Metadata Extraction and Image Analysis System

Mrs. P Sowjanya (Guide), Computer Science & Engineering (Cyber Security) Department, Raghu Engineering College, Visakhapatnam, Andhra Pradesh, India.

Bandaru Janusha, Computer Science & Engineering (Cyber Security) Department, Raghu Engineering College, Visakhapatnam, Andhra Pradesh, India.

Chelliboyina Hari Venkata Sai, Computer Science & Engineering (Cyber Security) Department, Raghu Engineering College, Visakhapatnam, Andhra Pradesh, India.

Kavvati Aakash, Computer Science & Engineering (Cyber Security) Department, Raghu Engineering College, Visakhapatnam, Andhra Pradesh, India.

Pangi Modha Sivamani, Computer Science & Engineering (Cyber Security) Department, Raghu Engineering College, Visakhapatnam, Andhra Pradesh, India.

Gandaboyina Jaswanth, Computer Science & Engineering (Cyber Security) Department, Raghu Engineering College, Visakhapatnam, Andhra Pradesh, India.

ABSTRACT

Digital images are widely used as evidence in legal, forensic, cybersecurity and social-media investigations. However, the reliability of digital images is increasingly challenged by metadata tampering, editing software, recompression, screenshots and AI-generated content. This research presents MetaForensicAI, a Metadata Extraction and Image Analysis System for Digital Forensics that performs batch image validation, internal metadata extraction, provenance analysis, authenticity-oriented assessment, timestamp checking, evidence correlation and structured report generation in a unified graphical workflow. The proposed system follows a modular Python-based architecture and integrates an interactive forensic chatbox, analysis modules and reporting components to support explainable decision-making. Instead of depending on system-level external tools or a single indicator, the system combines metadata strength, software traces, structural cues, screenshot signals and synthetic-content indicators to derive practical origin classes. The work is positioned as a systems-oriented research contribution and is intended for further strengthening with controlled experiments, validated metrics and formal citations before submission.

KEYWORDS:

Digital forensics, image provenance, metadata analysis, image authenticity, screenshot detection, AI-generated image detection, explainable AI, forensic reporting.

1. INTRODUCTION

Digital images have become one of the most important forms of digital evidence. They are used in cybercrime investigation, journalism, insurance verification, legal proceedings and law-enforcement analysis. A single image may influence a case decision, establish the timing of an event or support the authenticity of a claim. Because of this, image reliability has become a major research and practical concern.

Traditional image inspection methods often rely on visible content or limited metadata review. In real-world scenarios, this is insufficient. Images may pass through editing tools, messaging applications, social platforms or screen-capture workflows before reaching the investigator. In addition, modern generative AI tools can create photorealistic images with little or no trustworthy metadata. These changes create a need for a forensic system that can analyze multiple sources of evidence and explain its conclusion clearly.

The proposed project addresses this need by developing MetaForensicAI, a metadata extraction and image analysis system that combines extraction, analysis, scoring, interactive explanation and reporting in a single workflow. It features a streamlined graphical user interface (GUI) supporting batch folder processing, efficient caching for rapid re-analysis and an integrated forensic chatbox powered by natural language processing (NLP) to conversationalize metadata exploration. The system is intended to support digital investigators by offering practical provenance analysis rather than only raw metadata display.

2. LITERATURE SURVEY

1. Metadata-based image forensics has been widely studied for analyzing EXIF, XMP, GPS information, software identifiers and device details to determine the origin and history of digital images. Chen and Davis (2019) proposed metadata verification techniques using deep representation learning to improve reliability. These approaches are interpretable and useful for forensic analysis, but they become weak when metadata is missing, stripped or intentionally manipulated using editing tools.

2. Compression and structural forensic techniques have been explored to detect image manipulation through intrinsic properties such as JPEG quantization tables, double compression artifacts and file structure inconsistencies. Levecque et al. (2025) introduced dual JPEG compatibility analysis for reliable forensic detection, while Furushita et al. (2024) proposed lightweight double compression detection for HEIF images. These methods are effective when metadata is unreliable, but they may not always provide complete information about the original provenance of the image.

3. Source attribution and camera forensics research focuses on identifying acquisition-level characteristics such as sensor pattern noise, demosaicing traces and camera-specific imaging patterns. Shabala and Korniichuk (2025) highlighted the role of forensic analysis in verifying image authenticity in cybercrime investigations. These techniques help distinguish native camera images from edited or synthetic images, although they often require high-quality data and controlled conditions.

4. Recent studies have addressed emerging challenges such as screenshot detection, social media transformations and AI-generated image identification. Yang et al. (2025) proposed methods for smartphone screenshot traceability based on metadata, while Davis (2024) analyzed the impact of screenshot software on forensic detection. Additionally, Li et al. (2025) and Kumar et al. (2025) explored AI-generated image detection and deepfake analysis, emphasizing the difficulty of identifying synthetic images that mimic real-world photography but lack consistent forensic traces.

5. Existing forensic tools and systems typically focus on isolated aspects such as metadata analysis or manipulation detection and are often implemented using command-line interfaces. Mohit et al. (2026) proposed blockchain-based provenance verification using perceptual hashing, highlighting the need for stronger traceability mechanisms. However, current solutions lack integration, explainability and user-friendly workflows. Therefore, there is a clear research gap in developing a unified forensic system that combines metadata extraction, provenance analysis, authenticity assessment, explainability, batch processing and natural language interaction within a single graphical framework.

3. IMPLEMENTATION STUDY

3.1 EXISTING SYSTEM

Most image forensic systems use metadata to figure out if an image's real or not. They look at things like what device took the picture when it was taken and where it was taken. The problem is that this information can be easily changed or removed using tools that are easy to find. This makes it hard to trust the results in investigations. Also most systems have a time telling the difference between pictures taken with a camera pictures that have been edited, screenshots and pictures made by artificial intelligence. This makes it hard to be sure about the results.

Another big problem is that we do not really know how these systems come up with their answers. They just give us a yes or no or a score without explaining how they got that answer. This makes it hard for investigators to explain their findings in court or in settings. Current systems also do not have ways to keep track of the evidence like logging and tracking who has handled the images. They also do not have ways to process many images at once or to make reports

automatically. This means investigators have to look at each image one by one which takes a lot of time and effort. Additionally looking at metadata and checking if an image is real are usually two steps, which makes the process slower and less efficient.

3.2 PROPOSED SYSTEM

The MetaForensicAI system we are proposing solves these problems by creating a workflow that includes many different types of analysis. It has a user- interface that makes it easy for investigators to look at one image or many images at once. It also has a chatbox that uses natural language processing to help investigators understand the results.

The system first checks the images to make sure they are safe to use and have not been changed. Then it looks at the metadata in the image files. Uses a special approach that looks at many different signs to figure out where the image came from. This approach is better than systems because it looks at many different things like how much metadata is in the image what software was used to edit it and if it has any signs of being a screenshot or made by artificial intelligence.

The system then looks at the timestamps checks for any signs of tampering and compares the signs to see if they match. This all helps to give a score that shows how likely it is that the image has been changed or made by intelligence. The system also explains how it came up with its answers, which makes the results more transparent and easier to use in reports. Finally it makes a report that is easy to understand and use. By looking at all the signs together the proposed system gives more accurate and reliable results, than current systems.

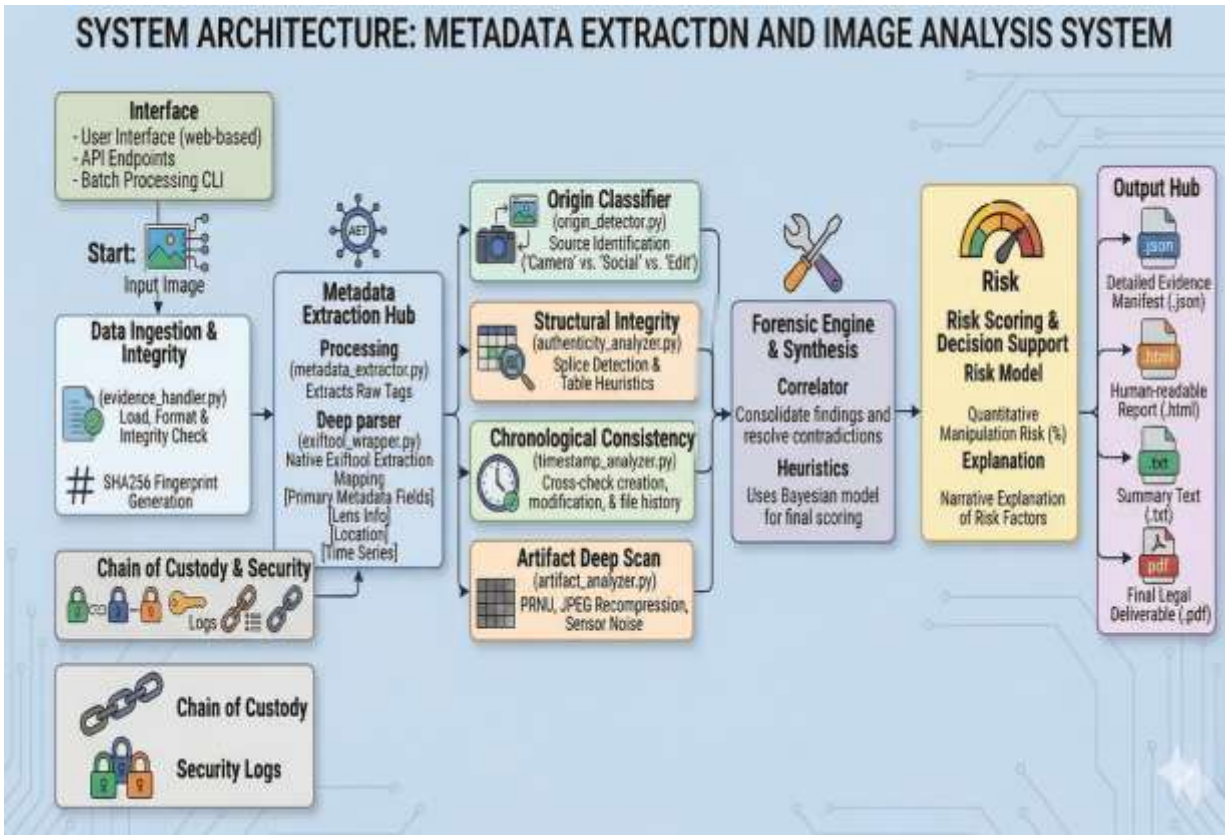
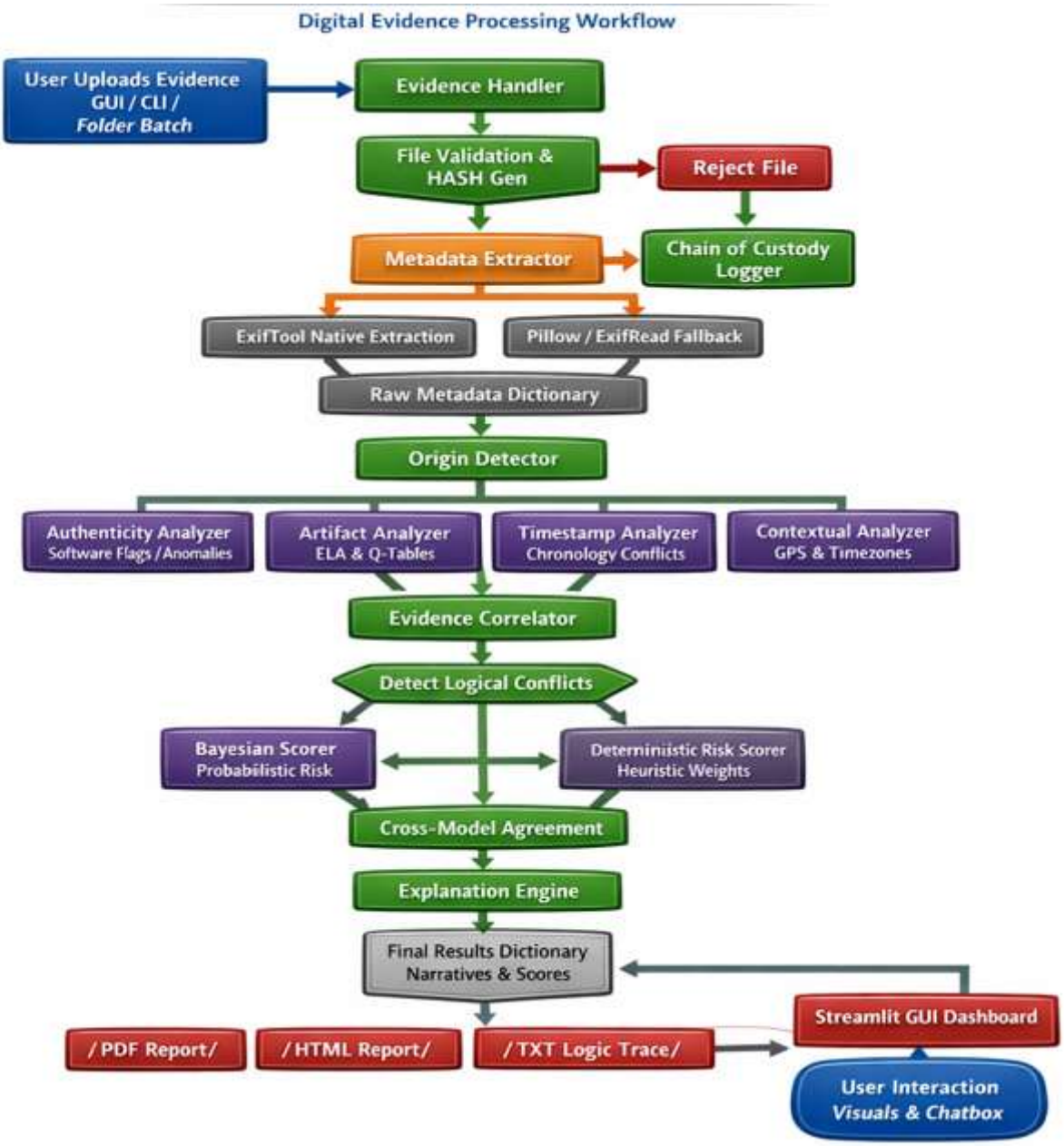


Fig. 1

Fig. 2

METADATA EXTRACTION AND IMAGE ANALYSIS SYSTEM FLOW CHART



4. SOFTWARES AND LIBRARIES DESCRIPTION

The system is implemented in Python and uses a modular repository structure. The major software technologies and libraries used in the project include the following.

PROGRAMMING LANGUAGE

Python is used as the primary programming language for developing the system due to its flexibility ease of use and strong ecosystem for cybersecurity digital forensics and image processing. It supports rapid development and seamless integration of multiple libraries required for metadata extraction analysis and reporting workflows.

CORE LIBRARIES AND TOOLS

The implementation of the MetaForensicAI system is supported by a comprehensive set of tools and technologies that enable efficient forensic analysis, visualization and reporting. At the core of the system is ExifTool, which acts as the primary external dependency for deep forensic metadata extraction. It provides extensive support for reading and analyzing metadata across a wide variety of file formats, making it a reliable engine for extracting raw evidentiary information.

The frontend of the system is built using Streamlit, which powers the interactive graphical user interface, including dashboards, analysis tabs and the integrated forensic chatbox. For visualization purposes, Plotly is used to generate dynamic and interactive charts such as risk scoring indicators, enhancing interpretability for investigators. Pandas plays a key role in handling and structuring metadata by converting complex outputs into organized dataframes that can be easily displayed within the interface.

For image processing and computer vision tasks, the system utilizes Pillow for basic image handling operations such as loading, preview generation and structural validation. To support modern image formats, pillow-heif extends this functionality to HEIC and HEIF files commonly produced by mobile devices. OpenCV is employed for advanced forensic analysis at the pixel level, including detection of manipulation artifacts and structural inconsistencies. Additionally, ImageHash is used to generate perceptual hashes, enabling the identification of duplicate or visually similar images across datasets.

To ensure robustness in metadata extraction, fallback libraries such as ExifRead and PieXif are incorporated. These libraries provide native Python-based mechanisms to extract and manage EXIF data when ExifTool is unavailable, ensuring continuity in analysis.

The system also integrates data science and machine learning capabilities through NumPy and SciPy, which support high-performance numerical computations required for forensic techniques such as error level analysis and quantization table examination. Scikit-learn is used to implement machine learning heuristics and probabilistic models that contribute to the risk scoring engine. For more advanced detection tasks, particularly identifying AI-generated or synthetic images, PyTorch and Torchvision are utilized to build and run deep learning models.

For report generation, ReportLab is used to create structured, court-ready PDF reports that summarize forensic findings in a professional format. Supporting system utilities include PyYAML for configuration management, psutil for monitoring system resources during intensive batch processing and tqdm for providing real-time progress tracking in command-line operations. Together, these technologies form a cohesive and scalable foundation for the MetaForensicAI system, enabling accurate, efficient and explainable digital image forensic analysis.

5. CONCLUSION

This research paper presented a Metadata Extraction and Image Analysis System for Digital Forensics designed to support practical image provenance and authenticity assessment. The system addresses the weaknesses of existing fragmented approaches by integrating metadata extraction, origin detection, timestamp validation, artifact analysis, evidence correlation, explanation generation and reporting in a single workflow. The proposed design emphasizes explainability, forensic traceability and modular implementation.

The project provides a strong systems-oriented foundation for publication. To strengthen the paper further, the next step is to add experimentally verified results, formal literature citations and finalized benchmark tables. Even in its current form, the system demonstrates a clear research contribution in the area of digital image forensics.

6. REFERENCES

[1] Yang Mengxuan, Li Shengnan, Qiu Xiulian, Zeng Jinhu. (2025). [Research on smartphone screenshot image traceability technology based on metadata]. [Forensic Science and Technology], 50(5), 482–488.

[2] Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, Shu Hu (2025). Detecting multimedia generated by large AI models: A survey. arXiv preprint arXiv:2402.00045v4. https://doi.org/10.48550/arXiv.2402.00045

[3] Dr. Mohit Kumar, Dr. Alexei Souri, Dr. Alvin Chan (2025). AI-based deepfake image detection: Robust and explainable approaches for ensuring digital image integrity. Acta Scientiae, 26(2), 390–411. https://www.periodicos.ulbra.org/index.php/acta/article/view/451

[4] Hao Wang, Xin Cheng, Hao Wu, Xiangyang Luo, Bin Ma, Hui Zong, Jiawei Zhang (2025). A GAN-based anti-forensics method by modifying the quantization table in JPEG header file. Journal of Visual Communication and Image Representation. https://www.sciencedirect.com/science/article/abs/pii/S1047320325000768

[5] Yevheniia Shabala, Borys Korniichuk. (2025). Application of forensic analysis to determine the authenticity of images in cybercrime investigations. Management of Development of Complex Systems, (63), 223–229. https://doi.org/10.32347/2412-9933.2025.63.223-229

[6] Bor-Chun Chen, Larry S. Davis (2019). Deep Representation Learning for Metadata Verification. In Proceedings of the IEEE/CVF International Conference on Computer Vision.

[7] Yoshihisa Furushita, Marco Fontani, Stefano Bianchi, Alessandro Piva, Giovanni Ramponi (2024). A Lightweight Double Compression Detector for HEIF Images Based on Encoding Information. *Sensors*, 24(16), 5103. https://doi.org/10.3390/s24165103

[8] Davis, Jacob Benjamin (2024). Thesis on Screenshot Software Effects. A Framework for Analyzing Changes to Manipulated Image Detection Introduced by Screenshot Software on Windows Computers. Auraria Library Digital Repository. https://digital.auraria.edu/work/sc/60b93b21-e0dc-4191-ad54-c2ef8bb69a7f

[9] Apoorv Mohit, Bhavya Aggarwal, Chinmay Gondhalekar (2026). Provenance Verification of AI-Generated Images via a Perceptual Hash Registry Anchored on Blockchain. *arXiv preprint arXiv:2602.02412*. https://doi.org/10.48550/arXiv.2602.02412

[10] Etienne Levecque, Jan Butora, Patrick Bas (2025). Dual JPEG Compatibility: a Reliable and Explainable Tool for Image Forensics. *arXiv preprint arXiv:2408.17106*. https://doi.org/10.48550/arXiv.2408.17106