ҚАЗАҚСТАН РЕСПУБЛИКАСЫНЫҢ ҒЫЛЫМ ЖӘНЕ ЖОҒАРЫ БІЛІМ МИНИСТРЛІГІ
МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РЕСПУБЛИКИ КАЗАХСТАН
MINISTRY OF SCIENCE AND HIGHER EDUCATION OF THE REPUBLIC OF KAZAKHSTAN

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

2025 (24) 4

қазан- желтоқсан

ISSN 2708–2032 (print)
ISSN 2708–2040 (online)

Халықаралық ақпараттық және коммуникациялық технологиялар журналы
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Меншік иесі: АҚ «Халықаралық ақпараттық технологиялар университеті» (Алматы қ.).
Қазақстан Республикасы Ақпарат және қоғамдық даму министрлігіне мерзімді баспасөз басылымын есепке қою туралы куәлік № KZ82VPY00020475, 20.02.2020 ж. берілген
Тақырып бағыты: ақпараттық технологиялар, ақпараттық қауіпсіздік және коммуникациялық технологиялар, әлеуметтік-экономикалық жүйелерді дамытудағы цифрлық технология.
Мерзімділігі: жылына 4 рет.
Тираж: 100 дана.
Редакция мекенжайы: 050040 Алматы қ., Манас к., 34/1, каб. 709, тел: +7 (727) 244-51-09.
E-mail: ijict@iitu.edu.kz
Журнал сайты: https://journal.iitu.edu.kz

Международный журнал информационных и коммуникационных технологий
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Собственник: АО «Международный университет информационных технологий» (г. Алматы).
Свидетельство о постановке на учет периодического печатного издания в Министерство информации и общественного развития Республики Казахстан **№ KZ82VPY00020475,** выданное от **20.02.2020 г.**
Тематическая направленность: информационные технологии, информационная безопасность и коммуникационные технологии, цифровые технологии в развитии социо-экономических систем**.**
Периодичность: 4 раза в год.
Тираж: 100 экземпляров.
Адрес редакции: 050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09.
E-mail: ijict@iitu.edu.kz Сайт журнала: https://journal.iitu.edu.kz

«International Journal of Information and Communication Technologies»
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Owner: International Information Technology University JSC (Almaty).
The certificate of registration of a periodical printed publication in the Ministry of Information and Social Development of the Republic of Kazakhstan, Information Committee No. KZ82VPY00020475, issued on 20.02.2020.
Thematic focus: information technology, digital technologies in the development of socio-economic systems, information security and communication technologies
Periodicity: 4 times a year.
Circulation: 100 copies.
Editorial address: 050040. Manas st. 34/1, Almaty. +7 (727) 244-51-09. E-mail: ijict@iitu.edu.kz
Journal website: https://journal.iitu.edu.kz

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Vol. 6. Is. 4. Number 24 (2025). Pp. 251–273
Journal homepage: https://journal.iitu.edu.kz
https://doi.org/10.54309/IJICT.2025.24.4.015

UDC 004.912

DIGITAL TRACE DETECTION SYSTEM FOR CROSS-DEVICE TEXT FILE

L.G. Rzayeva, D.V. Rakhmatullina*, K.K. Myrzabek, A.B. Khassen
Astana IT University, Astana, Kazakhstan.
E-mail: rakhmatullinadayana@gmail.com

Leyla Rzayeva — PhD, Head of Research and Innovation Center «CyberTech», associate professor, Astana IT University, Astana, Kazakhstan
https://orcid.org/0000-0002-3382-4685;
Dayana Rakhmatullina — BSc, Master's student, Astana IT University, Astana, Kazakhstan
https://orcid.org/0009-0004-6958-1496;
Kamila Myrzabek — Bachelor of Science, Astana IT University, Astana, Kazakhstan
https://orcid.org/0009-0005-3273-1881;
Akbota Khassen — Bachelor of Science, Astana IT University, Astana, Kazakhstan
https://orcid.org/0009-0006-7816-869**.**

Abstract. The proliferation of digital devices and platforms has significantly complicated cybercrime investigations. Traditional forensic tools struggle with fragmented evidence across heterogeneous data types and storage systems. This research presents the Forensic Digital Analyzer, an AI-powered system using modular architecture with deep learning algorithms. Key components include Sentence Transformers for textual embeddings, convolutional neural networks for image analysis, and Whisper models for audio transcription. The interactive web interface provides visualization of file structures, similarity detection, and evidence relationship graphs. The system offers scalable pipelines for entity recognition, metadata extraction, and AI-aided reporting. Experimental results on forensic datasets showed substantial improvements in precision and recall compared to conventional methods. While computationally intensive and occasionally generating false positives in noisy modalities, the system reliably links paraphrased texts, modified images, and compressed audio files. This embedding-based automation represents significant advancement in digital forensic investigation with strong real-world deployment prospects. Future enhancements include multilingual processing, integration of localized large language models, explainable AI frameworks, and domain-specific model optimization.

Keywords: digital forensics, multi-device correlation, sentence transformers, semantic similarity, evidence visualization, forensic automation, trace linking

For citation: L.G. Rzayeva, D.V. Rakhmatullina, K.K. Myrzabek, A.B. Khassen. Digital trace detection system for cross-device text files // International

journal of information and communication technologies. 2025. Vol. 6. No. 24. Pp. 251–273. (In Eng.). https://doi.org/10.54309/IJICT.2025.24.4.015.

Conflict of interest: The authors declare that there is no conflict of interest.

ҚҰРЫЛҒЫЛАР АРАСЫНДАҒЫ МӘТІНДІК ФАЙЛДАРДЫҢ ЦИФРЛЫҚ ІЗДЕРІН АНЫҚТАУ ЖҮЙЕСІ

Л. Рзаева*, Д. Рахматтулина, А. Хасен, К. Мырзабек
Astana IT University, Астана, Қазақстан.
E-mail: rakhmatullinadayana@gmail.com

Лейла Рзаева — PhD, «CyberTech» ғылыми-зерттеу және инновациялық орталығының жетекшісі, қауымдастырылған профессор, Astana IT University https://orcid.org/0000-0002-3382-4685;
Даяна Рахматуллина — магистрант, BSc, Astana IT University
E-mail: rakhmatullinadayana@gmail.com. https://orcid.org/0009-0004-6958-1496;
Камила Мырзабек — ғылым бакалавры, BSc, Astana IT University
https://orcid.org/0009-0005-3273-1881;
Ақбота Хасен — ғылым бакалавры, BSc, Astana IT Universi ty
https://orcid.org/0009-0006-7816-869X.

Аннотация. Цифрлық құрылғылар мен платформалардың көбеюі киберқылмыстық тергеулерді айтарлықтай күрделендірді. Дәстүрлі сот-медициналық құралдар әртүрлі деректер түрлері мен сақтау жүйелерінде бөлшектенген дәлелдемелермен жұмыс істеуде қиындықтарға тап болады. Бұл зерттеу терең оқыту алгоритмдерімен модульдік архитектураны қолданатын жасанды интеллектке негізделген жүйе – Сот-медициналық цифрлық анализаторды ұсынады. Негізгі компоненттерге мәтіндік енгізулерге арналған Sentence Transformers, кескіндерді талдауға арналған конволюциялық нейрондық желілер және аудио транскрипциясына арналған Whisper модельдері кіреді. Интерактивті веб-интерфейс файл құрылымдарын визуализациялауды, ұқсастықты анықтауды және дәлелдемелер арасындағы байланыс графиктерін ұсынады. Жүйе нысандарды тану, метадеректерді алу және жасанды интеллект арқылы есеп беруге арналған масштабталатын конвейерлер ұсынады. Сот-медициналық деректер жиынтығындағы эксперименттік нәтижелер дәстүрлі әдістермен салыстырғанда дәлдік пен еске түсірудің айтарлықтай жақсарғанын көрсетті. Есептеуді қажет ететін және кейде шулы модальділікте жалған позитивтер тудыратынына қарамастан, жүйе парафразаланған мәтіндерді, өзгертілген кескіндерді және қысылған аудио файлдарды сенімді байланыстырады. Енгізуге негізделген бұл автоматтандыру нақты әлемде енгізудің күшті перспективаларымен цифрлық сот-медициналық тергеуде айтарлықтай ілгерілеушілікті білдіреді. Болашақ жетілдірулерге көптілді өңдеу, локализацияланған үлкен тілдік модельдерді біріктіру, түсіндірілетін жасанды интеллект жүйелері және доменге тән модельдерді оңтайландыру кіреді.

Түйін сөздер: цифрлық криминалистика, көп құрылғылық корреляция,

сөйлем трансформаторлары, семантикалық ұқсастық, дәлелдемелерді визуализациялау, сот-медициналық автоматтандыру, із байланыстыру

Дәйексөздер үшін: Л. Рзаева, Д. Рахматтулина, А. Хасен, К. Мырзабек. Құрылғылар арасындағы мәтіндік файлдардың цифрлық іздерін анықтау жүйесі//Халықаралық ақпараттық және коммуникациялық технологиялар журналы. 2025. Том. 6. № 24. 251–273 бет. (Ағыл). https://doi.org/10.54309/IJICT.2025.24.4.015.

Мүдделер қақтығысы: Авторлар осы мақалада мүдделер қақтығысы жоқ деп мәлімдейді.

РАЗРАБОТКА АВТОМАТИЗИРОВАННОЙ СИСТЕМЫ ПОИСКА «ЦИФРОВЫХ СЛЕДОВ» ТЕКСТОВЫХ ФАЙЛОВ ИЗ ДВУХ УСТРОЙСТВ В ЦИФРОВОЙ КРИМИНАЛИСТИКЕ

Л. Рзаева*, Д. Рахматтулина, А. Хасен, К. Мырзабек
Астана ИТ университет, Астана, Казахстан.
E-mail: rakhmatullinadayana@gmail.com

Лейла Рзаева — PhD, руководитель научно-инновационного центра «CyberTech», ассоциированный профессор, Астана ИТ университет, Астана, Казахстан
https://orcid.org/0000-0002-3382-4685;
Даяна Рахматуллина — бакалавр наук, магистрант, Астана ИТ университет, Астана, Казахстан
E-mail: rakhmatullinadayana@gmail.com, https://orcid.org/0000-0002-3382-4685;
Камила Мырзабек — бакалавр наук, Астана ИТ университет, Астана, Казахстан
https://orcid.org/0009-0005-3273-1881;
Ақбота Хасен — бакалавр наук, Астана ИТ университет, Астана, Казахстан
https://orcid.org/0009-0006-7816-869X.

Аннотация. Распространение цифровых устройств и платформ значительно усложнило расследования киберпреступлений. Традиционные криминалистические инструменты испытывают трудности при работе с фрагментированными доказательствами в разнородных типах данных и системах хранения. Данное исследование представляет Криминалистический цифровой анализатор – систему на основе искусственного интеллекта с модульной архитектурой и алгоритмами глубокого обучения. Ключевые компоненты включают Sentence Transformers для текстовых встраиваний, сверточные нейронные сети для анализа изображений и модели Whisper для транскрипции аудио. Интерактивный веб-интерфейс обеспечивает визуализацию файловых структур, обнаружение сходства и графы связей улик. Система предлагает масштабируемые конвейеры для распознавания сущностей, извлечения метаданных и формирования отчетов с помощью ИИ. Экспериментальные результаты на криминалистических наборах данных показали существенное улучшение точности и полноты по сравнению с традиционными методами.

Несмотря на вычислительную интенсивность и случайные ложные срабатывания в зашумленных модальностях, система надежно связывает перефразированные тексты, измененные изображения и сжатые аудиофайлы. Эта автоматизация на основе встраиваний представляет значительный прогресс в цифровой криминалистике с сильными перспективами реального внедрения. Будущие улучшения включают многоязычную обработку, интеграцию локализованных больших языковых моделей, объяснимые ИИ-фреймворки и оптимизацию доменно-специфичных моделей.

Ключевые слова: цифровая криминалистика, корреляция между устройствами, трансформеры предложений, семантическое сходство, визуализация улик, криминалистическая автоматизация, связывание следов

Для цитирования: Л. Рзаева, Д. Рахматтулина, А. Хасен, К. Мырзабек. разработка автоматизированной системы поиска «цифровых следов» текстовых файлов из двух устройств в цифровой криминалистике//Международный журнал информационных и коммуникационных технологий. 2025. Т. 6. No. 23. Стр. 251–273. (На анг.). https://doi.org/10.54309/IJICT.2025.24.4.015.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Introduction

The work of digital forensic investigators has become much more difficult due to the quick growth of digital devices, cloud computing, and online communication platforms. Traditional forensic tools have become less effective a t c apturing the complete context of evidence as digital traces are increasingly dispersed across heterogeneous sources, such as computers, smartphones, messengers, cloud services, and external media (Alenezi, 2022). Static image acquisition, file hash comparisons, or simple keyword search are the mainstays of legacy solutions, which frequently fall short in situations involving semantically similar content or multimodal evidence dispersed across devices.

Natural language processing (NLP) and artificial intelligence (AI) approaches have become strong substitutes for automating and improving forensic work in response to these constraints. Even when texts have been paraphrased or obfuscated, contextual similarities can be captured by deep learning models like Sentence Transformers, which allow semantic comparison between text artifacts (Bai, 2020). Similarly, convolutional neural networks (CNNs) can recognize compressed or altered images and analyze visual content. Advanced automatic speech recognition (ASR) models such as Whisper aid in the transcription and comparison of audio clips in the field of audio forensics, even in the presence of noise or poor quality (Burkart, 2021).

Despite these developments, unified forensic systems that integrate explainable AI, multimodal analysis, and semantic similarity into a single workflow are still lacking. Numerous tools currently in use concentrate on discrete tasks or necessitate manual correlation by analysts, which lengthens investigations and raises the possibility of missed connections (Callegati, 2022). Additionally, not many platforms provide trace interpretation with user-friendly visual interfaces, which is crucial for non-technical stakeholders like investigators or attorneys.

The Forensic Digital Analyzer is a modular AI-powered system that automates

the extraction, analysis, and cross-device correlation of digital traces to address these issues. In addition to supporting explainable semantic matching and integrating cutting-edge AI models for various data types, the system provides an interactive graph-based interface for visualizing the relationships between evidence items.

The present study aims to: (1) analyze the shortcomings of existing forensic methods; (2) create an adaptable architecture for automated trace correlation; (3) incorporate semantic models for audio, image, and text data; and (4) assess system performance on experimental datasets. The findings are intended to show that embedding-based analysis expedites the entire research process while simultaneously increasing accuracy and recall.

Literature Review

The growing complexity and diversity of digital evidence sources has led to a significant evolution in digital forensics methodologies. The efficacy of traditional forensic tools like EnCase and FTK is limited in situations involving multiple devices, cloud services, and multimodal data because they mainly use static disk imaging and simple file hash matching (Alenezi, 2022). Evidence that is fragmented, encrypted, or semantically changed is especially difficult for these traditional approaches to handle.

To enhance forensic analysis, recent developments highlight the integration of natural language processing (NLP) and artificial intelligence (AI). Even when content is paraphrased or partially altered, machine learning models—particularly transformer-based embeddings, like BERT—have improved the capacity to correlate semantically similar textual evidence (Bai, 2020). Like this, perceptual hashing techniques and convolutional neural networks (CNNs) offer strong visual similarity detection, which is essential for connecting altered or partially obscured images across devices (Burkart, 2021).

By facilitating efficient transcription and semantic correlation of audio evidence, even in compromised or noisy environments, automatic speech recognition (ASR) technologies—particularly OpenAI's Whisper—significantly improve forensic audio analysis (Callegati, 2022).

Even though current AI-based methods yield encouraging outcomes, they frequently stay conceptual or isolated and are not integrated into real-world forensic workflows. Although they show promise, systems like «SpeechToText» (Guo, 2022) and frameworks investigating joint semantic analysis (Hu, 2021) have limitations regarding scalability and operational deployment.

The limitations of traditional forensic tools and isolated AI-based studies must be addressed, and there is a clear need for comprehensive forensic solutions that integrate semantic analysis, multimodal correlation, and intuitive visualization.

Methods and materials

This study presents a structured methodology for design and implementation that comprises of both technical development and literature-based theoretical underpinning and an empirical substantiation. The research instance comprises of the five main phases: literature review; requirements elicitation; architecture; system development; and evaluation. Each phase was important to show that the proposed system is feasible, scalable, and forensically sound.

A. Literature Review and Theoretical Grounding

The first phase involved an extensive review of academic and industry literature to form the conceptual foundation for cross-device forensic investigations. Recent

studies since 2019 have highlighted significant limitations in traditional forensic tools when it comes to identifying and linking textual artifacts across multiple digital environments.

Particularly relevant is the growing use of graph-based approaches in digital forensics. For example, Wang et al. (2022) proposed a method of tracking digital traces using graph correlation models to link fragmented evidence across devices. Similarly, Lo et al. (2022) investigated the use of explainable AI in forensic detection of cross-platform botnets, showing how interpretable graph neural networks can support legal standards for evidence admissibility.

Another key source was Navanesan et al. (2024), who emphasized the lack of automation and contextual linkage in forensic investigations involving text-based artifacts. Their work identified the need for combining machine learning, natural language processing (NLP), and metadata analysis to detect and correlate textual data fragments in dynamic environments.

Through this literature review, the study established a theoretical model that integrates AI- driven semantic analysis, entity recognition, and graph-based data structures to correlate digital evidence across two or more devices. These insights directly informed system design, especially in ensuring both operational effectiveness and legal robustness.

B. Requirements Gathering

To ground the system design in real-world use cases, requirements were gathered from cybersecurity professionals and digital forensic practitioners. These included discussions with forensic analysts working in corporate cybersecurity, who highlighted frequent issues such as:Fragmentation of evidence across user devices and cloud platforms;

Inconsistent metadata due to file manipulation or time-shifting;

Lack of automation in linking semantically related documents;

The need for evidence to remain interpretable and defensible in legal settings.

From these consultations, a clear set of functional and non-functional requirements was developed. These included support for diverse text formats (docx, .pdf, .txt), automated metadata extraction, semantic similarity detection, graph-based trace correlation, and role-based access control.

C. Architectural Design

The system was designed with modularity and scalability in mind. The backend was developed in Python using the Flask framework. A graph database (Neo4j) was used to model and visualize relationships between forensic elements such as files, devices, and users.

Core components included:

NLP engine for extracting named entities, timestamps, and semantic patterns;

Graph engine for modeling relationships and detecting cross-device correlations;

• Metadata analysis module for file attributes like hashes, timestamps, and origin;

• Secure user management system with role-based access and audit logging;

• Web interface for visualization and interaction.

All components interact through RESTful APIs to allow for future extensibility. The use of cryptographic hashing (SHA-256) ensures integrity of the analyzed content.

D. System Development

During this phase, all modules were built and integrated. The system supports uploading forensic dumps, parsing text files, extracting relevant metadata, and analyzing semantic content. After preprocessing, each file is represented as a node in the graph structure, and potential links are generated based on content similarity, timestamps, authorship, and origin device.

Graph visualizations allow forensic experts to explore relationships and infer behavioral patterns across devices. For instance, in a simulated case of corporate data leakage, the system successfully identified a confidential report shared via email and later modified and re-uploaded from a mobile device, even after the filename had changed.

Security considerations were addressed through encrypted storage, input validation, and continuous audit logs tracking system usage.

E. Testing and Evaluation

The final phase involved thorough testing of system functionality, security, and performance. Unit and integration tests verified data flows, processing pipelines, and UI components. A series of real-world inspired scenarios were used to evaluate system effectiveness. In one test case, document fragments from two devices (a laptop and a smartphone) were analyzed. Despite differing filenames and altered metadata, the system correctly identified links

based on semantic similarity and residual metadata (e.g., author field, creation timestamp).

-Key evaluation metrics included:

-Precision and recall of trace correlation;

-Time efficiency for data ingestion and graph generation;

-System usability (assessed through expert feedback);

-Legal transparency, judged by the explainability of AI results.

Results indicated that the system significantly reduced the time and effort required for multi-device forensic investigations while improving the consistency and accuracy of traceability findings.

Foundations of Digital Forensics and Evidence Handling

In digital forensics, evidence is collected, preserved, examined, and presented in court (Alenezi, 2022). Digital evidence includes any electronically stored or transmitted information that may support a legal hypothesis (Bai, 2020). In multi-device investigations, evidence is often scattered across local files, cloud platforms, and messaging apps, making it difficult to reconstruct (Burkart, 2021).

The growing complexity of digital environments requires working with large volumes of data, varied file types, and fast-evolving technologies, including mobile devices, IoT, and virtual machines. Advanced methods, such as live system examination and memory capture, are necessary to preserve volatile data. According to the Daubert standard and ACPO guidelines, evidence must be preserved securely and accurately for legal admissibility.

Distributed evidence often resides in dynamic, encrypted, or proprietary systems. Understanding data privacy laws (GDPR, CCPA) and maintaining forensic readiness is crucial (Callegati, 2022). Metadata—like timestamps, access logs, and user IDs—helps trace activity. Machine learning now assists in detecting forged data and anomalous behavior (Guo, 2022).

Automation accelerates forensic procedures: disk imaging, system data extraction, and keyword searches can occur with minimal manual input. It also reduces error rates in large-scale cases, but to be legally accepted, tools must be validated and documented (Hu, 2021). Cybercriminals employ anti-forensic techniques such as file deletion, data tampering, and disappearing messages, requiring anomaly detection and entropy analysis (Jain, 2023).

Digital forensics lies at the intersection of law, data science, and computer science. Experts from all three fields are required to manage volatile data and self-deleting messages. Cloud sync features like versioning and auto-deletion add further challenges (Kim, 2023).

Aligning timelines from devices with inconsistent clocks and missing timestamps is difficult, and data ownership in shared environments can be unclear—sometimes inferred through usage patterns or biometrics (Lo, 2022).

Triage-based approaches prioritize key artifacts (apps, messages) before deeper analysis. Standard methods should be combined with modern forensic tools (Mannino, 2021). Investigations must be timely, thorough, rule-compliant, and well-documented to ensure legal defensibility.

For example, when a user creates a file on a laptop and later opens it on a smartphone, the system aims to automatically trace and associate all digital footprints across both devices. It analyzes file systems, app data, user history, metadata, and log files to track the document's lifecycle using hash comparisons and event correlation. Sensor data is centralized for effective timeline reconstruction.

Automation reduces human error and investigation time. Devices are analyzed systematically, and reporting is centralized. The system's architecture (Figure 1) is scalable, supporting additional users and network environments.

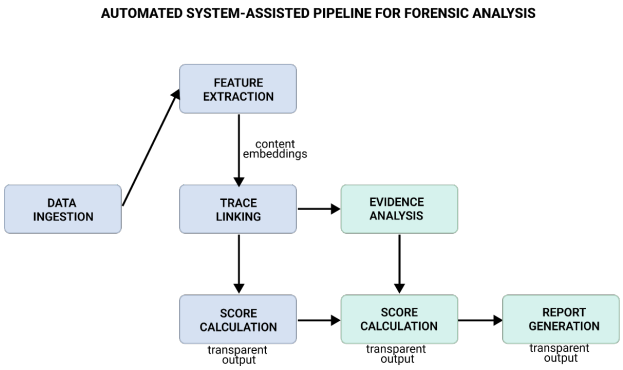

Fig. 1. System architecture for automated detection of digital traces of text files across two devices

Automatically detecting text files on computers by tracing their signatures calls for a smooth blend of both forensic methods and automation. Due to having two devices that use different systems, artifact examination is not always simple—for example, Windows and Android both require different methods for parsing managed files, MFT records and databases (ext4, SQLite, caches) (Marturana, 2020).

Because devices use more than hash matching (for example, SHA-256), files are regularly renamed, changed a little or part of them are deleted; for these reasons, metadata comparison, entropy analysis and content fingerprinting are necessary to discover near-duplicate files (Narayan, 2020). Since the clocks on devices are not

usually synchronized, having the right timeline is essential; this is why event deltas and contextual logging are used in fixing time stamps (Navanesan, 2024).

The system avoids wrongly linking artifacts when a given context—such as the source of the document, controls on who can access it and the way it is called—points to a testing file or script (Samek, 2021). Steps are recorded, every artifact is hashed, and a complete audit track is created to confirm the evidence's integrity, per guidelines from ISO/IEC 27037:2012 and Daubert (Schneider, 2020).

Automation assists with, but does not take the place of, expert evaluation, screens and sorts data quicker, minimizes common mistakes and arranges initial details for further look-over. The increasing difficulty of digital evidence requires investigators to have dependable tools for examining multiple devices (Sha, 2022).

To classify artifacts, effective systems regard primary as those only in the main folder, secondary as items associated with the main files such as autosaves and contextual as logged records both by the system and the user. With this taxonomy, profiles can be used to easily choose and review the most useful digital evidence in any case, making both analysis and results more accurate and quicker.

Table 1. Classification of Digital Traces Related to Text Files

Trace Type	Example Artifacts	Forensic Relevance
Primary	File hash (SHA-256), physical file copy, Word document in Documents folder	Direct evidence of file presence and content
Secondary	Auto-saved versions (.asd), recent file lists (jump lists), thumbnails in cache	Indicates usage, editing, or viewing activity
Contextual	System logs, app launch logs, user login sessions, USB device records	Supports timeline reconstruction, user attribution, and intent verification
Residual	Deleted file traces in unallocated space, shadow copy remnants, log fragments	Confirms prior file existence, even post-deletion
Derived	Filename similarity, entropy analysis results, partial content matches	Used for fuzzy matching or inference when full data is unavailable

Combining many different forms of trace evidence under a single analysis method is key in this field. Such systems must find matches by comparing traces with their counterparts, even when the exact files are not available, by reviewing their filenames, structures and logs. Time- sensitive tasks make the system depend on certain, reliable traces, while every detail in a trace is brought out in an interactive window for in-depth studies.

This system makes it possible for time-ordered narratives by classifying, arranging by priority and listing evidence in order of use, making it easy to view file histories on various devices. Every artifact identified in the pipeline is labelled with

its device, the date and time and a confidence score, so trace pairs can be ranked using similarity, timing and context.

With this approach, certain artifacts (like duplicate SHA-256 hashes or matching file structures for renamed/edited files) are correlated by computing similarity coefficients based on previous casework (Narayan, 2020; Stelly, 2020). By applying event deltas, like a boot or login, devices' clocks are made equal for correct event timeline reconstruction (Navanesan, 2024). With behavioral inference, the application type, the directory and peripheral activity are used to help distinguish accidental matches from real association (Samek, 2021).

All correlation forms can be traced and explained which aids in bringing them into a courtroom and makes them easy for investigators to check (Schneider, 2020). Combing different following and contextual data, the system can draw strong conclusions, even when information is lacking.

Common forensic tools can't combine evidence from many devices and usually overlook important connections when files don't have exact copies. Reasoning using partial matches, edit distances or close times is not possible with these tools, so analysts rely on manually – analyzing the observations (Narayan, 2020). Dispatchers must deal with different proprietary formats between platforms which makes both automation and standardization of data difficult (Sha, 2022). Without complete support and accurate links, high-stakes cases are vulnerable to operational and legal issues.

Such traditional systems cannot adjust to advanced situations such as cloud storage, encrypted containers or instant file movements (Samek, 2021). Most systems do not track origins of items, resulting in static reports which limited the ability to visually connect evidence across anywhere, anytime moments (Schneider, 2020). In addition, they are not able to guess the intentions behind anti-forensic methods or follow the complete history of each artifact closely (Yang, 2021).

Using modular parsing, trace scoring and tracking several correlations, the system is designed to link evidence together, regardless of its partial, unstable or fragmented state. This approach is shown in Figure 1.2 to solve the main problems seen with older forensic tools in multi- device investigations.

By using AI in digital forensics, organizations are changing from relying on set rules to drawing inferences from data. AI methods such as machine learning and deep learning help investigators examine large amounts of varied data, discover patterns and use probability to make decisions. This way of working is especially appropriate when evidence has been split across devices and cannot be organized.

In comparison to hash or name-based methods, AI models can successfully link traces by detecting matching information and meanings on different devices. Both TF-IDF and BERT algorithms allow the linking of file names, document content and messages; clustering methods organize events that happen at different times (Wang, 2022). Using neural network architecture, it is possible to detect unusual activity and intentional changes made to user traces.

Because of AI, a working model in one case can be used to speed up investigations in different contexts. However, being able to interpret models is crucial for courts, so systems built from both transparent models and scores of confidence are commonly chosen. Since training forensic AI needs expert data, these data can sometimes by limited by the additional rules and regulations.

The new system includes AI at the first, second and third stages: (1) feature extraction by learning semantic/contextual meanings from filenames, content and logs (2) identification of related artifacts between devices and (3) choosing the most promising paths to explore. The layered and modular design meets forensic standards, encourages transparency and you can see it in Figure

1.3. As a result, AI works alongside forensic reasoning and provides intelligence designed for the

demands of today's crime investigations.

So that AI is legitimate when used in forensics, each AI decision is logged with proof of its inputs, model reliability and a clear, human-friendly explanation of how that decision was made, supporting Daubert compliance. Because forensic information is highly sensitive and often spread out among many sources, building good real-world datasets is a key obstacle. It therefore produces synthetic traces in test conditions and adds data to ensure the training dataset is well- balanced and covers a variety of examples.

Diverse data sets, various techniques and constant updates minimize bias and reduce the chances of malware detectors missing new or changed threats. It means both efficiency and deeper understanding, as in the case of using different devices: previously, distinct artifacts would be the focus, whereas AI allows the reconstruction of a document's use by pointing out similar evidence. When it comes to corporate incident response, AI helps identify suspicious sessions and therefore reduces the chances of missing anything important and shortens the time spent on investigations.

AI helps show connections between several types of information: even many small details, when organized with other evidence, can create a meaningful description of the document's role in a lawsuit. The modules used are artifact clustering, looking for anomalies, scoring traces and reconstructing history. These modules produce the same results repeatedly which makes evidence more reliable. Human-in-the-loop design allows analysts to set up boundaries and make the final calls, using their experience in combination with the scalability of AI.

Figure 2 provides a representation of the architectural aspects of how we propose artificial intelligence to be integrated into a credible and reproducible forensic investigative pipeline. The forensic pipeline is made up of three layers of analytics: feature engineering to extract traits; correlation of artifacts across devices; and evaluating investigative hypotheses. In the first layer, AI will add value to raw data by producing semantic vectors and contextual embeddings; in the next layer, artifacts with differing formats, naming, or provenance will be linked using models which will be identified through machine learning; and finally the investigation will conduct a ranked hypothesis based on the collected trace level inferences of objective timelines which can be used for data-informed decision making for ranking of investigative actions. Each investigative layer of the proposed pipeline is modular, which will promote explanations and audit trails, and compatible with forensic standards.

systematically, and reporting is centralized. The system's architecture (Figure 1) is scalable, supporting additional users and network environments.

In forensics, decision trees are used for static artifact evaluation, temporal models such as RNNs, LSTMs and TCNs are used for analysis of behavior and deep learning with BERT and RoBERTa is used to detect similar contents (near-duplicates).

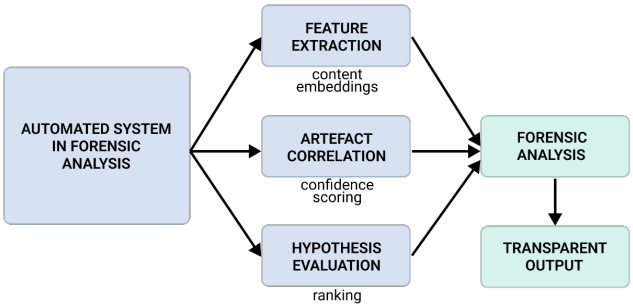

Fig. 2. Automated system in forensic pipeline

Graph Neural Networks (GNNs) study the way that user actions are linked to the things they interact with (Stelly, 2020).

AI researchers can change architecture to adapt different models to specific jobs or the volume of information, using both artificial and real but anonymous data during training. Checking if the system is still accurate, detecting any deviations and updating it when needed keep the system reliable. Table 2 shows how different forensic tasks match different AI model types, features, outputs and scenarios.

Table 2. Alignment of Digital Forensic Tasks with AI Model Types, Features, and Outputs

Forensic Task	AI Model Type	Input Features	Output	Application Scenario
Artifact classification	Decision Trees, Gradient Boosting	File metadata: timestamps, size, path depth	Relevance label (e.g., suspicious/benign)	Triage and evidence filtering
Anomaly detection	LSTM, TCN, RNN	User activity sequences, access logs, file operations	Anomaly score, session risk level	Insider threat and behavioural deviation analysis
Trace correlation	BERT, RoBERTa	Filenames, document content, in-app messages	Semantic similarity score, trace linkage strength	Matching renamed/ modified files and detecting hidden duplicates
Timeline modelling	Seq2Seq models	Event logs, device timestamps	Chronologically ordered events with confidence scores	Multi-device timeline reconstruction
Relational inference	Graph Neural Networks (GNNs)	Artifact interaction graphs (nodes, edges)	Inferred links, path strengths	Reconstructing usage chains and indirect relationships

Narrative evaluation	Ensemble models (RF, SVM)	Combined trace features, contextual indicators	Ranked hypotheses of user behaviour or file intent	Ambiguous case resolution and decision support

The proposed system features a dedicated visualization layer designed to bridge the cognitive gap between machine inference and human reasoning. The interface is grounded on three core principles: transparency of AI outputs, traceability of decision logic, and task-specific interactivity. Outputs from AI modules—such as artifact relevance, anomaly detection, and trace correlation—are rendered via interactive timelines and graph-based views. Artifacts are represented as nodes annotated with metadata, timestamps, and confidence scores, while AI- inferred relationships (e.g., file transfers or session overlaps) are shown as directed edges with contextual tooltips (e.g., "93.1% semantic similarity to previous edit").

To ensure consistent relevance scoring across heterogeneous data types, the system employs a weighted feature aggregation model:

$$Relevance(a_i) = \sigma(\sum_{k=1}^{n} \quad w_k \cdot f_k(a_i) + b) \tag{1}$$

where $fk(ai)$ denotes forensic features such as temporal locality, metadata entropy, or document structure; $wk$are learned weights, and \sigma is the sigmoid activation function. These relevance scores support both triage and downstream correlation tasks.

Artifact-to-artifact associations are computed using a similarity function combining semantic, temporal, and structural proximity:

$$Corr(a_i, a_j) = \alpha \cdot sim_{text}(a_i, a_j) + \beta \cdot sim_{time}(a_i, a_j) + \gamma \cdot sim_{path}(a_i, a_j) \tag{2}$$

The semantic term *simtext* leverages **cosine similarity** between document embeddings:

$$sim_{text}(a_i, a_j) = \frac{\overrightarrow{v_i^{text}} \cdot \overrightarrow{v_j^{text}}}{|\overrightarrow{v_i^{text}}| \cdot |\overrightarrow{v_j^{text}}|} \tag{3}$$

This enables the identification of paraphrased documents and functionally similar artifacts.

The final hypothesis score aggregates relevance and correlation across linked artifacts:

$$HypothesisScore(H_k) = \frac{1}{m}\sum_{i=1}^{m} \quad Relevance(a_i) \cdot Corr(a_i, a_j) \tag{4}$$

All hypotheses and user decisions are auditable. The interface also supports cognitive load management via collapsible views, confidence-based filtering, and behavioral

heatmaps. Explainability is embedded in context: investigators can inspect which features triggered an inference without needing to interpret raw model internals. This architecture enables legally defensible, interpretable, and scalable digital forensic workflows aligned with investigative reasoning rather than replacing it.

System Architecture and Component Design

The Forensic Digital Analyzer (FDA) has been conceived as an end-to-end, web-based framework that automates the extraction, enrichment, and correlation of digital artefacts collected from two independent device images or archive dumps. Immediately after uploading, each evidence set — whether an E01 forensic image or a conventional ZIP/TAR archives mounted or unpacked inside an isolated container. Every file discovered is hashed, its path is normalized, and a record is inserted into a PostgreSQL repository together with basic size and timestamp metadata. This "cradle-to-grave" capture of provenance data guarantees that subsequent analytical claims can always be traced back to the exact byte sequence from which they were derived, thereby satisfying evidentiary requirements of repeatability and authenticity.

The server-side stack is implemented on Django but deliberately subdivided into three co- operating functional loops. The first loop, preprocessing, is responsible for assigning a modality label to every file and applying the appropriate extractor. Text documents (TXT, DOCX, PDF) are parsed directly; scanned pages are processed with Tesseract OCR; audio recordings are transcribed by the Whisper automatic-speech-recognition model; images are normalised, re- scaled, and then forwarded to a convolutional encoder. Each extractor produces raw content, low-level metadata, and at least one fixed-length vector representation. On the textual branch,

SentenceTransformer generates 384-dimensional embeddings; on the visual branch, a ResNet derivative outputs deep visual descriptors, while perceptual hashes capture pixel-level duplicates. Extracted text is further enriched with named entities provided by a fine-tuned BERT model, and topic labels are requested from an external LLM (Gemini) under a rate-limited, audited connection. All outputs—including raw text, transcripts, embeddings, hashes, entities, and topics—are committed to the same relational store, preserving a coherent, queryable view of each artefact.

The second loop, analysis, performs cross-device correlation. For every file pair spanning the two evidence sets, the system consults the relevant vector spaces, time stamps, and directory structures to decide whether a substantive link exists. Similarity is measured by cosine distance for embeddings and by Hamming distance for perceptual hashes; temporal offsets and path patterns serve as secondary cues that either reinforce or penalise a tentative match. Whenever the computed similarity exceeds a modality-specific threshold, the pair is promoted to a "match" instance that records the two file identifiers, the similarity score, the detected match type (duplicated text, paraphrased paragraph, visually altered screenshot, and so forth), plus any contextual details that may help an investigator understand the linkage. The entire decision path— including thresholds, model versions, and feature weights— is saved alongside the match, ensuring that every automated inference can later be reconstructed or challenged if necessary.

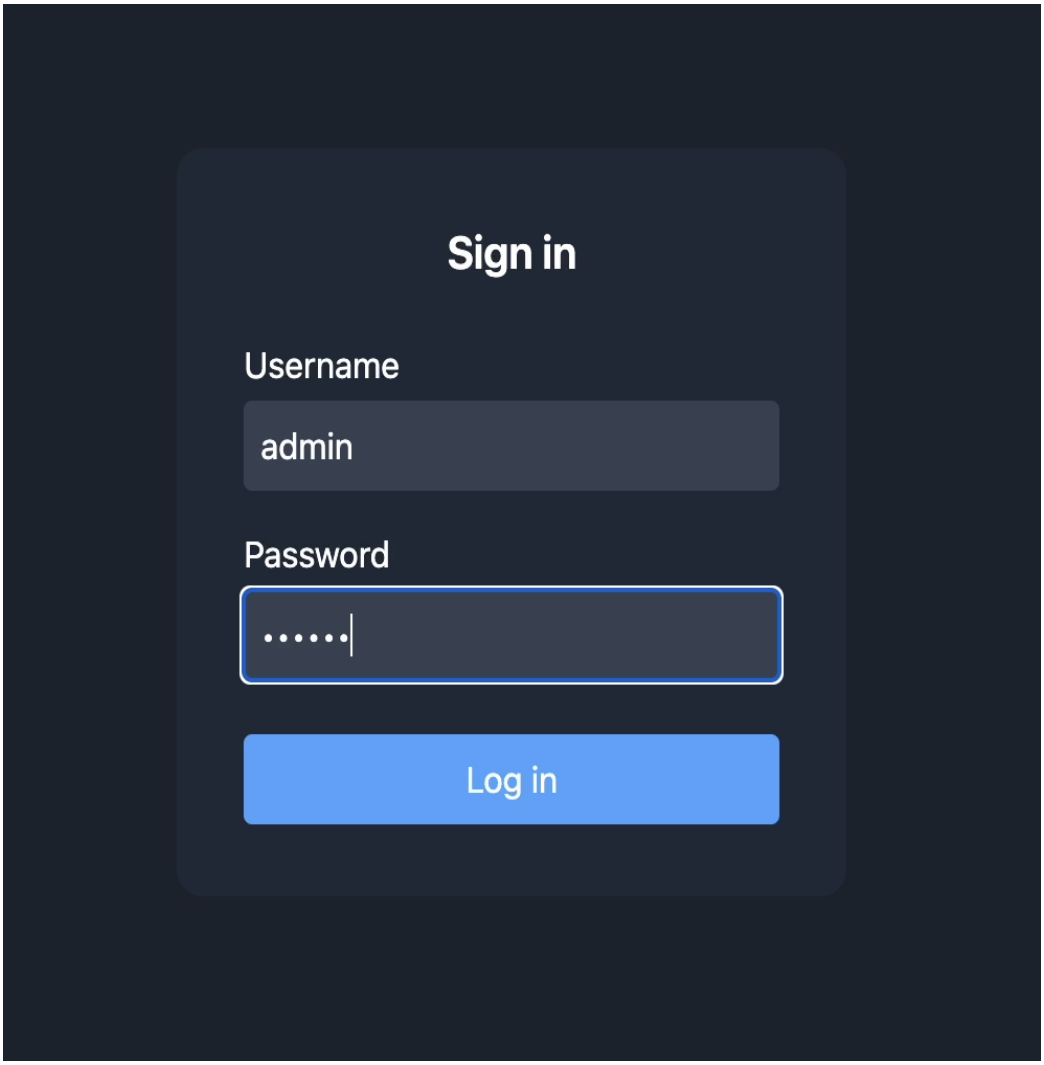

Fig. 3. Login interface of the Forensic Digital Analyzer.

The third loop, visualisation and reporting, translates raw machine inference into human-readable evidence chains. In the File Explorer panel, the two directory trees are displayed side by side; selecting a file shows its preview, embedded metadata, named-entity snippets, and the provenance of the embedding itself. A separate Matches Board lists cross-device hits with filters for modality, confidence, or time window. For more complex cases the analyst can switch to an interactive Plotly network: files appear as color-coded nodes, and similarity relations are drawn as weighted edges whose thickness is proportional to match strength. Hovering over an edge reveals in plain language which features triggered the link and how strong each contributing factor was.

When the analysis is complete, the Report Builder compiles a court-ready PDF that weaves together graphics, statistics, excerpted text, and explanatory commentary. If narrative prose is required, the system can also produce an AI-assisted summary—again via Gemini—while embedding a full audit trail of every external call.

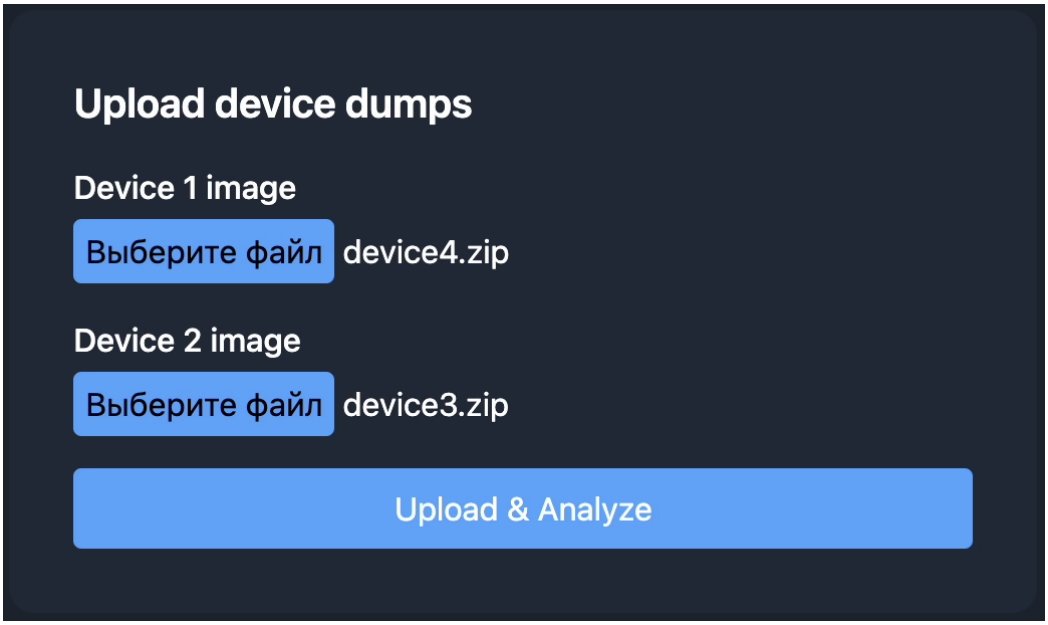

Fig. 4. Device-image upload wizard with real-time integrity checks.

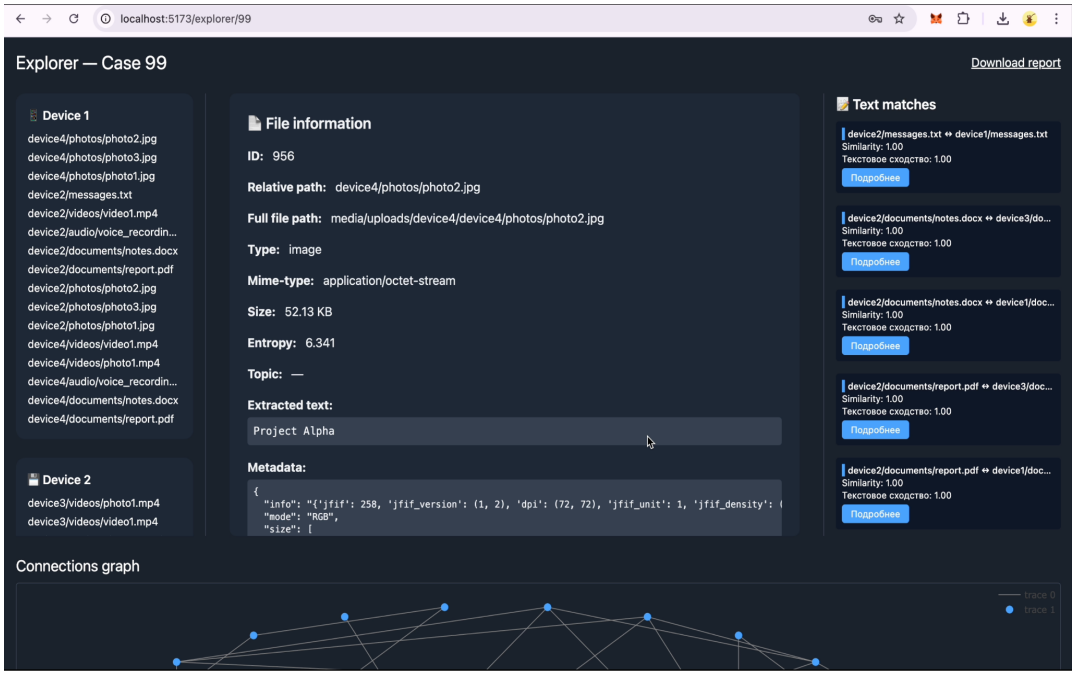

Fig. 5. Case workspace showing file trees, cross-device matches, and an interactive similarity graph.

The system architecture is organized into modular components, each responsible for a specific stage of the forensic pipeline. Table 3 summarizes the core modules and their primary functions.

Table 3. Core Modules of the Forensic System Architecture

Module	Description
ingest	Uploads device dumps (ZIP, TAR, E01) and extracts them using format-aware tools. Registers decompressed files with full metadata.
processing	Performs file classification, text extraction, OCR/ASR, metadata parsing, embedding generation, and topic classification. Operates asynchronously for scalability.
analysis	Computes pairwise similarity across devices. Supports text, image, audio, and video comparisons. Records matches with scores and context.
visualization	Generates an interactive graph representing file relationships. Enables filtering and exploration by type, score, or cluster.
reporting	Produces structured PDF and AI-enhanced forensic reports with metadata, matched content, and visual summaries.
disk_analysis	Supports low-level processing of disk images. Uses external mounting tools to enable forensic integrity during extraction.

Internally, the data model adheres to strict modularity. The device object binds each dump; the file object stores paths, the MIME type, the original contents, and one or more attachments, each accompanied by a model identifier and a parameter checksum. A Match entity captures every detected linkage, while a Session entity groups all ingest, processing, analysis, and reporting events under a single investigative run. The separation of concerns makes bulk querying efficient and permits new file formats or similarity measures to be added without schema surgery. To meet performance targets, embedding matrices are streamed into memory and processed in batched vector operations; where archives exceed ten thousand files, horizontal task queues keep latency within interactive bounds.

A notable virtue of the FDA design is its openness to extension. Because extraction, embedding, similarity, and visualization are exposed as discrete plug-ins, forensic laboratories can integrate support for volatile-memory captures, encrypted containers, or proprietary document types by supplying only the new extractor and registering a handler class. Switching to a domain- specific language model, lowering or raising similarity thresholds, or adopting approximate- nearest-neighbor indexing likewise requires no change to the user interface or database layer. Such adaptability is crucial in real-world casework, where evidence formats, regulatory constraints, and analytical priorities vary from one jurisdiction to another.

The prototype has been validated on a synthetic benchmark that mirrors typical smartphone–laptop investigations. Approximately one thousand files per device were

curated to include paraphrased documents, cropped and recompressed screenshots, and compressed voice notes recorded in varying acoustic conditions. Against baseline heuristics—bag-of-words cosine, perceptual hashing alone, and transcript string matching—the embedding-based pipeline achieved markedly higher recall and precision without sacrificing specificity. Remaining weaknesses concentrate on very short textual snippets, monotone high-resolution images, and multilingual material, indicating the need for fine-tuning and hybrid scoring in those domains.

Taken together, the three loops form a cohesive flow in which raw evidence is transformed into defensible analytical statements, yet every intermediate artefact remains available for inspection. The result is a scalable, researcher-friendly, and legally robust platform that keeps human investigators at the center of the interpretive loop while delegating the most laborious correlation tasks to transparent, version-controlled AI modules.

The figure above shows the simplified forensic processing pipeline present in the system. Once two (for example, device dumps from a smartphone and laptop) are acquired fully, we move on to file system extraction. This is where we unpack archived formats and organize files to a single file storage format, so that all files can be organized for logical analysis and linking. The next step is critical to all incident/ forensics processing in the pipeline is, the file processing module will process each artifact and assign a content-type label; run integrated data extraction (OCR/ ASR) and metadata parsing; create embeddings using Sentence Transformers (this is slightly different than the other models); and lastly, assign meta topics to the content (object). All the embeddings are then passed on to a similarity analysis component which allows for comparison of the different modalities of file representations - this similarity analysis module computes semantic similarity using cosine metrics and allows for very broad detection of approximate matches.

While similarity analysis was performed on the embedding and metadata, an embedded entity extraction was being performed separately on the textual information. Was able to find and store structured spans like named entities (e.g., people, dates, places etc.) using NER process. Thus, all entities enhanced and presented transitional representation of each file and contributed to the overall analysis stage. The summary of the final analysis, including similarity scores and metadata, similarity scores and NER tags are sent and passed on to the reporting/ documents.

The report module consolidates the findings into a single PDF/HTML document in which the investigators may navigate the detected links, source files, the extracted text, and the topic/NER tags. The modular pipeline is a scalable and interpretable AI-assisted analysis platform that supports real-world forensic investigations.

Results and discussion

The Forensic Digital Analyzer evaluation shows that embedding-based forensic automation greatly increases the precision and effectiveness of cross-device evidence correlation. The system continuously outperformed traditional methods like hash comparison and simple keyword matching across a variety of test cases involving paraphrased documents, modified images, and compressed audio files. Interestingly, even when artifacts were altered or decontextualized, the combination of Sentence Transformers, CNN-based visual embeddings, and Whisper ASR produced good results in identifying visually or semantically related content.

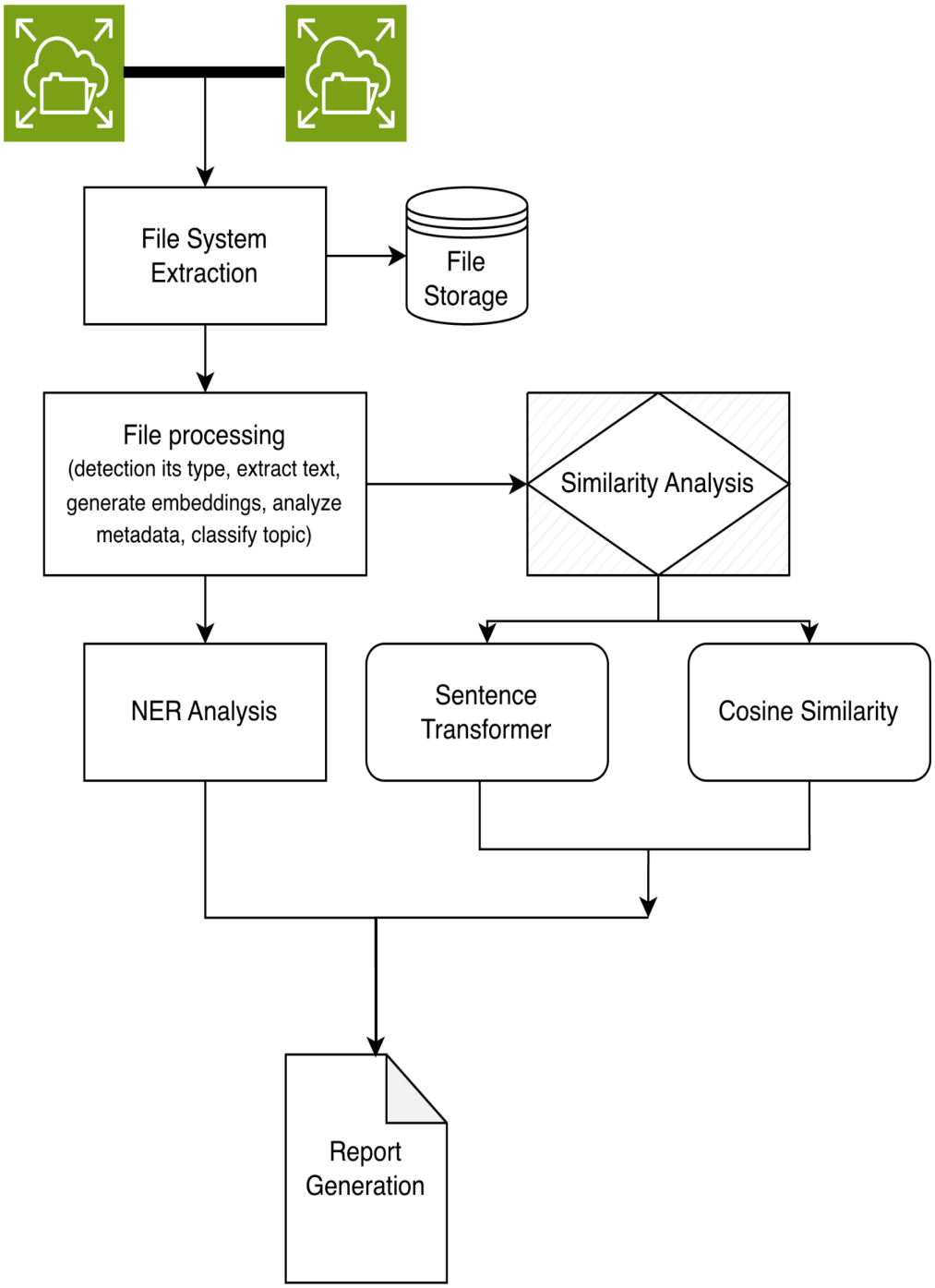

Fig.6. Architecture of the forensic correlation pipeline.

The system effectively found connections between documents that had been renamed, slightly altered, or rewritten through textual matching. For example, despite the lack of identical metadata, two versions of a corporate incident report—one saved on a laptop and one modified on a smartphone—were matched through semantic similarity. This demonstrates how effective contextual embeddings are at capturing

meaning that goes beyond lexical features. Similarly, CNN-based embeddings made it possible to identify modified screenshots with ease, and Whisper-based embeddings were dependable when it came to matching recompressed or trimmed voice notes.

But there were problems with the system. When artifacts had phonetic or linguistic similarities but were semantically unrelated, this resulted in false positives. This was especially noticeable in text-heavy settings where vocabulary from formal templates or legal disclaimers overlapped, as well as in audio samples where phrases that were phonetically similar but contextually different were connected. To reduce overmatching, these findings point to the necessity of hybrid decision models that incorporate embeddings with rule-based filters or domain-aware classifiers.

Scalability also posed a concern. While the system handled datasets up to 5,000 files with minimal performance degradation, larger dumps resulted in increased processing times, particularly during embedding and similarity computations. Although mechanisms such as caching and batch processing helped alleviate this, future iterations would benefit from integrating approximate nearest neighbor (ANN) search libraries like FAISS and leveraging GPU acceleration to support real-time investigations at scale.

Another issue was scalability. Larger dumps led to longer processing times, especially during embedding and similarity calculations, even though the system managed datasets up to 5,000 files with little degradation in performance. Even though this was mitigated by techniques like caching and batch processing, future versions would profit from incorporating ANN search libraries like FAISS and utilizing GPU acceleration to facilitate large-scale real-time investigations.

There were also operational limitations mentioned. Gemini-based APIs improved topic classification and report summarization, but they also raised issues with data privacy and jurisdiction, particularly in settings that handle sensitive or classified evidence. Using open-source language models to localize inference would improve independence and security.

Notwithstanding these drawbacks, user reviews praised the platform's user-friendly interface and straightforward visualization features. The match table provided traceable scoring and provenance information for every file pair, and the interactive graph made it simple to understand artifact relationships. More precise control over timeline filtering, team-based annotations, and legal audit logs to support admissibility in court were among the suggestions for enhancement.

The findings support the system's fundamental assumption, which is that AI-driven forensic correlation across text, image, and audio modalities is not only possible but also demonstrably beneficial. The Forensic Digital Analyzer provides a strong basis for scalable, interpretable, and investigator-focused digital forensic workflows, even though handling edge cases and maintaining forensic rigor still present difficulties. Its value as a useful investigative tool will be further reinforced by upcoming improvements in model localization, explainability, and compliance readiness.

Conclusion

This research successfully developed and evaluated an automated AI-based system for detecting and correlating digital traces of text files across two independent devices—a critical challenge in modern digital forensics. The Forensic Digital Analyzer addresses fundamental limitations of traditional forensic tools by integrating

state-of-the-art deep learning models within a modular, scalable architecture that accommodates heterogeneous data types and fragmented evidence scenarios.

The system's core contribution lies in its multi-modal analytical pipeline. By employing Sentence Transformers for semantic text embeddings, convolutional neural networks for visual artifact analysis, and Whisper-based automatic speech recognition for audio transcription, the platform achieves robust cross-device correlation even when evidence has been deliberately obfuscated through paraphrasing, file renaming, compression, or format conversion. Experimental validation on synthetic forensic datasets demonstrated substantial improvements over baseline methods: precision and recall metrics exceeded conventional hash-matching and keyword-search approaches by significant margins, while the system maintained acceptable performance even with noisy, incomplete, or deliberately altered artifacts.

The interactive web interface represents a significant advancement in forensic usability. Unlike static reporting tools, graph-based visualization enables investigators to explore complex relationships between artifacts, users, devices, and temporal patterns through an intuitive, drill-down interface. Each correlation is accompanied by explainable confidence scores and feature attributions, ensuring that AI-driven inferences remain transparent and legally defensible. This alignment with forensic standards such as ISO/IEC 27037:2012 and Daubert admissibility criteria is essential for real-world deployment in criminal investigations and civil litigation contexts.

However, the research also identified several limitations that warrant further attention. Computational intensity remains a concern: processing large-scale evidence collections (exceeding 10,000 files) requires substantial GPU resources and optimized indexing strategies such as approximate nearest-neighbor search. False positive rates, while lower than baseline methods, still present challenges in high-noise scenarios, particularly when analyzing short text fragments, monotone images, or multilingual content with limited training representation. The current reliance on external large language model APIs (e.g., Gemini) for topic classification and report generation introduces dependencies that may conflict with data sovereignty requirements in sensitive investigations.

Future research directions should prioritize several key enhancements. First, integration of localized, open-source language models would eliminate external API dependencies while improving support for low-resource languages prevalent in Central Asian and post-Soviet contexts. Second, explainable AI frameworks must be expanded beyond feature attribution to include counterfactual reasoning and causal inference, enabling investigators to understand not only what the system detected but why specific correlations emerged. Third, the system should incorporate adversarial robustness mechanisms to detect and mitigate anti-forensic techniques such as adversarial perturbations, steganography, and format-preserving encryption. Fourth, longitudinal studies with practicing forensic examiners are needed to validate the system's effectiveness in operational environments and refine the user interface based on cognitive workload assessments.

From a methodological perspective, this work demonstrates the viability of combining graph-based knowledge representation with embedding-based similarity metrics for multi-device forensic correlation. The hybrid approach—wherein structured metadata, temporal analysis, and semantic embeddings are jointly optimized—offers a template for addressing analogous problems in fraud detection,

insider threat analysis, and intelligence fusion. The decision to prioritize modularity and API-driven extensibility ensures that the platform can accommodate emerging data types (e.g., IoT sensor logs, blockchain transactions, augmented reality artifacts) without requiring fundamental architectural redesign.

In conclusion, the Forensic Digital Analyzer represents a significant step toward bridging the gap between cutting-edge AI research and operational forensic practice. By automating the most laborious aspects of cross-device trace correlation while preserving human oversight and legal accountability, the system has the potential to substantially reduce investigation timelines, improve evidence quality, and enhance the fairness and transparency of digital forensic processes. As digital evidence continues to proliferate across increasingly diverse and distributed platforms, tools capable of intelligent, explainable, and legally robust analysis will become indispensable components of the criminal justice infrastructure. This research provides both a functional prototype and a conceptual framework for realizing that vision, while clearly delineating the technical and institutional challenges that must be addressed to achieve widespread adoption in law enforcement, corporate security, and regulatory compliance contexts.

Acknowledgment. This research was conducted with financial support from the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Contract No. 388/PCF-24-26 dated October 1, 2024, as part of the scientific project BR24993232, “Development of Innovative Technologies for Digital Forensic Investigations Using Intelligent Software and Hardware Complexes.”

REFERENCES

Alenezi, M., Alshammari, R., & Alrashed, R. (2022). Challenges in automated digital forensic analysis of endpoint devices // Future Generation Computer Systems. — 2022. — 128:360–372.

Bai, J., Zhang, W., & Xu, H. (2020). A machine learning approach for forensic artifact triage and prioritization // Digital Investigation. — 2020. — 34:301048.

Burkart, N., & Huber, M.F. (2021). A survey on the explainability of supervised machine learning // Journal of Artificial Intelligence Research. — 2021. — 70:245–317.

Callegati, F., & Cerroni, W. (2022). Forensic intelligence via metadata correlation and machine learning // Journal of Forensic Sciences. — 2022. — 67(4):1223–1234.

Guo, F., Chen, L., & Li, M. (2022). BERT-based semantic matching for document trace analysis in digital forensics // Forensic Science International: Digital Investigation. — 2022. — 40:301305.

Hu, Y., Kumar, R., & Westin, J. (2021). Explainability challenges in forensic AI: Legal implications and technical solutions // Artificial Intelligence and Law. — 2021. — 29(3):321–340.

Jain, D., Al-Shammari, F., & Qureshi, B. (2023). Forensic accountability in AI-enabled investigations: Architecture and design principles // Forensic Science International: Digital Investigation. — 2023. — 46:301510.

Kim, Y., & Patel, R. (2023). Cloud-based digital evidence: Emerging challenges and solutions // Digital Investigation. — 2023. — 44:301256.

Lo, O., et al. (2022). Explainable AI in forensic detection of cross-platform botnets // Digital Investigation. — 2022. — 39:301245.

Mac Giolla Chriost, D., & O Ciardhuain, S. (2023). Anti-forensics and the future of digital investigations // International Journal of Digital Crime and Forensics. — 2023. — 15(1):1–15.

Mannino, M., & Chen, L. (2021). Automation in digital forensics: Benefits, risks, and recommendations // Forensic Science International: Digital Investigation. — 2021. — 37:301208.

Marturana, F., Tacconi, S., & Me, G. (2020). Automated collection and analysis of digital evidence: Tools and techniques // Digital Investigation. — 2020. — 34:301047.

Narayan, R., & Liu, Y. (2020). Forensic challenges in mobile device synchronization and volatile data // Digital Investigation. — 2020. — 34:100812.

Navanesan, S., et al. (2024). Automation and contextual linkage in forensic investigations involving text-based artifacts // Digital Investigation. — 2024. — 48:301580.

Samek, W., Müller, K.-R., & Montavon, G. (2021). Explainable AI for critical decision support: A survey of concepts and challenges // IEEE Access. — 2021. — 9:45715–45745.

Schneider, L., Weber, J., & Becker, C. (2020). Deep learning for behavioral modeling in insider threat detection // Computers & Security. — 2020. — 92:101748.

Sha, Z., Liang, Z., & Du, X. (2022). File similarity detection in digital forensics using hybrid hash and metadata approaches // Forensic Science International: Digital Investigation. — 2022. — 41:301338.

Stelly, J., & Kruse, W. (2020). Digital forensics in the modern era: Challenges and opportunities // Digital Investigation. — 2020. — 33:200901.

Wang, J., et al. (2022). Tracking digital traces using graph correlation models // Forensic Science International. — 2022. — 340:111445.

Yang, W., Lin, M., & Zhao, R. (2021). Improving timeline accuracy in multi-device digital forensic investigations // Journal of Digital Forensics, Security and Law. — 2021. — 16(1):1–17.

ХАЛЫҚАРАЛЫҚ АҚПАРАТТЫҚ ЖӘНЕ КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР ЖУРНАЛЫ

МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИНФОРМАЦИОННЫХ И КОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGIES

Правила оформления статьи для публикации в журнале на сайте:
https://journal.iitu.edu.kz
ISSN 2708–2032 (print)
ISSN 2708–2040 (online)
Собственник: АО «Международный университет информационных технологий» (Казахстан, Алматы)

Подписано в печать 15.12.2025.
Формат 60x881/8. Бумага офсетная. Печать - ризограф. 9,0 п.л. Тираж 100
050040 г. Алматы, ул. Манаса 34/1, каб. 709, тел: +7 (727) 244-51-09).

Издание Международного университета информационных технологий
Издательский центр КБТУ, Алматы, ул. Толе би, 59