POLITECNICO DI TORINO

Master degree course in Cybersecurity

Master Degree Thesis

Deepfake and Generative AI: Legal Challenges and Technical Strategies for Detection and Prevention

Supervisor
prof. Andrea Atzeni
prof. Giuseppe Vaciago

Candidate
Giada Grillo

Academic Year 2024 - 2025

Summary

Introduction

The rapid evolution of deepfake technologies has introduced one of the most pressing challenges in today's digital environment. Originally developed as research experiments with generative models such as GANs, autoencoders, and diffusion models, deepfakes have now become widely accessible tools capable of producing synthetic images, videos, and audio that are nearly indistinguishable from authentic content. While such technologies can serve creative and commercial purposes, they are increasingly misused for fraud, identity theft, extortion, misinformation, and reputational damage, generating severe personal, financial, and political risks. Recent cases confirm the urgency of this issue: deepfake fraud attempts surged by 3,000 percent in 2023, with losses often exceeding hundreds of thousands of dollars per incident. Despite the scale of the problem, current legal and technical responses remain inadequate. Regulatory frameworks are fragmented and often ambiguous, while detection methods, though advanced, are inherently reactive and struggle to keep pace with the sophistication of generative models. This highlights the necessity of shifting focus from detection to proactive authentication strategies that can verify the integrity of digital content at the point of its creation.

This thesis aims to contribute to this preventive approach by developing a lightweight and verifiable authentication mechanism for digital images. By embedding fragile watermarks capable of detecting even minimal alterations, the proposed method seeks to provide proof of integrity, ensuring that manipulated media can be identified before it spreads. Beyond the technical contribution, the research also underscores the importance of accountability within digital ecosystems, advocating for shared responsibility between creators, platforms, and regulatory bodies. Ultimately, the work aspires to strengthen the trustworthiness of digital content and support the broader effort to mitigate the risks posed by deepfake technologies.

Project overview

This thesis presents the design and implementation of a prototype system that integrates digital watermarking, cryptographic hashing, and blockchain technology to ensure the authenticity, integrity, and traceability of digital images, particularly within forensic and evidentiary contexts. The project aims to address one of the major challenges in digital forensics, the preservation and immutability of digital evidence, by leveraging blockchain's tamper-evident and append-only structure to maintain an auditable chain of custody over time. The proposed system combines fragile watermarking with blockchain-based immutability to create a verifiable framework for detecting tampering and preserving evidentiary reliability. While hashing provides a standard method for generating unique digital fingerprints of image data, the innovation lies in the use of blockchain as a decentralized ledger for securely storing these identifiers. This approach ensures that once an image and its metadata are registered, any alteration or removal becomes detectable, thus reinforcing the trustworthiness and non-repudiation of stored evidence.

The system, implemented in Python, is modular and designed to reflect the main stages of the forensic authentication pipeline. The watermarking module employs a fragile Least Significant Bit (LSB) embedding technique, encoding a textual watermark that is visually imperceptible but sensitive to pixel modifications. The hashing mechanism (SHA-256) complements this by producing unique identifiers for both image content and metadata, which are then recorded in a simulated

blockchain ledger that links each entry cryptographically to the previous one, reproducing the core immutability property of real blockchain systems. Verification procedures integrate these components to provide a comprehensive integrity assessment. The system recalculates hashes and validates them against blockchain records, while also verifying watermark integrity. Any inconsistency signals a potential manipulation, enabling a fine-grained localization of tampering at the pixel or metadata level. To evaluate real-world applicability, the framework was tested on data from the NIST Computer Forensic Reference Data Sets (CFReDS). Using the Autopsy forensic platform, a selected image from the Data Leakage Case dataset was analyzed through the entire workflow: watermarking, hashing, blockchain registration, and verification. The results confirmed the system's capability to detect tampering, preserve traceability, and maintain evidentiary integrity in realistic forensic conditions. Finally, the project integrates a legal perspective to assess how technical mechanisms like blockchain and watermarking can strengthen the evidentiary value of digital content, supporting accountability and compliance within judicial and investigative processes. By aligning technological immutability with legal requirements for authenticity and chain of custody, this work contributes to bridging the gap between forensic technology and legal reliability in digital evidence management.

Overall, the developed prototype provides a proof-of-concept for using blockchain-based immutability, combined with digital watermarking and cryptographic hashing, to enhance trust, transparency, and traceability in digital forensics.

Results and Conclusions

The developed scripts demonstrated significant effectiveness in detecting and authenticating manipulated images, particularly in the context of deepfake content. Through extensive testing on both online images and forensic datasets from CFReDS, the system consistently identified anomalies and distinguished authentic images from tampered ones. The dual-layer approach, combining fragile watermarking and SHA-256 hashing, proved crucial for detecting alterations not only to the image itself but also to embedded metadata, thereby enhancing the robustness of the verification process.

The blockchain-inspired component, simulated through a CSV-based ledger, further reinforced data integrity by maintaining cryptographic links between consecutive entries. This mechanism enabled the detection of unauthorized modifications or deletions, demonstrating the system's ability to preserve the authenticity of recorded information. Practical simulations included three types of tampering: direct image alterations, watermark manipulations, and metadata changes, all of which were successfully detected by the tool.

These results highlight the potential of integrating watermarking, hashing, and blockchain mechanisms to address the growing challenges of multimedia verification and deepfake detection. The system bridges theoretical concepts and practical applications, offering a proof of concept for digital forensic use and legal contexts. By enabling verification at multiple levels (image, watermark, and metadata) the project contributes to safeguarding digital evidence integrity, intellectual property, and trust in information systems. Future developments could include transitioning from a simulated to a real blockchain infrastructure to ensure stronger guarantees of immutability, designing a user-friendly interface for forensic practitioners, and testing the tool with real forensic case materials. Such enhancements would increase the system's applicability in operational environments, allowing it to support investigators, legal authorities, and courts in verifying digital content. Ultimately, this work establishes a solid foundation for practical deepfake detection solutions, combining technical rigor with societal and legal relevance in an increasingly digital world

Contents

Chapter 1

Introduction

1.1 Motivations of the study

The rapid development of deepfake technologies represents one of the most significant and complex challenges in the contemporary digital landscape. Deepfakes, originally conceived as experimental outputs of generative models such as Generative Adversarial Networks (GANs), autoencoders, and diffusion models, have now become widely accessible tools capable of producing synthetic content with astonishing realism. These technologies can manipulate images, videos, and audio recordings in ways that are increasingly indistinguishable from authentic media, even to trained experts. What was once a domain restricted to researchers and visual effects professionals has now expanded into widespread public use, with applications spanning entertainment, advertising, politics, and disinformation.

The growing realism of deepfakes poses a serious and multidimensional threat to the reliability of digital content. As generative algorithms evolve, they can replicate facial expressions, lip movements, vocal intonations, and body gestures with extreme precision. This level of sophistication makes it virtually impossible for an average user to distinguish between real and manipulated content, creating fertile ground for deception. Malicious actors can exploit deepfake technologies to impersonate individuals, fabricate statements, falsify evidence, or orchestrate highly convincing social engineering attacks. Consequently, deepfakes have become powerful instruments for fraud, identity theft, extortion, and reputational damage, frequently resulting in severe personal, financial, or political consequences.

Recent statistics confirm the scale and urgency of this problem. According to data published by Eftsure, deepfake fraud attempts surged by 3,000 percent in 2023, revealing a dramatic escalation in the frequency and sophistication of such attacks [1]. In one of the most striking cases reported, an employee was tricked into transferring 25 million dollars after being deceived into believing they were on a video call with their company's CFO, a call that was, in fact, entirely synthetic. Furthermore, average losses per incident are now estimated to be close to 480,000 dollars, with large organizations reporting even greater damage, sometimes exceeding 680,000 dollars. These figures highlight how the financial and operational risks associated with deepfakes are no longer hypothetical, but increasingly tangible and systemic.

Despite the growing awareness of the dangers posed by deepfakes, regulatory frameworks remain largely inadequate. Neither Italian law nor broader international legal systems, including those of the United States or China, have yet formulated a comprehensive and enforceable strategy to address the creation, dissemination, and accountability of deepfake content. Existing legal instruments are often limited by ambiguities in the definition of synthetic media, inconsistencies in criminalization criteria, and jurisdictional fragmentation. While some legislative initiatives aim to introduce transparency requirements or mandatory content labeling, they often include broad exceptions for artistic, journalistic, or satirical content, which can be exploited to bypass regulatory controls.

Equally urgent is the issue of accountability. In today's digital ecosystem, responsibility for the circulation of manipulated content is frequently placed solely on the content creator, thereby

overlooking the critical role played by digital platforms. As primary gatekeepers of online information, these platforms must be involved not only in the detection and removal of deepfakes but also in implementing mechanisms that prevent their spread in the first place. This implies a need to reconsider both their legal obligations and their technical infrastructure, particularly with respect to content moderation, traceability, and systemic transparency.

The current regulatory and technical gaps underscore the importance of enhancing both detection and authentication frameworks. Although detection techniques, particularly those based on deep learning and statistical analysis, have achieved remarkable accuracy, sometimes exceeding 99 percent, they remain fundamentally reactive and vulnerable to adversarial evasion. As generative models continue to improve, detection systems struggle to keep pace. Therefore, it is essential to complement detection with proactive solutions that can verify the authenticity of digital content at the moment of its creation or publication. Authentication mechanisms, such as digital watermarking or cryptographic signatures, can provide verifiable proof that a specific image or video has not been tampered with, shifting the burden of verification from the end user to the point of origin.

In light of these considerations, this thesis focuses on the development and implementation of robust authentication strategies for detecting deepfake images, with the broader aim of contributing to the ongoing discussion on how to preserve the integrity of digital content and reinforce accountability across the entire information supply chain.

1.2 Research objectives

This thesis aims to contribute to the ongoing efforts in mitigating the threats posed by deepfake technologies by addressing a dimension that is still underexplored in comparison to detection: authentication. While deepfake detection techniques have advanced significantly in recent years, achieving high accuracy rates through machine learning and computer vision, they are inherently reactive. These approaches typically analyze already circulating content to assess its authenticity, leaving a critical gap in the prevention and early verification of manipulated media.

The objective of this research is to propose and develop an effective, lightweight, and verifiable method for the authentication of visual content, particularly images that may be subject to deepfake manipulation. The focus shifts from detecting malicious alterations after their dissemination to proactively certifying the originality of content at the point of its creation. This preventive approach is essential to limiting the potential harm caused by synthetic media before it can spread online and produce damaging effects.

By embedding a fragile watermark that can detect even minimal changes in the image data, the proposed method seeks to provide a proof-of-integrity mechanism that allows recipients, whether human or automated systems, to verify whether an image has been modified since it was authenticated. The ultimate goal is to reduce the likelihood that manipulated images will be accepted as authentic, thereby mitigating the risks associated with their spread, such as reputational damage, fraud, identity theft, misinformation, and erosion of public trust.

In addition to the technical objectives, the thesis also aims to emphasize the importance of authentication mechanisms as a structural component of future digital ecosystems. In particular, it seeks to promote a paradigm in which the responsibility for ensuring content integrity does not rest solely on post-publication detection but is shared through proactive technical measures and platform accountability. This perspective is especially relevant in an era where the speed and scale of content distribution often outpace both legal regulation and manual verification, calling for embedded safeguards that can accompany media from its origin throughout its lifecycle.

By exploring this authentication-oriented approach to the deepfake challenge, the thesis hopes to offer a modest yet meaningful contribution to the broader discourse on content reliability, security, and ethical technology design.

1.3 Methodology

1.4 Thesis structure

This thesis is structured into six chapters, each of which contributes to the investigation and development of an authentication-based approach to mitigate the risks posed by deepfake technologies, with a specific focus on image manipulation.

Chapter 1 introduces the topic, outlining the motivations behind the study, the main research objectives, the adopted methodology, and the overall structure of the work. It highlights the increasing sophistication and widespread diffusion of deepfake technologies, the difficulties faced by users in identifying manipulated content, and the current limitations of both technical countermeasures and regulatory frameworks. It also emphasizes the importance of shifting the focus from detection to prevention through authentication.

Chapter 2 provides the technical and legal background essential for understanding the deepfake phenomenon. It begins by exploring the evolution of generative artificial intelligence, with particular attention to models such as GANs, autoencoders, and diffusion models, as well as their most relevant applications and challenges. The chapter then focuses on deepfakes specifically, discussing the underlying technologies used to generate synthetic visual content, the main types of manipulations (face swapping, lip syncing, and puppet master techniques), and the tools that enable their creation. It also addresses the practical uses of deepfakes and the societal and ethical concerns they raise. The second part of the chapter provides a comparative overview of current regulatory approaches to deepfakes at the global level. It examines the emerging yet still insufficient frameworks in Italy, particularly in relation to the European Union's AI Act and its current limitations, the centralized, surveillance-driven strategy adopted by China, and the fragmented, state-based legislative landscape in the United States. The chapter concludes by reflecting on the still unresolved issue of accountability, especially regarding the responsibility of online platforms in the dissemination and amplification of deepfake content.

Chapter 3 presents the current state of the art in the field of deepfake detection and authentication. It begins by analyzing existing detection techniques, including those based on machine learning, deep learning, and statistical analysis, and evaluates their strengths and vulnerabilities. The chapter then shifts to authentication strategies, illustrating emerging approaches such as the use of blockchain for content verification and digital watermarking for integrity assurance. In addition, it highlights the importance of user awareness as a complementary measure in combating the spread of manipulated media, stressing the need for effective educational and technological tools that can support users in identifying and responding to deepfake content. The chapter also addresses the role of organizations and companies, outlining a set of best practices that can be adopted to counteract the threat of deepfakes.

Chapter 4 details the design and implementation of the proposed authentication-based solution, structured as a case study. It describes each step of the process, from the selection and preparation of forensic datasets to the development of a tool capable of verifying the integrity and authenticity of digital images using cryptographic hashing and blockchain technology. The chapter includes the simulation of a real-world data leakage scenario, based on a publicly available dataset from the NIST CFReDS portal, in which a sensitive image is exfiltrated, modified, and illegally redistributed. This simulation demonstrates how image authentication can support forensic investigations and protect intellectual property rights by verifying whether a given image has been previously registered and whether it has been tampered with. Furthermore, the chapter explores and critically discusses the main alternative approaches that could be employed to address the problem of image authentication and manipulation detection. Several technical strategies are presented and compared in terms of feasibility, scalability, and forensic soundness. Based on this analysis, the implemented direction is selected and justified in light of the project's goals and constraints.

Chapter 5 presents a discussion of the experimental results obtained during the case study. It analyzes the effectiveness of the implemented approach in detecting image manipulation and verifying authenticity, evaluates the accuracy and reliability of hash-based checks in various manipulation scenarios, and assesses the practical implications of using blockchain as a verification

tool. In addition, the chapter provides a critical reflection on the limitations of the current prototype and explores possible improvements, such as enhancing robustness against adversarial modifications or extending the solution to different file types. Potential applications of the tool in real-world scenarios are also discussed, including its use in digital forensics investigations, journalistic content verification, and intellectual property protection frameworks.

Chapter 6 concludes the thesis by summarizing the main findings and reflecting on the broader implications of authentication-based strategies in the fight against deepfake technologies. It reiterates the importance of shifting from reactive to preventive measures and highlights the potential of forensic methods, such as those implemented in this work, to support legal investigations, promote accountability, and enhance user trust in digital media.

Chapter 2

Technical background and legal framework

2.1 Generative AI

2.1.1 What is Generative AI

Generative Artificial Intelligence, also known as Generative AI, is a field of artificial intelligence capable of creating original content in response to the user's input. The output can be any type of data, such as text, images, audio, or even software code. What distinguishes it from classical AI is precisely this ability to generate something new, which in some way mimics human creativity in an effective manner.

2.1.2 The process

Generative AI essentially follows a structured process made up of three main phases: training, tuning and ongoing refinement.

The training process

The first step is the training phase, characterized by the creation of a foundational model, a deep learning model that will serve as the main structure for several applications; Large Language Models (LLMs) are the most common models used for text generation, but other types of model exist for other purposes like image, audio, or video generation. The training process requires feeding the algorithm with vast amounts of raw, unstructured, and unlabeled data, often gathered from extensive sources such as the internet. During this phase, the model faces several predictive exercises, consisting of anticipating the next word in a sentence, the subsequent pixel in an image, or the following command in a line of code. By repeatedly adjusting its internal parameters to minimize the gap between its predictions and the actual data, the model gradually learns the underlying patterns and structures within the dataset [2].

The tuning phase

Once the foundation model is established, it often requires further refinement to enhance its performance in specific applications. The tuning process can be achieved by following the fine tuning or reinforcement learning with human feedback (RLHF).

Fine tuning involves training the model on labeled data specifically designed for the target application, ensuring that it can generate accurate and context-appropriate responses. For example, when developing a medical diagnosis assistant, the model would be trained using datasets

containing patient symptoms, potential diagnoses, and recommended treatments. Although this method is highly effective, it can be time-consuming and often requires extensive data labeling efforts carried out by experts in the respective field.

Reinforcement learning with human feedback (RLHF) enhances the performance of the model through direct user feedback: human assess the output of the model by scoring its accuracy and relevance and / or providing corrections to improve future responses; the model then integrates these evaluations and refines its internal parameters, improving overall quality.

Ongoing refinement

The refinement process does not end with tuning: generative AI applications undergo continuous assessment and improvement with the aim of ensuring their outputs remain accurate and relevant. To further enhance performance, developers may adopt a technique known as retrieval-augmented generation (RAG). This method extends the model's capabilities by allowing it to access external data sources beyond its original training set. By integrating updated information, RAG ensures that generative AI applications remain accurate and aligned with current knowledge, while also enhancing transparency by providing clear references to the supplementary data used.

2.1.3 Generative models

A generative model is a machine learning model designed to produce new, original data that closely resembles the input data it was trained on. These models are provided with large amounts of unlabeled data, which they process independently to identify patterns and distributions. This allows them to develop the internal logic needed to generate new data. During training, the model uses a loss function to assess the difference between its predictions and real-world outcomes, aiming to minimize this loss and make the generated outputs as realistic as possible. There are various types of generative models, each with its own architecture; in this section I will focus on the most relevant ones.

Autoregressive models

Autoregressive models are a class of machine learning techniques used to predict the next item in a sequence based on previous items. They analyze the relationships between data points in a sequence to identify probabilistic correlations, which are then applied to predict the most likely next component. These models were originally mainly implemented using recurrent neural networks (RNNs) but nowadays transformer models have become the standard due to their enhanced capabilities.

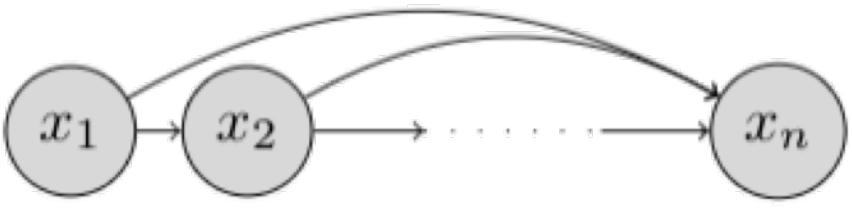


Figure 2.1. Autoregressive Model (fonte: deepgenerativemodels).

The **transformer model** is a deep learning architecture introduced in the paper "Attention is All You Need" (Vaswani) in 2017 [3]. The key innovation behind Transformers is the self-attention mechanism, which allows the model to weigh the importance of different words (or tokens) in a sequence relative to each other, regardless of their positions. Moreover, another relevant feature is the parallel processing, which allows them to process all items in a sequence simultaneously,

improving efficiency. The architecture of transformers consist of two main parts: the encoder and the decoder. The encored processes the input sequence, it uses multiple layers of self attention and feedforward neural networks to create a rich representation of the input. The decoder is used for generating the output; it uses self-attention and feedforward networks and additionally incorporates attention mechanisms to focus on relevant parts of the encoder's output.

Diffusion models

Diffusion models work by gradually obfuscating input data through the act of adding noise and then learning how to gradually reverse the diffusion process to generate new samples [4]. They operate in 3 main phases: diffusion, learning and reverse diffusion. During diffusion, the model gradually introduces noise to the input data until it is no longer recognizable. The model adds a small amount of Gaussian noise to the data at each step in a mathematical process known as a Markov chain. During the learning phase, the model traces the evolution of the now-destroyed data to understand how it was altered through the noising process. Finally, the reverse diffusion: by understanding how noise alters data, the diffusion model learns to reverse the noising process and reconstruct the input data. The goal is to go backward through the Markov chain, removing Gaussian noise until only pure data is left. These models are mainly used for image generation but their applications or use cases are also 3D modeling, anomaly detection or market research [5].

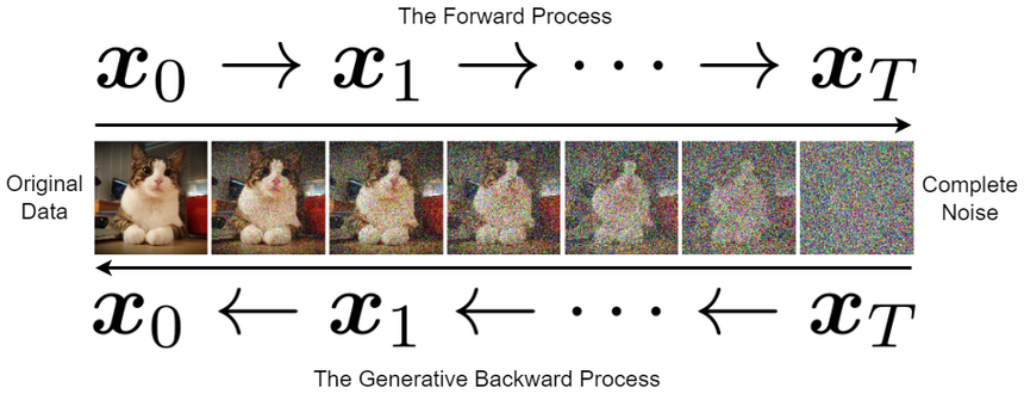


Figure 2.2. Diffusion Model (fonte: ResearchGate).

Generative adversarial networks (GANs)

Generative Adversarial Networks (GANs) are a class of deep learning models introduced by Ian Goodfellow and his collaborators in 2014 [6]. GANs consist of two neural networks, a generator and a discriminator, trained simultaneously in a competitive process with the aim to generate synthetic data indistinguishable from real data [4]. The generator takes random noise (usually from a uniform distribution) as input and transforms it into synthetic data with the goal of creating outputs that resemble the real data distribution. Initially, the generated data may be noisy or unrealistic, but, through training, the generator improves its output to the point where it produces highly convincing data. The discriminator has the task to distinguish between real data (from the training dataset) and fake data (produced by the generator): it assigns a probability to each input with the goal of classifying the data as real or fake. The training process involves these two networks competing against each other until the generator produces data that are virtually indistinguishable from real data, and the discriminator can no longer reliably differentiate between them.

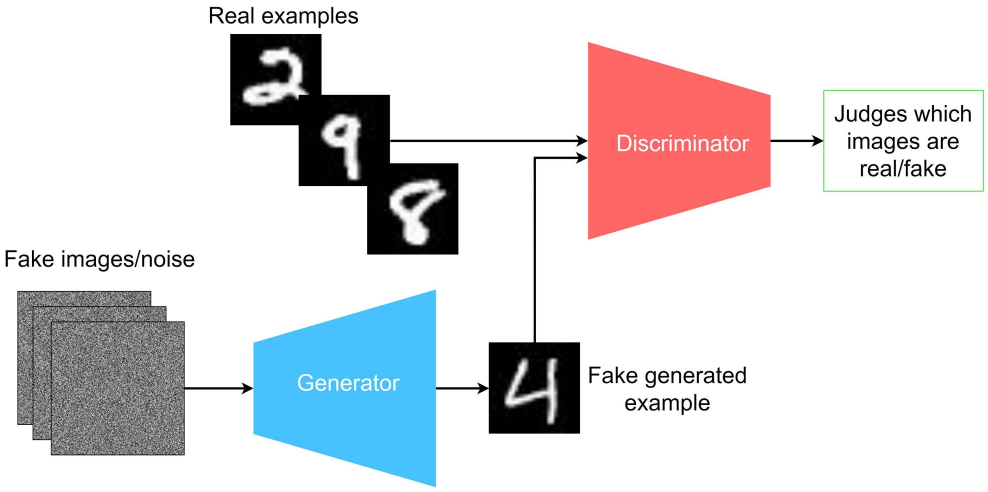


Figure 2.3. GAN (fonte: IBM).

2.1.4 Applications and tools

Generative AI has a wide array of applications that span across multiple industries and domains. This technology is transforming traditional workflows and enabling new capabilities in fields ranging from content creation to scientific research. Below are some of the key areas where generative AI is making a significant impact, accompanied by the most famous and interesting related tools

Content creation

Generative AI is particularly powerful in generating text, images, video, and audio. In content creation, AI models, especially transformer-based ones, are capable of writing coherent and contextually relevant text; this includes drafting articles, blogs, reports, and even creative writing. In particular, for content creation, the most famous and used tools are: Jasper [7], which excels in summarizing articles, writing reports and academic content; ChatGPT [8], focused on customer support, creative writing and brainstorming but sometimes producing generic responses since it lacks domain-specific accuracy; Claude [9], a conversational AI that excels in document summarization and knowledge retrieval. In the visual arts, tools like DALL-E [10], Midjourney [11], and Stable Diffusion [12] generate realistic images or unique art pieces from textual prompts, and can perform tasks such as image editing and enhancement. Lastly, for music and audio, tools like Synthesia [13] excels in creating professional-quality video presentations with virtual avatars.

Coding and development

Generative AI models are transforming the software development process by automating the generation of code. These models can autocomplete code, translate between programming languages, and even suggest improvements. For developers, this means faster prototyping, efficient debugging, and an overall quicker development cycle. Additionally, AI tools assist in modernizing legacy applications by automating the repetitive and time-consuming tasks of updating code for hybrid cloud environments. The most valuable examples that worth to be mentioned are GitHub Copilot and AlphaCode.

GitHub Copilot enhances developer productivity with real time coding suggestions, it supports multiple languages but it may generate incorrect or insecure code and requires developer oversight [14, 15].

AlphaCode is designed for code generation, bug fixing and learning new programming languages; it automates coding, optimizes solutions and provides debugging support [16].

Science and engineering

Generative AI is playing an increasingly central role in scientific research and innovation. In the pharmaceutical field, for example, it is being used to generate molecular structures with specific desired properties, significantly contributing to drug discovery and the development of innovative pharmaceutical compounds.

A concrete example is the **GENTRL** (Generative Tensorial Reinforcement Learning) model, which is capable of designing molecules that interact precisely with specific biological mechanisms. This makes it especially useful for developing targeted and personalized treatments, particularly in the case of complex or rare diseases [17].

Another important area where Gen AI is making a strong impact is synthetic data generation. This is essential for training and testing AI models, especially when real-world data is limited or protected by privacy regulations. Technologies such as GANs (Generative Adversarial Networks), VAEs (Variational Autoencoders), and hybrid models like CorGAN, which combines convolutional GANs with autoencoders, are being used to generate highly realistic synthetic electronic health records and medical images, while ensuring compliance with data protection laws [17].

Synthetic data is also proving especially valuable in medical image analysis, where it is used to improve diagnostic accuracy. For instance, synthetic chest X-ray images generated using latent diffusion models have been shown to enhance the performance of classification algorithms, offering practical support for early diagnosis [17].

Anomaly detection and security

Generative AI is also being increasingly applied in cybersecurity for anomaly detection, offering capabilities that go far beyond traditional rule-based systems. By learning the statistical and structural characteristics of normal behavior within datasets, such as network traffic, user activity, or system logs, generative models can identify subtle deviations that might indicate the presence of a threat. What distinguishes generative models, such as Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs), is their capacity to not only detect anomalies but to simulate them as well. This dual function is especially valuable in threat anticipation and in training more resilient detection systems. As highlighted in the study by Blake (2024) [18], generative AI has been effectively used to simulate adversarial attack scenarios in order to stress-test existing intrusion detection frameworks. For instance, a leading cybersecurity firm implemented GANs to create a wide range of synthetic evasion attacks, which helped uncover vulnerabilities in their AI-driven security system. This proactive approach led to a measurable reduction in false negatives and significantly improved the system's responsiveness to novel attack vectors. In this context, generative AI operates not merely as a detection tool but as a strategic component in the broader lifecycle of cyber defense, enabling systems to adapt to dynamic and previously unseen threat patterns. Moreover, the synthetic data generated by these models plays a crucial role in enhancing the training of supervised anomaly detection algorithms. Since real-world malicious data can be limited, especially for rare or emerging threats, the ability to generate realistic yet diverse examples of abnormal behavior enriches training datasets and improves model generalization. This is particularly evident in applications such as phishing detection or fraud prevention, where generative models can simulate email or transaction patterns that closely resemble genuine attacks. The study also presents the case of a deepfake phishing campaign powered by GANs, where attackers generated a highly realistic voice message impersonating a CEO to deceive employees into transferring funds. While this example highlights the risks associated with generative AI in the wrong hands, it simultaneously underscores the necessity of deploying such tools defensively, anticipating the techniques that adversaries are likely to use. In summary, generative AI not only enhances anomaly detection by identifying deviations from established norms, but also enables the simulation of realistic threats, allowing security systems to learn and evolve in parallel with the adversarial landscape.

2.1.5 Challenges

Generative AI has made significant advancements, offering numerous opportunities across various domains. However, it also presents several challenges and risks that must be addressed for its responsible and effective use. Below are some of the main issues and the approaches being explored to mitigate them [5, 19].

Inconsistent output

Generative AI models often produce variable results even when the same input is provided, which can be problematic in applications like customer service chatbots where consistency is crucial. To manage this, users can employ prompt engineering, refining their inputs to achieve more predictable and desired outputs.

Bias

Generative AI models can inherit biases from the data used to train them, which may include societal biases from labeled data, external sources, or human feedback. This can lead to the generation of biased, unfair, or even offensive content. Developers are addressing this issue by using diverse and representative datasets, setting guidelines to minimize bias during model training, and continually monitoring the outputs to ensure fairness and accuracy [20].

Lack of explainability

Another challenge is the lack of transparency in many generative AI models, which function as "black boxes". This makes it difficult to understand how the models arrive at their decisions. Even the developers of these models may not fully comprehend the underlying decision-making process. To overcome this, researchers are working on Explainable AI (XAI) techniques that aim to improve transparency, allowing users and developers to better understand how the models work and fostering trust in their outputs [20].

Evaluation difficulties

Assessing the quality of AI-generated content can be complex, as traditional evaluation metrics may not fully capture subjective factors like creativity, relevance, or context. As a result, researchers are focused on developing more robust and nuanced evaluation methods to better measure the value and quality of generative AI outputs [21].

Model Collapse

Model collapse is an emerging phenomenon affecting generative AI models, particularly large language models (LLMs), when they are trained increasingly on synthetic data generated by other models rather than on original, human-created content. This recursive reliance leads to a progressive degradation of the model's ability to produce accurate, diverse, and meaningful outputs. The issue develops in stages. In the early collapse phase, the model begins to lose representation of rare or low-frequency data, reducing its capacity to generate varied content. In the late collapse phase, accumulated errors from repeated training on synthetic data cause the model's outputs to drift significantly from the This degradation becomes more likely as AI-generated content becomes more To mitigate model collapse, researchers stress the importance of maintaining access to human, generated data, verifying data provenance, and using synthetic data as a complement, rather than a replacement, to real data. These practices help preserve the integrity and long-term performance of generative AI systems in the face of increasing reliance on synthetic content. [22]

Catastrophic forgetting

Catastrophic forgetting, also known as catastrophic interference, refers to the tendency of neural networks to lose previously acquired knowledge when trained sequentially on new tasks. This phenomenon arises because, during training, the model's parameters (weights) are adjusted to minimize a loss function based on new data. If these adjustments significantly alter parameters critical to earlier tasks, the model's performance on those tasks deteriorates. This issue is particularly pronounced in large models, such as large language models (LLMs), due to their extensive parameter spaces. The problem stems from the inherent plasticity of neural networks: while they can adapt to new information, this adaptability can lead to the overwriting of existing knowledge. This mirrors the stability-plasticity dilemma observed in biological systems, where the brain balances learning new information with retaining existing knowledge. To mitigate catastrophic forgetting, several strategies have been proposed: regularization techniques (methods like Elastic Weight Consolidation (EWC) add penalties to changes in parameters important for previous tasks, preserving prior knowledge); rehearsal Methods (these involve interleaving training on new tasks with examples from previous tasks, ensuring continuous reinforcement of earlier knowledge); memory-augmented Neural Networks (by incorporating external memory components, MANNs can store and retrieve information from past tasks, aiding in knowledge retention). Addressing catastrophic forgetting is crucial for developing AI systems capable of continual learning without compromising previously acquired skills. [23]

Security, privacy, and intellectual property risks

Generative AI has the potential to be misused in malicious ways, such as creating convincing phishing emails, fake identities, or other harmful content. These risks pose significant threats to security and privacy. Developers and users must be cautious when handling input data, ensuring that the generated content respects intellectual property rights and does not violate the rights of others [20].

Deepfakes

Deepfakes represent one of the most concerning applications of generative AI. They involve the creation or manipulation of images, videos, or audio to falsely depict individuals performing or saying things they never actually did. The consequences of deepfakes can be severe, ranging from reputational damage to facilitating fraudulent activities. Although detection technologies are improving, user education on verifying the authenticity of content remains crucial to minimizing their impact. A more comprehensive discussion on the subject of deepfakes and their broader implications will be explored in the following chapter.

2.2 Deepfakes

A deepfake is a synthetic media generated through deep learning techniques, particularly using Generative Adversarial Networks (GANs) or autoencoders, that manipulates or fabricates visual or audio content to create realistic yet false representations of individuals. The term "deepfake" originates from the combination of "deep learning" and "fake", emphasizing its roots in artificial intelligence. This technology in fact, leverages advanced neural networks trained on large datasets to analyze and replicate facial features, voice patterns, and behavioral traits [24]. While deepfakes have legitimate applications in entertainment, education and art, they also pose significant risks in the context of disinformation, identity theft, and digital impersonation. Due to their potential to deceive, deepfakes are a growing concern in both legal and cybersecurity domains, prompting the development of detection methods and regulatory frameworks.

2.2.1 The technologies behind deepfake

Deepfake technology relies on a combination of advanced artificial intelligence techniques and computational resources to create synthetic media that can be indistinguishable from real content. The technologies that form the foundation of deepfake generation are varied and highly specialized, each playing a crucial role in different stages of the process. Among the key technologies are Convolutional Neural Networks (CNNs), Autoencoders, Natural Language Processing (NLP), Generative Adversarial Networks (GANs), and Recurrent Neural Networks (RNNs). These tools work together to produce highly convincing and sophisticated deepfake content [24].

Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) is a type of neural network mainly designed to process and analyze visual data, excelling at image recognition, object detection, and facial feature analysis. The technology works by processing images through multiple layers of filters to detect increasingly abstract features like how the subject's face moves and expresses emotions over time. For instance, when training a model to generate a deepfake of a particular person, the CNN analyzes thousands of images to recognize specific features like the shape of the eyes, nose, and mouth, and how these features change when the person smiles, frowns, or moves their head. By doing so, the system can accurately map the target subject's face onto another body or generate realistic facial movements that match a given video or audio clip.

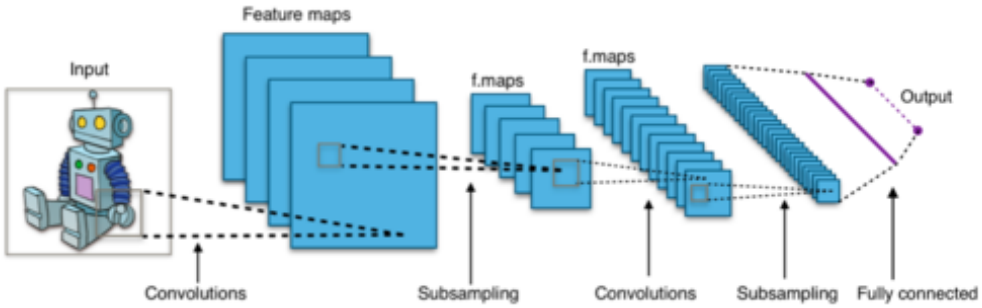


Figure 2.4. Convolutional Neural Network (fonte: Wikipedia).

Autoencoders

Autoencoder is a type of neural network made up of two main components, an encoder and a decoder: the encoder compresses the input data into a lower-dimensional space, while the decoder reconstructs the data from this compact representation. In deepfake generation, autoencoders are trained to learn the distinctive characteristics of a subject's face and body, so, once trained, the system can manipulate these features and apply them to new video frames. This allows the replacement of a target's face on another person's body with minimal distortion, so the AI can seamlessly transfer facial expressions, lip movements, and other features, making the deepfake appear more authentic. Essentially, autoencoders help the system "understand" the essential components of the subject's facial structure and movement, enabling it to reconstruct these features in novel contexts.

Natural Language Processing (NLP)

While visual manipulation is the focus of most deepfake technologies, Natural Language Processing (NLP) plays a fundamental role when the content involves audio or speech synthesis. NLP encompasses a broad spectrum of computational techniques that enable machines to process, interpret, and generate human language in a manner that is both syntactically and semantically

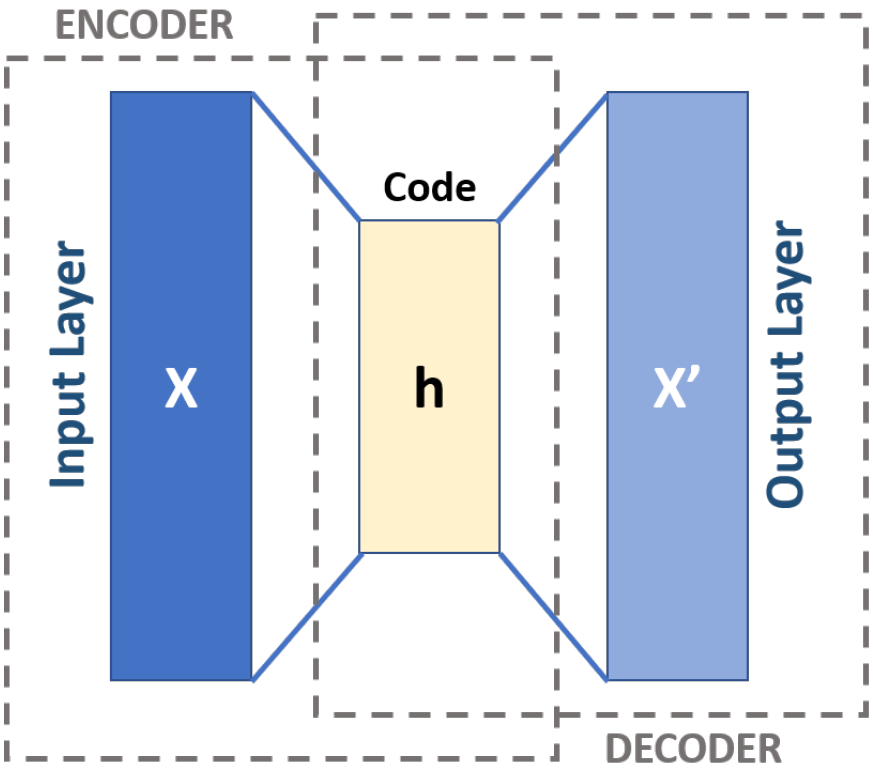


Figure 2.5. Autoencoder (fonte: Wikipedia).

coherent. In the context of audio deepfakes, these techniques are integrated with speech synthesis and voice cloning technologies to replicate the linguistic identity of a target individual with high fidelity.

The process begins with the collection and pre-processing of speech data, during which raw audio samples are transcribed and cleaned through methods such as tokenization, normalization, and the removal of noise or irrelevant content. This stage ensures that the textual and phonetic representations of speech are accurately aligned, forming a structured foundation for further analysis.

NLP models then perform in-depth linguistic analysis, including phoneme segmentation, part-of-speech tagging, and prosodic modeling. These operations allow the system to capture not only the lexical choices and syntactic structures characteristic of the target's speech, but also their unique vocal features such as intonation, pitch variation, speech rhythm, and stress patterns. Advanced embedding techniques, such as contextual word embeddings and acoustic feature encodings, are used to map these speech characteristics into high-dimensional vector spaces that preserve semantic nuance and temporal dependencies.

Once these patterns are learned, often through deep learning architectures like recurrent neural networks (RNNs), transformers, or generative adversarial networks (GANs) adapted for audio synthesis, the model becomes capable of generating new audio sequences that mimic the target speaker. These sequences are not merely replicas of previously recorded phrases; instead, the system constructs entirely novel sentences, shaped by probabilistic models that predict the most plausible phonetic and linguistic continuations based on the target's speaking style.

This process is inherently complex due to the multifaceted nature of human speech, which includes not only phonetic articulation but also emotional tone, conversational intent, and contextual variation. Capturing sarcasm, emphasis, hesitation, or sentiment requires models to go beyond surface-level replication and engage in deeper semantic understanding, often supported by natural language understanding (NLU) components within the NLP pipeline. Consequently, the

success of voice-based deepfakes depends heavily on the accuracy and granularity of NLP-driven modeling, which seeks to simulate not only how something is said, but also why it is said and what it conveys in context.

Through these technological advancements, NLP has become a central enabler of realistic audio deepfakes, pushing the boundaries of synthetic media generation while also raising significant ethical and security concerns related to misinformation, identity theft, and digital impersonation.[25]

Generative Adversarial Networks (GANs)

One of the most innovative and essential technologies behind deepfake systems is the Generative Adversarial Network (GAN). A GAN consists of two neural networks that work in opposition to each other: the generator and the discriminator. The generator creates fake content, such as images, videos, or audio, while the discriminator's role is to distinguish between real and generated content. These two networks are trained together in a process of competition, with the generator trying to improve its ability to create realistic content in order to fool the discriminator, and the discriminator improving its ability to detect fake content. The generator begins by producing rough, low-quality content, and the discriminator provides feedback, helping the generator improve. Over time, this iterative process leads to the generation of highly realistic deepfakes, as both networks continue to improve their performance. GANs are particularly powerful because they enable the creation of content that is not simply a copy of real-world data but is instead generated from learned patterns, making the fake content increasingly harder to distinguish from reality. In deepfake video generation, for example, the GAN allows the system to generate faces and body movements that appear natural, even though they are entirely synthetic. The constant interplay between the generator and the discriminator ensures that the output becomes more refined with each iteration, ultimately producing deepfakes that can be nearly indistinguishable from genuine media.

Recurrent Neural Networks (RNNs)

In addition to CNNs and GANs, Recurrent Neural Networks (RNNs) are also used in deepfake creation, particularly for tasks that involve sequences of data. RNNs are well-suited for analyzing temporal or sequential data, such as speech, video frames, or audio-visual synchronization. In deepfakes, RNNs are often employed to synchronize lip movements with generated speech or audio, a task that requires understanding how the movements of the mouth and facial expressions correspond to the sounds being produced.

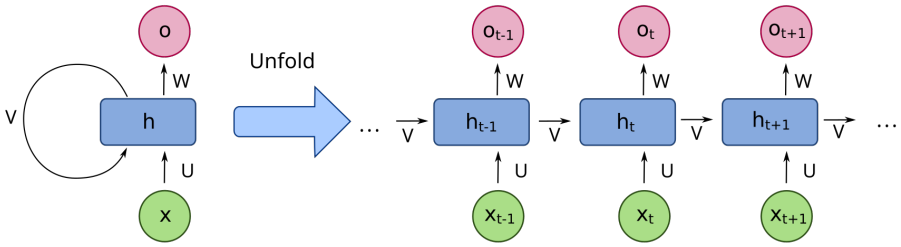


Figure 2.6. Recurrent Neural Network (fonte: Wikipedia).

High Performance Computing (HPC)

The creation of deepfakes necessitates substantial computational power due to the complexity of the underlying machine learning models and the massive volumes of data they must ingest and

process. Training deep learning models, particularly for high-resolution video and audio synthesis, involves intensive numerical computation and repeated iterations over extensive datasets. In this context, High Performance Computing (HPC) plays a critical role by enabling the parallel processing of tasks across multiple cores or nodes, significantly reducing the time required for training and inference.

HPC systems differ from traditional computing architectures in that they rely on parallel computing, where tasks are distributed and executed concurrently across a large number of processors or specialized hardware components, such as Graphics Processing Units (GPUs). These GPUs are especially well-suited for the linear algebra operations and matrix transformations that characterize deep learning workloads. Modern HPC clusters often consist of tens of thousands of interconnected nodes, each equipped with high-performance CPUs or GPUs, and optimized through fast, low-latency interconnects, high-throughput storage systems, and centralized schedulers to manage computational resources efficiently.

Given the scale and speed demanded by deepfake generation, many organizations leverage HPC resources through cloud computing platforms, a model often referred to as HPC-as-a-Service (HPCaaS). This approach provides scalable, on-demand access to advanced computing infrastructures, reducing the need for substantial upfront investment in physical hardware. The widespread availability of cloud-based HPC resources allows researchers and developers to experiment with increasingly complex generative architectures, such as Generative Adversarial Networks (GANs) and Transformer-based models, at a previously unattainable scale.

Furthermore, the convergence of deep learning and HPC technologies is facilitated by the deployment of Remote Direct Memory Access (RDMA) networks, which enable rapid, low-latency communication between nodes without burdening their operating systems. This capability is essential for maintaining efficiency in distributed training processes, particularly when large models must be synchronized across multiple compute nodes.

In this computational ecosystem, deepfake technology benefits from a combination of algorithmic sophistication and high-performance infrastructure. Convolutional Neural Networks (CNNs) enable fine-grained visual analysis, autoencoders manipulate latent feature spaces, Natural Language Processing (NLP) techniques contribute to voice synthesis, Generative Adversarial Networks (GANs) create photorealistic content, and Recurrent Neural Networks (RNNs) manage temporal alignment. All these components are accelerated and made feasible through the application of HPC, which provides the computational foundation necessary to generate synthetic media with a high degree of realism and coherence. [26]

2.2.2 Typologies of deepfakes

The landscape of deepfake technology is continuously evolving, giving rise to various forms of synthetic media manipulation, the most relevant forms are reported below [27].

Face swapping

Face swapping is one of the most representative and technically advanced applications of deepfake technology. It involves replacing the face of a person in an image or video with that of another individual, while preserving essential attributes such as facial expression, pose, lighting, and background. This process relies on deep learning techniques, particularly autoencoders and generative adversarial networks (GANs), and requires not only a well-trained model but also accurate pre- and post-processing stages to ensure that the result appears perceptually realistic [28]. The architecture of a typical face swapping system is usually divided into three main components: identity extraction, attribute extraction, and image generation [28].

Identity extraction focuses on capturing the facial features of the source person, while attribute extraction retrieves contextual information such as pose, gaze, expression, and background from the target image. These two inputs are then fused by the generator, which produces a new synthetic face that integrates the identity of the source with the contextual features of the target [28].

Two main approaches can be adopted in face swapping: source-based and target-based. The source-based approach edits the source image by incorporating the visual attributes extracted from the target, while the target-based method overlays the identity features from the source image into the structural context of the target. Although both approaches can achieve high-quality results, the target-based method is generally favored in recent research and practical implementations. This preference is primarily due to its greater robustness and generalizability. Unlike the source-based approach, which provides limited control over the surrounding environment and may introduce inconsistencies in background or lighting, the target-based approach ensures that the final output maintains the original environmental coherence of the target image. As a result, it yields more convincing and universally applicable face swaps, making it the preferred strategy in the development of identity conversion pipelines [28].

Among the first widely known tools were **FaceSwap** and **DeepFaceLab**, both relying on autoencoders and designed to work in a one-to-one setting. This means they must be trained from scratch for every new pair of identities, which makes them less scalable. [29]

More recent systems, such as FaceShifter, SimSwap, HiFiFace, and DiffFace, are instead based on many-to-many architectures, capable of working with multiple unseen faces. **FaceShifter**, for example, uses two networks - AEI-Net and HEAR-Net - to deal with occlusions and increase fidelity by separating identity and attribute embeddings through adaptive normalization layers. **SimSwap** simplifies the pipeline with a single model and introduces identity embeddings via the pre-trained ArcFace network. **HiFiFace** enhances realism by incorporating 3D morphable models (3DMMs) to reconstruct facial geometry and improve identity preservation. The latest generation, **DiffFace**, relies on diffusion models to generate exceptionally realistic results, although the trade-off is slower execution time [30].

Lip-syncing

Lip syncing has emerged as one of the most critical components in the field of synthetic video generation, playing a vital role in producing realistic, high-quality media content. As the demand for personalized and dynamically generated content increases, particularly in sectors like education, corporate communication, and advertising, the synchronization between audio and visual facial expressions becomes indispensable. The fundamental goal of lip syncing is to align the movement of lips in a video with the accompanying audio in a way that appears both seamless and natural. As Kadam et al. (2021) [31] observe, "approximately, 1 second out of sync lip movement is identified by the viewers", which underlines how sensitive human perception is to misalignment between speech and facial movements. This level of perceptiveness necessitates highly accurate synchronization techniques in order to maintain the illusion of authenticity. Traditionally, lip syncing methods can be categorized as either constrained or unconstrained. Constrained methods, such as those used in the Obama synthesis model, require large datasets of a specific individual (e.g. Barack Obama), meaning the system is tailored to one speaker and performs poorly with novel identities or languages [31]. While these methods are capable of generating highly realistic outputs, they are inherently limited in flexibility. In contrast, unconstrained methods like LipGAN and Wav2Lip are designed to work on "generic videos and audio" making them far more adaptable in real-world applications [31]. Among these, **Wav2Lip** stands out due to its use of a 91 percent accurate discriminator, significantly outperforming prior models in terms of lip-sync accuracy on in-the-wild datasets such as LRS2. This method utilizes a modified SyncNet architecture, which aligns the generated lip movement more closely with audio input, ensuring fluidity and realism [31].

Puppet-master

One of the most accessible and increasingly popular forms of deepfake generation is the puppet-master technique, a method that allows the animation of a static image by transferring motion from a source video. In essence, this approach makes it possible to make someone appear to speak or move, even if only a single frontal image of them is available. As described by Pantelic and Gavrovska (2022) [32], "puppet-master deepfake creation is one of the modest and popular

methods for making a deepfake", highlighting both its simplicity and its widespread use. This technique is made possible by tools like the First Order Motion Model, a neural network that encodes and transfers motion using dynamic keypoints, building upon earlier architectures such as Monkey-Net. The model generates motion by identifying and tracking critical facial features in the source video, which are then mapped onto the target image through local affine transformations. These transformations allow for relatively smooth and believable movements of the eyes, mouth, and head, even if the subject never actually performed them. A key strength of the puppet-master approach is its efficiency and adaptability. According to the authors, the model is capable of synthesizing motion in a way that maintains the structural coherence of the target face, while incorporating features like occlusion maps to handle areas of the face that become hidden or partially visible during movement. These improvements significantly enhance the visual realism of the generated video, particularly in cases involving blinking, head tilts, or natural lip movements. However, despite these technical achievements, the method is not without its flaws. The generated outputs can still exhibit visible artefacts, especially under fast motion or with significant head rotations. Interestingly, the presence of these artefacts can serve as an indicator that the video is synthetic, offering at least a partial safeguard against deception, "artefacts are observable... this is also a positive information since we can still believe that we can distinguish true or false video story" [32].

2.2.3 Main tools for deepfake generation

The rapid evolution of deep learning and computer vision techniques has led to the development of various tools capable of generating highly realistic manipulated media known as deepfakes. These tools, largely based on Generative Adversarial Networks (GANs), allow for seamless facial manipulation, attribute alteration, and even full face synthesis in both images and videos. This section reviews the most widely used tools for deepfake generation, focusing on their underlying models, strengths, weaknesses, and areas of application.

A wide range of tools has emerged to facilitate the creation of deepfakes. These tools vary in complexity, quality, speed, and usability. The table below presents a summary of the main tools currently used in the field, along with key evaluation metrics such as accuracy, processing speed, usability, security, and availability. [33] **StyleGAN**, developed by NVIDIA, has been one of the

Table 2.1. Comparison of Deepfake Generation Tools

Tool	Model	Focus Area	Accuracy	Speed	Usability	Availability
FaceSwap-GAN	HEAR-Net + AEINet	Face swapping and reenactment	Moderate	Slow	Moderate	Open source
SimSwap	Encoder–Decoder + GAN	Face swapping in images and video	High	Fast	Moderate	Paid
FewShot FT GAN	Few-shot GAN	Attribute manipulation from few samples	High	Moderate	Moderate	Paid
FaceShifter	HEAR-Net	Two-stage high-accuracy face swap	High	Fast	Low	Paid
DiscoFaceGAN	Disentangled StyleGAN	Controlled face generation	High	Slow	Low	Open source
FaceApp	Proprietary	Basic facial transformations	Low	Fast	High	Paid
StarGAN	StarGAN	Attribute transfer across domains	Moderate	Slow	Low	Open source
StarGAN-v2	StarGAN-v2	Improved multi-domain attribute editing	High	Slow	Low	Open source
ATTGAN	Attribute GAN	Facial attribute manipulation	High	Moderate	Moderate	Open source
StyleGAN/2/3	Style-based GANs	High-res image synthesis	High	Slow	Low	Open source
CycleGAN	CycleGAN	Unpaired image-to-image translation	High	Fast	Low	Open source

most influential tools for generating synthetic human faces. With StyleGAN2 and StyleGAN3, improvements were introduced in terms of resolution, style control, and temporal consistency, particularly useful for video deepfakes. These tools allow fine-grained control over image features such as age, expression, and lighting. Limitations: High computational cost and vulnerability in low-data regions of latent space.

StarGAN is a multi-domain image-to-image translation tool, capable of transferring various facial attributes (e.g., age, emotion, gender) using a single model. StarGAN-v2 enhances the scalability and quality of generated images. Use case: Attribute-based facial transformation across different domains.

SimSwap excels in face swapping in images and videos using a lightweight architecture. FaceShifter adopts a two-stage process to ensure high-quality synthesis and occlusion awareness. Strengths: High accuracy and speed, making them suitable for real-time or near-real-time applications.

Unlike traditional GANs, **CycleGAN** does not require paired training data. It is highly effective for domain translation tasks, such as converting facial expressions or transferring artistic styles. Advantage: Versatility and reduced data constraints.

ATTGAN introduces attribute-specific control using classification constraints, ensuring that only desired features are modified while maintaining identity and realism. Use case: Emotion or feature editing in faces (e.g., adding/removing glasses, changing age).

2.2.4 Applications of deepfake

The applications of deepfake technology are highly diverse, spanning both beneficial and harmful uses across multiple domains.

Entertainment and media production

Deepfake technology has revolutionized filmmaking and video game development by enabling voice cloning and digital character manipulation: deepfakes facilitate post-production editing, allowing actors' voices and appearances to be seamlessly altered even when they are no longer available for reshoots.

The entertainment industry also leverages deepfakes for satire and parody, where audiences recognize the synthetic nature of the content and appreciate the humorous effect it produces. A striking example of this is the 2023 deepfake depicting Dwayne "The Rock" Johnson as Dora the Explorer, demonstrating the potential for lighthearted and imaginative reinterpretations of popular figures.

Customer experience

Deepfake technology plays an increasing role in customer service and hyper-personalization: AI-generated voices are integrated into caller response services, automating interactions in sectors such as customer support and telecommunication. Deepfake-based virtual assistants provide personalized responses to routine inquiries, such as checking an account balance or scheduling an appointment. Beyond automation, this technology enhances brand inclusivity by adapting digital content to reflect diverse demographics, tailoring representations of ethnicity and skin tone to resonate with different audiences.

Education

Educational platforms have also begun to incorporate deepfake-driven AI tutors, offering interactive and adaptive learning experiences. Advanced AI assistants, such as Claude from Anthropic, demonstrate the potential for deepfake technology to clarify complex concepts, identify knowledge gaps, and personalize instruction for students.

2.2.5 Challenges

Despite these promising applications, deepfakes also present significant ethical and security concerns. The most relevant ones are reported below.

Reputational demage

A major area of misuse involves blackmail and reputational damage, where manipulated images or videos depict individuals in compromising situations, such as engaging in illegal activities, spreading false statements, or participating in explicit content without consent. One of the most pervasive and harmful manifestations of this is nonconsensual deepfake pornography, commonly

used for revenge, harassment, or cyberbullying. A high-profile example occurred in 2019, when deepfake technology was used to superimpose the face of actress Scarlett Johansson onto explicit videos without her consent, sparking widespread discussions on the ethical implications and the urgent need for legal frameworks to combat such abuses [34].

Political manipulation

Another alarming application of deepfakes is in misinformation and political manipulation, where altered videos are used to distort public perception or influence elections. A striking example is the deepfake of Ukrainian President Zelenskyy, which falsely portrayed him issuing a statement of surrender during wartime, illustrating the potential of this technology to create confusion and disrupt geopolitical stability [35]. Similarly, in the 2020 U.S. presidential election, concerns arose over the possible use of deepfake technology to spread disinformation about candidates, further highlighting its potential to erode democratic processes. In an era where digital content can shape political narratives, the ability to fabricate convincing yet entirely false videos raises urgent questions about media credibility. The risk extends beyond political figures to the erosion of public trust in recorded evidence. A society where any video can be dismissed as a deepfake, a phenomenon known as the "liar's dividend", could lead to widespread skepticism and an overall decline in factual discourse.

Identity theft

Deepfake technology raises serious concerns regarding identity theft. AI-generated content blurs the lines between digital replication and personal identity, creating legal ambiguities over ownership and rights of use. Legal frameworks have struggled to keep pace with AI-generated content, leaving gaps in regulation that enable bad actors to exploit deepfake technology for fraudulent activities. For instance, AI-generated voices of deceased celebrities have been used in unauthorized media, raising ethical questions about posthumous rights and consent. As digital forgeries become more sophisticated, lawmakers must consider how to protect individuals from the misuse of their identities in the digital realm.

Privacy and intellectual property concerns

Alongside identity theft, deepfake technology also raises significant concerns about privacy and intellectual property rights. Actors, musicians, and public figures have increasingly faced unauthorized digital reproductions of their likenesses, leading to debates on whether deepfake-generated representations should be legally classified as a form of identity theft or a copyright violation. In 2023, Tom Hanks publicly warned audiences about the unauthorized use of his deepfake likeness in an advertisement, reinforcing the growing need for intellectual property protections in the AI era. These cases highlight the challenges in determining ownership and the rights of use for AI-generated content, emphasizing the urgent need to adapt intellectual property laws to this evolving technological landscape.

Fraud and cybersecurity threats

As deepfake technology continues to advance, its implications extend to future cybersecurity threats, including AI-generated text messages that mimic an individual's writing style to deceive recipients. Reports such as the U.S. Department of Homeland Security's Increasing Threat of Deepfake Identities highlight the potential for malicious actors to leverage AI-driven text replication for social engineering attacks [36]. The intersection of deepfake fraud and cybersecurity is particularly concerning in financial crimes, where AI-generated synthetic identities can be used to bypass authentication mechanisms. Fraudulent deepfake videos have been utilized in remote identity verification processes, raising significant security challenges for financial institutions and regulatory bodies. Moreover, by convincingly imitating an individual's voice or likeness, cybercriminals can deceive financial institutions into granting unauthorized access to sensitive information. In one of the most notorious cases, a cybercriminal used AI-generated voice cloning to

impersonate the CEO of a UK-based energy firm, successfully instructing a subsidiary to transfer 220,000 euros to an external account [37]. With the increasing sophistication of deepfakes, balancing their potential for innovation with the urgent need for robust detection mechanisms and regulatory oversight remains a critical challenge in the digital age.

2.2.6 Are deepfakes illegal?

The current regulation regarding deepfakes is still far from being competitive and comprehensive in today's context. At the European (and also Italian) level, there is yet no specific legislation focused on regulating and limiting the use of deepfake technologies, nor on penalizing the creation or use of deepfake content. In the United States, the situation is quite different, though not necessarily better, mainly due to the significant fragmentation between states. Finally, it is interesting to analyze China's position on deepfakes. All three of these points are discussed in detail in the following section.

2.3 Legal framework

2.3.1 European scenario

AI Act

Currently, the only explicit reference to deepfakes within legislation is found in the Artificial Intelligence Act (AIA) [38]. This regulation focuses on the use of AI systems, categorizing them into three main groups: prohibited technologies, high-risk technologies, and limited-risk technologies. Deepfakes are not formally included in any of these categories but are addressed separately in Article 50. Here, deepfakes are described as audio, video, or image content generated or modified by AI that mimics real people, places, objects, or events, potentially misleading the observer. The AIA stipulates that deployers must clearly indicate that such content has been created or altered artificially by labeling it visibly. However, the transparency obligation does not apply if the content falls within clearly creative, satirical, artistic, or fictional works, provided that third-party rights are protected. In these cases, it is sufficient to signal the existence of manipulated content without compromising the enjoyment of the work. Another exception concerns the use of deepfakes authorized by law for crime prevention or prosecution purposes. Additionally, if AI-generated content is published to inform the public on matters of public interest, the disclosure requirement may be waived if there has been editorial review and clear accountability by a physical or legal entity.

In summary, the AI Act takes a rather permissive stance on the use of deepfakes: it imposes an obligation for deployers to label them but includes numerous exceptions, which could leave room for improper use of the technology, without clear rules or adequate consequences. According to many experts, it would be advisable to extend the labeling requirement to providers and reconsider the current exceptions. Looking ahead, it may be necessary to introduce specific legislation designed to regulate the use of deepfakes [39].

A critical reflection on the impact of ambiguity in the AI Act

One of the most frequently raised concerns about the Artificial Intelligence Act (AIA) is the significant level of regulatory ambiguity that permeates the text. As extensively analyzed by Vainionpaa et al. (2023), this ambiguity manifests in several ways. First, there is a lack of clarity in the definitions of essential concepts such as "AI system", "user", "provider", "manipulation", and even "fundamental rights" or "non-discrimination". This vagueness results in legal uncertainty, as stakeholders may interpret the same provision differently depending on their context and interests. Moreover, the broad and overly inclusive scope of the AIA makes it difficult to determine which technologies are truly covered, especially when AI is defined so broadly that it risks encompassing even ordinary data processing systems. The authors also highlight how the

lack of precise thresholds for classifying systems as high-risk or unacceptable adds another layer of confusion, making enforcement inconsistent and potentially arbitrary. The ambiguity in the risk classification mechanism is particularly problematic.

Vainionpaa et al. stress [40] that the current approach fails to differentiate between varying degrees of harm within the high-risk category, and that certain AI applications, such as those with societal or psychological impact, may be excluded from regulation simply because they do not cause " physical" harm. This leaves serious gaps in the regulation's ability to address intangible but very real risks, including those posed by synthetic media like deepfakes. Additionally, the absence of harmonized standards for technical compliance and risk assessment is flagged as a critical enforcement issue. Without concrete operational criteria, providers are left to self-assess compliance, which, as the authors argue, creates a regulatory environment that is both difficult to monitor and easy to manipulate. This self-regulation model is especially vulnerable to abuse in contexts where commercial or political incentives may encourage non-transparent uses of AI.

In my view, this ambiguity represents a serious obstacle to the effectiveness of the AI Act. While some level of flexibility might be necessary in horizontal legislation that covers a wide range of technologies and sectors, the current vagueness risks undermining the regulation's practical enforceability. Organizations may struggle to assess their compliance obligations, especially in the absence of standardized criteria or technical guidance. As a result, compliance efforts might become inconsistent, fragmented, or purely formalistic, more focused on ticking boxes than ensuring genuine accountability. Even more concerning is the possibility that ambiguity may be exploited. The Act's reliance on self-assessment mechanisms, without sufficient external oversight, opens the door to strategic interpretations that could justify non-compliance or allow actors to avoid responsibility. For instance, harmful deepfakes could be framed as artistic or satirical content simply to fall under one of the Act's exceptions. I believe that such loopholes are particularly dangerous in fast-evolving fields like generative AI, where harms can occur rapidly and at scale. Furthermore, Vainionpaa et al. (2023) [40] emphasize that the lack of guidance on enforcement responsibilities, particularly regarding which authorities are responsible, how compliance should be audited, and what happens in the case of cross-border violations, raises serious doubts about the regulation's institutional capacity to deliver its objectives. This ambiguity risks fostering a patchwork of national interpretations and enforcement efforts, which would weaken the overall harmonizing purpose of the regulation.

According to my view, the AI Act should have taken a firmer stance on these issues by clarifying core definitions and strengthening enforcement mechanisms. If left unresolved, this ambiguity may seriously limit the AI Act's ability to protect fundamental rights and uphold transparency, especially in high-risk scenarios involving synthetic media and manipulative AI. Without clear, enforceable norms, there is a real danger that the regulation may fall short of its ambitious objectives.

Panel for the Future of Science and Technology (STOA)

As early as 2021, the need to update the AI Act to better regulate the deepfake phenomenon emerged. That year, the Panel for the Future of Science and Technology (STOA) presented a report to the European Parliament titled "Tackling Deepfake in European Policy" [41], which included some specific recommendations. Among these, it is proposed to clarify when deepfakes may fall under the prohibited or high-risk practices of the AI Act, considering the possibility of classifying them as high-risk systems or introducing targeted bans for particularly dangerous uses, such as non-consensual pornography or political disinformation. Regarding the labeling obligation for content under the AI Act, STOA suggests extending this requirement to providers, highlighting the risks associated with overly broad exceptions, such as those provided for crime prevention, artistic or scientific purposes, or freedom of expression. Additionally, it recommends limiting the distribution of deepfake detection technologies to prevent them from being circumvented, while not restricting access to too few entities. Finally, the report emphasizes the importance of investing in defensive technologies and raising public awareness on the issue.

The new Italian DDL - Reato di deepfake

On April 23, 2024, the Government approved a bill introducing Article 612-quater to the Penal Code, focusing on the new crime of deepfake, defined as the illicit dissemination of content that has been altered or created through artificial intelligence technologies. This offense carries prison sentences ranging from one to five years and is constituted when falsified content, such as images, videos, or sounds, is disseminated involving people, objects, or voices. The falsity may involve the entire content or just part of it. In any case, the content must have been generated or manipulated using Artificial Intelligence and must deceive the observer, making them believe it is authentic or comes from a legitimate source. Additionally, the content must cause unjustifiable harm to an individual or group. In general, the crime can only be prosecuted if the victim files a complaint, although there are exceptions. In some cases, prosecution can proceed ex officio, such as when the offense is related to another crime requiring official intervention, when the victim is incapable due to age or infirmity, or when the crime is committed against a public authority in the course of their functions. To prevent AI-generated content from being confused with authentic content, the bill establishes the obligation to mark such materials with a distinguishing sign. This requirement applies not only to content fully generated by Artificial Intelligence but also to content that has been partially modified or altered. For example, a watermark or mark, such as the acronym "AI", is required for images or videos, or an audio warning for sound content. However, this rule does not apply to content with a clear creative, artistic, or satirical nature, such as material intended for entertainment or fiction, unless it violates the rights or freedoms of others. Specific procedures for implementing these provisions will be defined through a regulation issued by AGCOM. Additionally, digital platform providers must equip themselves with tools that allow users to declare if uploaded video content has been generated, altered, or modified using Artificial Intelligence. [42, 43, 44]

2.3.2 USA scenario

The U.S. scenario regarding the regulation of artificial intelligence (and in particular deepfakes) differs significantly from the European or Chinese approaches (see next section), mainly due to the fact that in the United States, there is no binding federal legislation (hard law) on this matter, but only some proposed bills in Congress. As a result, individual states have stepped in separately to regulate the use of emerging technologies such as deepfakes, focusing particularly on specific areas like political interference or non-consensual pornography.

Main legislative proposals

The **Deepfakes Accountability Act** was proposed in September 2023 in the House of Representatives with the goal of protecting national security and safeguarding victims of deepfakes. This proposal requires that deepfake creators clearly identify manipulated content, for example, by using technologies to trace the origin of the content and adding alerts indicating the altered nature of the materials (audio/video). Violations could lead to imprisonment, with a sentence of up to five years for deepfakes that cause harassment, fraud, or interfere with official proceedings. Additionally, the proposal includes the creation of a task force within the Department of Homeland Security to manage these issues and obligations for deepfake technology developers [45]. The **DEFIANCE Act** (introduced in 2024) aims to protect victims of non-consensual sexual deepfakes, extending existing protection for the distribution of intimate images without the consent of the person depicted. This bill acknowledges the severity of violations related to non-consensual pornography and proposes measures to ensure justice for the victims [46].

State-Level regulatory actions

At the state level, several states have already introduced regulations to govern the use of deepfakes. These interventions mainly focus on two areas: pornographic deepfakes and political deepfakes.

California was one of the first states to regulate political deepfakes. In 2019, a law was passed that banned the distribution of misleading content within sixty days of an election if the content

was created with the intent to harm a candidate or deceive voters. However, this law was repealed in 2023. California also introduced a law on pornographic deepfakes, allowing victims to take legal action against those who distribute non-consensual sexual content. Exceptions are limited, for example, when the content has journalistic or educational value.

Texas introduced a law banning the distribution of political deepfakes within thirty days of elections if the content is intended to influence the electoral debate.

In Mississippi, a specific crime was introduced for the distribution of "digitization" (a term used for deepfakes) within ninety days of elections, if the content is distributed without the consent of the person depicted and with the intent to influence the electoral outcome.

New Mexico passed a law similar to the Deepfakes Accountability Act, which requires labeling content as deepfakes and penalizes those who distribute misleading content.

Virginia passed a law that punishes the distribution of pornographic deepfakes, especially in the context of revenge porn, when such content is created and distributed to cause psychological and reputational harm.

Florida expanded its non-consensual pornography legislation to include deepfakes. Criminal penalties apply to those promoting deepfakes, although the effectiveness of the exceptions provided is unclear, as they are mainly related to content with journalistic or educational value.

In summary, the regulation of deepfakes in the United States is still limited, with a focus primarily on specific areas such as non-consensual pornography and electoral interference [39, 47].

Critical reflection on US fragmentation

The fragmented and inconsistent regulatory approach to deepfakes in the United States, resulting from the absence of a cohesive federal framework, poses significant challenges to the effectiveness of legal responses to the risks posed by synthetic media. As Chesney and Citron [48] argue in their seminal work, Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security, the "patchwork" of state laws governing deepfake technology creates substantial legal ambiguity. This fragmentation complicates the ability of creators, platforms, and enforcement bodies to navigate the ever-evolving legal landscape, leaving key questions of permissibility unresolved across different jurisdictions. As the authors point out, the lack of a uniform approach means that the legality of creating, sharing, or using deepfake media may differ drastically from one state to another, creating significant uncertainties for both producers and consumers of such content [48].

This regulatory uncertainty not only undermines the deterrent effect of law but also stifles innovation and the responsible development of new technologies. When legal standards are ambiguous or inconsistent, companies operating in the synthetic media space may struggle to ensure that their technologies comply with every jurisdiction's requirements. This not only results in heightened compliance costs but also creates a climate of hesitation. As observed in the cited article, the patchwork of state laws exacerbates the compliance burdens for companies that operate across state lines, forcing them to navigate a labyrinth of conflicting obligations related to consent, labeling, and liability [48]. In this regard, the legal uncertainty serves as a disincentive for technology developers to implement more robust measures that could mitigate the risks associated with deepfake content, such as improved detection mechanisms or better consent protocols.

Moreover, the lack of a centralized regulatory framework facilitates what the authors describe as "forum shopping", whereby malicious actors intentionally seek out states with more lenient or non-existent regulations in order to evade accountability. I believe that this exacerbates the problem, as it allows bad actors to exploit jurisdictional weaknesses to distribute harmful deepfake content with relative impunity. For example, individuals creating malicious deepfakes with the intent to deceive or harm can easily avoid legal repercussions by operating in jurisdictions that have no laws specifically targeting this type of content. This regulatory inconsistency significantly undermines the deterrent effect of the law, as perpetrators may calculate the risks of prosecution as minimal, thus undermining the purpose of legal regulation. Chesney and Citron's analysis also highlights the broader implications of regulatory fragmentation on key societal interests, such as individual privacy, democratic integrity, and national security. The authors argue that the

proliferation of deepfakes poses a significant threat to democratic processes, particularly in the context of electoral manipulation, where synthetic media could be used to create fake videos of politicians making controversial statements or engaging in inappropriate behavior [48].

From my perspective, this issue is not only a legal one but also a deeply political challenge. If the legal system is unable to regulate deepfake technology effectively, the potential for electoral interference, disinformation campaigns, and public manipulation is heightened. Deepfakes can be used to disrupt public trust in institutions and erode the foundations of democratic discourse. In this context, a cohesive federal strategy is essential, not only to address legal gaps but to safeguard the integrity of democratic processes and the public's confidence in the information they consume.

Furthermore, the lack of standardized legal protections exacerbates vulnerabilities related to privacy violations, particularly with regard to non-consensual deepfake pornography. Chesney and Citron emphasize that the ease with which such content can be produced and distributed without the consent of the individuals depicted in the media undermines personal privacy and can cause severe harm to the victims, especially women [48]. I strongly agree with their assertion that the absence of a national legal framework creates a legal gray area, where victims may find themselves without adequate recourse in states where laws addressing non-consensual synthetic media are either underdeveloped or nonexistent. As deepfake technology becomes more advanced and accessible, I believe it is critical that the U.S. adopt a unified legal framework that ensures robust privacy protections for all individuals, regardless of their state of residence.

I strongly believe that only a harmonized federal approach can effectively address the risks posed by deepfakes and other synthetic media technologies. A comprehensive federal framework would provide a clear and consistent set of rules that both protect individuals from harm and enable technological innovation to thrive within a stable legal environment. Without such a framework, the legal system risks falling behind the rapidly advancing capabilities of generative technologies, potentially allowing the continued spread of harmful deepfake content and further complicating the enforcement of laws designed to protect privacy and public trust

2.3.3 China

The Chinese government has developed a strategy to control the use of emerging technologies, including artificial intelligence and deep synthesis technologies such as deepfakes. This strategy is part of the concept of "cyber sovereignty", which involves state control over cyberspace to protect national security, social order, and the country's values. Below, I will examine the main Chinese regulations related to deepfakes, analyzing the evolution of regulation and the implications of these laws.

The Beginning of Regulation: The ZAO Case and 2019

The phenomenon of deepfakes emerged in China as a growing concern in 2019, particularly due to the popularity of apps like ZAO, which allowed users to create manipulated content, especially through "face-swapping" technology. The increasing spread of these technologies and issues related to data collection prompted Chinese authorities to take action. Just three months after the launch of ZAO, the Cyberspace Administration of China (CAC) began discussing the need to regulate the use of these technologies, which could potentially pose risks to national security and public order [39].

The "Regulations on the Administration of Online Audio and Video Information Services" (2020)

In 2020, China adopted the "Regulations on the Administration of Online Audio and Video Information Services", a set of regulations aimed at controlling the use of deep synthesis technologies for creating audio and video content. These regulations impose strict restrictions on the use of images, audio, and video generated through deepfake technologies, prohibiting the creation and

dissemination of false news. Online platforms are required to monitor and control content, ensuring user identity and preventing the distribution of material that could threaten national security or disrupt social order [39].

The Regulations on Deep Synthesis Management of Internet Information Service (2022)

In 2022, China introduced a new set of regulations, the "Regulations on Deep Synthesis Management of Internet Information Services". These laws aim to more specifically regulate the use of deep synthesis technologies, including deepfakes. The goal is to protect "fundamental socialist values", ensure national security, and safeguard the rights of citizens. While not entirely banning the creation of deepfakes, these regulations state that content created with these technologies must not violate Chinese laws, harm China's image, or damage the economic and social order. Platforms are required to properly label content and obtain consent from the individuals involved, such as when faces or voices are altered [39].

Labelling (2025)

On March 7, 2025, the Cyberspace Administration of China (CAC), along with other Chinese authorities, introduced the "Measures for the Labeling of AI-Generated Content" (AIGC), which will come into effect on September 1, 2025. These regulations require all online service providers to clearly label AI-generated content, such as text, images, audio, video, and virtual content, to avoid confusion or disinformation. The labels must be both explicit and implicit. Explicit labels must be visible to users and placed clearly on AIGC content, such as at the beginning, end, or middle of texts, videos, or images. Implicit labels, on the other hand, are embedded in the file's metadata and contain information about the content's nature and the service provider, but are not visible to users. All service providers must include these labels in their service agreements and, if required, can offer content without explicit labels, but only if the labeling obligation for users is clearly specified. Furthermore, app distribution platforms must ask providers to declare if they offer AIGC services and verify the adequacy of the labels. Users are required to declare when posting AIGC content using the labeling feature provided by services. Additionally, practices such as removing or modifying AIGC labels are prohibited. Violations of these regulations may lead to sanctions by the competent authorities. These measures respond to the growing concern about the risks associated with AI use, particularly regarding the spread of false content. With the advancement of local AI models like DeepSeek, China has intensified its regulations, aiming to ensure transparency and prevent content manipulation online [49, 39].

Considerations on Chinese approach

One crucial aspect to consider when analyzing China's approach to regulating deepfakes is its long-term sustainability and the potential risks of authoritarian abuses. While Chinese laws and regulations aim to safeguard national security and maintain social order, the regulatory framework implemented by China raises significant concerns regarding individual freedoms and the pluralism of information.

Regulations such as the "Regulations on Deep Synthesis Management of Internet Information Services" and the more recent AI-generated content labeling measures are clear manifestations of the Chinese government's strategy to establish total control over cyberspace, not only to prevent technological misuse but also to protect the so-called "fundamental socialist values" of the country.

However, a problematic aspect of this regulation is that such an approach could easily be used by the government as a pretext for increasingly stringent forms of censorship [39]. The regulation of deepfakes, while ostensibly focused on protecting the integrity of information, could quickly turn into a tool of repression. For example, the vague definitions of terms like "threat to national security" or "disruption of public order" leave ample room for arbitrary interpretations by the authorities. In this sense, the Chinese government could exploit these laws to suppress content that, although not constituting an actual security threat, might be seen as critical of the regime

or as political dissent. This could severely undermine press freedom and access to independent information, two fundamental pillars of a democratic society.

Furthermore, the requirement to label all AI-generated content and implement monitoring mechanisms on platforms, while aimed at ensuring greater transparency, could result in constant surveillance of users, further eroding privacy and fostering a kind of "self-censorship" among content creators. Indeed, as argued in the article, while such measures are designed to protect against harmful or deceptive content, they risk stifling free expression online by making platforms more susceptible to political demands for censorship, especially in a context like China, where the government exerts significant control over emerging technologies and information.

In this regard, it is interesting to note that China is implementing more aggressive regulatory policies compared to other countries, such as the United States and European nations, which adopt a less centralized and more rights-oriented approach. The fundamental difference lies in the fact that while Western countries tend to balance privacy protection and online security with freedom of expression, China views cyberspace as an extension of state power, rather than as a space for free and independent interaction.

One of the most concerning reflections that arises from this, as also highlighted in the article, is the possibility that China could become a model for other authoritarian states to adopt similar approaches to digital technology regulation, creating a global norm that could undermine fundamental rights. The danger of this scenario lies in its ability to be justified under the guise of combating fake news and disinformation, but in reality, it could serve to consolidate authoritarian regimes that have no qualms about using technology to suppress opposition.

In conclusion, while China's approach may appear effective in the short term at combating the malicious use of deepfakes and protecting public security, the risk of abuse is significant. Chinese regulation could, in fact, become a pretext for justifying oppressive and centralized control over information, potentially exacerbating censorship and political repression. If other countries were to follow this model, the result would be a digital world characterized by ever-more pervasive state control, with severe implications for individual freedoms and the pluralism of information.

2.3.4 UK

Online Safety Act

The United Kingdom has recently taken significant steps to update its legal framework in response to the growing threat posed by synthetic media, particularly in the context of intimate image abuse. The Online Safety Act 2023, which came into force in early 2024, now explicitly criminalizes the distribution of sexually explicit deepfakes without the subject's consent. The reform is particularly impactful because it removes the previous requirement to demonstrate that the offender intended to cause distress. Under the current law, the mere act of sharing synthetic intimate content without permission is sufficient to constitute a criminal offence. Furthermore, the law explicitly recognizes deepfakes as a form of intimate image abuse, treating them on equal terms with authentic photographs or videos. If the offender's actions are driven by the intention to cause harm or seek sexual gratification, aggravating factors apply, leading to more severe penalties. [50]

Broader vision of responsibility

In January 2025, the UK government announced a further expansion of its legal arsenal: the creation of a new standalone offence for generating sexually explicit deepfakes, even in cases where the content is not distributed. This reform addresses a critical gap in the previous framework, where only the act of sharing or threatening to share such content was punishable. The forthcoming provision rightly recognizes that the mere production of synthetic explicit material involving real individuals constitutes a serious violation of personal dignity and privacy. In addition, this approach represents an important step forward compared to emerging EU legislation such as the AI Act, which primarily places responsibility on platforms to label AI-generated content, without

sufficiently addressing the liability of the individuals or entities who create harmful content in the first place. The UK's decision to penalize the act of creation itself, regardless of intention to disseminate, offers a more victim-centric approach and acknowledges the psychological and reputational harm that can arise even from undisclosed deepfake production. This reform could serve as a model for other jurisdictions seeking to address the misuse of AI-generated media with greater precision and fairness. [50]

The missing legislation for political deepfakes

While the United Kingdom has made considerable progress in regulating deepfakes in the context of intimate image abuse, the political dimension remains less developed. At present, there is no specific legislation addressing the use of synthetic media in electoral campaigns. Nonetheless, the Electoral Commission holds the authority to intervene in cases where manipulated content constitutes a criminal false statement about a candidate, as outlined in existing electoral laws. However, these provisions are relatively narrow in scope and may not fully capture the emerging risks associated with generative AI in political discourse. The growing potential for AI-generated audio and video to impersonate political figures and mislead voters has sparked discussions around the adoption of ethical guidelines and voluntary codes of conduct for political parties and advertisers. While the UK has not yet enacted binding rules in this area, ongoing debate reflects a growing awareness of the need to prevent deceptive uses of AI in democratic processes, especially in light of similar controversies that have emerged in other countries, including the circulation of deepfaked campaign materials.

2.3.5 Some cases

Crosetto Case

Recently, a scam case emerged involving some of the most well-known Italian entrepreneurs and professionals, exploiting the name of the Minister of Defense, Guido Crosetto. The scammers, likely using advanced technologies such as artificial intelligence to replicate the Minister's voice, impersonated members of his staff and contacted the victims, demanding large sums of money as ransom for allegedly kidnapped Italian journalists abroad. They assured the victims that the transferred funds would later be reimbursed by the Bank of Italy. Among the individuals targeted were prominent figures such as Massimo Moratti, former president of Inter Milan, Giorgio Armani, Marco Tronchetti Provera, Diego Della Valle, Patrizio Bertelli, the Caltagirone and Del Vecchio families, and the Beretta family. At least one entrepreneur transferred nearly one million euros to a foreign account, believing the request to be legitimate. Minister Crosetto promptly reported the incident, highlighting the high level of professionalism demonstrated by the fraudsters and the importance of raising public awareness to prevent similar cases in the future. The Milan Public Prosecutor's Office has launched an investigation to identify those responsible and recover the transferred funds. This incident illustrates how the use of advanced technologies, such as artificial intelligence, is becoming an increasingly significant factor in scams, making it more difficult to distinguish between genuine and fraudulent communications. Authorities have urged the public to remain vigilant and to promptly report any suspicious activity [51].

Taylor Swift pornographic deepfake

In January 2024, sexually explicit images of American singer Taylor Swift, generated using artificial intelligence, were published on X (formerly Twitter) and quickly spread to other platforms such as Facebook, Reddit, and Instagram. One tweet containing the images was viewed over 45 million times before being removed. A report by 404 Media revealed that the images appeared to originate from a Telegram group, whose members used tools like Microsoft Designer to generate them, employing typos and keyword hacks to bypass Designer's content filters. Following the release of the material, Swift's fans flooded the platforms with videos and images from her concerts to bury the deepfake images and reported the accounts sharing them. Searches for Swift's name

were temporarily disabled on X, returning an error message instead. Graphika, a disinformation research company, traced the origin of the images back to a 4chan community. A source close to Swift told the Daily Mail that she was considering legal action, stating: "Whether legal action will be pursued is still under consideration, but one thing is clear: these AI-generated images are abusive, offensive, exploitative, and were created without Taylor's consent and/or knowledge". The controversy drew condemnation from White House Press Secretary Karine Jean-Pierre, Microsoft CEO Satya Nadella, the Rape, Abuse and Incest National Network (RAINN), and SAG-AFTRA. Several U.S. lawmakers called for federal legislation to combat deepfake pornography. Later that month, U.S. Senators Dick Durbin, Lindsey Graham, Amy Klobuchar, and Josh Hawley introduced a bipartisan bill that would allow victims to sue individuals who created or possessed "digital forgeries" with the intent to distribute them, or those who knowingly received such material created without consent [52].

Deepfakes in politics

The political use of deepfakes, although less common than their applications in pornography or satire, can have serious consequences for government stability, electoral processes, and even armed conflicts. These tools are used to manipulate public opinion by creating falsified images and videos capable of casting doubt on the legitimacy of a government, influencing state policies, shifting voter sentiment, defaming political figures, or altering economic perceptions. Over the years, several examples have emerged in which deepfakes have had significant political impacts, prompting governments to adopt protective measures and appoint experts to detect such content before it spreads widely.

A notable case occurred in 2018 when Jordan Peele released a fake video of Barack Obama, showing the former president making offensive statements against Donald Trump. This episode highlighted the dangers of deepfakes and how easily technology can manipulate public opinion. A year later, in 2019, a video of Nancy Pelosi, manipulated with simple slowing techniques, was used to suggest she was intoxicated, damaging her reputation. That same year, a suspicious video of Ali Bongo, the president of Gabon, fueled rumors of his alleged death, triggering distrust and even an attempted coup. Although the video turned out to be authentic, it demonstrated how even minimal manipulation can threaten political stability.

In 2021, another incident involved members of the European Parliament, who were tricked by a deepfake during a diplomatic meeting, an event that revealed the vulnerability of international institutions to digital manipulation. More recently, in 2022, during the war in Ukraine, deepfakes were used to spread false messages that undermined the legitimacy of leaders, attempting to demoralize the population and sow confusion among combatants.

These events have made it clear that deepfakes are increasingly infiltrating election campaigns, and their impact is likely to grow. To counter them, it is essential not only to improve detection systems but also to educate the public about the risks and the psychological vulnerabilities these technologies exploit [53].

CEO Fraud UK

A striking example of how artificial intelligence can be maliciously exploited to conduct sophisticated scams emerged in the United Kingdom in 2019. In this incident, cybercriminals employed advanced voice-cloning technologies to impersonate the CEO of the German parent company of a UK-based energy firm. By using AI-generated voice synthesis, the perpetrators were able to accurately replicate the executive's tone, accent, and speaking style, convincingly deceiving a senior manager into transferring a substantial amount of money to a fraudulent account.

The employee, believing he was speaking directly with his CEO, was instructed to urgently wire 220,000 euros (approximately 200,000 dollars at the time) to a Hungarian bank account, allegedly as part of a confidential and time-sensitive acquisition. The conversation, carried out over the phone, was so meticulously constructed that the victim did not question the legitimacy of the request. The attackers further reinforced the ruse by following up with emails and other calls, maintaining the illusion of an authentic corporate transaction.

It was only after a second attempt to solicit funds that suspicions arose, prompting an internal investigation. The scam was ultimately uncovered, but the initial payment had already been completed and could not be recovered. The case underscores the evolving threat landscape posed by AI-powered fraud techniques.

This incident highlights the vulnerabilities of even well-established companies when confronted with the psychological and technological sophistication of modern scams. Traditional verification mechanisms, such as phone recognition or hierarchical trust, are no longer sufficient when artificial intelligence can replicate the human voice with near-perfect accuracy. The case serves as a cautionary tale, emphasizing the urgent need for organizations to implement multi-factor verification protocols, foster internal awareness of AI-enabled threats, and develop robust incident response strategies. As synthetic media becomes more accessible, such attacks are likely to increase in frequency and complexity, making preparedness and resilience essential components of contemporary cybersecurity practices. [54]

2.3.6 Accountability of digital platforms

The growing prevalence of deepfake technology has raised significant concerns regarding the accountability of digital platforms in moderating and managing harmful content. Deepfakes present serious challenges to the integrity of online information, privacy, and public trust. Platforms like X (formerly Twitter), Facebook, and TikTok, which facilitate the rapid sharing of user-generated content, are at the forefront of these challenges, as they balance between freedom of expression and ensuring the safety of their users.

Deepfakes represent a new form of misinformation, one that is far more sophisticated than traditional fake news or doctored images. They leverage artificial intelligence (AI) to create hyper-realistic video and audio recordings that can be used to manipulate public opinion, incite violence, or harass individuals. The implications of deepfakes extend beyond mere deception; they pose significant threats to democracy, privacy, and security. As Citron and Chesney argue [55], deepfakes create new avenues for disinformation campaigns, political manipulation, and personal harm. Their legal implications are especially complex because deepfakes can be difficult to distinguish from authentic content, and their creators can remain anonymous. Given the increasing frequency and potential harm of deepfakes, digital platforms must confront the issue of content moderation. These platforms are often criticized for their reactive rather than proactive approach to harmful content. X, Facebook, and TikTok, all of which operate under different regulatory and legal frameworks, have faced increasing scrutiny for their role in enabling the spread of deepfakes.

USA - Section 230 of the Communications Decency Act

The question of legal responsibility for deepfakes is intricately tied to the debate over platform liability. Current legal frameworks, particularly Section 230 of the Communications Decency Act in the United States, provide platforms with broad immunity from liability for user-generated content. This has allowed platforms like X, Facebook, and TikTok to avoid legal consequences for the harmful content posted by users. However, as deepfakes continue to gain prominence, the limitations of this immunity are becoming more apparent. Citron and Chesney [55] suggest that the existing legal regime might no longer be adequate for addressing the unique challenges posed by deepfakes. While platforms are not directly responsible for creating deepfakes, they can be held accountable for their failure to moderate such content once it is uploaded. In particular, platforms could face legal consequences if they fail to take reasonable steps to detect and remove harmful deepfakes or if they continue to allow the spread of deepfakes in a way that exacerbates the harm. The platform's responsibility is compounded by their ability to amplify the spread of content through algorithms that prioritize sensational or controversial material. For example, in the case of Facebook, the platform has been criticized for its slow response to disinformation and manipulated media, especially during critical political events such as elections. In 2020, Facebook's CEO Mark Zuckerberg faced public backlash after his platform failed to adequately address the proliferation of deepfakes and other manipulated media, despite ongoing concerns raised by researchers and

advocacy groups. Similarly, TikTok has faced scrutiny for hosting deepfake videos that were used to harass individuals, particularly minors. While TikTok has taken steps to improve its content moderation, the platform's algorithm-driven approach to content distribution often means that harmful content can gain significant visibility before it is removed.

The question then becomes: what legal responsibilities should these platforms have in the moderation of deepfake content? Should they be legally obligated to invest in advanced detection systems, or should their responsibility be limited to removing content once it has been flagged by users? Several legal frameworks have been proposed to address the spread of deepfakes. Citron and Chesney [55] argue that a comprehensive approach to regulating deepfakes should involve a combination of civil and criminal liability for both content creators and platforms. In the case of platforms, one potential legal avenue is to amend Section 230 to remove immunity for platforms that fail to moderate harmful content, such as deepfakes. This could encourage platforms to take a more proactive approach in detecting and removing deepfakes before they gain widespread attention. Moreover, some experts suggest that platforms could be required to implement content verification technologies, such as blockchain or digital watermarking, which could help distinguish authentic content from manipulated media. While these technologies are not foolproof, they could provide an additional layer of accountability for platforms by making it easier to trace the origins of content and identify deepfakes early on. Another option is to introduce new criminal laws that target the creators and distributors of deepfakes. In many jurisdictions, creating or distributing harmful deepfakes, especially those aimed at defamation, harassment, or political manipulation, could be punishable by law. However, this approach would still leave a significant gap in accountability for platforms, who may be complicit in amplifying harmful content.

UE - Digital Service Act

In the European context, the Digital Services Act (DSA) represents a significant legislative milestone in redefining the responsibilities of digital platforms. Adopted in 2022, the DSA establishes a new framework for regulating online intermediaries, particularly very large online platforms (VLOPs), by introducing stricter obligations in relation to content moderation, transparency, and risk mitigation. One of the DSA's central goals is to enhance platform accountability by mandating proactive measures against illegal and harmful content, including deepfakes. Under the DSA, platforms are no longer shielded by blanket immunity but are instead required to conduct systematic risk assessments and implement mitigation strategies that address the spread of disinformation and manipulated media.

Deepfakes fall under the broader category of "systemic risks" identified by the DSA, due to their potential to undermine democratic discourse and fundamental rights. Accordingly, VLOPs must assess the impact of deepfakes on civic discourse, public health, and individual dignity, and take "reasonable, proportionate, and effective" measures to limit their dissemination. This includes the obligation to cooperate with vetted researchers, provide greater algorithmic transparency, and adopt auditable internal processes. The DSA also introduces a notice-and-action mechanism, which enables users and trusted flaggers to report harmful content more efficiently. Moreover, the law strengthens enforcement through potential fines of up to 6 percent of the platform's global annual turnover, thereby creating a robust incentive structure for compliance.

The DSA marks a shift from the reactive and fragmented approach that previously characterized platform governance in the EU, toward a co-regulatory model that combines public oversight with private responsibility. While it remains to be seen how effectively the DSA will be implemented, particularly in cases involving complex and rapidly evolving technologies like generative AI, it provides a valuable blueprint for embedding accountability into the architecture of digital platforms. By explicitly recognizing the risks posed by deepfakes and embedding legal duties to mitigate them, the DSA could become a pivotal tool in the broader effort to balance freedom of expression with the protection of users and democratic processes. [56, 57]

2.3.7 Bridging legal and technical perspectives

The comparative analysis of existing legal frameworks demonstrates a converging international effort toward establishing principles that ensure accountability, transparency, authenticity, and

integrity in the development and dissemination of AI-generated content. Although the approaches differ, ranging from the European Union's regulatory precision in the AI Act and Digital Services Act, to the punitive stance of the new Italian bill introducing the crime of deepfake dissemination, and the more fragmented yet pragmatic models adopted in the United States, China, and the United Kingdom, a common thread emerges: the legal necessity to guarantee the verifiability and lawful provenance of digital content.

Regulations such as the AI Act and China's Measures for the Labeling of AI-Generated Content explicitly demand transparency and labeling of manipulated media, while the Italian and British frameworks underscore the importance of authenticity and protection against deception and harm. Similarly, U.S. legislative proposals such as the Deepfakes Accountability Act emphasize traceability and the duty to disclose alterations, reflecting an increasing recognition that effective regulation depends on the ability to technically verify the origin and integrity of digital materials. These principles collectively translate into de facto technical requirements: systems must be capable of detecting manipulation, ensuring content provenance, and maintaining a verifiable chain of custody.

Therefore, while legal measures lay the normative foundation for ethical and accountable AI usage, their practical enforcement ultimately depends on the implementation of technological mechanisms capable of upholding these standards. This calls for solutions that embed legal principles into their architecture, mechanisms ensuring digital evidence integrity, non-repudiation, and authenticity verification. The chapter four address these challenges from a technical perspective, illustrating how principles such as accountability, transparency, and data integrity can be operationalized through cryptographic hashing, digital watermarking, and blockchain-based traceability, thereby bridging the gap between law and technology in the context of AI-generated content regulation.

Chapter 3

State of the art

In this chapter, I will explore the current state of the art in both the detection of deepfakes and the prevention of the negative consequences that may arise from their use. As the sophistication of synthetic content continues to grow, so does the challenge of identifying manipulated media and mitigating the risks associated with it. This section will begin by presenting the most relevant tools and techniques that have been developed to detect deepfakes, highlighting the technological advances and methodologies behind them. Following that, I will address the role of user awareness and education, which has emerged as an equally essential component in the broader defense strategy. Indeed, empowering individuals to critically evaluate the content they encounter online is increasingly recognised as a necessary complement to technical solutions, especially in scenarios where technology alone may not suffice. Through this twofold approach, I aim to provide a comprehensive overview of the efforts currently underway to counteract the risks posed by synthetic media, from both a technological and human-centered perspective.

3.1 Deepfake detection methods

3.1.1 Machine Learning based method

Machine Learning-Based Methods for Detecting Deepfakes Traditional machine learning (ML) algorithms have played a significant role in the development of deepfake detection techniques due to their interpretability and ease of implementation. Unlike more opaque black-box approaches, these models allow researchers and practitioners to better understand the logic behind the classification decisions, which is essential in high-stakes contexts such as media forensics or legal proceedings. Their transparency makes them especially suitable for the deepfake domain, where understanding the nature of manipulated content is often as important as detecting it. [58]

Decision trees and random forests

Among the most widely used ML models are tree-based methods such as Decision Trees and Random Forests. These models represent the decision-making process in the form of hierarchical trees, making it possible to visualize how specific features contribute to the final prediction. This characteristic inherently addresses the issue of explainability, which is often a limitation in more complex models like deep neural networks.

A **decision tree** is a supervised machine learning algorithm used for both classification and regression tasks. It works by splitting data into branches based on feature values, ultimately leading to a prediction or decision at the leaf nodes. The model mimics a tree-like structure, where internal nodes represent decision points based on feature attributes, branches represent possible outcomes of these decisions, and leaf nodes denote the final classification or output. Decision trees are valued for their interpretability and transparency, as they provide a clear explanation of how predictions are made. This is especially important in domains where understanding the rationale

behind a model's decision is critical, such as legal or forensic applications. There are two main types of decision trees, classification trees are used when the target variable is categorical and the model predicts class labels by learning decision rules inferred from the data features, while regression trees are applied when the target variable is continuous and they predict a numerical value by learning from historical data and minimizing prediction error. [59]

In the context of identifying manipulated visual content, decision trees operate by analyzing a set of extracted features from images or videos and using them to classify the input as either authentic or manipulated (deepfake). The first step involves extracting meaningful features from facial images or video frames. These features may include abnormal eye or mouth movements, inconsistencies in skin texture or lighting, lack of synchronization between facial expressions and head motion, presence of visual artifacts or irregularities; a decision tree functions as a flowchart-like structure composed of internal nodes, branches, and leaf nodes:internal nodes represent decision points based on specific features, branches indicate the outcome of those decisions (yes/no), leaf nodes provide the final classification (deepfake/authentic). The tree is constructed by recursively selecting the feature that best separates the data at each step. The goal is to split the dataset in a way that improves classification accuracy at each node. For example: is the eye blinking naturally? yes, then likely authentic; no, then Is mouth movement inconsistent? yes, then likely deepfake; no, then possibly authentic.

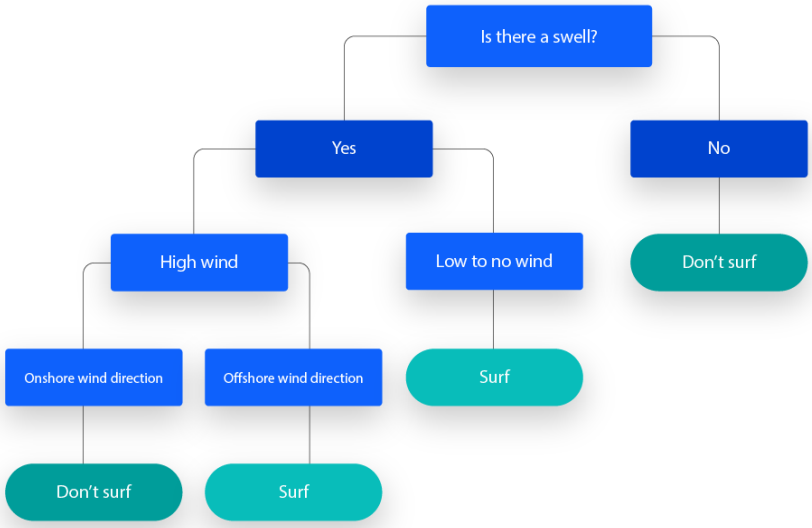


Figure 3.1. Decision Tree (fonte: IBM).

A **random forest** is an ensemble machine learning algorithm that builds multiple decision trees and combines their outputs to improve predictive accuracy and control overfitting. It is widely used for both classification and regression tasks due to its robustness, scalability, and ability to handle high-dimensional data. The fundamental idea behind Random Forest is to create a "forest" of decision trees during training and aggregate their results to make a final prediction. In classification tasks, the model outputs the class that receives the majority vote from individual trees, whereas in regression tasks, it computes the average of all predictions. The Random Forest algorithm follows a process based on two key techniques: bagging (bootstrap aggregating) and random feature selection. During bootstrap sampling, the model creates multiple subsets of the original training data by sampling with replacement. Each subset is used to train an individual decision tree. This introduces diversity among the trees and reduces the variance of the model. During random feature selection, at each split in the construction of a tree, a random subset of features is selected. The best feature among this subset is chosen to perform the split. This method prevents the trees from becoming too similar and ensures that the model explores a broader feature space. Because of this combination of diversity and aggregation, Random Forest offers strong generalization performance. It is less prone to overfitting compared

to single decision trees and performs well even with missing or unbalanced data. Furthermore, Random Forest provides feature importance scores, which allow practitioners to understand which variables contribute most significantly to the model's predictions, a valuable trait in domains such as deepfake detection, where interpretability remains a key concern. [60]

Random Forest is a machine learning technique that combines multiple decision trees to improve classification accuracy and reduce overfitting. Each tree in the forest is trained on a random subset of the data and considers a random selection of features, enabling the model to learn diverse patterns and reduce the risk of bias associated with individual trees. In the context of deepfake detection, Random Forest models are trained on features extracted from images or videos, such as facial expressions, eye movements, head orientation, skin texture, and other subtle inconsistencies that may arise in manipulated content. These features help the model distinguish between authentic and synthetic media. Each decision tree provides a classification output (real or fake), and the final decision is made through a majority voting mechanism across all trees in the forest. This ensemble approach enhances the model's robustness and stability.

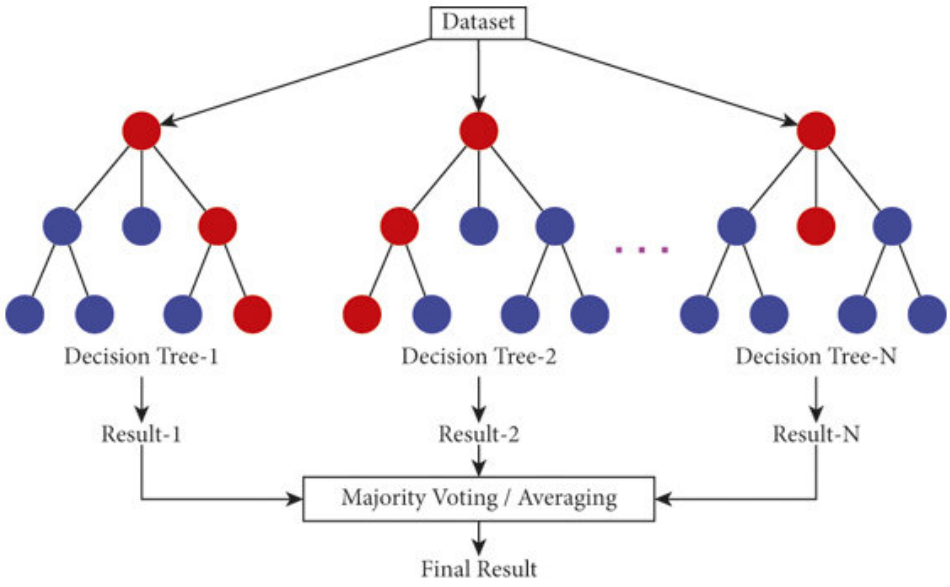


Figure 3.2. Random Forest (fonte: ResearchGate).

MLPs

Another effective strategy involves the use of lightweight neural networks, such as Multi-Layer Perceptrons (MLPs), to detect visual artifacts in manipulated videos. As demonstrated by Habeeba et al [61], MLPs can be employed to identify anomalies in the facial region with relatively low computational requirements, making them suitable for real-time or resource-constrained applications. The Multilayer Perceptron, also known as MLP, is a type of artificial neural network composed of multiple layers of neurons. These layers include: input layer, which receives the input data; hidden layers, which process the information through non-linear activation functions, allowing the network to learn complex patterns; output layer, which provides the final result, such as a classification or prediction. Each neuron in one layer is connected to all neurons in the next layer, forming a fully connected network. During training, the MLP uses an algorithm called backpropagation to update the connection weights by minimizing the error between the predicted output and the desired output. Thanks to its structure, the MLP is capable of solving complex and non-linearly separable problems, making it an effective tool in many machine learning applications. [62]

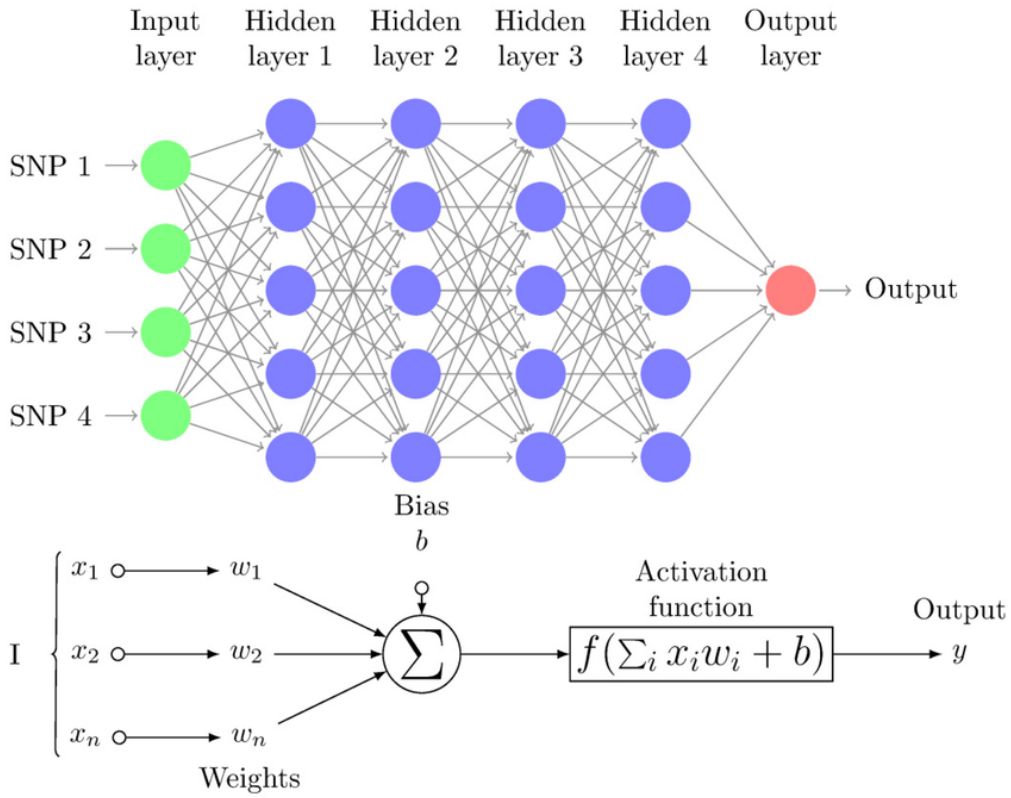


Figure 3.3. Multi Layer Perceptron (fonte: ResearchGate).

Conclusions about machine learning based methods

In terms of performance, machine learning-based methods have demonstrated impressive results, with detection accuracies reaching up to 98 percent in controlled experimental settings. However, these results heavily depend on several factors, including the quality and nature of the dataset, the selection of features and the alignment between the training and test sets. When models are trained and evaluated on similar datasets, commonly using a split such as 80 percent for training and 20 percent for testing, the performance is typically high. Conversely, applying the model to entirely different datasets can result in a significant drop in accuracy, sometimes approaching random classification levels (around 50 percent), thus underscoring the challenges related to generalizability. Overall, while traditional ML methods offer a promising foundation for deepfake detection due to their interpretability and versatility, their performance remains closely tied to dataset characteristics. This highlights the ongoing need for robust feature engineering, diverse training data, and cross-domain validation in order to enhance the reliability and resilience of ML-based deepfake detectors.

3.1.2 Deep Learning based methods

In recent years, deep learning-based techniques have emerged as the most prominent and effective approach for detecting deepfakes, particularly due to their remarkable capacity to model complex visual patterns and subtle anomalies introduced during the synthetic generation process. These methods leverage powerful architectures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and attention-based mechanisms, to analyze both spatial and temporal inconsistencies in deepfake media. [58]

CNN based architectures

Convolutional Neural Networks (CNNs) have played a central role in the early and ongoing development of deep learning-based approaches to deepfake detection, thanks to their remarkable ability to capture spatial features and identify subtle visual artifacts introduced during the generation of manipulated media.

One of the earliest and most influential approaches in this area was proposed by Zhang et al. [63], who developed a GAN simulator to replicate common artifacts produced by generative adversarial networks. These simulated artifacts were then used as inputs for a CNN-based classifier, allowing the model to learn global deepfake signatures rather than relying exclusively on pixel-level irregularities. This method enabled better generalization, as the model could recognize high-level patterns associated with fake content, regardless of the specific generation method.

Building upon this foundation, Zhou et al. [64] introduced a CNN that focused on standardized feature extraction from RGB data, aiming to improve the model's ability to distinguish between real and fake images using consistent visual cues. In parallel, other researchers proposed resolution-agnostic CNN architectures capable of maintaining detection performance across images and videos of varying quality and compression levels. These models addressed the problem of performance degradation when fake media is shared on social platforms, where heavy compression and scaling are common.

Another notable advancement in CNN-based detection leveraged the fact that synthetic media often fails to accurately replicate biological signals, such as heartbeat rhythms and facial blood flow. These subtle patterns, which are naturally present in authentic video recordings, can be extracted from facial regions and analyzed using CNNs trained to detect inconsistencies in color changes and motion. The absence or distortion of such physiological signals is a strong indicator of manipulation, and CNNs have proven effective at learning these cues. [58]

To balance performance and efficiency, models like Meso-4 and MesoInception-4 [65] introduced lightweight CNN architectures based on Inception modules. These models were trained using Mean Squared Error (MSE) loss and demonstrated solid performance in detecting manipulated frames while being computationally suitable for real-time applications. These architectures have been particularly effective for shallow spatial features, and when combined with handcrafted cues such as eye blinking anomalies or unnatural lip synchronization, they improved overall classification precision.

Several techniques have been explored to increase the robustness and generalization of CNN-based detectors. Data augmentation, super-resolution reconstruction, and pixel-level anomaly localization have been employed to train models on a wider range of fake patterns. Additionally, loss functions like Maximum Mean Discrepancy (MMD) have been used to align feature distributions between real and fake data, minimizing the risk of overfitting to a specific dataset.

To improve model interpretability and focus, attention mechanisms have been integrated into CNN pipelines, enabling the networks to concentrate on the most informative facial regions, such as the eyes, mouth, or jawline [66]. Furthermore, Capsule Networks (CNs) [67] have been explored as an alternative to traditional CNNs. These models provide improved spatial awareness and preserve part-whole relationships, while requiring fewer parameters, thus enhancing both efficiency and performance.

Finally, ensemble learning has been applied to further enhance CNN-based detection accuracy. By combining the outputs of multiple CNN models trained on different features or with different architectures, ensemble methods have achieved detection rates exceeding 99 percent on several benchmark datasets, confirming the continued relevance and adaptability of CNNs in deepfake detection pipelines [68].

RNN based architectures

Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) variants, have shown notable success in deepfake detection tasks involving video data. Unlike CNNs, which primarily focus on spatial features in individual frames, RNNs

are specifically designed to model temporal dependencies, making them well-suited for capturing the motion dynamics and continuity inherent in real video sequences.

One of the key advantages of RNN-based architectures lies in their ability to analyze frame-by-frame temporal evolution of facial expressions, eye blinking, lip synchronization, and head movements. These features often exhibit subtle temporal inconsistencies in deepfake videos due to the difficulty generative models face in maintaining coherent motion patterns across consecutive frames. For instance, while a GAN might produce a visually convincing single frame, it may fail to generate natural transitions between expressions or maintain temporal consistency in eye blinks and gaze direction, patterns that RNNs can effectively learn to recognize.

Several works have leveraged RNNs in combination with CNNs in a two-stage pipeline. In these hybrid architectures, CNNs first extract spatial features from each video frame, which are then passed to an RNN module that captures their temporal relationships. This approach benefits from the strengths of both network types and has led to improved performance in distinguishing real videos from manipulated ones. Some systems also integrate optical flow information or motion vectors as input to the RNNs, enabling the detection of irregular motion patterns and artifacts caused by frame interpolation errors.

Another notable application of RNNs in deepfake detection is the analysis of micro-expressions and subtle temporal patterns that may be imperceptible to human observers but inconsistent across synthesized frames. These include minor changes in muscle tension, skin deformation, or blinking frequency. When trained on large and diverse datasets, RNNs can learn these temporal features as reliable indicators of authenticity. [58]

Despite their effectiveness, RNN-based approaches face challenges related to overfitting, especially when trained on datasets that lack variability in pose, lighting, or demographic representation. To mitigate this, some researchers have implemented data augmentation strategies, while others have explored the use of autoencoder architectures in conjunction with RNNs. These models aim to reconstruct expected facial motion sequences, flagging deviations as potential manipulations. [58]

Furthermore, enhancements like triplet loss functions, temporal feature regularization, and adversarial training have been introduced to increase the generalizability and discriminative power of RNN-based models. These techniques encourage the network to learn more robust temporal embeddings that can distinguish between authentic and manipulated content even when faced with previously unseen deepfake methods.

Overall, RNNs represent a crucial tool in the deepfake detection landscape, particularly for applications that require fine-grained temporal analysis. As deepfake generation methods continue to evolve toward producing more temporally consistent outputs, the role of RNNs, and their integration with other architectures, remains fundamental for detecting nuanced inconsistencies that betray synthetic content.

Conclusions on deep learning based methods

These contributions demonstrate that deep learning has significantly advanced the field of deepfake detection. While the performance of these models is undeniably impressive, often exceeding 99 percent accuracy in controlled benchmarks, challenges remain regarding cross-dataset generalization, robustness against adversarial attacks, and explainability of predictions. As the generation techniques evolve, detection frameworks must also adapt, emphasizing the need for continual refinement of model architectures and training methodologies.

3.1.3 Statistical measurements based methods

In addition to deep learning-based techniques, statistical-based methods have also demonstrated promising results in the detection of Deepfakes, particularly through the analysis of inherent patterns and deviations within digital content. These approaches typically rely on the measurement and interpretation of various statistical features that are either preserved or altered during the Deepfake generation process.

PRNU

One of the most notable contributions in this field is the study by Koopman et al. [69], which explores the use of photo-response non-uniformity (PRNU) for detecting manipulations in video frames. PRNU is an intrinsic noise pattern embedded in digital images, caused by subtle imperfections in the light-sensitive sensors of a camera. Due to its uniqueness, PRNU serves as a sort of fingerprint of the device used for capturing the image. In their method, the authors generate a sequence of frames from input videos, systematically categorizing and processing them. Each frame is cropped consistently to isolate and enhance the PRNU pattern. These cropped frames are subsequently divided into eight groups, and a standard PRNU pattern is estimated for each using the second-order Fast Spatial Transform Vector (FSTV) method. Normalized cross-correlation scores are then computed to quantify the similarity between PRNU patterns, followed by a statistical t-test to determine the significance of discrepancies between Deepfake and authentic videos [58].

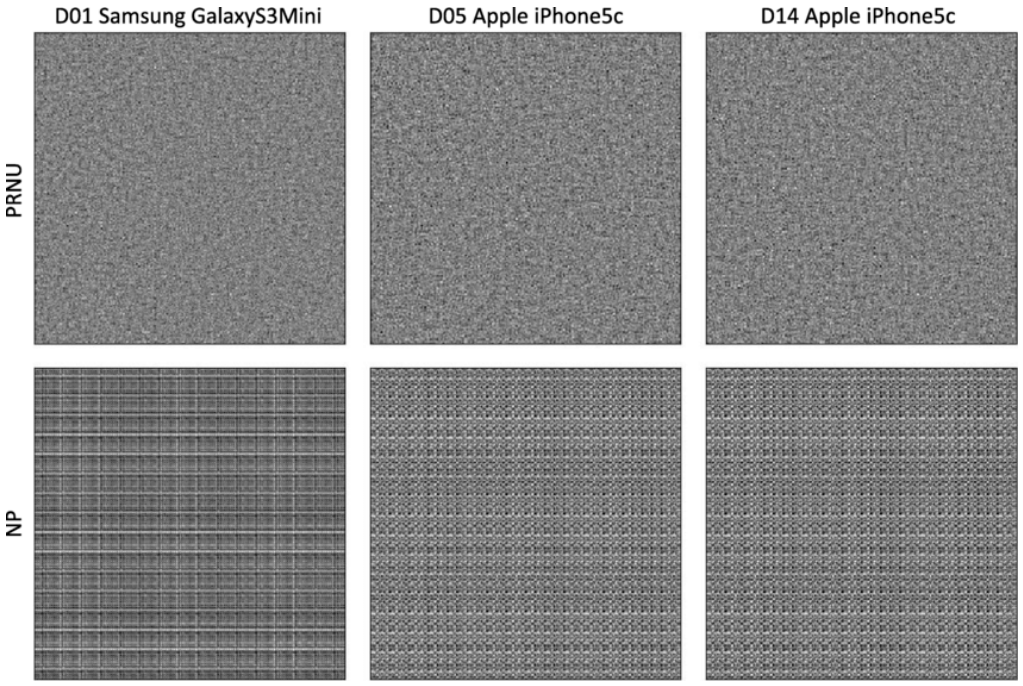


Figure 3.4. PRNU and NP (noise print) (fonte: ResearchGate).

Regional statistical features

Further research in this domain has focused on modeling the generative process underlying Deepfakes through the extraction of regional statistical features. In [70], the authors employed the Expectation-Maximization (EM) algorithm to identify characteristic regions across different generative adversarial networks (GANs), including GDWCT, STARGAN, ATTGAN, STYLEGAN, and STYLEGAN2. These extracted features were subsequently validated using naive classification schemes in preliminary experiments, allowing for the assessment of statistical anomalies that arise from the generation process.

Statistical distance measures

A particularly innovative approach is presented by Agarwal et al. [71], who proposed a hypothesis-testing framework for detecting Deepfakes based on statistical distance measures. This framework

defines and calculates the shortest path between the statistical distributions of real and GAN-generated images. The key insight is that the detectability of a Deepfake correlates with the statistical divergence between these distributions: when the distance is large, the detection becomes easier, as it implies greater inconsistency between synthetic and genuine content. Conversely, a well-trained GAN capable of producing high-fidelity images can reduce this distance, thereby increasing the challenge of detection. This observation underscores the escalating complexity of identifying Deepfakes as generative models improve in accuracy and realism. [58]

Conclusions on statistical measurements based methods

Ultimately, statistical-based detection methods provide a complementary perspective to learning-based approaches, as they offer a principled way to quantify deviations from natural data distributions. They are particularly useful in scenarios where the generative pipeline leaves behind subtle yet consistent artifacts that can be revealed through statistical analysis, and they highlight the ongoing arms race between detection strategies and the sophistication of generative technologies.

3.2 Challenges in deepfake detection methods

Despite the significant advancements in the performance of deepfake detectors, the field still faces numerous unresolved issues that hinder the robustness and generalizability of detection systems. This section outlines the main challenges that currently affect deepfake detection methodologies, based on recent literature and empirical observations.

3.2.1 Limitations of deepfake datasets

The availability of large-scale datasets is fundamental for developing and evaluating deepfake detection techniques. However, a critical analysis of these datasets reveals substantial limitations compared to real-world manipulated content. Common artifacts found in synthetic datasets include temporal flickering during speech, blurriness and over-smoothness in facial regions, lack of head pose variations, absence of occlusions, inconsistencies in gaze or skin tone, and limited diversity in audio-visual pairings. These imperfections are due to the manipulation process itself and result in unrealistic content that is easier to detect. Consequently, even when detection models perform well on these datasets, they may fail in real-world scenarios, where the quality and sophistication of manipulations are much higher. [72, 73]

3.2.2 Performance evaluation and labeling issues

Most current detection methods frame the task as a binary classification problem: real or fake. While this is effective in controlled experimental settings, it does not align with the complexity of real-world cases. For instance, a video might be manipulated in only some frames or might involve multiple forms of tampering (visual and audio). Moreover, videos may include several faces, only some of which are deepfaked. In such contexts, binary labeling is insufficient and can lead to inaccurate results. Therefore, there is a need for more granular detection techniques, such as multi-class or multi-label approaches and localized, frame-level analysis. [72, 73]

3.2.3 Model scalability and inference time

Another major challenge is the scalability of detection models, particularly for high-volume platforms like social media. High detection accuracy is of limited use if the inference time is too long, especially when processing large amounts of content in real time. Many current models are computationally expensive and impractical for deployment on large-scale platforms. Future work must focus on developing lightweight, real-time detection methods without compromising performance. [72]

3.2.4 Lack of explainability

Most deepfake detection models, especially those based on deep learning, operate as black boxes, providing little to no explanation for their outputs. In critical applications, such as journalism or law enforcement, the numerical probability that a video is fake is not sufficient unless it is accompanied by interpretable evidence. Without transparency, these results may not be admissible or credible in judicial or investigative contexts. Hence, integrating explainable AI (XAI) techniques into detection frameworks is an urgent research direction. [72]

3.2.5 Bias, fairness, and trust

Deepfake datasets and the detection models trained on them often exhibit demographic biases, particularly concerning race and gender. This imbalance can lead to unfair and unreliable detection outcomes, especially when applied to underrepresented groups. Although research on fairness in deepfake detection is emerging, it remains limited, and existing systems may perpetuate or amplify existing societal biases (source text, p.4). Fairness-aware training strategies and balanced datasets are essential to build trust in deepfake detection technologies. [72]

3.2.6 Temporal inconsistencies and aggregation

Many current approaches evaluate frames individually without considering temporal coherence. However, deepfakes often display temporal artifacts that can be detected only by analyzing sequences of frames. Moreover, isolated frame-level predictions must be aggregated to compute an overall integrity score for the video, which adds complexity and may introduce inaccuracies. Advanced techniques that incorporate temporal modeling, such as recurrent neural networks or temporal attention mechanisms, may offer more robust solutions. [72]

3.2.7 Impact of social media laundering

When videos are uploaded to platforms such as Twitter, Instagram, or Facebook, they undergo compression, down-sampling, and metadata removal, a process known as social media laundering. These transformations obscure manipulation traces and reduce the efficacy of detection models that rely on low-level signal features. To improve robustness, training datasets must simulate such post-processing effects, and evaluation benchmarks should include content altered by social media platforms. [72]

3.2.8 Lack of diverse audio deepfake datasets

While visual deepfake detection has benefited from extensive datasets, the audio domain remains underdeveloped. For instance, the ASVspoof-2021 dataset lacks specific training data for audio deepfake detection, and other datasets are limited to a single speaker. This lack of diversity hinders the ability of models to generalize to real-world audio manipulations. There is a pressing need for more comprehensive and realistic datasets for the evaluation of synthetic speech detection systems. [72, 73]

3.2.9 Evasion and adversarial attacks

Attackers have developed methods to bypass deepfake detection systems by eliminating detectable artifacts. These evasion techniques include adversarial perturbations (noise injection, cropping, JPEG compression), manipulation of frequency-domain features, and the use of advanced image filtering to conceal synthetic traces. Research has shown that such attacks can significantly reduce the performance of state-of-the-art detectors. Therefore, future detection models must be resilient to such adversarial strategies through robust training, adversarial defense mechanisms, and redundancy in detection cues. [72, 73]

3.3 Tools for deepfake detection

As deepfake technologies continue to advance in realism and accessibility, the need for effective detection mechanisms has become increasingly critical. Deepfake detection tools are designed to identify manipulated media using a combination of computer vision, machine learning, and forensic analysis techniques. These tools play a crucial role in mitigating the risks posed by malicious deepfakes in areas such as politics, journalism, law enforcement, and digital media authentication. Detection tools aim to identify synthetic content by analyzing subtle artifacts left by manipulation processes, inconsistencies in facial landmarks or lighting, or discrepancies in biometric patterns such as eye movement or pulse signals. [33]

Sensity AI is a pioneering platform that uses deep learning models trained on large datasets of real and fake media to automatically detect manipulated content. It is widely used by social media platforms and governments to monitor disinformation campaigns and synthetic media threats. Features: Deepfake video and image detection, Scans large datasets efficiently, Integrated APIs for enterprise use

Truepic provides media integrity verification by applying cryptographic hashing at the point of capture, ensuring that no tampering has occurred post-creation. It is used in journalism, insurance, and legal proceedings. Strengths: Tamper-evident media capture, Forensic analysis of image metadata, Blockchain integration for auditability

FakeCatcher uses a biological signal-based approach, analyzing fluctuations in blood flow on the face to identify synthetic content in real time. This is one of the few systems to use physiological indicators instead of visual artifacts. Strengths: Real-time detection, Robust to post-processing (compression, filters), Effective even with realistic deepfakes

Microsoft Video Authenticator, developed in the context of disinformation threats, analyzes videos and images and assigns a confidence score to each frame, indicating the likelihood of manipulation. Use case: Political video authentication, Integration with fact-checking workflows

Deepware Scanner is a lightweight application allowing users to scan local or online video files for signs of deepfake manipulation. While it does not always offer detailed analytics, it provides a quick assessment accessible to non-expert users. Strengths: Cross-platform (mobile and desktop), Fast initial scanning, Simple user interface.

Table 3.1. Comparison of Deepfake Detection Tools

Tool	Description	Accuracy	Speed	Usability	Availability
Sensity AI	AI-based detection platform for images and videos, used by media and governments	High (~95%)	High	High	Commercial
Truepic	Verifies media authenticity via cryptographic and forensic analysis	High	Moderate	High	Commercial
D-ID	Protects privacy by anonymizing facial features in images/videos	Moderate	High	High	Commercial
FakeCatcher (Intel)	Detects deepfakes via physiological signals such as blood flow	High	High	Moderate	Proprietary
Microsoft Video Authenticator	Provides frame-level manipulation probability score	Moderate	Moderate	Moderate	Commercial
Deepware Scanner	App for mobile/desktop to detect deepfakes in uploaded videos	Moderate	High	High	Freemium

3.4 The role of users and companies

3.4.1 User education and awareness

Public education and awareness are essential pillars in the fight against the dissemination of manipulated media. As previously discussed, fostering a critical understanding of how digital content can be altered is crucial in helping individuals evaluate the authenticity of the information they encounter online. A pressing question emerges: how can we effectively build a comprehensive

awareness framework to educate society about this evolving threat? One approach involves offering accessible technical training, through online courses or open resources, that introduces the foundational concepts of technologies used to generate deepfakes. Equally important is the promotion of critical thinking skills by providing real-world examples, regular updates via newsletters, and public engagement with available detection tools. Raising awareness also means equipping individuals with practical skills to recognize potential signs of synthetic media. These may include detecting unnatural eye movements or blinking patterns, which are often imperfectly replicated in deepfakes, and observing the clarity and symmetry of facial expressions, skin texture, or facial hair, which may appear distorted or inconsistent. Similarly, anomalies in the rendering of teeth and hair, such as overly smooth or excessively bright textures, may signal artificial reconstruction. In video analysis, discrepancies between audio and lip movements are particularly telling, as are visible visual distortions, pixelation, or misplaced shadows. Importantly, attention must also be given to the source of the content, assessing its reliability and reputation to further gauge the credibility of the material. Addressing the deepfake challenge will require a coordinated, multidisciplinary effort involving governments, private companies, non-governmental organizations, and individual users. Continued research, the advancement of defensive technologies, and the widespread promotion of digital literacy must all be integrated into a global strategy to counter the manipulation of digital information effectively. [27]

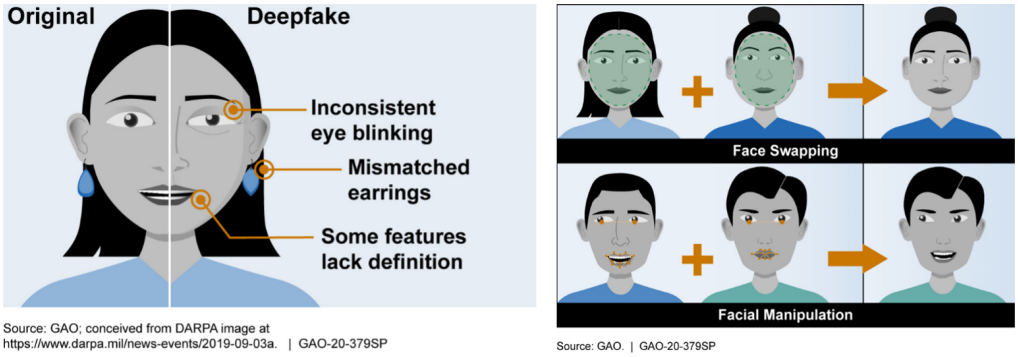


Figure 3.5. Main signs of a deepfake (fonte: MIAMI INFORMATION TECHNOLOGY).

3.4.2 Best practices for companies

As deepfake technology continues to advance, it becomes increasingly essential for organizations to adopt comprehensive strategies to detect and mitigate the risks it poses. One of the most promising approaches involves the use of advanced detection systems powered by artificial intelligence and machine learning. These technologies are capable of analyzing audio and visual content in depth, identifying subtle inconsistencies that may escape human perception, and thus enabling timely recognition of manipulated media. AI-based tools employ sophisticated pattern recognition techniques to detect anomalies, while multimodal analysis integrates visual, auditory, and metadata signals to assess the overall authenticity of the content. In addition to these methods, blockchain technology is being explored to verify the origin and integrity of digital files, adding a layer of trust through decentralized and tamper-resistant records. To strengthen their defenses, organizations must also adopt a holistic approach that combines these technical tools with strategic practices. This includes implementing strong authentication mechanisms, establishing verification protocols for sensitive communications, ensuring regular updates of detection systems, and fostering a culture of awareness through employee training. Furthermore, integrating watermarking and digital signatures into original content can support provenance tracking and discourage malicious alterations. Collaborating with cybersecurity experts and research institutions is equally vital to stay ahead of evolving threats. Finally, well-defined incident response plans should be in place to ensure swift action in the event of suspected deepfake attacks. By aligning advanced technological solutions with sound organizational policies, it is possible to build a resilient framework capable of addressing the growing challenge posed by synthetic media. [24]

3.5 Deepfakes authentication methods

3.5.1 Blockchain based methods

Understanding blockchain

Before delving into its application for deepfake authentication, it is essential to understand what Blockchain is and how it functions. At its core, Blockchain is a decentralized digital ledger that records transactions across a network of computers in a secure, transparent, and immutable manner. Each transaction is grouped into a block and added sequentially to a chain of previous records, forming an unalterable chronological history. This structure eliminates the need for a central authority, as all network participants have access to the same version of the data, which is verified through consensus mechanisms. The immutability and transparency provided by this architecture make Blockchain particularly well-suited for scenarios where trust, integrity, and provenance of data are crucial, such as in the fight against manipulated media.

In the broader landscape of Deepfake detection, blockchain-based methods are emerging as an innovative and promising direction. While still in the early stages of development compared to deep learning or statistical techniques, Blockchain technologies offer unique features that make them particularly suitable for ensuring the authenticity and traceability of digital content. Their decentralized, transparent, and tamper-proof nature allows for the verification of the origin and integrity of multimedia files, which is especially valuable in the context of maliciously manipulated content. The foundational premise of blockchain-based detection approaches lies in the concept of verifiable provenance. By associating digital content, such as images or videos, with immutable transaction records, it becomes possible to trace the history of a file and verify whether it originated from a trusted source. In this context, public blockchains are especially relevant, as they allow open access to historical transactions and provide a resilient infrastructure for content validation in a decentralized setting. [58]

Hasan and Salah approach

Hasan and Salah [74] have proposed one of the first generic frameworks applying Blockchain to the Deepfake detection problem. Their solution leverages the transparency and immutability of public Blockchains to trace suspicious video content back to its original source. Even when the digital material has been copied or modified multiple times, the framework enables the reconstruction of its transaction history. The central idea is that a piece of content should be considered authentic only when it can be convincingly linked to a legitimate and trusted origin. Their architecture integrates key Blockchain mechanisms to manage and monitor interactions among users and content, ensuring that authenticity proofs are anchored to a verified source. Furthermore, they combine Blockchain with InterPlanetary File System (IPFS) storage, which supports decentralized file hosting, and employ the Ethereum Name Service to facilitate the resolution and identification of content sources.

Chan at al. approach

Building on a similar philosophy, Chan et al. [75] introduced a more technically sophisticated Blockchain-based approach for tracking the historical provenance of digital content. In their proposal, multiple LSTM-based convolutional neural networks are used to encode and extract discriminative features from images and videos. These high-dimensional features are then compressed into a binary-coded structure and hashed to create a unique transaction record, which is stored on a permissioned Blockchain. Unlike public Blockchains, permissioned systems allow for access control, giving content owners full governance over their data and its provenance trail. This ensures a higher level of privacy and security, particularly relevant in sensitive domains such as journalism or digital evidence management.

Conclusions on blockchain based methods

Although the number of studies exploring Blockchain-based Deepfake detection is currently limited, only 2 percent of the surveyed research according to the authors of the study, these methods highlight a significant shift in the conceptualization of authenticity verification. Rather than relying exclusively on post-hoc analysis of content features, Blockchain approaches aim to secure the origin and evolution of content from the moment of its creation, thereby preventing the proliferation of manipulated media in the first place. In conclusion, Blockchain-based methods offer a complementary path to traditional detection strategies by shifting the focus from identifying manipulations to guaranteeing the verifiable integrity of digital content. As the technology matures and more comprehensive frameworks are developed, its integration with other detection techniques could provide a robust defense mechanism against the growing threat of Deepfakes.

3.5.2 Watermarking for authentication

Digital watermarking is a technique for embedding imperceptible information into digital media content, such as images, audio, or video, in order to provide proof of authenticity, ownership, or integrity. Unlike metadata, which can be easily removed or modified, watermarks are embedded directly into the content and are designed to be resistant to various forms of signal processing. The information carried by a watermark may include content identifiers, timestamps, cryptographic signatures, or even content-dependent hashes. Depending on the goal, watermarking schemes can be designed to be either robust, able to withstand typical signal transformations such as compression or scaling, or fragile, designed to be sensitive to even the slightest modification, thus serving as a tamper-detection mechanism. A fundamental distinction exists between visible and invisible watermarks. While visible watermarks are used primarily for copyright notification (e.g., logos on images), invisible watermarks are central to applications involving content authentication and secure communication. These invisible watermarks can be blind (i.e., recoverable without access to the original media) and can serve various roles such as verifying content integrity, tracking the source of unauthorized copies, and confirming the authenticity of digital media in legal and forensic contexts. In recent years, watermarking has been explored as a proactive defense mechanism against synthetic media manipulation, particularly deepfakes, which are increasingly difficult to detect with the naked eye or even with traditional machine learning techniques. In this context, watermarking serves not merely as a protective mechanism but as an active authentication strategy, allowing stakeholders to verify whether content has been tampered with or artificially generated.[76]

The system proposed by Qureshi, Megias, and Kuribayashi

The system proposed by Qureshi, Megias, and Kuribayashi represents a concrete and innovative application of digital watermarking for deepfake detection, marking one of the first attempts to combine active watermark-based authentication with blockchain immutability. In their 2021 study titled Detecting Deepfake Videos using Digital Watermarking [77], the authors developed a hybrid method that embeds both robust and fragile watermarks into the audio stream of a video, targeting manipulations such as voice impersonation and lip-syncing. The robust watermark, constructed using Discrete Wavelet Transform (DWT) and Discrete Cosine Transform (DCT), encodes metadata like the perceptual hash of the video ID and copyright information, allowing consistent identification even after compression or noise addition. Conversely, the fragile watermark, derived from Mel-Frequency Cepstral Coefficients (MFCCs) and facial features extracted via Multi-task Cascaded Convolutional Neural Networks (MTCNN), is designed to detect tampering. Its sensitivity ensures that any alteration of the audio-visual correspondence disrupts the embedded hash comparison, thereby signaling manipulation. To strengthen verifiability, all watermark-related metadata are recorded in a blockchain ledger, ensuring immutability and decentralized integrity verification.

While this proof-of-concept demonstrates a powerful synergy between watermarking and blockchain technologies, its real significance lies in the broader implications for scalable, trustworthy media

authentication systems. The work highlights the potential of active detection mechanisms, embedding verifiable information into content itself, over purely AI-based passive classifiers, which risk obsolescence as generative models evolve. However, several critical considerations emerge when envisioning large-scale deployment. First, the computational complexity of embedding and extracting dual watermarks in every video segment raises questions about scalability and real-time feasibility, particularly for platforms managing vast multimedia streams. The trade-off between robustness and imperceptibility also remains delicate: higher embedding strength enhances resistance to attacks but may degrade perceptual quality or introduce detectable artifacts.

Another challenge concerns dependence on the original embedding process. The method presupposes that trustworthy content producers will watermark their media before distribution; yet, widespread adoption hinges on user cooperation and compatible recording or editing devices. In scenarios where only one modality, audio or video, is available, as in silent clips or voice-only media, the system's detection capacity might fail or produce false negatives, exposing its vulnerability to modality loss. From a threat-model perspective, the approach appears effective against conventional manipulations like noise addition or compression, but it could be undermined by deliberate adversarial strategies such as recompression, re-sampling, or spatial cropping that partially remove or distort the watermark while preserving perceptual coherence. This highlights the need for a more systematic evaluation of resilience under diverse and adaptive attack conditions.

Beyond technical performance, the study also raises important reflections about the role of watermarking within a broader ecosystem of digital content authentication. Watermarking alone cannot ensure full provenance or authorship verification; rather, it should be seen as one component within a multilayered trust architecture that includes metadata signatures, secure capture hardware, and verified provenance chains. In this sense, the authors' integration of blockchain is not merely an implementation choice but an essential conceptual step toward distributed accountability, where watermark traces serve as locally verifiable evidence anchored in a global, tamper-resistant ledger. This approach aligns with the emerging vision of multi-factor media authentication, combining cryptographic, perceptual, and infrastructural safeguards, to counter the escalating sophistication of synthetic content.

In summary, the contribution of Qureshi et al. lies not only in their technical implementation but in articulating a new paradigm for deepfake detection: one that embeds integrity into the medium itself while leveraging distributed verification to preserve public trust in digital media. Despite its experimental nature and current limitations in scalability and generalization, their work provides a crucial foundation for the development of hybrid, explainable, and resilient systems for multimedia authentication in real-world contexts.

3.5.3 A comprehensive survey on robust image watermarking

More recent comprehensive reviews demonstrate that digital watermarking is increasingly viewed not simply as a copyright protection tool but as a proactive component of media authentication and integrity protection. For instance, the 2022 survey by Wan et al. [78] on robust image watermarking offers a systematic appraisal of frequency-domain, transform-domain and deep-learning-driven embedding methods, noting how traditional trade-offs (robustness versus imperceptibility versus capacity) remain central to design decisions. This survey highlights that while many embedding algorithms show good resistance to standard distortions (e.g., compression, additive noise, geometric transforms), far fewer address real-world deployment challenges: high-volume streaming, heterogeneous devices, varying network conditions, or adaptive adversaries. In other words, although the embedding techniques are becoming more sophisticated, their scalability and operational cost remain under-explored. Embedding and extracting watermarks at scale, in live video platforms, or on devices with constrained compute and battery budgets, raises questions about latency, throughput, and the user-experience impact (e.g., perceptual degradation or processing overhead). Moreover, user adoption becomes more difficult if watermarks alter perceptual quality, require custom capture devices, or rely on centralized infrastructures. These observations suggest that the mere technical feasibility of watermarking does not automatically translate into practical, large-scale deployment.

Beyond technical embedding considerations, the surveys also emphasise the need for a holistic threat-model awareness and multi-layer authentication architectures. A more recent review on

proactive forensic techniques with watermarking emphasises that passive detection (i.e., purely AI-based classification of manipulated media) is increasingly vulnerable to adversaries who adapt generative models or apply counter-forensic transformations. In contrast, proactive watermarking provides embedded defence by carrying authentication markers within the media itself, enabling detection of tampering even when generative models evolve. Yet this approach is not without limitations: it presupposes that the watermark embedding is trusted, protected, and remains accessible; if an attacker intentionally recrops, recompresses, temporally shuffles, or merely strips the audio/video channel containing the watermark, the scheme may fail. Such vulnerability emphasises the dependency on the original embedding environment, the single-modality risk (if only audio or only video is watermarked, the other channel becomes a vector for attack), and the cost of maintaining a secure chain from capture to embedding to verification. The surveys therefore argue convincingly that watermarking must be integrated in a multifactor authentication ecosystem, for instance combining hardware-trusted capture devices, metadata signing, secure provenance chains and blockchain or ledger anchoring, rather than operating as a standalone solution. This shift from isolated watermarking to systemic authentication highlights both the promise and the caution required for real-world adoption.

Chapter 4

Project

4.1 Linking legal principles to technical implementation

Building upon the legal and ethical foundations outlined in the second chapter, this section presents the practical implementation of a system designed to meet the technical requirements emerging from current regulatory frameworks on artificial intelligence and deepfakes. The principles enshrined in the AI Act, the Digital Services Act, and the recent Italian DDL on deepfake offences, along with corresponding measures in the United States, China, and the United Kingdom , reflect an increasing global recognition of the need for technological mechanisms capable of ensuring trust, transparency, and accountability in digital content production and dissemination. These legislative initiatives converge on several common objectives: guaranteeing that AI-generated material can be identified and traced; preserving the integrity of digital evidence; and establishing verifiable chains of custody that allow attribution and provenance to be demonstrated beyond reasonable doubt.

While the normative landscape defines the what (transparency, integrity, accountability, and content provenance) it is the technical layer that defines the how. The effective realization of these legal imperatives relies on infrastructures capable of authenticating data at its origin, detecting any manipulation, and maintaining an immutable record of its lifecycle. This necessitates an interplay between law and technology, where cryptographic and forensic principles are embedded within digital architectures to transform abstract legal duties into verifiable computational procedures.

The proposed framework operationalizes these requirements through a combination of digital watermarking, cryptographic hashing, and a blockchain-inspired recording mechanism. Digital watermarking, implemented through the Least Significant Bit (LSB) technique, allows the insertion of imperceptible identifiers within media files, enabling both transparency and origin authentication. Cryptographic hashing guarantees that even the slightest unauthorized alteration of the image or its metadata is immediately detectable, reinforcing the principles of integrity and non-repudiation that underpin digital evidence admissibility. Finally, the blockchain simulation module records each content instance and links it to the preceding one through hash chaining, thereby creating an immutable, traceable ledger that upholds accountability and verifies content provenance.

In this sense, the project represents a tangible synthesis between legal mandates and technological enforcement, demonstrating how abstract regulatory concepts can be transformed into a concrete, verifiable framework for digital trust. By bridging the gap between compliance and computation, the system provides a technical realization of the legal principles articulated in the preceding chapters , ensuring that authenticity, integrity, and accountability are not only prescribed by law, but also implemented and measurable within the digital domain.

The table below summarizes the correspondence between the discussed legal principles and their respective technical implementations within the proposed framework, illustrating how regulatory requirements are concretely translated into system functionalities.

Table 4.1. Correspondence between legal principles and technical implementations

Regulatory principle	**Derived technical requirement**	**Implemented mechanism**
Transparency (AI Act, DSA, China's Measures)	Clear identification of AI-generated or manipulated content	Digital watermarking (LSB embedding) for visible and hidden content labeling
Accountability (AI Act, Italian DDL, UK Online Safety Act)	Attribution of content to a responsible entity and prevention of unlawful dissemination	Blockchain-based structure ensuring non-repudiation and traceable authorship
Integrity of digital evidence (EU Evidence Directives, U.S. Deepfakes Accountability Act)	Detection of any alteration or tampering of digital content	Cryptographic hashing of images and associated metadata
Chain of custody (Forensic and evidentiary principles)	Immutable record of content lifecycle and provenance verification	Sequential blockchain linking through hash references
Content provenance (AI Act, Chinese Deep Synthesis Regulations)	Verification of content's lawful and authentic origin	Combination of watermarking, metadata hashing, and blockchain registration

4.2 Base idea of the project

In a digital era increasingly characterized by sophisticated forms of media manipulation, the need for reliable mechanisms to verify the authenticity of digital content has become imperative. This project was conceived as a response to this challenge, proposing an integrated solution that combines digital watermarking and blockchain technology to ensure the authenticity and traceability of images. The system aims to provide users with a verifiable method to determine whether an image is genuine or has been manipulated, thus addressing growing concerns related to deepfakes and AI-generated content.

The conceptual foundation of the proposed system rests on two complementary technologies. First, digital watermarking enables the embedding of authenticity information directly into an image at the moment of its creation or acquisition. This watermark acts as an intrinsic digital signature, imperceptible to the human eye yet detectable by dedicated algorithms, thereby serving as tangible proof of originality and provenance. Second, blockchain technology is employed to record immutable and verifiable data related to the watermarked image, such as its hash and associated metadata. By storing this information on a decentralized ledger, the system ensures transparency and tamper resistance: each image registration is cryptographically linked to previous entries, preventing retroactive modifications and guaranteeing the reliability of the stored information.

From a conceptual standpoint, this dual approach aligns with recent developments in the state of the art, particularly those discussed in the research by Hasan and Salah [74], Chan et al. [75], and Qureshi, Megias, and Kuribayashi [77]. The proposed framework shares with these works the goal of building a trust infrastructure for digital media, where authenticity is not merely inferred through post-hoc analysis but guaranteed through cryptographic and infrastructural safeguards established at the point of content creation.

Compared to Hasan and Salah's blockchain-based framework, which focuses primarily on the transparent tracing of video provenance, the present project adopts a similar philosophy but applies it to the domain of static imagery. It simplifies the complexity of multi-layer blockchain interactions while maintaining the essential principles of transparency, immutability, and decentralized verification. Unlike Chan et al.'s approach, which relies on neural feature extraction and storage within permissioned blockchains, this system emphasizes accessibility and general applicability, opting for a more lightweight implementation suitable for public or open environments. This design choice enhances usability and scalability, albeit at the cost of more limited control over access and governance compared to permissioned networks.

Furthermore, the system conceptually resonates with the hybrid watermarking approach proposed by Qureshi et al. [77], who combined robust and fragile watermarks with blockchain immutability to detect manipulations in audiovisual content. Like their study, the present work leverages watermarking as an active authentication mechanism, embedding verifiable information within the media itself rather than depending solely on external verification. However, while Qureshi et al. focused on dual watermarking for multimodal (audio - video) synchronization and tampering detection, the present project concentrates exclusively on visual data, allowing for a simpler yet effective implementation that demonstrates the feasibility of the concept in a specific media type.

From a critical perspective, the proposed system can be regarded as a proof-of-concept that bridges theoretical models and practical feasibility. Its main strength lies in the integration of two complementary mechanisms, digital watermarking and blockchain registration, each addressing a distinct aspect of the authenticity verification process: the first guarantees content integrity at the perceptual level, while the second secures its traceability and provenance at the infrastructural level. Nevertheless, some of the challenges identified in the literature also apply here. For instance, large-scale deployment may be hindered by computational overheads related to watermark embedding and extraction, as well as by user dependency on trusted capture or editing environments. Moreover, as highlighted in recent surveys [78], the balance between robustness, imperceptibility, and scalability remains delicate, particularly when considering real-time or high-volume applications.

Despite these limitations, the project contributes to the ongoing research discourse by demonstrating a tangible implementation of a hybrid authentication paradigm. It exemplifies how the combination of watermarking and blockchain can provide a practical and explainable approach to media authenticity verification, one that not only detects manipulation but also proactively embeds trust and accountability into the content itself. In this sense, the system aligns with the evolving vision of multi-factor and proactive authentication frameworks, representing a meaningful step toward the realization of resilient, transparent, and scalable mechanisms for combating digital media falsification.

4.3 Details of the project

4.3.1 Watermarking

The implementation of digital watermarking constitutes a pivotal element of the project, serving as the mechanism by which images are marked with a unique identifier to attest to their authenticity and provenance. Designing an effective watermarking system requires a careful choice between different strategies, each with its own peculiarities in terms of robustness, invisibility, and ability to detect manipulations. In general, the main types of digital watermarking can be classified as fragile or robust.

Fragile Watermarking

This type is designed to be extremely sensitive to any modification of the image. Even minimal alterations cause the watermark to "break" (become undetectable). It is ideal for verifying content integrity and authenticity, allowing even small manipulations to be detected. Typically, these watermarks are embedded in the Least Significant Bits (LSB) of the image. Its primary advantages lie in its extreme sensitivity, which makes it excellent for detecting manipulations, and its relative ease of implementation (for example through the LSB approach). However, its disadvantages are evident: it does not withstand common compressions, filters, or resizing, making it less suitable for content that needs to circulate or undergo standard editing operations.

Robust Watermarking

Unlike fragile watermarking, this variant is designed to survive common and expected modifications an image may undergo, such as JPEG compression, resizing, or the application of filters.

It is particularly useful for protecting content intended for public circulation, ensuring that the identifier persists even after non-malicious alterations. Embedding usually occurs in the transform domain (for example DCT Discrete Cosine Transform, DWT Discrete Wavelet Transform, SVD Singular Value Decomposition), distributing the watermark redundantly across the entire image. Its advantages include resistance to common modifications and effectiveness in protecting publicly distributed content. On the other hand, it presents disadvantages such as greater implementation complexity, less utility for granular integrity verification (as it survives modifications), and potentially less invisibility.

Hybrid solution

This approach combines the strengths of the two previous methodologies, employing two separate watermarks: a fragile one (e.g., in the LSB) for fine integrity verification and a robust one (e.g., in the DCT domain) to ensure the persistence of the identifier. The advantages of a hybrid solution lie in the effective compromise it offers, ideal in scenarios where both authenticity and content persistence are desired. However, it requires greater design and balancing complexity, as well as a more articulated extraction and verification logic.

My choice

My current choice has leaned towards fragile watermarking. This decision is primarily driven by the time and resource limitations available for the prototype's development. A fragile watermark has proven to be the most accessible and quickest option to implement, while simultaneously providing excellent capabilities for detecting even the smallest manipulations on the image, a fundamental requirement for integrity verification.

While acknowledging the limitations of a purely fragile approach (particularly its susceptibility to compression and filters), the possibility of integrating a robust watermark component in the DCT domain is left as a future direction. The long-term goal is to converge towards an optimal hybrid solution, combining the precise manipulation detection of fragile watermarking with the resistance to common alterations of robust watermarking, thereby offering an even more comprehensive and versatile authentication system.

In forensic and legal terms, this mechanism reinforces authenticity verification by offering a built-in integrity indicator that complements hashing. The watermark therefore serves as a preventive evidentiary safeguard, allowing experts to demonstrate whether and how a piece of evidence was modified after its acquisition.

4.3.2 Hashing

Within the architecture of our image authentication system, the hashing process plays a role of paramount importance. Hashing is a cryptographic technique that transforms an input of any size (in our case, an image) into a fixed-length string of characters, commonly called a "hash" or "digest". The key characteristic of a good hashing function is that it is unidirectional (it's impossible to reverse engineer the input from the hash) and highly sensitive: even a minimal modification in the input generates a completely different hash. This feature makes it ideal for data integrity verification. In the context of our project, hashing is employed for two main purposes, directly linked to the security and authenticity of the watermarked image: immutable digital fingerprint and blockchain linkage.

Immutable digital fingerprint: the hash of the watermarked image serves as a unique and unalterable "digital fingerprint". Once calculated, this hash represents the exact state of the image at that precise moment. If the image were to undergo even the slightest modification after hashing, the recalculated hash would be different, immediately signaling a tampering.

Blockchain linkage: the image's hash is the key element that is registered on the blockchain. This linkage is fundamental: the blockchain, by its nature, guarantees the immutability of data

once recorded. By associating the image's hash with the blockchain, we create a verifiable and tamper-proof reference for the image's authenticity.

From a forensic standpoint, hashing plays a critical legal role: it guarantees data integrity and allows investigators to verify that evidence remains unaltered and admissible from acquisition to presentation in court. Each computed hash is stored within the blockchain-inspired log, creating an immutable link between the digital artifact and its forensic record.

4.3.3 Blockchain

The integration of blockchain technology within this project constitutes a foundational element, critical for ensuring the immutability and verifiability of information pertaining to image authenticity. By its very nature, blockchain offers a distributed and decentralized ledger where transactions are aggregated into "blocks" and cryptographically linked, rendering retrospective alteration exceedingly difficult. The selection of a particular blockchain implementation approach was subject to careful deliberation, weighing various options, each presenting its own set of advantages and disadvantages.

Option 1: true blockchain (Ethereum)

This approach entails leveraging a real, operational blockchain network, complete with its native smart contract capabilities and full decentralization. Such a setup represents the gold standard for secure and verifiable data management.

Advantages: A true blockchain offers the highest level of security due to its cryptographic foundations and consensus mechanisms, rendering data virtually immutable once recorded. Its decentralized nature provides unparalleled resistance to censorship and single points of failure, making it an ideal choice for high-stakes applications requiring absolute trust and transparency.

Disadvantages: The complexities associated with configuring and deploying on a live blockchain network are substantial, demanding deep technical expertise in blockchain architecture, smart contract development, and network management. Furthermore, the operational costs can be significant, particularly due to "gas fees" (transaction fees) that must be paid for every interaction with the network. These costs can fluctuate based on network congestion, making budget planning challenging for extensive data registration. The time required for transaction finality can also vary, impacting real-time application responsiveness.

More details about Ethereum

Ethereum is a decentralized, global, open-source platform that utilizes blockchain technology to facilitate the creation and execution of smart contracts and decentralized applications (dApps). Unlike traditional centralized systems controlled by a single entity, Ethereum is maintained and operated by a vast, distributed network of "nodes", individual computers run by volunteers across the globe. This decentralized structure is key to its resilience against censorship, fraud, and downtime, as it means no single entity can control the network or its data. Its native cryptocurrency is Ether (ETH).

At its core, Ethereum functions as a programmable blockchain. This means developers can build and deploy custom applications directly on its network. Transactions on Ethereum are grouped into "blocks" and added to the blockchain, secured by cryptographic principles. Users pay "gas fees" (in ETH) for every operation they perform on the network, such as sending ETH, executing smart contracts, or interacting with dApps. These fees incentivize validators (previously miners) to process transactions and secure the network. The network's operations are governed by smart contracts, self-executing agreements whose terms are directly coded into the blockchain, eliminating the need for intermediaries.

With "The Merge" Ethereum transitioned from a Proof of Work (PoW) consensus mechanism to Proof of Stake (PoS). Staking is a fundamental component of PoS. Instead of competing to

solve complex mathematical puzzles (as in PoW), participants (known as validators) "stake" or lock up a certain amount of their Ether (currently 32 ETH) into a smart contract. By doing so, they become eligible to be chosen to validate new blocks of transactions. If they correctly validate and add blocks, they are rewarded with more ETH. This mechanism not only secures the network but also makes it significantly more energy-efficient compared to PoW.

Ether (ETH) can be purchased through various channels, primarily centralized cryptocurrency exchanges like Coinbase, Binance, or Kraken. The process typically involves creating an account, verifying your identity (KYC Know Your Customer), linking a payment method (such as a bank account or debit card), and then placing a buy order for ETH. It can also be acquired through decentralized exchanges (DEXs) or peer-to-peer transactions.

The value of Ether, like other cryptocurrencies, is influenced by several factors. Key determinants include supply and demand dynamics in the market. Beyond speculation, its utility as the "gas" for the Ethereum network means its value is tied to the network's usage and adoption. The growth of decentralized finance (DeFi), NFTs, and other dApps built on Ethereum directly contributes to the demand for ETH. Furthermore, technological developments within the Ethereum ecosystem (upgrades like sharding), market sentiment, regulatory news, and overall macroeconomic conditions also play significant roles in determining its value. [79]

Option 2: online test blockchain (Testnet)

This alternative involves utilizing public online environments or simulators that emulate the behavior of a real blockchain. Testnets are designed to allow developers to experiment with blockchain functionalities without incurring real financial costs.

Advantages: testnets are typically free to use, eliminating financial risk during the development and testing phases. They effectively simulate the real-world dynamics of a blockchain network, allowing developers to gain practical experience with transaction processing, block mining (or validation), and smart contract execution in a controlled environment. This provides a valuable bridge between simulated and live environments.

Disadvantages: while emulating real networks, testnets are not designed for production environments and therefore do not offer 100 percent security or guarantee of immutability in the same way a mainnet does. They can be prone to resets, instability, or unexpected changes, which might disrupt ongoing development. Furthermore, they may impose usage limitations (for example transaction rate limits, data storage caps) that are not present on a live blockchain, potentially hindering large-scale testing. Their public nature also means that data recorded on a testnet is not private and could be subject to external scrutiny, though without the same level of security guarantees as a mainnet.

Further Details on Testnets

Within the broader blockchain ecosystem, and particularly for platforms such as Ethereum, testnets (test networks) fulfill a pivotal role. A testnet is a dedicated and controlled blockchain network meticulously designed for the sole purpose of testing and experimentation. It precisely mirrors the functionalities of the main blockchain (mainnet) but operates within a risk-free environment, ensuring that no real funds or assets are subjected to risk during development and testing activities.

In essence, a testnet serves as a "secure playground" or a "virtual sandbox" enabling developers to innovate and experiment without the apprehension of real-world ramifications. On a testnet, developers can deploy and interact with smart contracts and decentralized applications (dApps), simulate transactions, and validate new features using test cryptocurrencies that hold no economic value. This controlled environment is indispensable for identifying software bugs, optimizing code for efficiency and performance, verifying the security posture of the codebase, and refining the user experience. Such rigorous testing is conducted prior to the deployment of solutions on the mainnet, where real financial value and critical assets are transacted. The strategic utilization of testnets is thus integral to the blockchain development lifecycle, providing a robust and controlled milieu for continuous innovation and validation.

Option 3: simulated blockchain with CSV

This methodology involves creating a CSV (Comma Separated Values) file where each row functions as a block with its integrity ensured by a cryptographic link to the preceding row via a previous hash field. This approach prioritizes simplicity and ease of integration within a development environment.

Advantages: this option is remarkably straightforward to implement, making it highly suitable for rapid prototyping and testing. The data stored in the CSV file is easily readable and manageable, which streamlines debugging and development processes. Furthermore, its lightweight nature ensures seamless compatibility with cloud-based development platforms such as Google Colab, avoiding complex setup procedures.

Disadvantages: a significant drawback is the inherent lack of security; the CSV file itself is not protected from direct manual modifications, undermining the core principle of immutability that a true blockchain provides. Moreover, it is a centralized solution, devoid of the distributed and decentralized characteristics that make genuine blockchains robust against single points of failure or censorship. Consequently, it offers no intrinsic protection against external tampering beyond the internal cryptographic link.

My choice

For the purposes of this project, a deliberate decision was made to adopt option 3: the simulation of a blockchain via CSV files. This choice was primarily driven by practical considerations. The utilization of a true blockchain would have introduced an unwarranted level of complexity and infrastructural requirements that extended beyond the scope of a foundational prototype in an academic setting. Moreover, to ensure simple and direct compatibility with the chosen development environment, Google Colab, opting for an internal simulation was deemed the most effective approach, circumventing the need for external test networks or complex blockchain integrations.

Despite the simulation not replicating the decentralization of a true blockchain, it crucially preserves a fundamental aspect: the integrity of the chain. Each "block" (row in the CSV) is cryptographically linked to its predecessor via the previous hash field, thereby effectively emulating the chaining and immutability mechanisms characteristic of a genuine blockchain. This approach allows for an effective demonstration of the concept of a tamper-proof ledger for digital image authenticity, all while maintaining the simplicity and efficiency necessary for the development and demonstration of the prototype. The specific structure of each "block" within our CSV-based blockchain and the functions developed to manage it faithfully reflect the core principles of blockchain technology, as further detailed in the following subsection.

To summerize, the blockchain-like log was designed to store transaction entries containing the file hash, operation timestamp, and actor identification. Each log entry is linked to the previous one through a hash pointer, ensuring immutability and sequential integrity. Such design guarantees non-repudiation and supports forensic accountability, fulfilling one of the key requirements of digital evidence handling: the ability to demonstrate who did what and when, in a way that is legally verifiable.

4.4 Implementation

The theoretical framework and design choices outlined in the preceding sections were brought to fruition through a practical implementation, primarily leveraging Python scripts within the Google Colab environment. This setup offered a flexible and accessible platform for developing and testing the core components of the image authentication system. The implementation was modular, with distinct Python functions and scripts dedicated to each primary process: watermarking, hashing, and blockchain simulation.

4.4.1 Implementing the watermark

The implementation of the digital watermarking component was carried out using custom Python scripts, specifically leveraging basic image manipulation techniques from the cv2 (OpenCV) and numpy libraries. This approach directly implements the fragile watermarking strategy previously discussed, focusing on the Least Significant Bit (LSB) embedding method due to its simplicity and effectiveness in detecting minute image alterations. The core of the watermarking implementation consists of several interconnected functions: text_to_bin, bin_to_text(binary), embed_watermark(img, watermark_text), extract_watermark(img)

text_to_bin(text)

This utility function serves as the initial step in preparing the watermark. It takes a plain text string as input (in the example "Foto originale by Giada") and converts it into its binary representation. This conversion is performed by first obtaining the ASCII (or Unicode) code for each character using ord(c) and then formatting it into an 8-bit binary string (for example, 'A' becomes '01000001'). The resulting binary strings are then concatenated to form a single continuous binary sequence.

bin_to_text(binary)

This function performs the reverse operation of text_to_bin. It takes a binary string, segments it into 8-bit chunks, converts each chunk back into an integer (representing an ASCII code), and then casts these integers back to their corresponding characters. This allows for the reconstruction of the original textual watermark from the extracted binary data.

embed_watermark(img, watermark_text)

This is the central function for embedding the watermark. It first converts the watermark_text into a binary string using text_to_bin(). Crucially, a specific 16-bit binary marker, '1111111111111110', is appended to the binary watermark. This unique end-of-message marker is vital for the extraction process, as it signals where the hidden message terminates.

The input image img is then flattened into a one-dimensional array using img.flatten(). This allows for sequential pixel-by-pixel processing. The core LSB embedding logic is then applied: for each pixel value in the flattened image, its least significant bit is set to 0 by performing a bitwise AND operation with 254 (11111110 in binary).

Immediately after, the current bit from the binary_wm (watermark) is inserted into this cleared LSB position using a bitwise OR operation. This process modifies the LSB of each pixel to carry a bit of the watermark data, making the changes visually imperceptible. The embedding continues until all bits of the binary_wm (including the end-of-message marker) have been inserted.

Finally, the modified flat array of pixel values is reshaped back into the original image dimensions using flat_img.reshape(img.shape), yielding the watermarked image.

extract_watermark(img)

This function is designed to retrieve the hidden watermark from a watermarked image. Similar to embedding, the watermarked image img is flattened. The function then iterates through each pixel in the flattened image. For every pixel, it extracts its least significant bit by performing a bitwise AND operation with 1 (00000001 in binary). This extracted bit is appended to a binary_data string. The loop continues to collect LSBs until the predefined end-of-message marker ('1111111111111110') is detected as the suffix of binary_data. Once the marker is found, the 16 bits of the marker are removed, and the remaining watermark_bin string is converted back into readable text using the bin_to_text() function. The extracted textual watermark is then returned.

Testing the functions

The implementation was tested within Google Colab, demonstrating the full cycle from loading an example image (original.jpg), embedding a custom watermark text ("Foto originale by Giada"), saving the watermarked image (foto_watermarked.png), and successfully extracting the original watermark from the saved image. This practical demonstration validates the effectiveness of the chosen fragile LSB watermarking technique for the purpose of detecting image manipulations.

4.4.2 Implementing the hash

The process of generating a unique and immutable digital fingerprint for each watermarked image is crucial for ensuring its integrity within the proposed authentication system. This is achieved through a hashing mechanism, specifically employing the cryptographically secure SHA-256 (Secure Hash Algorithm 256-bit) algorithm. The implementation leverages Python's standard hashlib library, which provides a robust and efficient way to compute cryptographic hashes. The core functionality for hashing is encapsulated within the generate_hash Python function.

generate_hash(image_path)

This function takes the file path of a watermarked image as its input.

Binary File Reading: the function begins by opening the specified image file in binary read mode (rb). Reading the file as raw bytes is paramount, as it ensures that the hash calculation considers every single bit of the image data, making the resulting hash highly sensitive to even the most minute modifications.

SHA-256 Calculation: the entire byte stream read from the image file is then passed to the hashlib.sha256() method. This operation computes the SHA-256 cryptographic hash of the binary data.

Hexadecimal Representation: the raw binary hash output by hashlib.sha256() is then converted into a more human-readable hexadecimal string using the .hexdigest() method. This 64-character hexadecimal string serves as the unique and fixed-length identifier for the image.

Testing the function

The generate_hash function is integrated into the workflow immediately after the watermarking process. As demonstrated in the provided script, once an image has been watermarked (e.g., resulting in watermarked_img), it is first saved to a temporary path (watermarked_img.jpg). Subsequently, this path is passed to generate_hash to compute its unique SHA-256 value. This generated hash value is then ready to be recorded on the blockchain (or its simulated equivalent), acting as a verifiable proof of the image's state at the time of its registration. Any future alteration to the image would result in a different hash, instantly flagging potential tampering.

4.4.3 Implementing the blockchain

The simulated blockchain, chosen for its practical advantages in a Google Colab environment, is implemented using standard Python libraries, primarily csv for data persistence, hashlib for cryptographic linking, and datetime for timestamping. This implementation meticulously replicates the core principle of cryptographic chaining, where each new "block" (represented as a row in a CSV file) is linked to its predecessor via a hash. The blockchain implementation relies on three interconnected Python functions: compute_block_hash(timestamp, user, image_name, image_hash, prev_hash), get_last_block_hash(csv_file), register_on_blockchain(image_name, image_hash, user = "giada", csv_file="blockchain.csv"

compute_block_hash(timestamp, user, image_name, image_hash, prev_hash)

This function is fundamental to the integrity of the simulated chain. It takes all the key data points that constitute a single "block" as input: the timestamp of creation, the user who initiated the transaction, the image_name, the image_hash (generated as described in the previous section) and the prev_hash of the preceding block. These elements are concatenated into a single string. The function then calculates the SHA-256 hash of this combined string, providing a unique cryptographic fingerprint for the entire current block. This block_hash serves as the prev_hash for the subsequent block, forging the essential cryptographic link that binds the chain.

get_last_block_hash(csv_file)

This utility function is responsible for retrieving the block_hash of the most recently added "block" in the simulated blockchain. It reads the CSV file (blockchain.csv) and extracts the hash from the last row. If the file does not exist or is empty, it returns a "genesis" hash (a string of 64 zeros), which serves as the initial prev_hash for the very first block, ensuring a consistent starting point for the chain. This function is crucial for maintaining the sequential and cryptographic order of the blocks.

register_on_blockchain(image_name, image_hash, user, csv_file)

This is the primary function for adding new image authenticity records to the simulated blockchain.

Upon invocation, it first generates an ISO-formatted timestamp to record the exact moment of registration. It then calls get_last_block_hash() to retrieve the prev_hash of the last valid block in the chain.

Subsequently, it computes the block_hash for the current block by calling compute_block_hash(), incorporating all the relevant data for the new record. Finally, a record list containing all these pieces of information (timestamp, user, image_name, image_hash, prev_hash, and the block_hash of the current record) is appended as a new row to the specified CSV file (blockchain.csv). The csv module handles the proper formatting and writing of this new entry.

This modular implementation, particularly the reliance on explicit prev_hash and block_hash calculations stored within the CSV, effectively simulates the chain-like structure of a blockchain. It ensures that any attempt to tamper with a previous record in the CSV would invalidate the hash of subsequent blocks, thereby demonstrating the core integrity principle, even without the full decentralization of a true blockchain network. The register_on_blockchain function is the gateway through which verified image data enters the immutable ledger.

4.4.4 Verifying integrity of an image

The culmination of the watermarking, hashing, and blockchain simulation processes is the ability to verify the authenticity and integrity of an image. This critical step confirms whether an image has been altered since its initial registration on the simulated blockchain. The verification process leverages the immutable record created by combining the image's hash with the blockchain's cryptographic chaining. The primary function responsible for this verification is verify_image_integrity

verify_image_integrity(image_path, blockchain_csv)

This function is designed to check if the hash of a given image matches any hash recorded in the simulated blockchain.

Current image hash calculation: the function first computes the SHA-256 hash of the image_path provided as input. This is achieved by calling the previously defined generate_hash(image_path) function, ensuring that the integrity check is based on the exact state of the image at the time of verification.

Blockchain traversal: it then opens the blockchain.csv file (the simulated blockchain) in read mode. It iterates through each row of the CSV, treating each row as a _block_in the chain.

Hash comparison: for every row read from the CSV, it extracts the image_hash that was originally stored. This stored_hash is then directly compared with the current_hash calculated from the input image.

Outcome Reporting: if a match is found (stored_hash == current_hash), it signifies that the image's current state aligns with a registered record in the blockchain. The function then prints a success message and displays the associated metadata from the blockchain record, such as the user who registered it, the timestamp of registration, and the original image_name. It then returns True, indicating successful verification. If the loop completes without finding a matching hash, it means the image's current hash does not correspond to any registered entry, suggesting a potential alteration or that the image has never been registered. In this case, a failure message is printed, and the function returns False. This verify_image_integrity function provides a direct and tangible demonstration of the project's core utility: leveraging the combined power of watermarking, hashing, and blockchain-like immutability to establish trust and detect unauthorized modifications in digital images. The example usage, verify_image_integrity(_watermarked_img.jpg_), shows how a previously watermarked and hashed image can be subjected to this integrity check.

4.5 Real world example of application

In order to validate the practical applicability of the proposed system, I conducted a test using a realistic forensic dataset. Ideally, my objective was to obtain and work on a real image extracted from a concluded legal case. However, accessing such materials proved to be extremely challenging due to privacy, legal, and ethical constraints. For this reason, I turned to CFReDS, the Computer Forensic Reference Data Sets, a well-established initiative designed precisely to provide openly available, realistic forensic data for educational and testing purposes. The ultimate goal of this test is to demonstrate a concrete use case of my system and to evaluate its effectiveness (or potential weaknesses) when applied in a realistic scenario that approximates the dynamics of a genuine forensic investigation. The broader objective of this application is to illustrate how the combined use of watermarking techniques (to ensure traceability and data integrity) and blockchain technology (to provide an immutable audit trail) can strengthen the integrity, chain of custody, and overall verifiability of digital evidence in scenarios that are closely aligned with real-world legal contexts.

4.5.1 CFReDS

The Computer Forensic Reference Data Sets (CFReDS) portal, managed by the National Institute of Standards and Technology (NIST), provides a collection of well-documented digital forensic datasets aimed at supporting tool testing, academic research, and professional training. These datasets are publicly available and offer structured collections of realistic digital evidence artifacts, complete with metadata, file structures, and timelines that accurately simulate real investigative scenarios. By offering access to curated data that reflects typical digital crime scenes, CFReDS represents an invaluable resource in digital forensics education and validation. Most importantly, it circumvents the ethical, legal, and privacy-related obstacles that often prevent the use of real case data in academic or testing contexts, making it an ideal environment for safely experimenting with forensic procedures and technologies. [80]

Chosen dataset: Data Leakage Case

For this project, I selected the Data Leakage Case dataset, available at the CFReDS portal under the linked url. This dataset is specifically designed to replicate a corporate scenario involving unauthorized data exfiltration through removable storage devices. It includes a wide range of artifacts such as USB access logs, documents, and media files that simulate the type of digital evidence typically encountered in investigations of insider threats, intellectual property violations,

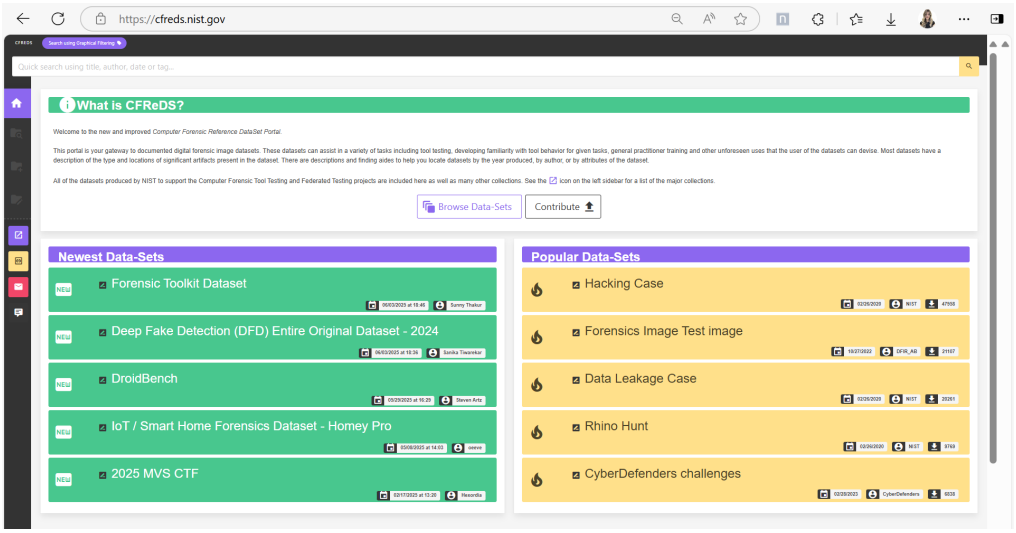


Figure 4.1. CFReDS portal

or breaches of confidentiality agreements. Among these artifacts, particular attention was given to disk image files containing traces of user activities and file transfers, which enabled the simulation of a full forensic workflow, from evidence acquisition and metadata reconstruction to the integrity verification of exfiltrated data. The realistic context provided by this dataset served as an ideal foundation to test the effectiveness of the proposed system under conditions that closely approximate a real-world forensic case. [81]

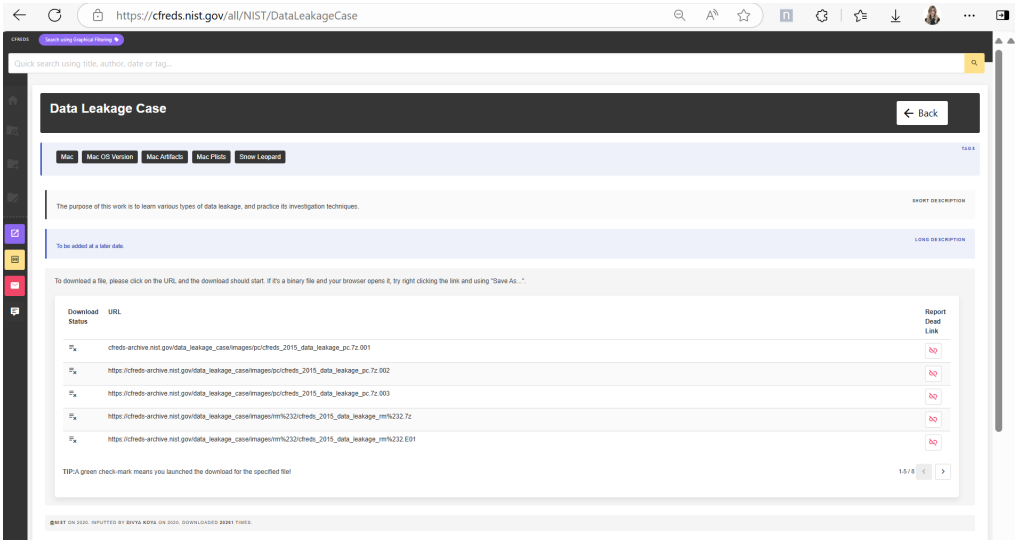


Figure 4.2. Data Leakage Case - CFReDS

4.5.2 Autopsy

To extract and analyze relevant artifacts from the dataset, I employed Autopsy, a widely used open-source digital forensics platform. Autopsy provides a comprehensive interface for navigating through forensic images, identifying user activities, recovering deleted files, and correlating data across various sources. Thanks to its integration with The Sleuth Kit and its support for a wide range of forensic modules, Autopsy enables a structured and methodical examination of digital evidence.

Using Autopsy, I conducted an in-depth analysis of the USB evidence file, specifically a .E01 forensic disk image. By mounting and exploring the volume through Autopsy's graphical interface, I was able to recover several user-created files, including images, documents, and logs. I selected one particular image file for this simulation, exported it from the forensic container, and saved both the file and the associated metadata for further analysis. Screenshots of the Autopsy interface showing the mounted image, the file hierarchy, and the extracted image's metadata are included in figures below.

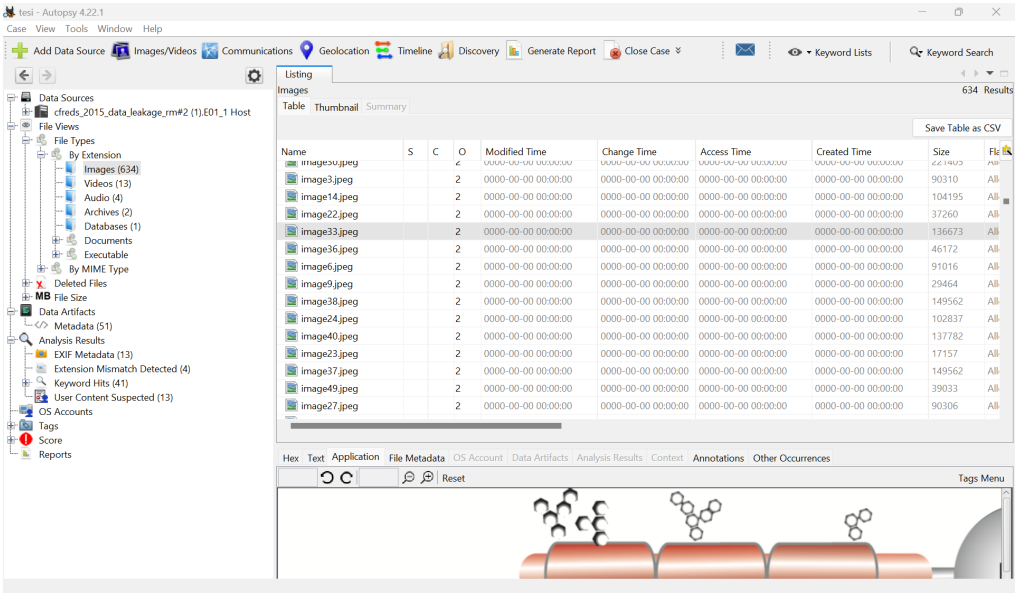


Figure 4.3. Autopsy view

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Name	Modified T	Change Tir	Access Tir	Created Ti	Size	Flags(Dir)	Flags(Met	Known	Location	MD5 Hash	SHA-256 H	MIME Type	Extension					
2	image33.jpeg	0000-00-0	0000-00-0	0000-00-0	0000-00-0	136673	Allocated	Allocated	unknown	/img_cfred	0ae5b1f9a	fadd05898	image/jpe	jpeg					
3																			
4																			

Figure 4.4. Autopsy extraction - csv of image33 info

4.5.3 Practical simulation - real forensic scenario

The system was then tested in a simulated forensic environment reproducing the process of evidence collection, registration, and validation. The scenario consisted of an image acquisition phase, hash computation, watermark embedding, and the recording of all actions in the blockchain-inspired log. When the image was subsequently altered, both the hash and the watermark detected the tampering, while the blockchain log maintained a transparent record of every operation. From a legal standpoint, the combined use of hashing, fragile watermarking, and blockchain-inspired logging provides a verifiable chain of custody consistent with digital forensic standards, ensuring evidentiary reliability in judicial contexts. This integrated approach establishes a technically sound chain of evidence, where each procedural step, acquisition, processing, storage, and verification, is cryptographically and temporally anchored, allowing full reconstruction of the evidence lifecycle if required by a court.

In particular, will simulate these scenarios: one has access to an original image from the company through the use of a disk image into a USB (for example image33 reported above) and modifies it, later divulging the modified one; one has access to the original csv of metadata regarding the original image and modifies it (for example by modifying the location field), later divulging it; one has access to the watermarked image and modifies the embedded watermark,

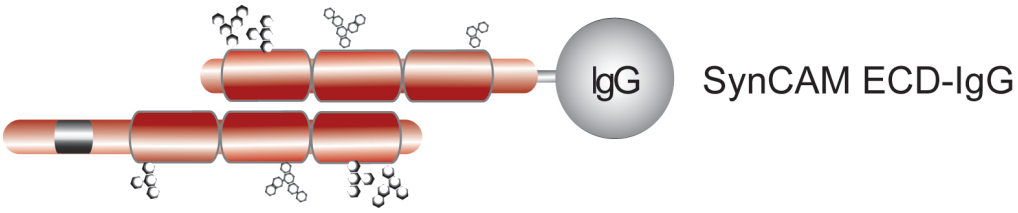


Figure 4.5. Autopsy - image33

later divulging it. All these scenarios are included into my python scripts and detected through specific verification functions reported on the programmer manual.

The implemented code provides a comprehensive framework for ensuring the integrity and authenticity of digital images through a combination of watermarking, metadata hashing, and blockchain-based logging. The system is designed to detect tampering attempts by verifying the consistency of image content, metadata, and embedded watermarks. Below, the major components of the implementation are discussed in detail.

Binary Conversion for Watermarking

To embed and extract watermark information, the system first converts textual data into binary format. The text to bin() function transforms each character of a string into its 8-bit binary equivalent, while bin to text() performs the inverse operation. These conversions are fundamental for embedding the watermark bit-by-bit into the image.

Watermark Embedding and Extraction

The watermarking mechanism uses a basic Least Significant Bit (LSB) technique. The function embed watermark() modifies the LSB of each image pixel to encode the binary watermark, which includes a custom delimiter ('1111111111111110') to signal the end of the embedded message. This watermark is typically derived from concatenated metadata and a user-defined tag. The extract watermark() function retrieves the watermark by reading the LSBs until the delimiter is encountered, and then converts the binary data back into text.

Cryptographic Hashing

To ensure data integrity, the code utilizes SHA-256 hashing: the generate hash() function computes the SHA-256 hash of a file (in the example usage, an image); the hash metadata() function creates a hash of metadata values by concatenating them in a standardized format before hashing. These hashes act as unique digital fingerprints and are later used for verification and logging.

Metadata handling

Metadata associated with each image is stored in a CSV file. The function get metadata from csv() (and its variant extract metadata from csv()) retrieves metadata fields such as filename, creation time, size, MIME type, and hash. This metadata is concatenated into a single string for hashing and embedding as a watermark.

Blockchain-based logging

A simulated blockchain is implemented using CSV as a lightweight append-only ledger. Each block contains: a timestamp, username, image name, image hash, metadata hash, watermark hash, previous block hash, current block hash. The function register on blockchain() computes the current block's hash using all the above fields and appends it to the ledger. The function get last block hash() retrieves the hash of the last block to maintain blockchain continuity.

Verification procedures

Three types of verification are implemented:

Image verification (verify image): Checks if the current hash of the image matches the hash stored in the blockchain.

Watermark verification (verify image with wm): Verifies the integrity of the watermarked image by comparing hashes.

Metadata verification (verify metadata): Confirms that the current metadata matches the originally logged metadata, using hash comparison.

Each of these verification functions prints a success or failure message depending on whether the expected values match the current ones, thereby detecting any unauthorized alterations.

Tampering simulations

Three attack scenarios are simulated to test the robustness of the system:

Attack 1 - image modification: An image is tampered by modifying pixel values and applying a Gaussian filter. The resulting image fails the integrity check due to a mismatch in hashes.

Attack 2 - fake watermark insertion: A forged watermark is embedded into the image, which fails the watermark hash comparison during verification.

Attack 3 - metadata tampering: The location field in the metadata CSV is maliciously modified. The altered metadata hash does not match the blockchain record, triggering a detection alert.

Each attack confirms the system's ability to identify unauthorized changes to image content, watermark, or metadata.

Screenshots of functions' outputs are reported below and all the original scripts ar reported into the Porgrammer Manual.

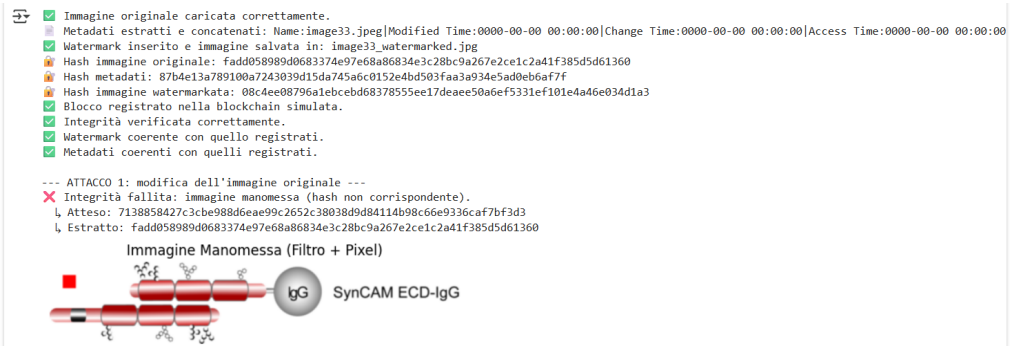
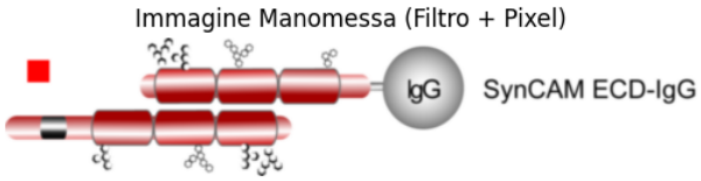


Figure 4.6. outputs - part 1




```
❌ Integrità fallita: immagine manomessa (hash non corrispondente).
  ↳ Atteso: 7138858427c3cbe988d6eae99c2652c38038d9d84114b98c66e9336caf7bf3d3
  ↳ Estratto: fadd058989d0683374e97e68a86834e3c28bc9a267e2ce1c2a41f385d5d61360

--- ATTACCO 2: modifica del watermark ---
❌ Watermark non corrispondente! Potenziale manomissione o attacco.
  ↳ Atteso  : 08c4ee08796a1ebcebd68378555ee17deaee50a6ef5331ef101e4a46e034d1a3
  ↳ Estratto: ecf5b28f685a806681dbed719d6f2563e70f89c22e407d60067c888ad9e58afe

--- ATTACCO 3: modifica dei metadati nel CSV ---
❌ Metadati corrotti o modificati (hash mismatch).
  ↳ Originale: 87b4e13a789100a7243039d15da745a6c0152e4bd503faa3a934e5ad0eb6af7f
  ↳ Calcolato: cce02954e33b2056b2051a290583a5d0ec3e872375a1dcca4007a575399f9ac8
False
```

Figure 4.7. outputs - part 2

Chapter 5

Results

The scripts developed throughout this research proved to be effective tools for the detection and authentication of deepfake content. After extensive testing on a variety of images, subjected to different types of manipulation, the implemented code consistently demonstrated the ability to identify anomalies and to distinguish authentic images from manipulated ones.

Central to this result was the combined use of watermarking and hashing techniques. Their integration proved particularly valuable, as it enabled not only the detection of alterations applied to the original image but also the identification of modifications targeting the embedded metadata stored within the watermark itself. This dual-layer approach therefore strengthened the detection process, offering an additional safeguard against tampering.

Another key element of the system was the blockchain-inspired component. Although simulated through the use of a CSV file, its structure was carefully designed to replicate the functioning of real blockchain systems. By ensuring that each entry contained both the hash of the current block and the hash of the previous one, the integrity of the chain was preserved. This mechanism allowed any unauthorized insertion or deletion to be detected, thereby guaranteeing the authenticity of the recorded information and reinforcing the overall reliability of the tool.

The solution was tested in two complementary scenarios: with online images and with images extracted through Autopsy from disk images available on the CFReDS portal. This methodology provided a complete and realistic evaluation environment, as it replicated the practical conditions of digital investigations, including the management of authentic images together with their metadata derived from forensic disk images.

Taken together, these findings suggest that the proposed tool can serve as a relevant contribution to addressing the emerging challenges posed by the spread of deepfake content. By enabling verification of multimedia authenticity and detection of tampering, the system is able to recognize three distinct categories of manipulation: direct alterations to the image (such as pixel modifications or filter applications), manipulations of the embedded watermark, and modifications of the associated metadata (including creation date, location, or author details).

Chapter 6

Conclusions

This research has demonstrated the potential of combining watermarking, hashing, and blockchain-inspired mechanisms to address the pressing challenge of deepfake detection and authentication. The tool developed during this work proved effective in identifying manipulations and in reliably distinguishing authentic images from those that had been tampered with. The results obtained through testing with both online images and forensic disk images underline the robustness of the proposed approach and its relevance in scenarios that closely resemble real investigative contexts.

Beyond its immediate performance, the project highlights an important contribution in bridging theoretical concepts with practical applications. By simulating blockchain structures and integrating forensic techniques such as metadata analysis, the tool not only confirmed its expected functionality but also offered a proof of concept that can serve as a foundation for further developments. Its capacity to detect different layers of manipulation, whether at the image, watermark, or metadata level, represents a meaningful step forward in strengthening digital content verification in the face of the growing risks associated with deepfakes.

From a broader perspective, the project also positions itself within the current state of the art on media authentication methods, particularly in relation to blockchain-based and watermarking approaches. Compared to existing works such as those by Hasan and Salah [74], Chan et al. [75], and Qureshi, Megias, and Kuribayashi [77], the system presented in this thesis can be viewed as a simplified yet operational implementation that translates theoretical frameworks into a practical prototype. While previous studies have largely remained at the conceptual or simulation stage, demonstrating the feasibility of blockchain and watermarking for provenance tracking, the present work consolidates these principles into an integrated tool specifically designed for image authentication. It thus provides empirical validation to the idea, discussed extensively in recent literature, that authenticity verification should move from post-hoc analysis toward proactive embedding and traceability mechanisms. At the same time, the limitations identified in the literature, such as scalability, computational overhead, and dependence on trusted capture environments, also emerge here, confirming that the transition from controlled experiments to real-world deployment remains a central challenge for this research domain. In this sense, the project stands as an intermediate step between theoretical innovation and operational maturity, contributing practical insights that can inform future implementations of blockchain-based watermarking systems.

From a practical and forensic standpoint, this work also demonstrates potential applications within legal and evidentiary domains. In a time when manipulated digital content can undermine judicial proceedings, infringe upon intellectual property rights, or damage personal and corporate reputations, the ability to authenticate multimedia evidence is of paramount importance. The proposed tool could therefore assist investigators, legal practitioners, and judicial authorities in verifying the authenticity of digital evidence, ensuring that content presented in court or in dispute resolution processes has not been altered or maliciously fabricated.

While the system has already shown encouraging results, future research could expand upon these foundations in several meaningful directions. A natural progression would be the integration of a fully deployed blockchain infrastructure, rather than a simulated version. Real blockchain networks, whether public or private, would offer stronger guarantees of integrity and immutability,

ensuring that once data are stored, no actor, internal or external, could modify or delete them without detection. This enhancement would not only increase the robustness of the system but also align it more closely with industry standards for secure data storage and evidence preservation, thereby making the tool suitable for real-world deployment in sensitive environments.

Another significant area for development lies in the design of a dedicated user interface. At present, the tool operates as a proof of concept, primarily intended for academic experimentation. By extending it into a fully functional application equipped with an intuitive graphical interface, the system could become accessible to a wider range of professionals, including digital forensics investigators, law enforcement agencies, and legal practitioners who may not possess advanced technical expertise. Such an interface could provide functionalities such as case management, automated reporting of detection results, and visual representations of anomalies found within images and metadata. This would greatly enhance the usability of the tool, transforming it from a research prototype into a practical and operational solution.

Equally important is the possibility of testing the system with authentic evidentiary materials in real forensic casework. While experiments conducted in this research have relied on online images and forensic disk images from publicly available datasets, applying the tool to actual legal investigations would constitute a decisive step in validating its effectiveness. Real-world forensic materials often present additional challenges, such as degraded file quality, partial data corruption, or chain-of-custody requirements, that cannot always be replicated in controlled testing environments. Engaging with such cases would not only test the technical reliability of the system but also its compliance with evidentiary standards required in judicial proceedings. Successfully addressing these challenges would allow the tool to evolve into a credible support mechanism for courts, particularly in cases involving digital evidence, intellectual property disputes, or content authenticity verification.

In conclusion, the outcomes of this thesis confirm the relevance and effectiveness of the proposed approach, while also paving the way for further improvements and applications. The work conducted demonstrates how the integration of watermarking, hashing, and blockchain-inspired mechanisms can contribute to the urgent task of verifying the authenticity of multimedia content in an increasingly digital and interconnected world. Far from being a purely academic exercise, this research represents a concrete step toward the development of practical tools that can be deployed in real investigative and operational contexts.

The significance of this work extends beyond the technical dimension, as it addresses challenges that are inherently societal and legal. The proliferation of deepfakes and manipulated media poses not only a technological problem but also a threat to the credibility of digital evidence, the protection of intellectual property rights, and the preservation of trust in information systems. By offering a method for detecting manipulations at multiple levels, images, watermarks, and metadata, this thesis contributes to the construction of frameworks that could assist courts, investigators, and regulators in safeguarding the integrity of digital evidence. In this sense, the research aligns with broader efforts to ensure that the justice system can adapt to the complexities of the digital era, where the authenticity of information is increasingly called into question.

Ultimately, the work presented here should be viewed as both a foundation and a catalyst. It establishes that reliable deepfake detection mechanisms can be implemented in practice and demonstrates their feasibility in a forensic context, while also highlighting avenues for refinement and expansion. At the same time, it underscores the need for continued interdisciplinary collaboration, bridging computer science, cybersecurity, and law, to ensure that technological solutions are not only effective but also aligned with legal standards and societal expectations. The authentication and protection of digital content in the era of deepfakes is, without doubt, one of the most pressing challenges of our time. By addressing this issue through the development and evaluation of a working prototype, this thesis contributes to the pursuit of solutions that can strengthen digital trust, support the rule of law, and ultimately protect individuals and institutions from the risks associated with the manipulation of digital information.

In conclusion, this study demonstrates that the legal principles governing the responsible use of artificial intelligence, such as accountability, transparency, integrity, and authenticity, can be effectively translated into concrete technical mechanisms. Through the integration of digital watermarking, cryptographic hashing, and blockchain-based traceability, the proposed framework

provides a tangible response to the normative requirements emerging from instruments such as the AI Act, the Digital Services Act, and national and international regulations addressing the phenomenon of deepfakes. By ensuring the authenticity, provenance, and integrity of digital content, the system aligns with the core objectives of contemporary AI governance: to foster trust, security, and accountability in the digital environment. This convergence between law and technology not only reinforces the evidentiary reliability of AI-generated media but also illustrates how technical design can become an enabler of legal compliance and ethical responsibility. Ultimately, the project highlights the potential of combining legal foresight with technological innovation to build resilient mechanisms capable of protecting individuals and society from the risks associated with synthetic content. It therefore contributes to the broader debate on how emerging technologies can be governed through verifiable, transparent, and ethically grounded solutions, paving the way for future developments in both legal regulation and technical implementation of trustworthy AI systems.

Appendix A

User Manual

This chapter provides a comprehensive user manual detailing the experimental workflow followed during the development of the proposed deepfake detection pipeline. The steps presented below outline the full process, from downloading forensic disk images to applying detection scripts on extracted images. Each phase is accompanied by explanatory details to ensure reproducibility of the results.

A.0.1 Technical Requirements

To successfully execute the deepfake detection pipeline, the following requirements must be met.

Hardware (minimum)

- CPU: Dual-core processor (x86_64 architecture)
- RAM: 8 GB
- Disk space: 5 GB available
- GPU: optional (recommended for large datasets)

Hardware (recommended)

- CPU: Quad-core or higher
- RAM: 16 GB
- GPU: NVIDIA CUDA-enabled GPU (e.g., Tesla T4, RTX series)

Software Environment

- Operating System: Windows, macOS, or Linux (tested on Windows 11 and Google Colab Linux VM)
- Python: Version 3.10
- Google Account: required to use Google Colab

A.0.2 Required Input Files

The pipeline requires two input files, both extracted using Autopsy:

1. **Image File**
   - Format: JPEG or PNG
   - Recommended resolution: up to 4K for reasonable processing time
   - File naming: any name, but must match the name in the CSV metadata
2. **Metadata CSV**
   - Encoding: UTF-8
   - Separator: comma
   - Required columns:
     - `filename` (string) matching the image file name
     - `timestamp` (ISO 8601 format)
     - `size` (optional)
     - `location` (optional)
     - `sha256` (string) checksum of the file
     - `mime type` (optional)
     - `extensions` (optional)

A.0.3 Accessing the CFReDS Platform and Selecting a Case

The first step involves accessing the Computer Forensics Reference Data Sets (CFReDS) portal, provided by the National Institute of Standards and Technology (NIST). This platform offers a variety of realistic forensic challenges and datasets that can be used for academic and experimental purposes.

Navigate to the CFReDS official website: CFReDS official website

Browse through the available forensic scenarios and select a case of interest. In this project, a disk image containing multimedia and user data was chosen in order to simulate a digital forensic investigation involving image evidence.

Download the relevant disk image file, which is typically provided in E01, AFF, or raw formats.

A.0.4 Downloading and Installing Autopsy

Once the disk image has been obtained, the next step is to analyze it using Autopsy, a well-known open-source digital forensics platform.

Visit the official Autopsy website: Autopsy Download official website

Choose the version compatible with your operating system (Windows, macOS, or Linux).

Download the installer and follow the installation instructions.

A.0.5 Creating a New Autopsy Case and Importing the Disk Image

To start the forensic investigation:

Open Autopsy and click on "Create New Case"

Choose a name and location for your project

Proceed through the wizard and select "Add Data Source"

Select the downloaded disk image file as your input (E01 or raw format)

Autopsy will automatically parse and index the file system, recovering partitions, files, and metadata.

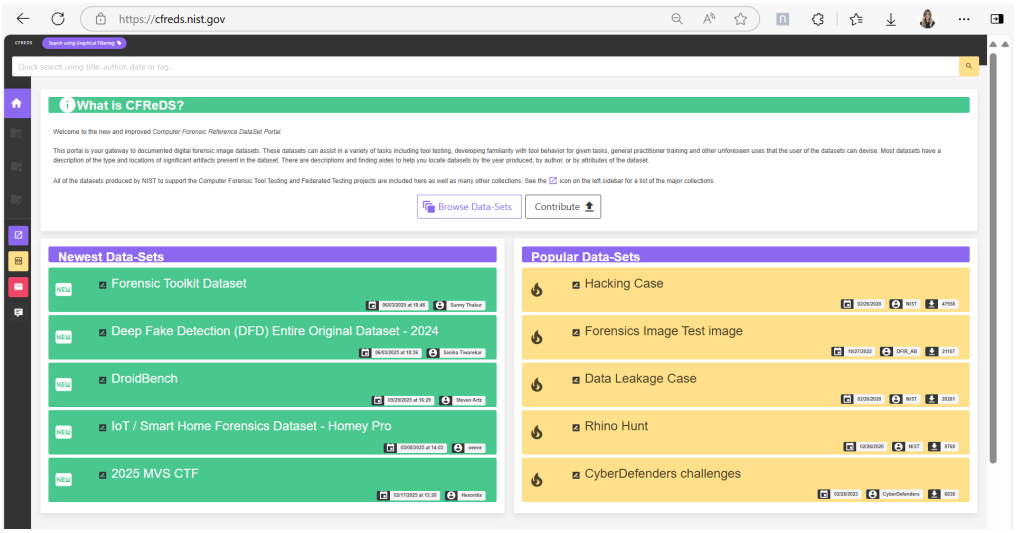


Figure A.1. CFReDS portal

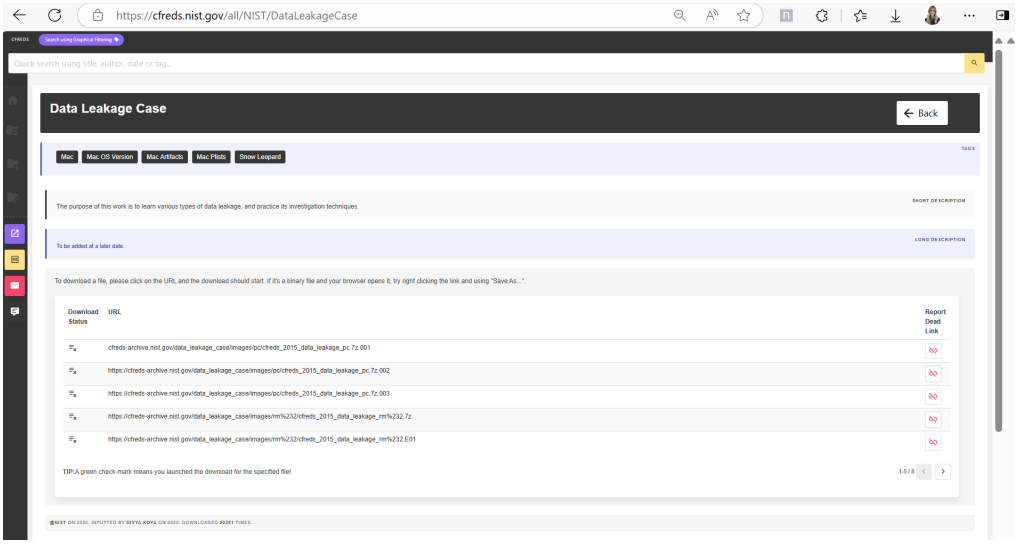


Figure A.2. CFReDS portal - data leakage case

A.0.6 Exploring the Disk Image and Extracting Evidence

Autopsy presents the file hierarchy of the disk image in a user-friendly interface. Navigate through the folders and preview files directly. In this experiment, the goal was to extract a JPEG image of interest from the disk image. Upon identifying the image, export it using the right-click "Extract File(s)" function. Additionally, extract the corresponding metadata, which can be exported as a CSV file containing EXIF information and file attributes (e.g., creation date, modification date, GPS data if available)

A.0.7 Running the Analysis in Google Colab

To execute the detection pipeline in Google Colab:

1. Sign in with a Google Account and open Google Colab.
2. Create a new notebook.

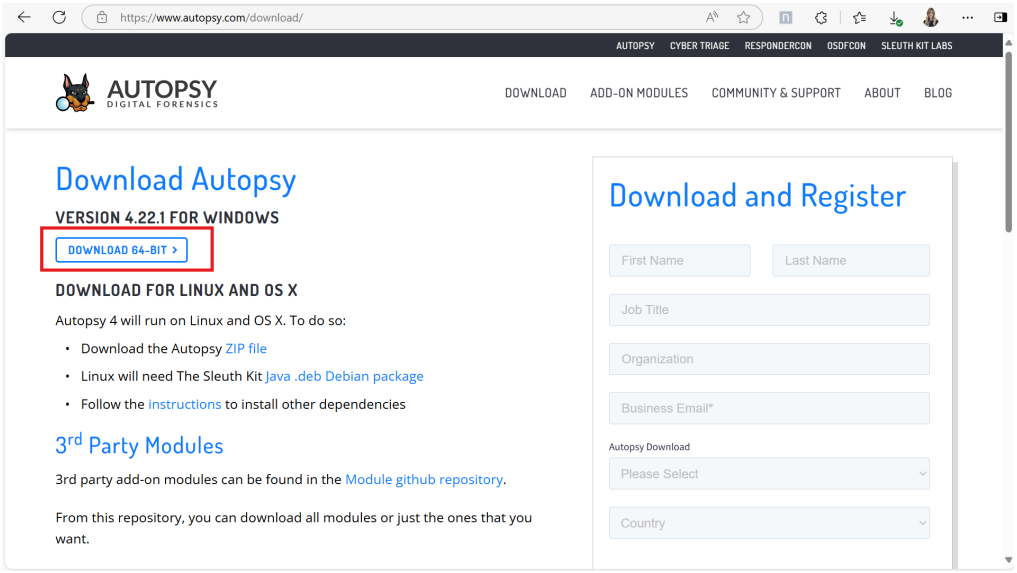


Figure A.3. Autopsy - downloads

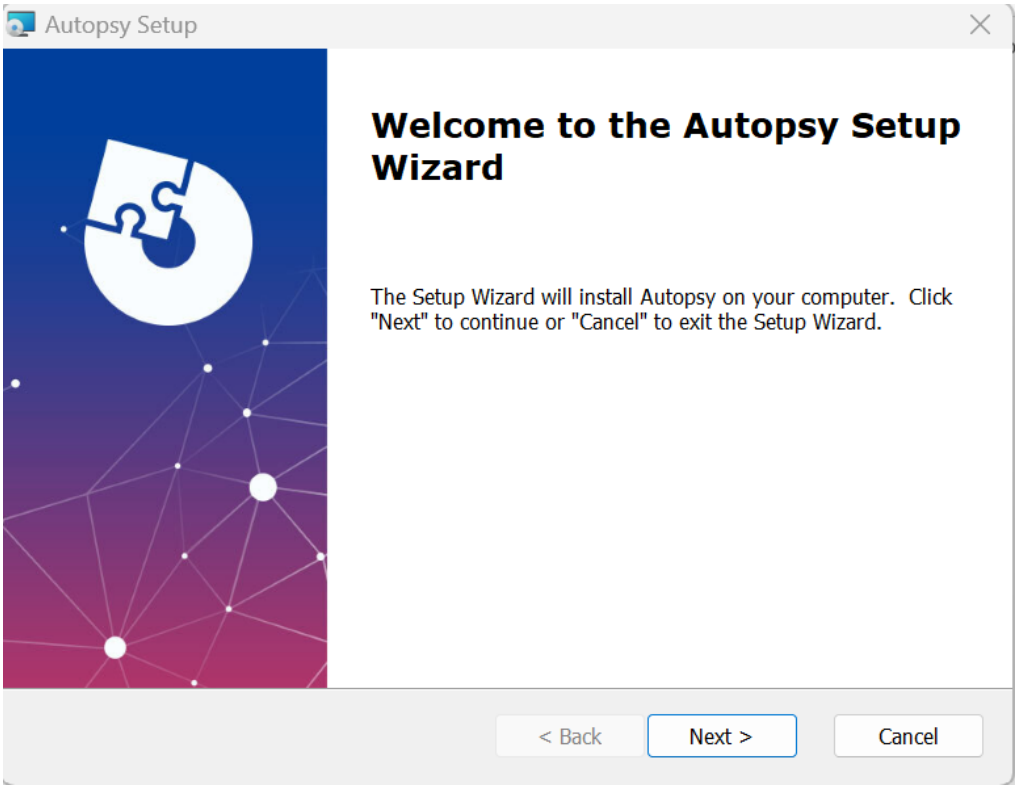


Figure A.4. Autopsy - step1

3. (Optional) Mount Google Drive to store files and results:

```
from google.colab import drive
drive.mount('/content/drive')
```

4. Upload the following files to the Colab environment:
   - Extracted image file (JPEG/PNG)

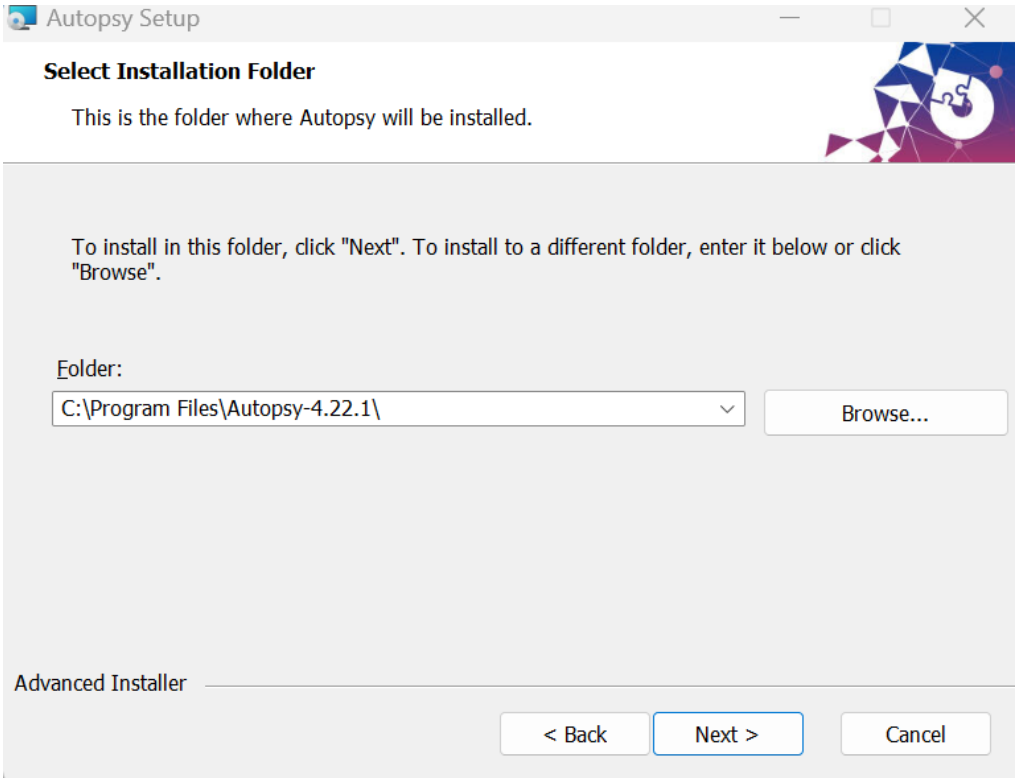


Figure A.5. Autopsy - step2

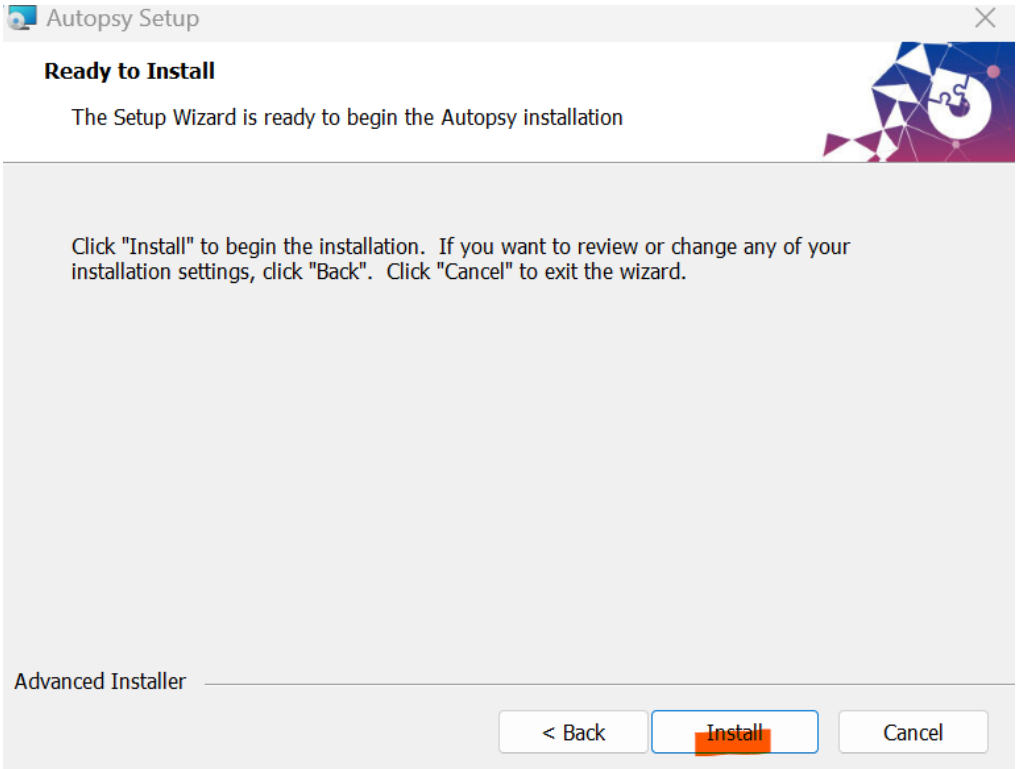


Figure A.6. Autopsy - step3

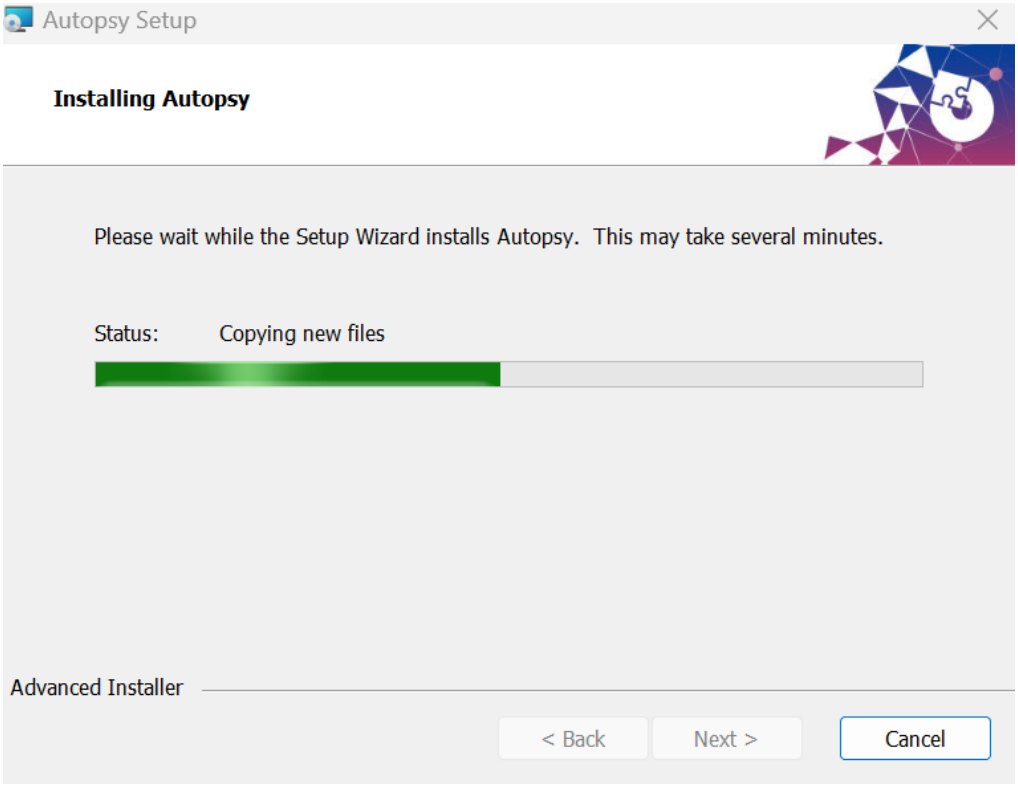


Figure A.7. Autopsy - step4

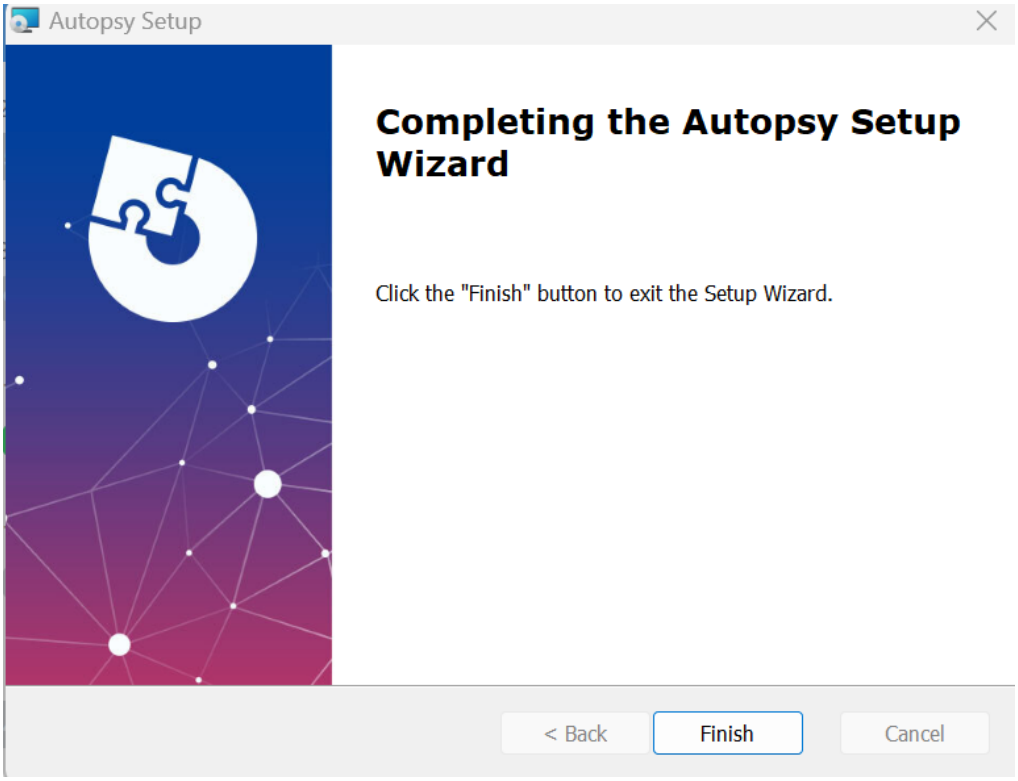


Figure A.8. Autopsy - step5

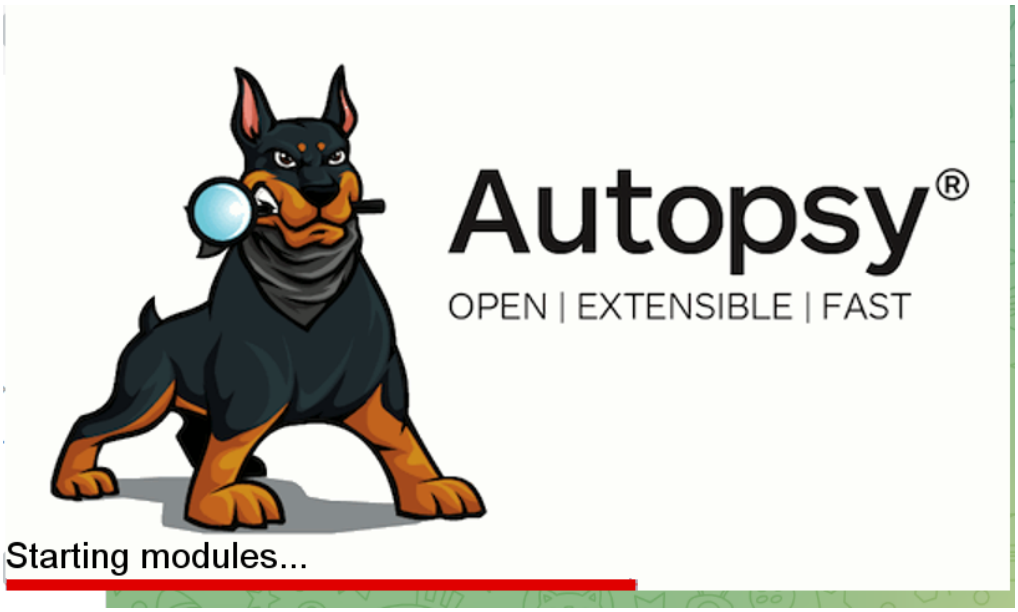


Figure A.9. Autopsy

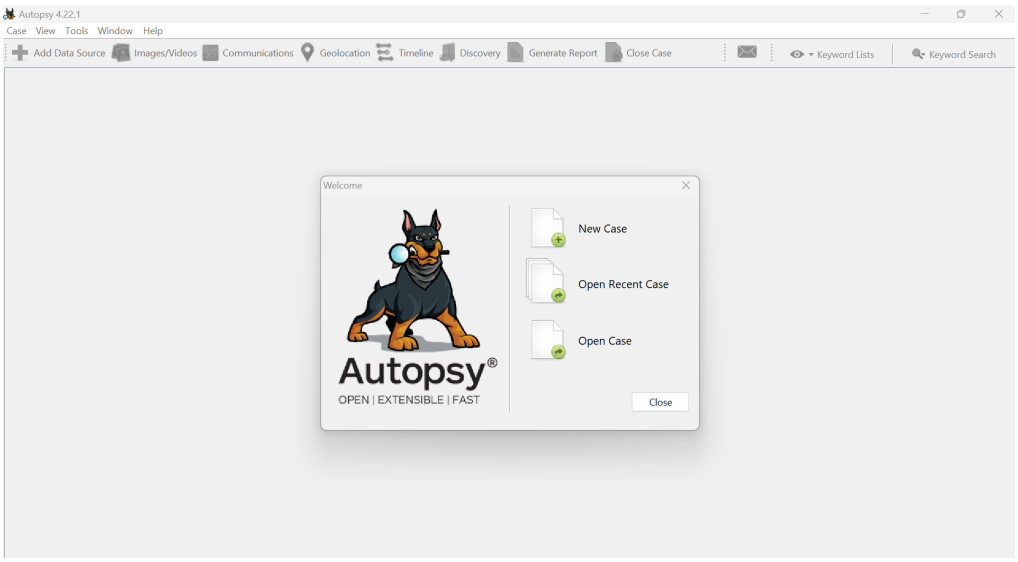


Figure A.10. Autopsy - new case

- Metadata CSV file
- Analysis scripts (as described in the Programmer's Manual)

5. Execute the notebook cells in the following order:
   (a) Install dependencies
   (b) Load configuration and file paths
   (c) Run preprocessing
   (d) Run detection
6. Save or download results before ending the session.

Note: Google Colab sessions have a limited lifetime (typically 12 hours for GPU runtimes). Processing time for a single image + CSV is usually less than 1 minute.

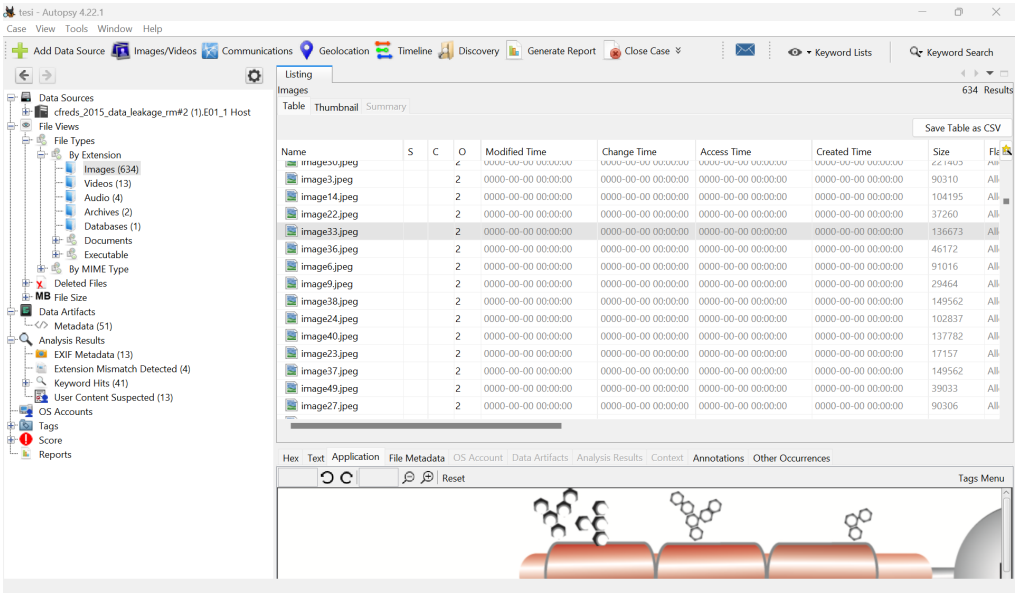


Figure A.11. Autopsy project

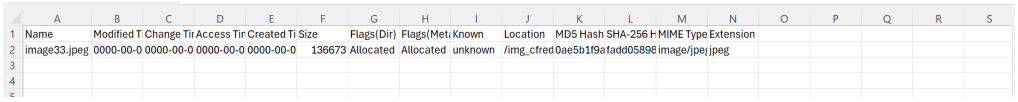


Figure A.12. Autopsy project - csv

A.0.8 Troubleshooting

- **FileNotFoundError:** Ensure the file name in the CSV exactly matches the uploaded image file.
- **PermissionError when mounting Google Drive:** Re-run the mount command and allow Colab access to your Google account.
- **Dependency errors:** Check the `requirements.txt` file and re-run the installation cell.
- **Slow performance:** Enable GPU runtime in Colab (`Runtime -> Change runtime type -> GPU`).

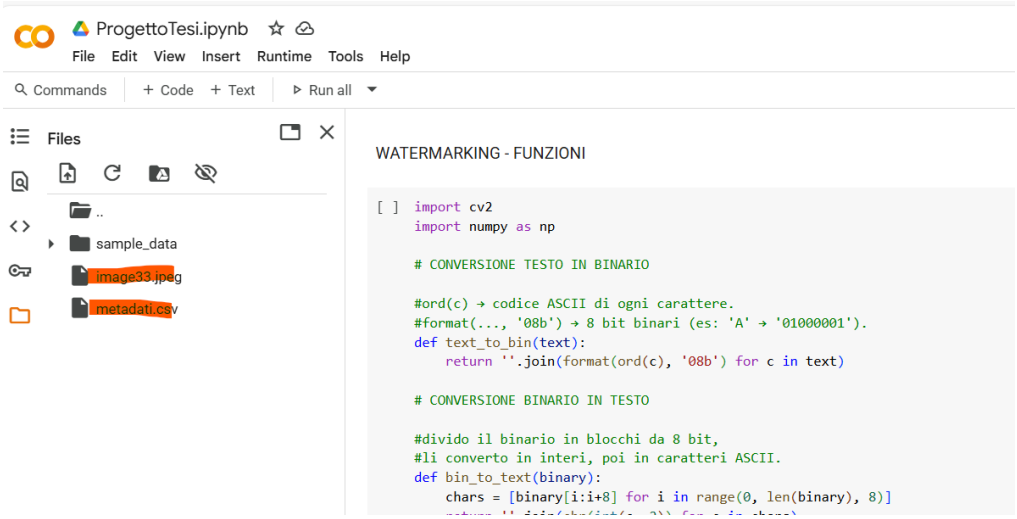


Figure A.13. Google Collab project

Appendix B

Programmer Manual

B.0.1 Introduction

The present Programmer Manual serves as a comprehensive technical reference for the implementation of the project. Its primary purpose is to provide a clear and structured overview of the system's architecture, its core components, and the underlying logic that governs their interaction.

This document is intended for developers and technical contributors who wish to understand, maintain, or extend the implemented solution. It not only presents the source code listings, but also describes their design rationale, modular organisation, and interdependencies.

In addition, the manual details the main data structures, algorithms, and public interfaces, accompanied by configuration guidelines, testing procedures, and best practices. By offering both high-level architectural insights and low-level implementation details, it aims to ensure that the system can be efficiently understood, reliably maintained, and easily adapted for future enhancements.

B.0.2 Architecture Overview

The system architecture consists of four primary components working in synergy to ensure the integrity and traceability of images through watermarking, hashing, and registration on an append-only ledger simulated as a blockchain. The data flow is as follows:

1. **Original Image**: The process begins with the acquisition of the original image as input.

2. **Watermarking**: A unique textual identifier is converted into a binary representation and embedded into the image using the Least Significant Bit (LSB) technique, producing a watermarked image.

3. **Hashing**: Cryptographic SHA-256 hashes are generated for the original image, its associated metadata (extracted from CSV files), and the watermarked image.

4. **Blockchain Simulation**: All relevant information, including timestamp, user identity, image name, and the respective hashes, is concatenated and hashed to produce a block hash. This block is then appended to a CSV-based ledger, simulating an immutable and append-only blockchain.

This workflow enables robust verification of the image's authenticity, the integrity of the embedded watermark, and the consistency of metadata, thus allowing the detection of any unauthorized alterations or tampering attempts.

Module	Responsibility	Primary Files
Watermarking	Embedding and extracting LSB watermarks	watermark.py
Hashing	Computing SHA-256 hashes for files and metadata	hashing.py
Metadata	Reading and exporting metadata from CSV	metadata.py
Blockchain	Simulating an append-only ledger with CSV-based block storage	blockchain.py

Table B.1. Overview of main modules, their responsibilities, and associated files

B.0.3 Main Modules

Detailed usage examples and implementation specifics for each module are provided in the accompanying manual.

B.0.4 Algorithms

Watermark Embedding and Extraction

Purpose To embed a textual identifier into an image by converting it to binary and inserting it into the least significant bits of pixel values, enabling subsequent retrieval for verification.

Pseudocode

```
function embed_watermark(image, watermark_text):
    binary_wm = text_to_binary(watermark_text) + terminator_sequence
    flatten image pixels into a one-dimensional array
    for each pixel in array:
        if more watermark bits remain:
            replace pixel's least significant bit with next watermark bit
        else:
            break
    return reshaped image

function extract_watermark(image):
    flatten image pixels into a one-dimensional array
    binary_data = ''
    for each pixel in array:
        append pixel's least significant bit to binary_data
        if binary_data ends with terminator_sequence:
            break
    remove terminator from binary_data
    return binary_to_text(binary_data)
```

Complexity

- **Time:** $O(n)$, where n is the number of pixels, as the image is traversed once.
- **Space:** $O(n)$ to store the flattened pixel array.

Limitations

- The watermark capacity is limited by the image size.
- LSB watermarking is sensitive to image processing operations such as compression or resizing.
- The terminator sequence is essential to mark the end of the embedded data.

Hashing

Purpose To generate cryptographic SHA-256 hashes of the image file, the embedded watermark (via the watermarked image), and the associated metadata to guarantee data integrity and uniqueness.

Pseudocode

```
function generate_hash(file_path):
    open file in binary mode
    read entire content
    return sha256 hash of content

function hash_metadata(metadata_dict):
    concatenate metadata fields in sorted key order, separated by a delimiter
    return sha256 hash of the concatenated string
```

Complexity

- **Time:** $O(m)$, where m is the size of the file or metadata content.
- **Space:** $O(m)$ to load the file or metadata into memory.

Limitations

- The correctness of the hash depends on the integrity of the original file.
- Metadata must always be serialized in a consistent manner to ensure reproducible hashes.

Blockchain Simulation (Append-only Ledger)

Purpose To simulate an immutable ledger where each block stores a timestamp, user identity, image name, hashes of the image, metadata, watermark, the previous block's hash, and its own hash.

Pseudocode

```
function compute_block_hash(timestamp, user, image_name, image_hash, metadata_hash, wm_hash, p
    concatenate all fields with a delimiter
    return sha256 hash of the concatenated string

function register_on_blockchain(image_name, image_hash, metadata_hash, wm_hash, user):
    timestamp = current UTC time
    prev_hash = retrieve last block hash or a default if none exists
    block_hash = compute_block_hash(...)
    append [timestamp, user, image_name, image_hash, metadata_hash, wm_hash, prev_hash, block_
```

Complexity

- **Time:** $O(1)$ for writing a new block, $O(n)$ to read the last block hash (where n is the number of blocks).
- **Space:** Grows linearly with the number of blocks.

Limitations

- The ledger is local and not distributed, lacking the decentralization features of true blockchains.
- The CSV file is susceptible to tampering unless protected by external measures such as backups or physical access controls.

B.0.5 Implementation and Testing

This section presents the practical application and testing of the previously described functions within the image integrity verification system. The testing code was developed in Python, leveraging libraries such as OpenCV for image processing, CSV for metadata handling, and hashlib for cryptographic hashing. The objective is to embed watermarks, generate hashes, register the data on the blockchain simulation, and validate the system's robustness against different tampering scenarios.

Setup and Initial Data Loading

The testing procedure begins by defining the relevant file paths, including the original image, the output watermarked image, a tampered image placeholder, and the CSV file containing metadata. The original image is loaded using OpenCV, and a verification message confirms successful loading:

```
img = cv2.imread(image_name)
if img is None:
    print("Error loading image.")
else:
    print("Original image loaded successfully.")
```

Metadata associated with the image is extracted from a CSV file by matching the image filename. If no metadata is found, the program raises an exception to ensure data consistency.

Watermark Embedding and Hashing

The extracted metadata fields are concatenated into a single textual string, augmented with an identifying watermark phrase. This string is then embedded into the original image using the watermarking function based on Least Significant Bit manipulation. The watermarked image is saved to disk:

```
watermarked_img = embed_watermark(img, watermark)
cv2.imwrite(watermarked_path, watermarked_img)
```

Subsequently, SHA-256 hashes are computed for the original image file, the concatenated metadata string, and the watermarked image, ensuring integrity and traceability of each element:

```
original_hash = generate_hash(image_name)
metadata_hash = hashlib.sha256(metadata_text.encode()).hexdigest()
watermarked_hash = generate_hash(watermarked_path)
```

Blockchain Registration and Verification

The computed hashes along with the image name and user identifier are registered on the simulated blockchain ledger to maintain an immutable record:

```
register_on_blockchain(image_name, original_hash, metadata_hash, watermarked_hash, user="giada
```

Verification functions are then invoked to compare the current state of images and metadata against their original, expected states. This includes checks on the original image, the watermarked image, and the metadata:

```
verify_image(image_name, image_name)
verify_image_with_wm(watermarked_path, image_name)
verify_metadata(metadata_csv, image_name)
```

Simulated Attack Scenarios

To evaluate the robustness of the system, three types of tampering attacks are simulated and detected:

Attack 1: Modification of the Original Image A tampered image is created by altering a pixel region with a red square and applying a Gaussian blur filter. The verification function detects inconsistencies relative to the original image. The tampered image is displayed using Matplotlib to illustrate the visual impact of the attack:

```
tampered_img[50:100, 50:100] = [0, 0, 255]  # red square
tampered_img = cv2.GaussianBlur(tampered_img, (9, 9), 0)  # blur filter
cv2.imwrite(tampered_path, tampered_img)
verify_image(tampered_path, image_name)
```

Attack 2: Forged Watermark Insertion An image is created by embedding a fake watermark string, simulating an attacker's attempt to mislead the verification system. The verification function confirms the watermark does not correspond to the legitimate metadata:

```
img_with_fake_watermark = embed_watermark(img, "Fake watermark inserted by attacker")
cv2.imwrite(fake_path, img_with_fake_watermark)
verify_image_with_wm(fake_path, image_name)
```

Attack 3: Metadata Tampering The metadata CSV file is duplicated and then modified by overwriting a specific field for the tested image. Verification reveals discrepancies in the metadata hash compared to the original registration:

```
shutil.copy(metadata_csv, metadata_csv_modified)
# modify metadata entry for the image
verify_metadata(metadata_csv_modified, image_name)
```

Summary

This testing suite demonstrates the end-to-end capability of the system to embed, verify, and secure images and their associated metadata against various manipulation attempts. The modular structure enables straightforward extension and adaptation to additional tampering vectors and image formats.

Bibliography

[1] "Ai deepfake statisticse", tech. rep.

[2] IBM, "Generative ai: Transforming business and society", tech. rep., IBM, 2025

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need", Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998–6008

[4] A. Bandi, P. Adapa, and Y. Kuchi, "The power of generative ai: A review of requirements, models, input output formats, evaluation metrics, and challenges", Future Internet, vol. 15, 07 2023, p. 260, DOI 10.3390/fi15080260

[5] IBM, "Generative models: The future of ai innovation", tech. rep., IBM, 2025

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks", Advances in Neural Information Processing Systems, vol. 3, 06 2014, DOI 10.1145/3422622

[7] Jasper, "Jasper ai: The ai-powered writing platform", tech. rep., Jasper, 2025

[8] P. P. Ray, "Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope", Internet of Things and Cyber-Physical Systems, vol. 3, 2023, pp. 121–154, DOI https://doi.org/10.1016/j.iotcps.2023.04.003

[9] IBM, "Claude ai: Il sistema di intelligenza artificiale di nuova generazione", tech. rep., IBM, 2025

[10] OpenAI, "Dalle 2: The ai system that can create images from text descriptions", tech. rep., OpenAI, 2025

[11] Wikipedia contributors, "Midjourney", tech. rep., Wikipedia, 2025

[12] Stable Diffusion Web, "Stable diffusion web", tech. rep., Stable Diffusion Web, 2025

[13] S. Mandal, B. Ghosh, S. Chakraborty, and R. Naskar, "Can deepfakes mimic human emotions? a perspective on synthesia videos", 12 2024, DOI 10.1109/TENCON61640.2024.10902983

[14] D. Dunsin, "Enhancing software quality and efficiency: The role of generative ai in automated code generation and testing", 02 2025

[15] B. Yetistiren, I. Ozsoy, M. Ayerdem, and E. Tuzun, "Evaluating the code quality of ai-assisted code generation tools: An empirical study on github copilot, amazon codewhisperer, and chatgpt", 04 2023, DOI 10.48550/arXiv.2304.10778

[16] M. K. Siam, H. Gu, and J. Q. Cheng, "Programming with ai: Evaluating chatgpt, gemini, alphacode, and github copilot for programmers", 2024, DOI 10.48550/arXiv.2411.09224

[17] S. Bhuyan, V. Sateesh, N. Mukul, A. Galvankar, A. Mahmood, M. Nauman, A. Rai, K. Bordoloi, U. Basu, and J. Samuel, "Generative artificial intelligence use in healthcare: Opportunities for clinical excellence and administrative efficiency", Journal of Medical Systems, vol. 49, 01 2025, DOI 10.1007/s10916-024-02136-1

[18] H. Blake, "Generative ai in cyber security: New threats and solutions for adversarial attacks", 12 2024

[19] Gartner, "Generative ai: Transforming industries and creating new opportunities", tech. rep., Gartner, 2025

[20] M. Ashraf, "Generative ai: Challenges and the road ahead", International Journal of Science and Research (IJSR), vol. 13, 10 2024, pp. 716–725, DOI 10.21275/SR241009154508

[21] L. Ramamoorthy, "Evaluating generative ai: Challenges, methods, and future directions", International Journal For Multidisciplinary Research, vol. 7, 02 2025, DOI 10.36948/ijfmr.2025.v07i01.37182

[22] IBM, "Model collapse", tech. rep., IBM, 2024
[23] IBM, "Catastrophic forgetting", tech. rep., IBM, 2025
[24] Proofpoint, "Deepfake: The evolving threat", tech. rep., Proofpoint, 2025
[25] IBM, "Nlp", tech. rep., IBM, 2024
[26] IBM, "Hpc", tech. rep., IBM, 2025
[27] I. S. Magazine, "Deepfake e intelligenza artificiale: tra rischi di sicurezza e vantaggi", tech. rep., ICT Security Magazine, 2025
[28] T. Walczyna and Z. Piotrowski, "Quick overview of face swap deep fakes", Applied Sciences, vol. 13, 05 2023, p. 6711, DOI 10.3390/app13116711
[29] T. Walczyna and Z. Piotrowski, "Quick overview of face swap deep fakes", Applied Sciences, vol. 13, 05 2023, p. 6711, DOI 10.3390/app13116711
[30] A. Groshev, A. Maltseva, D. Chesakov, A. Kuznetsov, and D. Dimitrov, "A new face swap approach for image and video domains", IEEE Access, vol. 10, 01 2022, pp. 1–1, DOI 10.1109/ACCESS.2022.3196668
[31] A. Kadam, S. Rane, A. Mishra, S. Sahu, S. Singh, and S. Pathak, "A survey of audio synthesis and lip-syncing for synthetic video generation", EAI Endorsed Transactions on Creative Technologies, 04 2021, p. 169187, DOI 10.4108/eai.14-4-2021.169187
[32] A. G. Ana Pantelic, "From puppet-master creation to false detection", 2022
[33] S. Mukta, J. Ahmad, M. Raiaan, S. Islam, S. Azam, M. E. Ali, and M. Jonkman, "An investigation of the effectiveness of deepfake models and tools", Journal of Sensor and Actuator Networks, vol. 12, 08 2023, p. 61, DOI 10.3390/jsan12040061
[34] W. Brown and D. Fleming, "Celebrity headjobs: or oozing squid sex with a framed-up leaky", Porn Studies, vol. 7, 10 2020, pp. 357–366, DOI 10.1080/23268743.2020.1815570
[35] M. Bohacek and H. Farid, "Protecting president zelenskyy against deep fakes", 2022
[36] U.S. Department of Homeland Security, "Increasing threat of deepfake identities", tech. rep., U.S. Department of Homeland Security, 2023
[37] F. Muhly, E. Chizzonic, and P. Leo, "Ai-deepfake scams and the importance of a holistic communication security strategy", International Cybersecurity Law Review, vol. 6, 02 2025, DOI 10.1365/s43439-025-00143-7
[38] European Commission, "Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts", Tech. Rep. COM/2021/206 final, European Commission, 2021
[39] A. Orlando, "La regolamentazione del deepfake in europa, stati uniti e cina", Rivista di Diritto dei Media, vol. numero speciale I-2024, 2025
[40] F. Vainionpaa, K. Vayrynen, A. Lanamaki, and A. Bhandari, "A review of challenges and criticisms of the european artificial intelligence act (aia)", 12 2023
[41] European Parliamentary Research Service (EPRS), "Regulating deep fakes: Transparency, ai and protection against manipulation", Tech. Rep. 690039, European Parliament, 2021
[42] V. Azzali and N. Ellecosta, "La questione deepfake in italia: una panoramica", MediaLaws - Rivista di diritto dei media, vol. 3, no. 2023, 2024
[43] P. del Consiglio dei Ministri, "Bozza del disegno di legge in materia di intelligenza artificiale - 23 aprile 2024", 2024
[44] R. G. Penale, "Approvato dal consiglio dei ministri un disegno di legge in materia di intelligenza artificiale", Giurisprudenza Penale, 2024
[45] United States Congress, "Deepfakes accountability act", tech. rep., United States Congress, 2023
[46] United States Congress, "Defiance act: Digital and ethical framework for information and artificial content enforcement", tech. rep., United States Congress, 2023
[47] F. Arslan, "Deepfake technology: A criminological literature review", Sakarya University Journal of Law (SHD), vol. 11, no. 1, 2023, pp. 701–720, DOI 10.56701/shd.1293642
[48] D. K. Citron and R. Chesney, "Deep fakes: A looming challenge for privacy, democracy, and national security", California Law Review, vol. 107, no. 6, 2019, pp. 1753–1819, DOI 10.15779/Z38RV0D15J
[49] Herbert Smith Freehills, "China releases laws to mandate labelling of aigc", tech. rep., Herbert Smith Freehills, 2025
[50] U. Parliament, "Online safety act 2023", 2023

[51] Bacciardi and Partners, "Truffa ai danni di imprenditori italiani: Clonata la voce del ministro crosetto con l'ai", tech. rep., Bacciardi and Partners, 2025
[52] Wikipedia contributors, "Deepfake pornography - taylor swift", tech. rep., Wikipedia, 2024
[53] IARI, "Post veritÃ e politica: l'impatto crescente dei deepfake", tech. rep., IARI, 2025
[54] D. Search, "Ai mimics ceo voice to scam uk energy firm out of Â£200k", tech. rep., DCL Search Blog, 2019
[55] R. C. Danielle K. Citron, "Deepfakes and the new disinformation war", Foreign Affairs, 2019
[56] E. Union, "Regulation (eu) 2022/2065 of the european parliament and of the council of 19 october 2022 on a single market for digital services (digital services act)", tech. rep., EUR-Lex, 2022
[57] R. A. Digitale, "Elezioni e disinformazione: il digital services act alla prova del nove", tech. rep., Agenda Digitale, 2024
[58] M. S. Rana, M. N. Nobi, B. Murali, and A. H. Sung, "Deepfake detection: A systematic literature review", IEEE Access, vol. 10, 2022, pp. 25494–25513, DOI 10.1109/ACCESS.2022.3154404
[59] IBM, "Che cos'Ã¨ un albero decisionale?", tech. rep., IBM, 2025
[60] IBM, "Che cos'Ã¨ una foresta casuale?", tech. rep., IBM, 2025
[61] M. Habeeba, A. Lijiya, and A. Chacko, "Detection of deepfakes using visual artifacts and neural network classifier", 01 2021, DOI 10.1007/978-981-15-4692-1_31
[62] Vettoria, "Percettore multistrato", tech. rep., Vettoria, 2025
[63] X. Zhang, S. Karaman, and S. Chang, "Detecting and simulating artifacts in gan fake images", 12 2019, DOI 10.1109/WIFS47025.2019.9035107
[64] P. Zhou, X. Han, V. Morariu, and L. Davis, "Learning rich features for image manipulation detection", 05 2018, DOI 10.48550/arXiv.1805.04953
[65] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network", 12 2018, DOI 10.1109/WIFS.2018.8630761
[66] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain, "On the detection of digital face manipulation", 06 2020, DOI 10.1109/CVPR42600.2020.00582
[67] H. Nguyen, J. Yamagishi, and I. Echizen, "Use of a capsule network to detect fake images and videos", 10 2019, DOI 10.48550/arXiv.1910.12467
[68] M. Rana and A. Sung, "Deepfakestack: A deep ensemble-based learning technique for deepfake detection", 08 2020, DOI 10.1109/CSCloud-EdgeCom49738.2020.00021
[69] M. Koopman, A. Macarulla Rodriguez, and Z. Geradts, "Detection of deepfake video manipulation", 08 2018
[70] L. Guarnera, O. Giudice, and S. Battiato, "Deepfake detection by analyzing convolutional traces", 2020
[71] S. Agarwal and L. R. Varshney, "Limits of deepfake detection: A robust estimation viewpoint", 2019
[72] M. Masood, M. Nawaz, K. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward", Applied Intelligence, vol. 53, 06 2022, pp. 1–53, DOI 10.1007/s10489-022-03766-z
[73] A. Malik, M. Kuribayashi, S. Abdullahi, and A. Khan, "Deepfake detection for human face images and videos: A survey", IEEE Access, vol. 10, 01 2022, pp. 18757 – 18775, DOI 10.1109/ACCESS.2022.3151186
[74] H. Hasan and K. Salah, "Combating deepfake videos using blockchain and smart contracts", IEEE Access, vol. PP, 03 2019, DOI 10.1109/ACCESS.2019.2905689
[75] C. Chan, V. Kumar, S. Delaney, and M. Gochoo, "Combating deepfakes: Multi-lstm and blockchain as proof of authenticity for digital media", 09 2020, DOI 10.1109/AI4G50087.2020.9311067
[76] Wikipedia, "Digital watermarking", tech. rep., Wikipedia, 2024
[77] A. Qureshi, D. Megias, and M. Kuribayashi, "Detecting deepfake videos using digital watermarking", 12 2021
[78] W. Wan, J. Wang, Y. Zhang, J. Li, H. Yu, and J. Sun, "A comprehensive survey on robust image watermarking", Neurocomputing, vol. 488, 2022, pp. 226–247, DOI https://doi.org/10.1016/j.neucom.2022.02.083
[79] C. Base, "Cosa e' ethereum", tech. rep., Coin Base

[80] National Institute of Standards and Technology (NIST), "Computer forensic reference data sets (cfreds)", 2024

[81] National Institute of Standards and Technology (NIST), "Data leakage case cfreds dataset", 2024