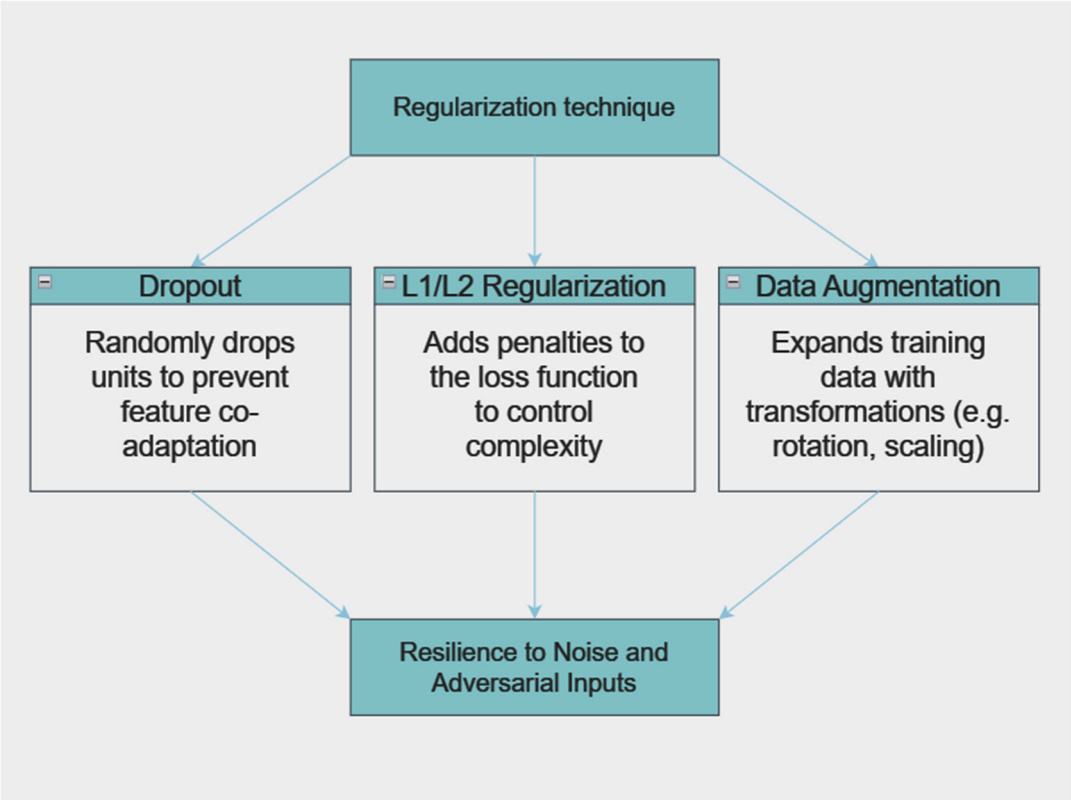


Figure 15. Regularization technique.

Robust optimization is another strategy to ensure model robustness. This approach focuses on minimizing the worst-case loss across all possible adversarial examples within a specified perturbation range of the training data. By optimizing the model to perform well under these worst-case scenarios, robust optimization techniques enhance the model's ability to withstand adversarial attacks. This method involves formulating and solving optimization problems that account for potential adversarial perturbations, leading to models that maintain reliable performance even in the presence of adversarial inputs.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

Techniques such as randomized smoothing and provable defenses based on the geometric properties of data offer good defense against adversarial attacks. Randomized smoothing involves adding random noise to the input and averaging the model's predictions, creating a smoothed classifier that is robust to small adversarial perturbations. Provable defenses rely on the mathematical properties of the data and model to ensure robustness. These certified defenses offer a high level of assurance that the model will remain effective even under adversarial conditions [65].

Integrating adversarial and robust detection techniques can provide a comprehensive defense mechanism for machine learning systems, particularly those deployed in security-sensitive environments. By preparing models to both recognize adversarial attempts and withstand them without degradation in performance, these integrated approaches significantly enhance the resilience of AI systems.

4.4. Machine Learning Models

Many detection methods rely on machine learning algorithms that are trained on datasets of human-written and AI-generated texts. These models learn to identify subtle differences in writing style, patterns, and other linguistic features that may not be immediately apparent to human readers.

Some detection tools leverage pre-trained language models, such as GPT-3, GPT-4, Electra or BERT, which are fine-tuned on datasets of AI-generated and human-written texts to enhance their detection capabilities [8,77–79]. These models are particularly effective in zero-shot or few-shot learning scenarios, where they can make accurate predictions based on minimal examples. In zero-shot learning, a model performs a task or identifies classes it has never seen before during training. Instead of relying on labeled data for the specific task, the model uses general knowledge learned from other tasks or data to make predictions. Few-shot learning allows a model to learn a new task or identify new classes from just a few labeled examples. Unlike zero-shot learning, this approach involves a small amount of task-specific data to help the model understand the new context or task requirements.

By utilizing the extensive knowledge embedded in pre-trained models, detection tools can achieve higher accuracy and reliability in distinguishing between human and AI-generated content and depending on the model it can be compared as in Table 16.

Table 16. Comparison of machine learning models [77,80–84].

Model	Accuracy	Reliability	Efficiency	Use Case Highlights	Adaptability to Few-shot Learning
GPT-3 [77,80,81]	High accuracy in certain generative and comprehension tasks, but lower in domain-specific accuracy	Moderate reliability, struggles with overfitting in limited domain-specific tasks	Resource-intensive, especially for fine-tuning on specialized hardware	Effective in generative tasks and large-scale NLP, limited adaptability for specialized tasks	Strong in few-shot learning, especially in generative tasks
BERT [80,81,85]	High accuracy in classification and NLU (Natural Language understanding) tasks, including	Generally reliable for structured NLP tasks, especially classification and sentiment analysis	More efficient than GPT-3, effective in many NLP applications without extensive resources	Best for information extraction, sentiment analysis, and question answering	Moderate, few-shot limited by pre-training on specific masked language tasks

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

	entity recognition				
BERT (BioBERT) [77,85]	High in biomedical domain	Reliable for biomedical text, high precision and recall in medical and research contexts	Moderate resource needs, optimized for biomedical domains	Protein interactions, biomedical text mining, disease detection	Limited, specialized for biomedical NLP, limited few-shot learning applicability
ELECTRA [82]	Comparable to BERT with higher efficiency in classification tasks	Consistently reliable, high performance in GLUE (General Language Understanding Evaluation) benchmark with reduced computation	High efficiency due to pre-training as discriminator rather than generator	Token-based NLP tasks, question answering, and tasks with constrained resources	Moderate, not optimized for few-shot, but robust for token detection tasks
CodeT5 [83]	High accuracy in code generation and defect detection tasks	Reliable in tasks with code semantics and identifier differentiation	Efficient for programming language tasks, leverages code-specific pretraining	Code generation, defect detection, multi-language programming tasks	Limited, focused on programming language and code understanding
GPT-4 (general NLP) [80,86]	Not domain-specific but achieves high accuracy on text generation and NLU	Higher reliability with human-aligned feedback, yet prone to bias	High compute needs similar to GPT-3 but more efficient at scale	Complex language modeling, context-driven generation	Strong, surpasses GPT-3 with fine-tuning on varied datasets

Combining multiple detection strategies can significantly improve the accuracy and robustness of identifying AI-generated content. For instance, a detection tool might employ both stylometric analysis and machine learning models to cross-verify the likelihood that content is AI-generated. This hybrid approach leverages the strengths of different methodologies, reducing the risk of false positives and negatives [18]. By integrating diverse detection techniques, it is possible to create more comprehensive and reliable detection systems.

The development of community-driven and open-source tools plays a crucial role in advancing the field of adversarial and robust detection. These efforts bring together researchers and practitioners to collaborate on the latest research and methodologies. Open-source projects provide access to cutting-edge tools and techniques, fostering innovation and facilitating the widespread adoption of effective detection strategies. Community-driven initiatives also ensure that the tools remain up-to-date with the latest advancements and are accessible to a broader audience.

Each of these options has its strengths and limitations, and their effectiveness can vary based on the context in which they are used and the sophistication of the AI models generating the content. As AI technology continues to advance, the development of detection methods remains an active area of research and innovation.

4.5. Blockchain

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

Blockchain technology, when combined with crowdsourcing, offers a promising approach to enhance the trustworthiness of content by providing a decentralized and secure method for validating their authenticity. The integration of these technologies can significantly reduce dependency on single sources and mitigate issues related to misinformation and biased content [72].

Blockchain operates on a distributed ledger system where data is stored across multiple nodes, ensuring no single point of control. This decentralization makes it difficult for any one entity to manipulate information. Once data is recorded on a blockchain, it cannot be altered or deleted [75,76]. This feature ensures the integrity and permanence of the information, making it a reliable source for historical data. Due to that, the blockchain technology has a high level of transparency and therefore all transactions and data entries on a blockchain are visible to all participants. This transparency allows for greater accountability and traceability of information sources.

Crowdsourcing leverages the collective knowledge and expertise of a large group of people to verify the authenticity of the content. This approach can help identify synthetic information more effectively than relying on a single entity. In a crowdsourcing model, individuals can review and validate content. This peer review process helps filter out inaccurate information through consensus. Crowdsourcing involves contributions from diverse individuals, providing multiple perspectives and reducing the risk of bias.

Using this technology the applications in the area of detection for synthetic content is unlimited, news articles or pieces of information can be submitted to a blockchain platform for validation, reduces the risk of manipulation and fake news (Figure 17). A preliminary verification is conducted by automated algorithms or trusted nodes to check for obvious signs of misinformation. The news is then reviewed by a distributed network of individuals who assess its credibility based on various criteria such as source reliability, factual accuracy, and consistency with known information. The collective assessments are recorded on the blockchain. A consensus mechanism, such as *Proof of Stake* or *Proof of Authority*, is used to validate the majority opinion [44,87–89].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

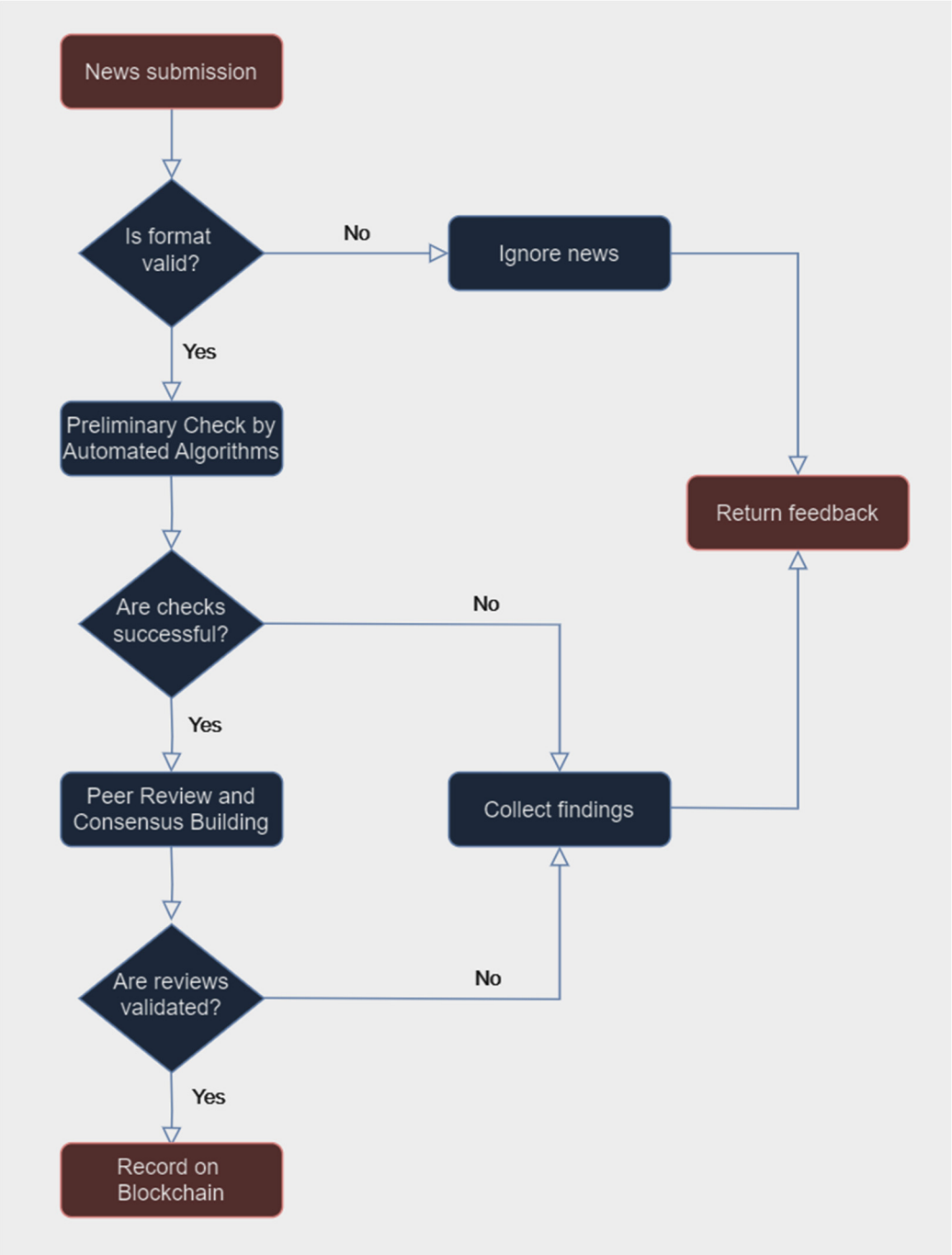


Figure 17. News submission on a blockchain ledger.

One big advantage in using this would be the reduced dependency of centralized authorities by decentralizing the validation process, reliance on single sources or central authorities is minimized. Also the blockchain technology has a set of cryptographic features that can protect against tampering and unauthorized modifications, therefore increasing the trust due to the transparency, immutability, and collective validation.

Several studies [72,87,88] have explored the integration of blockchain and crowdsourcing for validating news authenticity, the study [72] highlights the role of blockchain in establishing trusted information systems across various domains, highlighting the technology's ability to ensure data provenance, verify identities, and maintain trust among collaborating entities. On top of the previous research, there is a document verification on blockchain paper [74] that proposes also a blockchain-based model for document verification, which enhances security, reliability, and efficiency in the

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

verification process by enabling decentralized sharing of authenticated documents between government bodies and private organizations.

In regards of certificate verification using blockchain the most in depth research paper is [73] which describes a prototype for verifying the authenticity of academic certificates using blockchain, which ensures that third parties can independently verify certificates even if the issuing institution is no longer operational.

Incorporating blockchain technology with crowdsourcing can significantly enhance the trustworthiness of news by providing a decentralized, transparent, and secure method for validating authenticity. This approach not only reduces dependency on single sources but also leverages the collective intelligence of diverse individuals to filter out misinformation effectively.

5. Directions of Innovation

The identification and differentiation of AI-generated content, including videos, photos, and text, are pivotal areas of research, given the significant advancements in artificial intelligence. AI-generated content encompasses a wide range of outputs such as images, text, audios, and videos. While AIGC has numerous benefits, it also presents several risks, including issues related to privacy, bias, misinformation, and intellectual property. Identifying AIGC is crucial to mitigating these risks and ensuring responsible usage [4].

Research focused on distinguishing AI-generated images from traditional computer graphics has led to the development of a dual-stream network. This network utilizes texture information and low-frequency forged traces to identify images generated by AI, showing superiority over conventional detection methods [38]. Studies [49,58] on the robustness of watermark-based AI-generated content detection reveals that attackers can evade detection by adding small, imperceptible perturbations to watermarked images.

Advancements in AI-generated images necessitate ongoing detection efforts. The study [15] explores the generalization of detection methods across different AI generators and introduces pixel prediction techniques for identifying inpainted images, indicating progress in online detection capabilities.

Research [69] proposes a framework for distinguishing between human and AI-generated text, particularly in scientific writings, using a model trained on predefined datasets to assess accuracy and reliability.

For synthetic video content researchers proposed a framework [25] that employs instruction-tuned large language models to generate frame-by-frame descriptions from a single user prompt, enhancing the consistency and coherence of AI-generated videos. Paper [8] proposes using BERT and CNN models for detecting AI-generated text and images by analyzing vocabulary, syntactic, semantic, and stylistic features, alongside the use of a CNN model trained on specific datasets for image detection.

This directions showcase the breadth of approaches being explored to identify AI-generated content, from watermarking and dual-stream networks for images to advanced text analysis frameworks and the use of large language models for video generation. The ongoing development of detection methods is critical for addressing the challenges posed by the increasing sophistication of AI-generated content.

6. Conclusions and Ethical Concerns

The proliferation of AI-generated content, encompassing audio, video, text, and images, represents a transformative development in current society. This technological advancement has opened new horizons in creative expression, communication, and information dissemination [90].

The importance of these technologies is underscored by their wide-ranging impact across various domains. In media and entertainment, AI-generated content offers unprecedented creative possibilities, enabling the production of high-quality media at scale and reducing the barriers to content creation. In education and training, AI can generate personalized and interactive learning

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

materials, enhancing the educational experience. In business and marketing, AI-driven tools can create targeted and engaging content, driving customer engagement and brand loyalty.

Despite these benefits, the potential for misuse of AI-generated content poses substantial risks. Adversarial attacks, deepfakes, and other forms of content manipulation can undermine trust in digital media, spread misinformation, and cause significant social and economic harm. This underscores the critical need for advanced detection techniques to differentiate between genuine and synthetic content.

There are concerns about synthetic media in regard to authenticity and misrepresentation because it is being used to create convincing yet entirely fabricated pieces of content, which can mislead, manipulate public opinion, or defame individuals. The ease of generating realistic synthetic media raises challenges in maintaining authenticity and trust in digital content [20,91,92].

The use of AI to generate synthetic media often involves training models on large datasets, which may include personal data collected without explicit consent [93,94]. This raises privacy issues, particularly when personal characteristics are replicated or manipulated without permission. The generation of synthetic media can infringe on intellectual property rights, as AI can produce content that closely mimics the style or substance of copyrighted works. This poses legal and ethical challenges regarding the ownership and copyright of AI-generated content.

AI systems can perpetuate or even exacerbate existing social and cultural biases if they are not carefully designed. The data used to train these systems can contain biases, which the AI may then learn and replicate in its outputs, leading to stereotyping and discrimination in synthetic media [94].

Transparent and interpretable AI models in news feeds can help combat misinformation by clarifying why certain content is flagged as false, which can also help mitigate the effects of echo chambers and filter bubbles [87].

Deploying AI and crowd-sourced systems to label false news enhances discernment among users concerning what content to share on social media. Explanations on how AI-generated warnings are produced can further improve their effectiveness, although they may not increase trust in the labels [89].

Combining linguistic and knowledge-based approaches can significantly improve the accuracy of fake news detection systems. Features like the reputation of the news source and fact-check verifications are critical in this process [88,94].

In conclusion, as AIGC continues to integrate into the fabric of daily life, its impact on society will only grow. Ensuring the authenticity and integrity of this content is paramount to harnessing its potential benefits while mitigating associated risks. The advancement of adversarial detection and robust detection techniques is critical to achieving this balance, fostering a digital environment where AI-generated content enhances rather than detracts from societal well-being. The collective effort of researchers, technologists, and policymakers will be essential in navigating this complex landscape, ensuring that the transformative power of AI serves the best interests of society maintaining the integrity and trustworthiness of information disseminated through different platforms.

Author Contributions: Conceptualization, C.D.M.; methodology, C.D.M.; formal analysis, C.D.M.; investigation, C.D.M.; writing—original draft preparation, C.D.M.; writing—review and editing, C.D.M.; visualization, C.D.M.; supervision, D.E.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Spector, L. Evolution of Artificial Intelligence. *Artificial Intelligence* **2006**, *170*, 1251–1253, doi:10.1016/J.ARTINT.2006.10.009.
2. Anantrasirichai, N.; Bull, D. Artificial Intelligence in the Creative Industries: A Review. *Artificial Intelligence Review 2021 55:1* **2021**, *55*, 589–656, doi:10.1007/S10462-021-10039-7.
3. Wu, J.; Gan, W.; Chen, Z.; Wan, S.; Lin, H. AI-Generated Content (AIGC): A Survey. **2023**.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

4. Chen, C.; Fu, J.; Lyu, L. A Pathway Towards Responsible AI Generated Content. *IJCAI International Joint Conference on Artificial Intelligence* **2023**, *2023-August*, 7033–7038, doi:10.24963/ijcai.2023/803.
5. Belgodere, B.; Dognin, P.; Ivankay, A.; Melnyk, I.; Mroueh, Y.; Mojsilovic, A.; Navratil, J.; Nitsure, A.; Padhi, I.; Rigotti, M.; et al. Auditing and Generating Synthetic Data with Controllable Trust Trade-Offs. **2023**, doi:10.1109/JETCAS.2024.3477976.
6. Georgiev, G. Has Interest in Data Science Peaked Already? | by Georgi Georgiev | Towards Data Science Available online: https://towardsdatascience.com/has-interest-in-data-science-peaked-already-437648d7f408 (accessed on 3 June 2024).
7. Salvi, D.; Hosler, B.; Bestagini, P.; Stamm, M.C.; Tubaro, S. TIMIT-TTS: A Text-to-Speech Dataset for Multimodal Synthetic Media Detection. *IEEE Access* **2023**, *11*, 50851–50866, doi:10.1109/ACCESS.2023.3276480.
8. Vora, et al. V. A Multimodal Approach for Detecting AI Generated Content Using BERT and CNN. *International Journal on Recent and Innovation Trends in Computing and Communication* **2023**, *11*, 691–701, doi:10.17762/IJRITCC.V11I9.8861.
9. Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Ferrer, C.C. The DeepFake Detection Challenge (DFDC) Dataset. **2020**.
10. Nguyen, T.T.; Nguyen, Q.V.H.; Nguyen, D.T.; Nguyen, D.T.; Huynh-The, T.; Nahavandi, S.; Nguyen, T.T.; Pham, Q.V.; Nguyen, C.M. Deep Learning for Deepfakes Creation and Detection: A Survey. *Computer Vision and Image Understanding* **2022**, *223*, 103525, doi:10.1016/J.CVIU.2022.103525.
11. Agarwal, S.; Varshney, L.R. Limits of Deepfake Detection: A Robust Estimation Viewpoint. **2019**.
12. Shah, A.; Ranka, P.; Dedhia, U.; Prasad, S.; Muni, S.; Bhowmick, K. Detecting and Unmasking AI-Generated Texts through Explainable Artificial Intelligence Using Stylistic Features. *International Journal of Advanced Computer Science and Applications* **2023**, *14*, 1043–1053, doi:10.14569/IJACSA.2023.01410110.
13. Sadasivan, V.S.; Kumar, A.; Balasubramanian, S.; Wang, W.; Feizi, S. Can AI-Generated Text Be Reliably Detected? **2023**.
14. Rodriguez, J.D.; Hay, T.; Gros, D.; Shamsi, Z.; Srinivasan, R. Cross-Domain Detection of GPT-2-Generated Technical Text. NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference **2022**, 1213–1233, doi:10.18653/V1/2022.NAACL-MAIN.88.
15. Epstein, D.C.; Jain, I.; Wang, O.; Zhang, R. Online Detection of AI-Generated Images. *Proceedings - 2023 IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2023* **2023**, 382–392, doi:10.1109/ICCVW60793.2023.00045.
16. Corvi, R.; Cozzolino, D.; Zingarini, G.; Poggi, G.; Nagano, K.; Verdoliva, L. On The Detection of Synthetic Images Generated by Diffusion Models. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* **2023**, *2023-June*, doi:10.1109/ICASSP49357.2023.10095167.
17. Krishna, K.; Song, Y.; Karpinska, M.; Wieting, J.; Iyyer, M. Paraphrasing Evades Detectors of AI-Generated Text, but Retrieval Is an Effective Defense. *Adv Neural Inf Process Syst* **2023**, *36*.
18. Alamleh, H.; Alqahtani, A.A.S.; Elsaid, A. Distinguishing Human-Written and ChatGPT-Generated Text Using Machine Learning. *2023 Systems and Information Engineering Design Symposium, SIEDS 2023* **2023**, 154–158, doi:10.1109/SIEDS58326.2023.10137767.
19. Kumar, S.; Kumar, S. AI Generated Music. *International Journal of Research in Science & Engineering* **2024**, *4*, 10–12, doi:10.55529/IJRISE.41.10.12.
20. Kadam, A.; Rane, S.; Mishra, A.; Sahu, S.; Singh, S.; Pathak, S. A Survey of Audio Synthesis and Lip-Syncing for Synthetic Video Generation. *EAI Endorsed Transactions on Creative Technologies* **2021**, *8*, 169187, doi:10.4108/EAI.14-4-2021.169187.
21. Galbally, J.; Marcel, S. Face Anti-Spoofing Based on General Image Quality Assessment. *Proceedings - International Conference on Pattern Recognition* **2014**, 1173–1178, doi:10.1109/ICPR.2014.211.
22. Korshunov, P.; Marcel, S. DeepFakes: A New Threat to Face Recognition? Assessment and Detection. **2018**.
23. Giudice, O.; Guarnera, L.; Battiato, S. Fighting Deepfakes by Detecting Gan Dct Anomalies. *J Imaging* **2021**, *7*, doi:10.3390/jimaging7080128.
24. Pu, W.; Hu, J.; Wang, X.; Li, Y.; Hu, S.; Zhu, B.; Song, R.; Song, Q.; Wu, X.; Lyu, S. Learning a Deep Dual-Level Network for Robust DeepFake Detection. *Pattern Recognit* **2022**, *130*, doi:10.1016/j.patcog.2022.108832.
25. Hong, S.; Seo, J.; Shin, H.; Hong, S.; Kim, S. DirecT2V: Large Language Models Are Frame-Level Directors for Zero-Shot Text-to-Video Generation. **2023**.
26. Jonathan, B. Additional Challenges to Detecting AI Writing - Plagiarism Today Available online: https://www.plagiarismtoday.com/2023/07/31/additional-challenges-to-detecting-ai-writing/ (accessed on 11 June 2024).
27. Gillham, J. AI Content Detector Accuracy Review + Open Source Dataset and Research Tool – Originality.AI Available online: https://originality.ai/blog/ai-content-detection-accuracy (accessed on 8 June 2024).

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

28. Barshay, J. Proof Points: It's Easy to Fool ChatGPT Detectors Available online: https://hechingerreport.org/proof-points-its-easy-to-fool-chatgpt-detectors/ (accessed on 11 June 2024).
29. Pop P. ChatGPT and AI Detectors Available online: https://www.popautomation.com/post/chatgpt-and-ai-detectors (accessed on 17 June 2024).
30. Juhasz, B. How to Avoid Being Flagged by GPT Detectors! The Expert Strategies for Content Writers – Service Lifter Available online: https://servicelifter.com/guides/how-to-avoid-being-flagged-by-gpt-detectors-the-expert-strategies-for-content-writers/ (accessed on 10 July 2024).
31. Christian, P. How to Detect ChatGPT: Tools and Tips for Detection Available online: https://undetectable.ai/blog/how-to-detect-chatgpt/ (accessed on 14 July 2024).
32. Hanrahan, G. Computational Neural Networks Driving Complex Analytical Problem Solving. *Anal Chem* **2010**, *82*, 4307–4313, doi:10.1021/AC902636Q.
33. Ranade, P.; Piplai, A.; Mittal, S.; Joshi, A.; Finin, T. Generating Fake Cyber Threat Intelligence Using Transformer-Based Models. *Proceedings of the International Joint Conference on Neural Networks* **2021**, *2021-July*, doi:10.1109/IJCNN52387.2021.9534192.
34. Talaei Khoei, T.; Ould Slimane, H.; Kaabouch, N. Deep Learning: Systematic Review, Models, Challenges, and Research Directions. *Neural Comput Appl* **2023**, *35*, 23103–23124, doi:10.1007/S00521-023-08957-4.
35. Park, S.; Moon, S.; Kim, J. Ensuring Visual Commonsense Morality for Text-to-Image Generation. **2022**.
36. Welsh, A.P.; Edwards, M. Text Generation for Dataset Augmentation in Security Classification Tasks. **2023**.
37. Orenstrakh, M.S.; Karnalim, O.; Suarez, C.A.; Liut, M. Detecting LLM-Generated Text in Computing Education: A Comparative Study for ChatGPT Cases. **2023**, 121–126, doi:10.1109/compsac61105.2024.00027.
38. Xi, Z.; Huang, W.; Wei, K.; Luo, W.; Zheng, P. AI-Generated Image Detection Using a Cross-Attention Enhanced Dual-Stream Network. *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2023* **2023**, 1463–1470, doi:10.1109/APSIPAASC58517.2023.10317126.
39. Carlini, N.; Wagner, D. Towards Evaluating the Robustness of Neural Networks. *Proc IEEE Symp Secur Priv* **2017**, 39–57, doi:10.1109/SP.2017.49.
40. Marra, F.; Saltori, C.; Boato, G.; Verdoliva, L. Incremental Learning for the Detection and Classification of GAN-Generated Images. *2019 IEEE International Workshop on Information Forensics and Security, WIFS 2019* **2019**, doi:10.1109/WIFS47025.2019.9035099.
41. Neal, T.; Sundararajan, K.; Fatima, A.; Yan, Y.; Xiang, Y.; Woodard, D. Surveying Stylometry Techniques and Applications. *ACM Comput Surv* **2017**, *50*, doi:10.1145/3132039.
42. Brennan, M.; Afroz, S.; Greenstadt, R. Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity. *ACM Transactions on Information and System Security* **2012**, *15*, doi:10.1145/2382448.2382450.
43. Eder, M.; Rybicki, J.; Kestemont, M. Stylometry with R: A Package for Computational Text Analysis. *R Journal* **2016**, *8*, 107–121, doi:10.32614/RJ-2016-007.
44. Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; Stein, B. A Stylometric Inquiry into Hyperpartisan and Fake News. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* **2018**, *1*, 231–240, doi:10.18653/V1/P18-1022.
45. Michailidis, P.D. A Scientometric Study of the Stylometric Research Field. *Informatics 2022, Vol. 9, Page 60* **2022**, *9*, 60, doi:10.3390/INFORMATICS9030060.
46. Abbasi, A.; Chen, H. Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace. *ACM Trans Inf Syst* **2008**, *26*, doi:10.1145/1344411.1344413.
47. Quiring, E.; Arp, D.; Rieck, K. Fraternal Twins: Unifying Attacks on Machine Learning and Digital Watermarking. **2017**.
48. Boujerfaoui, S.; Riad, R.; Douzi, H.; Ros, F.; Harba, R. Image Watermarking between Conventional and Learning-Based Techniques: A Literature Review. *Electronics (Switzerland)* **2023**, *12*, doi:10.3390/ELECTRONICS12010074.
49. Jiang, Z.; Zhang, J.; Gong, N.Z. Evading Watermark Based Detection of AI-Generated Content. *CCS 2023 - Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security* **2023**, 1168–1181, doi:10.1145/3576915.3623189.
50. Makhrib, Z.F.; Karim, A.A. Digital Watermark Technique: A Review. *J Phys Conf Ser* **2021**, *1999*, doi:10.1088/1742-6596/1999/1/012118.
51. Kirchenbauer, J.; Geiping, J.; Wen, Y.; Katz, J.; Miers, I.; Goldstein, T. A Watermark for Large Language Models. *Proc Mach Learn Res* **2023**, *202*, 17061–17084.
52. Wen, Y.; Kirchenbauer, J.; Geiping, J.; Goldstein, T. Tree-Ring Watermarks: Fingerprints for Diffusion Images That Are Invisible and Robust. **2023**.
53. Frattolillo, F. A Watermarking Protocol Based on Blockchain. *Applied Sciences (Switzerland)* **2020**, *10*, 1–18, doi:10.3390/APP10217746.
54. Frattolillo, F. A Multiparty Watermarking Protocol for Cloud Environments. *Journal of Information Security and Applications* **2019**, *47*, 246–257, doi:10.1016/J.JISA.2019.05.011.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

55. Harika, D.; Noorullah, S. Implementation of Image Authentication Using Digital Watermarking with Biometric. *International Journal of Engineering Technology and Management Sciences* **2023**, *7*, 154–167, doi:10.46647/IJETMS.2023.V07I01.023.
56. Kelkoul, H.; Zaz, Y.; Mantoro, T. Countering Audiovisual Content Piracy: A Hybrid Watermarking and Fingerprinting Technology. *7th International Conference on Computing, Engineering and Design, ICCED 2021* **2021**, doi:10.1109/ICCED53389.2021.9664855.
57. Ren, N.; Wang, H.; Chen, Z.; Zhu, C.; Gu, J. A Multilevel Digital Watermarking Protocol for Vector Geographic Data Based on Blockchain. *Journal of Geovisualization and Spatial Analysis* **2023**, *7*, doi:10.1007/S41651-023-00162-0.
58. Liu, X.; Zhu, Y.; Sun, Z.; Diao, M.; Zhang, L. A Novel Robust Video Fingerprinting-Watermarking Hybrid Scheme Based on Visual Secret Sharing. *Multimed Tools Appl* **2015**, *74*, 9157–9174, doi:10.1007/S11042-014-2073-4/METRICS.
59. Wang, Cliff.; Gerdes, R.M..; Guan, Yong.; Kasera, S.Kumar. Digital Fingerprinting. **2016**, 189.
60. Yu, P.L.; Sadler, B.M.; Verma, G.; Baras, J.S. Fingerprinting by Design: Embedding and Authentication. *Digital Fingerprinting* **2016**, 69–88, doi:10.1007/978-1-4939-6601-1_5.
61. Ametefe, D.S.; Sarnin, S.S.; Ali, D.M.; Muhamad, W.N.W.; Ametefe, G.D.; John, D.; Aliu, A.A. Enhancing Fingerprint Authentication: A Systematic Review of Liveness Detection Methods Against Presentation Attacks. *Journal of The Institution of Engineers (India): Series B* **2024**, doi:10.1007/S40031-024-01066-3.
62. Ren, K.; Zheng, T.; Qin, Z.; Liu, X. Adversarial Attacks and Defenses in Deep Learning. *Engineering* **2020**, *6*, 346–360, doi:10.1016/J.ENG.2019.12.012.
63. Bai, T.; Luo, J.; Zhao, J.; Wen, B.; Wang, Q. Recent Advances in Adversarial Training for Adversarial Robustness. **2021**.
64. Gibert, D.; Zizzo, G.; Le, Q.; Planes, J. A Robust Defense against Adversarial Attacks on Deep Learning-Based Malware Detectors via (De)Randomized Smoothing. *IEEE Access* **2024**, *12*, 61152–61162, doi:10.1109/access.2024.3392391.
65. Ren, K.; Zheng, T.; Qin, Z.; Liu, X. Adversarial Attacks and Defenses in Deep Learning. *Engineering* **2020**, *6*, 346–360, doi:10.1016/J.ENG.2019.12.012.
66. Kong, Z.; Xue, J.; Wang, Y.; Huang, L.; Niu, Z.; Li, F. A Survey on Adversarial Attack in the Age of Artificial Intelligence. *Wirel Commun Mob Comput* **2021**, *2021*, doi:10.1155/2021/4907754.
67. Salehin, I.; Kang, D.K. A Review on Dropout Regularization Approaches for Deep Neural Networks within the Scholarly Domain. *Electronics 2023, Vol. 12, Page 3106* **2023**, *12*, 3106, doi:10.3390/ELECTRONICS12143106.
68. Jedrzejewski, F.V.; Thode, L.; Fischbach, J.; Gorschek, T.; Mendez, D.; Lavesson, N. Adversarial Machine Learning in Industry: A Systematic Literature Review. *Comput Secur* **2024**, *145*, 103988, doi:10.1016/J.COSE.2024.103988.
69. Paria Sarzaeim; Aarya Maturpalsingh Doshi View of A Framework for Detecting AI-Generated Text in Research Publications. *Internation Conference of Advanced Technologies* **2023**.
70. Wang, Z.; Liu, Y.; He, D.; Chan, S. Intrusion Detection Methods Based on Integrated Deep Learning Model. *Comput Secur* **2021**, *103*, doi:10.1016/J.COSE.2021.102177.
71. Hinton, G.E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R.R. Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors. **2012**.
72. Meroni, G.; Comuzzi, M.; Köpke, J. Editorial: Blockchain for Trusted Information Systems. *Frontiers in Blockchain* **2023**, *6*, 1235704, doi:10.3389/FBLOC.2023.1235704/BIBTEX.
73. Curmi, A.; Inguanez, F. BlockChain Based Certificate Verification Platform. *Lecture Notes in Business Information Processing* **2019**, *339*, 211–216, doi:10.1007/978-3-030-04849-5_18.
74. Malik, G.; Parasrampuria, K.; Reddy, S.P.; Shah, S. Blockchain Based Identity Verification Model. *Proceedings - International Conference on Vision Towards Emerging Trends in Communication and Networking, ViTECoN 2019* **2019**, doi:10.1109/VITECON.2019.8899569.
75. Adere, E.M. Blockchain in Healthcare and IoT: A Systematic Literature Review. *Array* **2022**, *14*, doi:10.1016/J.ARRAY.2022.100139.
76. Morar, C.D.; Popescu, D.E. A Survey of Blockchain Applicability, Challenges, and Key Threats. *Computers 2024, Vol. 13, Page 223* **2024**, *13*, 223, doi:10.3390/COMPUTERS13090223.
77. Zheng, X.; Zhang, C.; Woodland, P.C. Adapting GPT, GPT-2 and BERT Language Models for Speech Recognition. *2021 IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021 - Proceedings* **2021**, 162–168, doi:10.1109/ASRU51503.2021.9688232.
78. Hang, C.N.; Yu, P.D.; Morabito, R.; Tan, C.W. Large Language Models Meet Next-Generation Networking Technologies: A Review. *Future Internet* **2024**, *16*, doi:10.3390/FI16100365.
79. Gao, M. The Advance of GPTs and Language Model in Cyber Security. *Highlights in Science, Engineering and Technology* **2023**, *57*, 195–202, doi:10.54097/HSET.V57I.10001.
80. Rehana, H.; Çam, N.B.; Basmaci, M.; Zheng, J.; Jemiyo, C.; He, Y.; Özgür, A.; Hur, J. Evaluation of GPT and BERT-Based Models on Identifying Protein-Protein Interactions in Biomedical Text. **2023**.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2024 doi:10.20944/preprints202411.1219.v1

81. Grishina, A.; Kyrychenko, R. GPT-3 vs. BERT - Which Is Best? This Article Compares Both in Depth Available online: https://softteco.com/blog/bert-vs-chatgpt?WPACRandom=1731316845796 (accessed on 11 November 2024).
82. Clark, K.; Luong, M.T.; Le, Q. V.; Manning, C.D. ELECTRA: Pre-Training Text Encoders as Discriminators Rather Than Generators. *8th International Conference on Learning Representations, ICLR 2020* **2020**.
83. Wang, Y.; Wang, W.; Joty, S.; Hoi, S.C.H. CodeT5: Identifier-Aware Unified Pre-Trained Encoder-Decoder Models for Code Understanding and Generation. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings* **2021**, 8696–8708, doi:10.18653/V1/2021.EMNLP-MAIN.685.
84. Desai, S.; Durrett, G. Calibration of Pre-Trained Transformers. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* **2020**, 295–302, doi:10.18653/V1/2020.EMNLP-MAIN.21.
85. Vora, et al. V. A Multimodal Approach for Detecting AI Generated Content Using BERT and CNN. *International Journal on Recent and Innovation Trends in Computing and Communication* **2023**, *11*, 691–701, doi:10.17762/IJRITCC.V11I9.8861.
86. Zhang, Y.; Chen, D.Z. GPT4MIA: Utilizing Generative Pre-Trained Transformer (GPT-3) as A Plug-and-Play Transductive Model for Medical Image Analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **2023**, *14393*, 151–160, doi:10.1007/978-3-031-47401-9_15.
87. Mohseni, S.; Ragan, E. Combating Fake News with Interpretable News Feed Algorithms. **2018**.
88. Seddari, N.; Derhab, A.; Belaoued, M.; Halboob, W.; Al-Muhtadi, J.; Bouras, A. A Hybrid Linguistic and Knowledge-Based Analysis Approach for Fake News Detection on Social Media. *IEEE Access* **2022**, *10*, 62097–62109, doi:10.1109/ACCESS.2022.3181184.
89. Epstein, Z.; Foppiani, N.; Hilgard, S.; Sharma, S.; Glassman, E.; Rand, D. Do Explanations Increase the Effectiveness of AI-Crowd Generated Fake News Warnings? *Proceedings of the International AAAI Conference on Web and Social Media* **2022**, *16*, 183–193, doi:10.1609/ICWSM.V16I1.19283.
90. Chen, C.; Fu, J.; Lyu, L. A Pathway Towards Responsible AI Generated Content. *IJCAI International Joint Conference on Artificial Intelligence* **2023**, *2023-August*, 7033–7038, doi:10.24963/ijcai.2023/803.
91. Yang, X.; Pan, L.; Zhao, X.; Chen, H.; Petzold, L.; Wang, W.Y.; Cheng, W. A Survey on Detection of LLMs-Generated Content. **2023**.
92. Mao, H.; Nie, T.; Sun, H.; Shen, D.; Yu, G. A Survey on Cross-Chain Technology: Challenges, Development, and Prospect. *IEEE Access* **2023**, *11*, 45527–45546, doi:10.1109/ACCESS.2022.3228535.
93. Koulu, R.; Hirvonen, H.; Sankari, S.; Heikkinen, T. Artificial Intelligence and the Law: Can and Should We Regulate AI Systems? *SSRN Electronic Journal* **2023**, doi:10.2139/SSRN.4256539.
94. Kunda, I.; Kunda, I. Regulating the Use of Generative AI in Academic Research and Publications. *PUBMET* **2023**, doi:10.15291/PUBMET.4274.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.