



POLITECNICO DI TORINO

MASTER DEGREE COURSE IN CYBERSECURITY

MASTER DEGREE THESIS

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

Supervisors: prof. Andrea Atzeni, prof. Giuseppe Vaciago

Candidate: Giada Grillo

December 2025

Introduction

The rapid evolution of generative artificial intelligence has brought forth one of the most pressing challenges in today's digital landscape: the proliferation of deepfakes. Originally conceived as experimental outputs of GANs, autoencoders, and diffusion models, these synthetic media now enable the seamless fabrication of hyper-realistic images, videos, and voices that are often indistinguishable from genuine content. While such technologies hold creative and commercial potential, they have increasingly been exploited for fraud, identity theft, disinformation, extortion, and reputational harm, leading to severe social, financial, and political consequences.

According to recent reports, deepfake-related fraud attempts surged by over 3,000% in 2023, highlighting the growing scale and sophistication of this phenomenon.

Despite the gravity of the threat, existing legal and technical responses remain fragmented. Legislative efforts from the EU Artificial Intelligence Act, the Digital Services Act, and the recent Italian DDL introducing the crime of deepfake, to U.S. state laws, the Chinese Deep Synthesis Regulations, and the UK Online Safety Act, reveal an emerging global consensus on the need for transparency, accountability, integrity, and content provenance. Yet, regulatory measures alone are insufficient without corresponding technological infrastructures capable of enforcing these principles in practice. Similarly, most detection methods are reactive, struggling to keep pace with generative AI’s exponential progress.

This thesis therefore advocates for a proactive, preventive approach to authenticity verification, aiming to secure digital media at the point of creation rather than merely detecting manipulation post hoc. The research proposes a lightweight authentication framework that embeds verifiable traces of origin within images through fragile watermarking and ensures data integrity through cryptographic hashing and blockchain-based traceability. In doing so, it bridges the gap between legal theory and technical enforcement, promoting a digital ecosystem founded on transparency, accountability, and trust.

Project Overview

The project presents the design and implementation of a prototype system that operationalizes the legal principles identified in current regulatory frameworks, namely transparency, integrity, and provenance, through a combination of digital watermarking, cryptographic hashing, and a blockchain-inspired ledger.

The system addresses one of digital forensics’ primary challenges: the preservation and immutability of digital evidence. By leveraging blockchain’s tamper-evident and append-only structure, the framework maintains an auditable chain of custody for digital images and their metadata. The watermarking module, based on a fragile Least Significant Bit (LSB) embedding method, inserts imperceptible textual marks capable of detecting even minor alterations. A SHA-256 hashing process generates unique digital fingerprints for both image content and metadata, which are then registered in a simulated blockchain ledger that cryptographically links each entry to the previous one, reproducing the immutability and traceability properties of real distributed ledgers.

Verification procedures combine these components to perform comprehensive integrity checks: any discrepancy between recalculated hashes, stored blockchain values, or embedded watermarks signals possible tampering. The system was tested on datasets from NIST’s Computer Forensic Reference Data Sets, specifically the Data Leakage Case dataset analyzed through the Autopsy forensic platform. Experimental results confirmed the system’s ability to detect pixel-level manipulations, metadata alterations, and unauthorized edits while maintaining the evidentiary integrity required in forensic and judicial contexts.

Beyond its technical implementation, the project also incorporates a legal analysis, mapping each functional component to corresponding legal principles. For in-

stance, watermarking ensures transparency (as mandated by the AI Act and China’s AIGC Measures), hashing preserves integrity (consistent with evidentiary standards and the U.S. Deepfakes Accountability Act), and blockchain provides accountability and non-repudiation (reflecting obligations under the Italian DDL and UK Online Safety Act). This interdisciplinary perspective highlights how emerging AI legislation can be effectively translated into technical compliance mechanisms.

Results and Conclusions

The developed prototype demonstrated strong performance in detecting and authenticating manipulated images, particularly deepfake-related content. Through tests on both online data and forensic datasets, the system consistently identified inconsistencies and differentiated authentic media from tampered ones. The dual-layer mechanism, combining fragile watermarking with cryptographic hashing, proved effective in capturing even subtle modifications to image pixels or metadata, significantly enhancing verification robustness.

The blockchain-inspired ledger, implemented through a CSV-based simulation, successfully maintained the sequential integrity of entries via cryptographic hash chaining, allowing the detection of unauthorized deletions or alterations. Simulations involving diverse tampering types, including image modifications, watermark removal, and metadata corruption, all triggered successful alerts, validating the framework’s practical reliability.

By integrating legal reasoning with technical enforcement, this research demonstrates how the abstract concepts of transparency, authenticity, and accountability can be technically instantiated within digital infrastructures. The system provides a proof of concept for secure, auditable, and regulation-aligned content verification, with potential applications in digital forensics, media authenticity certification, and judicial evidence management.

Future developments could extend this work toward a fully decentralized blockchain deployment, a user-friendly forensic interface, and integration with AI-generated content labeling systems (as envisaged by the AI Act and China’s 2025 AIGC Measures). These improvements would enhance its applicability across investigative, journalistic, and legal contexts.

Ultimately, this thesis contributes to building a more trustworthy digital environment, where technological design actively enforces legal and ethical principles, ensuring that authenticity is not only verified but inherent to digital creation itself.