Towards a Quality Assessment Framework for LLM-Based Decision Support in Digital Forensics

Fausto Galvan[1], Pier Luca Montessoro[2], Federico Costantini[3], Francesco Crisci[4] and Luca Baron[5]

[1]Dipartimento di Scienze Giuridiche, Università degli Studi di Udine, Italy
[2]Dipartimento Politecnico di Ingegneria e Architettura, Università degli Studi di Udine, Italy
[3]Dipartimento di Scienze Giuridiche, Università degli Studi di Udine, Italy
[4]Dipartimento di Scienze Economiche e Statistiche, Università degli Studi di Udine, Italy
[5]Dipartimento di Scienze Giuridiche, Università degli Studi di Udine, Italy

Abstract

The use of Artificial Intelligence in investigative contexts has significantly enhanced the analysis of heterogeneous digital artefacts, including textual records, documents, images, and network-related data. While current analytical tools are effective in correlating low-level indicators, they often struggle with higher-level attribution reasoning across incomplete, noisy, and multimodal evidence.

Recently, Large Language Models (LLMs) have attracted growing interest for their reasoning and abstraction capabilities; however, their application in investigative settings raises critical concerns related to reliability, explainability, and the assessment of uncertainty, particularly in light of the constraints introduced by Article 5 of the EU Artificial Intelligence Act, which prohibits certain forms of automated decision-making and profiling in law enforcement contexts.

This paper focuses on the first step of a broader research project aimed at evaluating the suitability of LLM-based agents as decision-support tools for investigative analysis.

Keywords

Large Language Models, Digital Forensics, Decision Support, AI Quality Assessment, Investigative Analysis

1. Introduction

The increasing digitisation of criminal activities has led to the pervasive generation of heterogeneous electronic artefacts, including communication records, financial transactions, digital documents, images, network traces, and online infrastructures. In investigative contexts, such artefacts are typically collected incrementally, under conditions of partial observability and uncertainty, and originate from diverse technical and social sources. As a result, investigators are required not only to correlate individual indicators, but also to reason at a higher level about patterns of behaviour, modus operandi, and potential attribution hypotheses. Several recent studies have highlighted limitations in the robustness, bias, and reliability of AI-based tools currently used in digital forensic analysis, reinforcing the need for legally compliant, human-in-the-loop approaches.

Artificial Intelligence techniques have long been employed to support investigative analysis, particularly in the correlation and aggregation of low-level indicators of compromise. However, traditional rule-based systems and classical machine learning approaches often struggle to integrate heterogeneous, multimodal, and weakly structured evidence into coherent analytical hypotheses. While such tools are effective in correlating low-level indicators and evidentiary artefacts, they remain limited in supporting higher-level attribution reasoning, which requires the integration of technical traces with contextual, behavioural, and organisational patterns, and is inherently affected by uncertainty [1]. More recently, Large Language Models (LLMs) have emerged as promising candidates for supporting complex reasoning tasks, owing to their capacity to process diverse data modalities and to operate over abstract representations of knowledge. At the same time, their adoption in investigative and law

1st Workshop on Digital Forensics (WDF 2026), co-located with the 2nd Joint National Conference on Cybersecurity (ITASEC & SERICS 2026), February 09–13, 2026, Cagliari, Italy

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-14

CEUR-WS.org/Vol-4329/paper3.pdf

enforcement settings raises well-known concerns related to reliability, explainability, bias, and hallucinations, especially in domains where erroneous inferences may have significant legal consequences. Despite their impressive reasoning capabilities, LLMs are known to exhibit well-documented limitations in terms of hallucinations, opacity, and lack of calibrated uncertainty, which pose significant challenges in high-stakes domains such as law enforcement and judicial decision-making [2, 3].

These concerns are further reinforced by the regulatory framework introduced by the European Union Artificial Intelligence Act[1]. In particular, Article 5 explicitly prohibits certain forms of automated decision-making and profiling in law enforcement contexts, thereby imposing strict constraints on the design and deployment of AI-based investigative tools. Within this regulatory landscape, the role of LLM-based systems must be carefully framed as decision-support mechanisms, aimed at assisting human investigators rather than replacing investigative judgment or producing deterministic attribution outcomes.

Against this background, this paper presents the first step of a broader research project whose long-term objective is to assess the feasibility and limitations of LLM-based agents as analytical decision-support tools in investigative contexts. Rather than proposing a complete operational system, this work deliberately focuses on the foundational problem of dataset design, which is a critical yet often underexplored prerequisite for any systematic evaluation of AI-driven analytical capabilities.

Specifically, we propose a methodology for constructing a synthetic and unbiased dataset of electronic evidence, modeled on a fictitious criminal scenario, and structured to reflect the complexity and heterogeneity of real investigative cases. Each case is represented as a collection of feature vectors corresponding to different categories of digital artefacts (e.g., communication channels, financial instruments, digital infrastructure, documentary evidence, and social engineering techniques). Crucially, these feature vectors are annotated using a hierarchical labeling scheme that captures different levels of attribution granularity.

By formalising the structure, representation, and annotation of investigative data, this paper aims to provide a solid methodological foundation for subsequent stages of the project, including the development of LLM-based analytical agents, the definition of appropriate evaluation metrics, and the assessment of their performance under regulatory and ethical constraints. The remaining steps of the research programme are not implemented in this work, but are outlined and discussed in the concluding section, where future perspectives and research directions are explicitly articulated.

Accordingly, this work does not aim to assess the quality or performance of specific LLM architectures. Instead, it focuses on defining the methodological prerequisites—at the level of data representation and annotation—required for a meaningful, reproducible, and legally compliant quality assessment, which is deferred to subsequent phases of the project. Nevertheless, to ground the proposed framework in a concrete investigative setting, a fully instantiated reference implementation of the synthetic dataset is provided in Appendix A.

The remainder of this paper is structured as follows. Section 2 reviews relevant prior work on AI-assisted digital forensics, bias and legal constraints, and the use of synthetic datasets, positioning the present contribution within the existing literature. Section 4 introduces the proposed methodological framework and dataset design principles, while Section 5 describes the investigative workflow. Finally, Section 6 concludes the paper and outlines future research directions.

2. Related Work

The application of Artificial Intelligence techniques to digital forensic investigations has received increasing attention in recent years, driven by the growing volume, heterogeneity, and complexity of digital evidence. Existing research spans a wide range of approaches, from automated artefact

[1]Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), OJ L, 2024/1689, 12.7.2024, ELI: http://data.europa.eu/eli/reg/2024/1689/ojhttp://data.europa.eu/eli/reg/2024/1689/oj.

classification to decision-support systems for investigators, as well as methodological studies on dataset availability, bias, and legal compliance. This section reviews the most relevant strands of prior work and positions the present contribution within this landscape.

2.1. Artificial Intelligence as Decision Support in Digital Forensics

A substantial body of work investigates the use of Artificial Intelligence techniques to assist investigators in analysing large-scale digital evidence, particularly through automated classification, clustering, and prioritization of artefacts [4, 5]. Commercial forensic tools increasingly integrate AI-based components to support triage and content categorization, with the goal of reducing manual workload and accelerating investigative workflows.

However, several recent studies have highlighted limitations in the accuracy, robustness, and generalizability of such systems when applied to complex or adversarial forensic scenarios [6, 7, 8]. These findings reinforce the consensus that current AI-based forensic tools should be employed as assistive technologies rather than as autonomous decision-makers, particularly in judicial contexts where accountability, explainability, and evidentiary standards are paramount.

In line with this perspective, the present work explicitly frames AI-based agents as human-in-the-loop decision-support components, excluding any form of automated attribution or profiling.

2.2. Bias, Reliability, and Legal Constraints

The use of Artificial Intelligence in law enforcement and judicial investigations raises well-documented concerns related to bias, fairness, transparency, and legal compliance [9, 10]. Empirical studies have shown that biased or unrepresentative training data can lead to discriminatory outcomes, undermining the reliability and admissibility of AI-assisted analyses.

These issues are particularly salient in the context of the European Union Artificial Intelligence Act, which imposes strict limitations on automated decision-making and profiling in law enforcement settings. Recent research emphasizes the need for legally compliant AI-by-design approaches that incorporate human oversight, explicit uncertainty modeling, and traceable decision processes [11].

Rather than addressing model-level mitigation strategies, this paper focuses on the upstream problem of dataset design, arguing that bias control, attribution granularity, and uncertainty representation must be embedded at the data and annotation level to enable compliant and trustworthy AI-assisted forensic analysis.

2.3. Synthetic Datasets for Digital Forensics Research

The availability of realistic and legally usable datasets remains a persistent challenge in digital forensics research. Privacy constraints, ongoing investigations, and evidentiary sensitivity severely limit access to real-world data, motivating the use of synthetic datasets for research, training, and evaluation purposes [12, 13].

Prior work on synthetic forensic datasets has explored the generation of isolated artefact types, such as file system images, network traces, or communication logs [14, 15, 16]. While these efforts provide valuable resources, they often lack integrated representations of multi-modal evidence, hierarchical attribution structures, or explicit modeling of investigative uncertainty.

The present contribution advances this line of research by proposing a principled framework for constructing synthetic, heterogeneous, and hierarchically annotated datasets grounded in coherent fictitious criminal scenarios. By combining methodological design principles with a fully instantiated reference implementation (Appendix A), the work aims to support future systematic evaluation of AI-based decision-support agents under realistic forensic and legal constraints.

3. Towards an AIQA (Artificial Intelligence Quality Assessment)

The profound transformations brought about by the development and diffusion of algorithmic systems and Artificial Intelligence are affecting an ever-increasing number of areas of human activity[17]. It is therefore only natural that these systems are assuming an increasingly central role in the administration of justice, raising formidable questions and necessitating a reconsideration of some of the criminal justice system's fundamental pillars — both substantive and procedural. Within this perspective, the notion of Artificial Intelligence Quality Assessment (AIQA) is addressed in a progressive manner, with the present work deliberately limited to its foundational data-related dimension. At the heart of this issue is the potential use of AI-based tools for extracting and managing evidences. The stakes are exceedingly high, as results from academic discussion. Indeed, since the adoption of AI system in the judicial sector is included in Annex III as "High Risk" activity of the AI ACT, experts insist in restricting the role of AI to limited activities [18], while others highlight the potentials of such tools [19].

In scientific literature, we find the expression "AI-driven", "AI-enhanced" or "AI-powered", meaning that a wide scope of tasks can be performed by AI, yet not completely substituting the human operator. Instead, the role assigned to the artificial agent, although sensitive, remains one of support and assistance such as to identify offensive or harmful language [20], to scan images [21] or the dark web [22], or networks [23], perform other complex analysis [24, 25], or assist in legal drafting [26].

In our previous research, we investigated the problem of "Information Quality" in the field of Digital Forensics, developing a formal approach with the aim of developing a theoretically grounded operational model [27]. If we consider the concept of "quality" in a technical sense, understood as the ability of an object to fulfil a specific purpose, it becomes possible to identify a set of requirements that, in this case, the artifact produced through forensic acquisition must satisfy. This approach allows us to assess whether, and to what extent, the expectations of the actors involved, primarily the parties to the legal proceedings, are met. In this sense, a general formula was defined that allowed one to measure the reliability of the general information contained in a heterogeneous set of devices [28].

On this regard, it is important to remember that international standards have been developed for the sharing of digital evidence, as well as legal ontologies specifically dedicated to this sector, such as CASE[2], which contribute to spreading the adoption of AI in the forensic field and increasing the level of reliability.

However, if it is widely accepted that "data" and "algorithms" are two sides of the same coin, which we can generically name "information", their "quality" must be assessed using different criteria. First of all – but we will see this later on – very often assessments can only be made indirectly, because the technologies used do not offer transparency in their processing. Secondly, it should be noted that AI is, on the one hand, an "object", similar to the digital data being observed, and on the other hand, an "agent", like the human being who observes it. The automatism of the activity that AI is called upon to perform reveals not only the essence of its instrumentality – the fact that it is used to make data processing more efficient – but also the seed of issues concerning the balance between the operation of the artificial agent and the surveillance of the human being. In other words, AI can perform its tasks with "varying degrees of autonomy" – as also found in the regulatory definition of Art. 3(1) of the "AI ACT" Regulation – and, consequently, external expectations – in terms of "quality", which is what interests us most specifically here – depend on a very heterogeneous set of factors, including, in particular, the level of "trustworthiness" established by the human being with whom it interacts or in the reference context. In this sense, it becomes crucial to establish criteria for determining also the "quality" of algorithms.

In this work, we intend to present a research proposal that addresses the "Information Quality" in the implementation of Artificial Intelligence in Digital Forensics, with the aim of establishing an AIQA (Artificial Intelligence Quality Assessment) framework. To this end, we created a pilot experiment in

[2]The CASE (Cyber-investigation Analysis Standard Expression) ontology (https://caseontology.org) allows artifacts and their properties to be thoroughly described, facilitating sharing, transmission, and automatic processing.

which a "criminal scenario" is created based on official reports[3][4], and it is used to generate a synthetic dataset, which has been fed to LLM agents in order to assess their performances according to given benchmarks.

4. Experimental methodology

Recent contributions have explored the possibility of using LLMs in the forensic field for purposes other than strictly evidentiary or investigative ones, especially for training and educational purposes [29]. In some cases, they have been used for creating quasi-real criminal scenarios to be used as a “test bed” to evaluate artificial agent performances [30]. In our research, we intend to follow similar paths. The methodological principles introduced in this section are concretely instantiated in a reference synthetic dataset, whose structure, artefacts, and validation mechanisms are detailed in Appendix A.

4.1. Scenario description

The narrative used as a starting point for this work refers to online fraud. This is a constantly growing phenomenon in which fraud is perpetrated by offering the opportunity to invest in funds or virtual currencies [31, 32]. In the most common *modus operandi*, victims are “lured”, for example, by believing the promises of lucrative investments, via physical contacts, phone calls, direct messages on social networks, sometimes with aggressive campaigns involving "deep fake" videos that can feature the appearances of celebrities or even politicians, as happened recently with the Italian Prime Minister[5].

Fraudulent activities are perpetrated using a variety of methods, such as fake phone numbers (phone spoofing), forged identity documents or professional qualification certificates (e.g. “financial advisor”), sometimes issued by non-existent organizations, as well as web browsing techniques that allow them to mask their IP connection while interacting with their victims and each other(TOR, VPN). In addition, victims are often persuaded to grant remote access to their "consultants" to their computers, allowing them to transfer funds overseas very easily. In similar cases, law enforcement officers are expected to perform several tasks, such as:

1. **Forensic acquisition of digital evidence**. Cloning the victim’s devices, acquiring chats, emails, screenshots and files), while preserving data in accordance with international forensic standards (e.g. ISO/IEC 27037)[6].
2. **Verification of the integrity of communications**. Extracting and decoding data, verifying links to the fraudulent platform, analysing metadata and recovering deleted items).
3. **Tracing cryptocurrency flows**. Following the path of cryptocurrency funds and identify any wallets linked to exchanges or mixers, using blockchain forensic tools and specialised software (e.g. Chainalysis, Elliptic).
4. **OSINT**. Linking suspicious wallets, web domains, and digital identities.

4.2. Narrative, conventions adopted and dataset availability

The fictitious criminal organization we created is composed of four individuals located in the European Union (Italy, Austria, Croatia, Slovenia) with high operational capabilities and advanced technological skills, capable of creating shell companies and counterfeited ID cards, or to use real documents stolen from previous victims, managing spoofing techniques (both for emails and for telephone numbers), as well as of setting up fraudulent websites, producing "deep fakes", and generating international financial flows through digital platforms and cryptocurrencies. They supposedly operated in a specific time

[3] https://www.ic3.gov/annualreport/reports/2023_ic3cryptocurrencyreport.pdf?utm_source=chatgpt.com
[4] https://www.europol.europa.eu/cms/sites/default/files/documents/Spotlight-Report_Online-fraud-schemes.pdf?utm_source=chatgpt.com
[5] https://www.milanofinanza.it/news/truffe-online-falsi-video-di-giorgia-meloni-promettono-un-guadagno-di-50-000-euro-con-un-investimento-202412121252369462.
[6] https://www.iso.org/standard/44381.html.

period (from 1 January 2024 to 30 June 2025), luring one hundred victims and recruiting five “straw men” – some unaware, others accomplices – spread in their respective Countries, generating a total revenue of at least five million euros. The use of synthetic datasets is a common and accepted practice in security- and privacy-sensitive domains, where access to real data is constrained, provided that structural properties and inter-feature relationships are preserved [33].

The LLM was instructed to generate various kinds of digital artefacts produced by this activity, i.e. records of telephone conversations, texts (in different languages), and metadata from electronic communications (among criminals, or towards victims and with "straw men"), as well as financial transactions.

In creating the dataset, some simplifications were introduced to optimize Prompt and output lengths in accordance with the limitations of current LLM providers: telephone numbers are not falsified; the IP addresses of the connections are not masked; it was assumed that the exchanges complied with KYC (Know Your Customer) regulation and cooperated with the police authorities.

The prompt used, the dataset containing the “synthetic data” generated by the LLM and the outcomes are published in open format on a freely accessible dataset[7].

It is worth emphasising that the operational plausibility of the proposed dataset does not rely on the realism of individual data instances, but rather on the structure of the dataset, the types of artefacts represented, and the relationships among heterogeneous features, which are designed to reflect realistic investigative constraints. Nevertheless, to enhance operational plausibility, the design of the criminal scenario and the prompting process were also supervised by a domain expert with professional experience in law enforcement. While no real investigative data were used, qualitative validation was performed through expert-in-the-loop supervision and spot checks on generated artefacts, focusing on consistency, coherence, and realism at the level of investigative workflows. The involvement of domain specialists during the data generation and validation process is widely recognised as an effective strategy to enhance the plausibility and relevance of synthetic or weakly supervised datasets, particularly in domains requiring strong domain knowledge [34]. Although the primary contribution of this work is methodological, the proposed framework has already been instantiated in a full-scale synthetic implementation that covers the entire workflow described in this paper. In particular, the criminal scenario, data representation model, artifact typology, and labeling principles have been operationalized to generate a complete and internally consistent dataset of digital evidence, including telecommunication records, electronic communications, financial transactions, and blockchain artifacts. This implementation is not intended as an empirical evaluation of LLM performance, nor as a benchmark, but rather as a proof-of-feasibility demonstrating that the proposed data model, quality constraints, and legal-by-design assumptions can be concretely realized within a realistic forensic workflow. The resulting dataset adheres to established forensic standards (e.g., ETSI CDR formats and CASE ontology mappings), incorporates explicit validation and integrity mechanisms, and is made publicly available to support reproducibility and future empirical investigations. Detailed technical documentation of this reference implementation, including dataset structure, artifact formats, and validation procedures, is provided in Appendix A.

5. Workflow overview

As shown in Figure 1, our workflow can be divided into three phases: (1) Dataset preparation; (2) LLMs testing; (3) LLMs benchmarking. In the following paragraphs, we offer a few clarifications for each of them.

5.1. Dataset preparation

In the training phase, the choice of examples that are as realistic as possible to provide to the model is a critical aspect, since the quality of the outcome will depend on how the data are selected, classified,

[7] the LLM adopted was ChatGPT5, the dataset is available at https://zenodo.org/records/16927274.

labeled, and processed. Biases can also present themselves in the forensic field and are very dangerous.

The methodology proposed here takes into account that inputs can be biased, and tries to reduce such risks by adopting separate measures for each step.

1. **Criminal scenario design**. As explained above, we created a simulation. In other words, our intention is not to incriminate real individuals using predictive policing approaches but to generate fictitious criminals enacting consolidated strategies.
2. **Prompt design**. The purpose is to generate a synthetic data set that can represent - as a whole - the narrative behind the data.
3. **LLM1 prompting**. Using an AI tool (preferably different from the one being evaluated), synthetic data are generated. For example, for an investigation into a cryptocurrency scam conducted via email and WhatsApp, sequences of messages will be produced that are consistent, appropriately distributed over time, and similar to those actually received by real victims.
4. **Criminal scenario data set**. The data set based on the criminal scenario designed. Being fictitious, it is by design anonymous.
5. **Reference data set**. This data set contains public domain, hence neutral, references[8]. This ensures also that the format of the evidences generated by the agent (e.g. bank account reports, telephone call logs, Whatspp messages) is consistent with that used in real-life law enforcement (e.g. ETSI Charging Data Records).
6. **Synthetic data set**. The synthetic data is added to the reference dataset to construct what we will call the synthetic dataset. This represents a realistic and consistent example of a set of data that could be found on one or more devices belonging to a possible perpetrator.

5.2. LLMs Testing

Detection, classification and analytical functions are applied to our datasets by the LLM to be evaluated, and the results obtained are compared with those expected (i.e. with the information available *a priori*).

1. **Prompt design.** By applying appropriate metrics, the reliability of the tool in the specific context is then determined objectively.
2. **LLM2 - LLMn prompting.** By applying the same procedure to different LLMs, it is possible to compare them, to evaluate the best performer, or to obtain confidence indices with which to measure the results, or even to deploy multiple LLMs for the same case.
3. **LLM2-LLMn extracted criminal dataset.** This sequence of operations can be repeated with different synthetic datasets to average the measurements and obtain a higher level of confidence.

5.3. LLMs benchmarking

Inevitably, discrepancies will emerge between the results obtained and the original synthetic data; therefore, the metrics described below are applied to quantitatively measure the performance of the tool in the context under consideration.

To define the metrics, it is necessary to consider the objectives of using the AI tool. Since, in the course of judicial investigations, law enforcers cannot avoid directly analysing the original data, the AI tool can be used to identify data of interest within device memories, which may contain – as typically do – amounts of information that make manual searching and selection difficult or impossible.

The metrics for evaluating the effectiveness of the AI tool in question must be defined in such a way as to measure how close the automatic processing is to the ideal result.

Some criteria for their definition are as follows:

1. **Efficiency.** For each type of message, the ratio between the number of relevant messages found and the total number of messages of the same type in the synthetic data.

[8] E.g. Computer Forensic Reference DataSet Portal, NIST, https://cfreds.nist.gov.

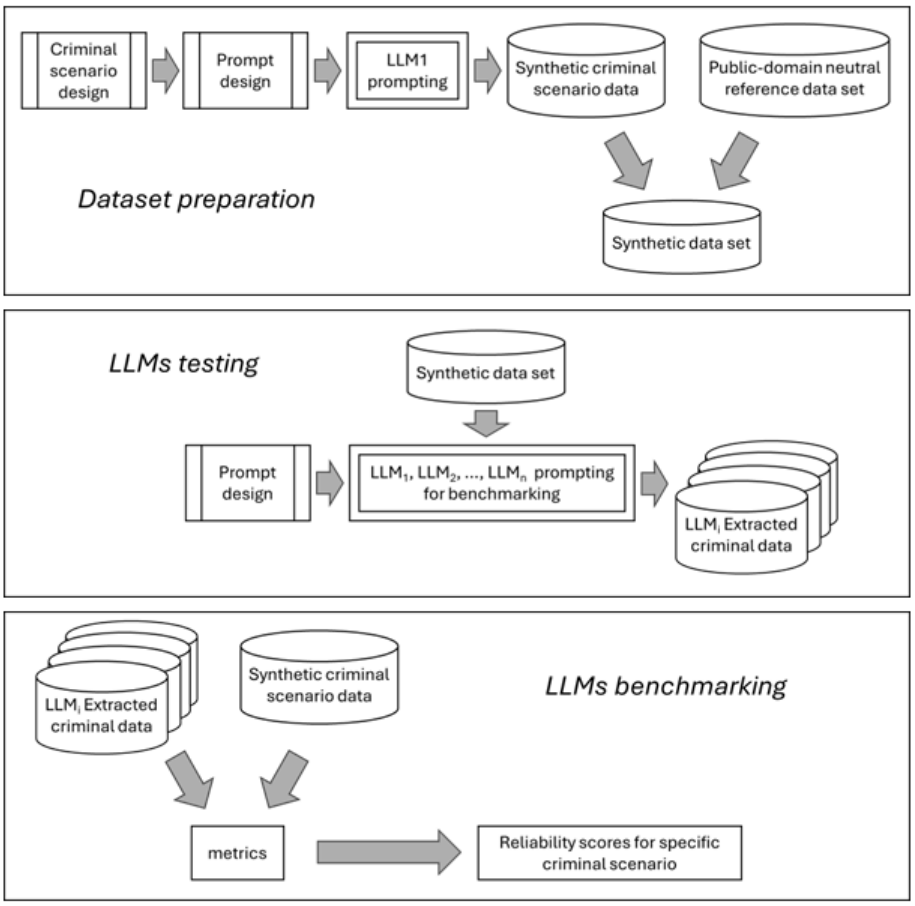


Figure 1: Workflow for a proper Dataset definition.

2. **False negatives.** For each type of message, the ratio between the number of relevant messages not found and the total number of messages of the same type in the synthetic data.
3. **False positives.** For each type of message, the ratio between the number of irrelevant messages classified as relevant and the number of messages of the same type in the summary data.
4. **Summary quality.** The LLM can provide a summary of the content of messages identified as relevant and recognise regular patterns, timing, operating modes, etc. The evaluation in this case will be represented by a score, for example obtained as the average of the scores assigned by the participants in the evaluation.

6. Concluding remarks and research prospects

This paper presented the first step of a broader research project aimed at investigating the suitability of Large Language Model–based agents as decision-support tools in investigative contexts. Rather than focusing on model architectures or performance claims, the contribution of this work lies in the methodological design of a synthetic, unbiased, and hierarchically annotated dataset of electronic evidence, explicitly tailored to the constraints and requirements of investigative analysis under regulatory and ethical considerations. The actual assessment of LLM quality and performance, grounded on this dataset design, is intentionally left to future phases of the project.

The proposed dataset structure reflects the intrinsic complexity of real-world investigations, where heterogeneous and multimodal digital artefacts are collected incrementally and under conditions of uncertainty. By representing investigative cases as collections of feature vectors corresponding to

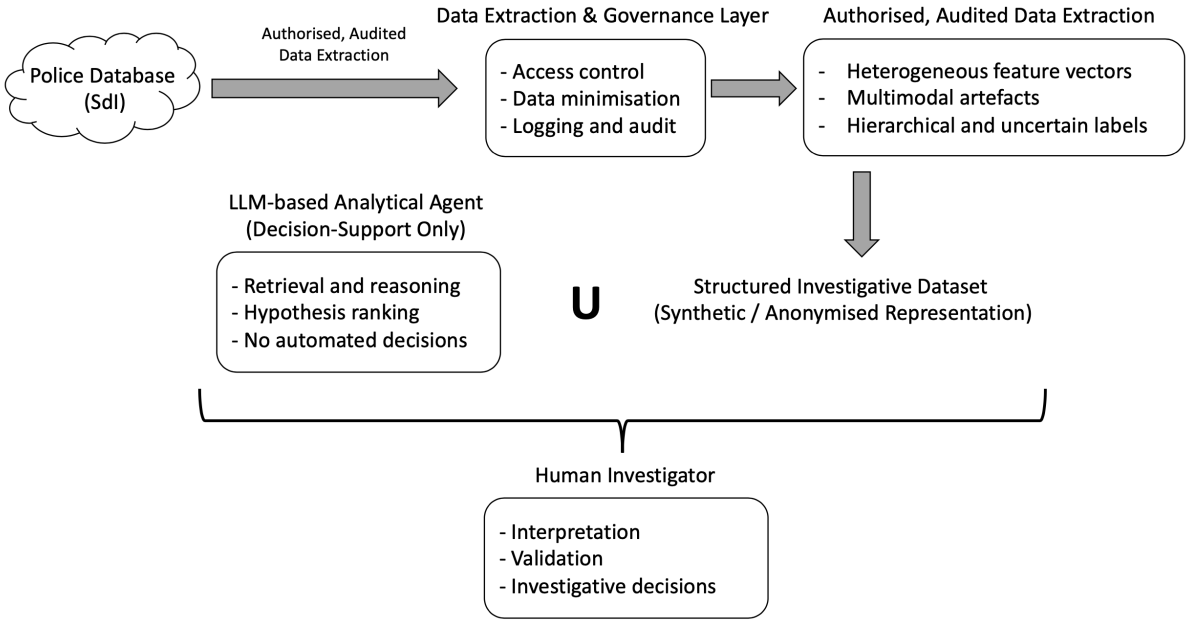


Figure 2: High-level architecture of the entire analytical framework, of which this paper constitutes the initial step.

different categories of artefacts, and by adopting a hierarchical labeling scheme that captures varying levels of attribution granularity, this work provides a principled foundation for reasoning about patterns of criminal behaviour without enforcing deterministic or legally binding conclusions. Importantly, uncertainty and incomplete knowledge are treated as first-class elements of the data representation, rather than as noise to be eliminated.

Building on this foundation, the subsequent phases of the research project will focus on the development of LLM-based analytical agents capable of processing the proposed dataset. A second step will address the design of agents that integrate retrieval, embedding, and reasoning mechanisms to generate ranked attribution hypotheses while maintaining traceability and explainability of the analytical process. Particular attention will be devoted to ensuring that such agents operate strictly within a decision-support paradigm, in line with the constraints imposed by Article 5 of the EU Artificial Intelligence Act.

A third step will concern the definition of appropriate evaluation metrics to assess the distance between the hypotheses produced by LLM-based agents and reference investigative outcomes. Given the probabilistic and hierarchical nature of the proposed labels, classical accuracy-based metrics are expected to be insufficient, as several studies have shown, evaluate systems operating over hierarchical labels and uncertain ground truth, motivating the adoption of alternative, structure-aware evaluation approaches [35]. Future work will therefore explore metrics capable of capturing partial matches, hierarchical consistency, and uncertainty-aware similarity between attribution hypotheses and ground-truth annotations.

Beyond its conceptual contribution, the work is supported by a complete reference implementation of the proposed dataset design, described in Appendix A, which serves as a proof of feasibility for future empirical evaluations.

Subsequent phases will investigate supervised and semi-supervised training strategies based on the proposed dataset and evaluation framework, with the objective of analysing how different model configurations, prompting strategies, and training regimes affect analytical performance, robustness, and bias. These experiments will be conducted exclusively on synthetic or fully anonymised data, and their results will be interpreted in close collaboration with domain experts to ensure methodological soundness and operational plausibility.

From an infrastructural perspective, a further research direction concerns the controlled interfacing of the proposed analytical agent with existing law enforcement information systems, such as the national Police Forces Database (Sistema di Indagine, SdI). In this context, the agent would not directly access or modify operational databases, but would operate on structured and authorised data extractions, transformed into feature representations consistent with the dataset model described in this work. Such an approach would allow the agent to assist investigators in organising and analysing available information, while preserving data governance requirements, access controls, and human oversight. The proposed scheme in Figure 2 summarizes the structure intended to be developed[9]. Framing AI-based systems as decision-support tools, rather than automated decision-makers, is a well-established requirement to ensure accountability, auditability, and human oversight in legally relevant contexts [36].

Finally, the long-term perspective of this research includes the controlled application of the proposed methodology to real investigative cases, subject to strict legal, ethical, and organisational safeguards. Any potential integration with operational systems would be limited to offline analytical support and hypothesis generation, explicitly excluding automated decision-making or attribution. Human investigators would retain full responsibility for interpretation, validation, and use of the analytical outputs within the investigative and judicial process.

The proposed methodology does not aim to address explainability challenges per se, but rather to provide a structured tool for assessing the trustworthiness of artificial agents employed in investigative and judicial contexts. Within this perspective, AIQA—possibly in synergy with established Information Quality Assessment (IQA) approaches and sector-specific ontologies such as CASE—may support a wide range of applications. In particular, it could enable not only the optimisation and training of individual artificial agents, but also the systematic comparison of the quality of different agents involved in the same proceeding, such as those adopted by the prosecution, the defence, or the civil party, in a manner that is transparent, communicable, and accessible to non-technical stakeholders.

A. Reference Implementation of the Synthetic Dataset

This appendix documents the reference implementation of the synthetic dataset used to instantiate the methodological framework proposed in this paper. The purpose of this implementation is not to evaluate the performance of Large Language Models, nor to provide a benchmark, but to demonstrate the feasibility, internal coherence, and forensic plausibility of the proposed data representation, annotation strategy, and workflow under realistic investigative constraints.

The implementation provides a concrete realization of the concepts discussed in Sections 4 and 5, covering the full lifecycle from criminal scenario definition to the generation, validation, and structuring of heterogeneous digital evidence.

A.1. Scope and Design Objectives

The reference implementation was designed to satisfy the following objectives:

- instantiate a complete fictitious criminal scenario consistent with real-world online fraud investigations;
- generate heterogeneous digital artefacts reflecting those typically encountered in digital forensic casework;
- preserve structural realism and inter-artifact relationships while avoiding the use of any real investigative data;
- ensure compatibility with established forensic standards and data models;

[9] Authors want to point out that the LLM-based agent will have to be explicitly separated from operational law enforcement systems, such as the Police Forces Database (SdI), and operates exclusively on structured and authorised data extractions. The agent provides decision-support through hypothesis generation and ranking, while all investigative decisions remain under human control.

- enable reproducibility, traceability, and independent verification of dataset integrity.

All data included in the dataset are synthetic by construction, and no real persons, identifiers, or transactions are represented.

A.2. Criminal Scenario Overview

The dataset is based on a fictitious cross-border online fraud scenario involving a structured criminal organization operating within the European Union over the period from 1 January 2024 to 30 June 2025. The scenario includes:

- four primary offenders operating from different EU Member States;
- five intermediaries ("straw men"), partially complicit or unaware;
- one hundred victims targeted through social engineering techniques;
- multiple communication channels, financial instruments, and digital infrastructures;
- a total illicit revenue exceeding five million euros.

The scenario was designed to reflect consolidated fraud patterns documented in official reports and threat assessments, which were used exclusively as narrative and structural references to ensure plausibility, without reusing or adapting real case data.

A.3. Generated Artefact Categories

The dataset includes multiple categories of digital artefacts, generated in accordance with forensic conventions and investigative practice.

A.3.1. Telecommunication Records

Telephony data are provided in the form of call detail records compliant with ETSI CDR specifications (version 9). Each record includes standard forensic fields such as timestamps with time zones, call direction, IMSI, IMEI, cell identifiers, call duration, and termination causes.

The generated records cover:

- regular communications among criminal actors;
- contacts between offenders and victims;
- periodic communications with intermediaries.

Temporal distributions and communication frequencies were constrained to ensure consistency with the defined scenario.

A.3.2. Electronic Communications

Electronic communications include email messages and short-form messaging (SMS and social messaging platforms). Email artefacts include full message headers with routing information, authentication results, encoding details, and geographic annotations. Message contents are multilingual, reflecting the geographical distribution of victims.

Short-form messages are represented using a unified law-enforcement-style log format, including sender and recipient identifiers, device metadata, IP addresses, geolocation fields, and message content.

A.3.3. Financial and Cryptocurrency Transactions

Financial flows are represented through both traditional payment summaries and cryptocurrency transactions. Cryptocurrency data include syntactically valid Bitcoin and Ethereum transaction records, structured to reflect realistic transaction chains with multiple hops, intermediate wallets, and final aggregation points.

Transaction metadata include block heights, timestamps, fees, confirmation counts, and indicators of mixing services. All identifiers are synthetic and non-spendable, and no real blockchain transactions are referenced.

A.3.4. Supporting Artefacts and Metadata

Additional artefacts are included to support forensic analysis and cross-correlation, such as:

- device and identifier inventories mapping SIMs, devices, and network attributes;
- domain and website listings associated with fraudulent activity;
- payment aggregation tables summarizing victim-level transactions across channels.

These artefacts enable realistic investigative workflows involving correlation across heterogeneous evidence sources.

A.4. Validation, Integrity, and Chain of Custody

To support traceability and internal validation, the dataset includes explicit integrity and consistency mechanisms:

- cryptographic hashes (SHA-256) for all generated files;
- manifest files documenting file inventories, timestamps, and generation metadata;
- automated validation summaries verifying compliance with scenario constraints, including temporal coverage, communication frequencies, transaction counts, and revenue thresholds.

These measures mirror standard chain-of-custody practices and facilitate independent verification of dataset completeness and internal coherence.

A.5. CASE Ontology Mapping

In addition to raw artefacts, the dataset is provided in a structured representation compliant with the Cyber-investigation Analysis Standard Expression (CASE) ontology. The CASE dataset includes:

- entities representing persons, organizations, and digital objects;
- explicit links between artefacts and actors;
- cryptographic wallets and blockchain transactions modeled as first-class objects;
- hash-based references to high-volume textual artefacts to preserve scalability and integrity.

This dual representation—raw artefacts and ontology-based structures—supports both traditional forensic analysis and AI-assisted analytical pipelines.

A.6. Availability and Intended Use

The complete dataset, including raw artefacts, CASE representations, manifests, and validation files, is made publicly available through an online repository. The dataset is intended for:

- methodological validation of forensic data models;
- development and testing of AI-based decision-support tools;
- training and educational activities in digital forensics;
- future empirical evaluation of analytical agents under controlled and legally compliant conditions.

The dataset is not intended for operational deployment and does not support automated attribution or decision-making.

A.7. Relation to the Main Contribution

This reference implementation serves as a proof-of-feasibility for the framework proposed in this paper. It demonstrates that the methodological principles—synthetic scenario design, heterogeneous artefact modeling, hierarchical representation, and legal-by-design constraints—can be concretely realized within a realistic forensic workflow.

By providing both the conceptual framework and a fully instantiated example, the paper aims to bridge the gap between abstract methodological proposals and empirically grounded future research, while remaining compliant with legal, ethical, and regulatory requirements.

Acknowledgements

This contribution is the result of joint research conducted by the authors as part of the Project "Best Practices in Cross Border Digital Forensics", Department of Law, University of Udine. The individual sections are attributed as follows: Fausto Galvan: 4,4.1, 6, A; Pier Luca Montessoro: 5; Federico Costantini: 1,3, 4.2; Francesco Crisci: 2,3; Luca Baron: 2;

Online Resources

The dataset generated is available at: https:zenodo.org/records/16927274.

Declaration on Generative AI

The authors employed Generative AI tools (ChatGPT 5.1 Plus, Overleaf Writefull and Deepl) in supporting text polishing and translation. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] T. Rid, B. Buchanan, Attributing cyber attacks, Journal of strategic studies 38 (2015) 4–37.

[2] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 2021, pp. 610–623.

[3] B. Mittelstadt, C. Russell, S. Wachter, Explaining explanations in ai, in: Proceedings of the conference on fairness, accountability, and transparency, 2019, pp. 279–288.

[4] S. L. Sanna, L. Regano, D. Maiorca, G. Giacinto, Exploring the robustness of ai-driven tools in digital forensics: A preliminary study, 2024. URL: https://arxiv.org/abs/2412.01363. arXiv:2412.01363.

[5] S. L. Sanna, A. Panzino, S. M. La Cava, S. Concas, L. Regano, D. Maiorca, G. L. Marcialis, G. Giacinto, Pixel perfect or perfectly fake? exploring the robustness of digital forensic tools against facial deepfakes and morphed images, in: 2025 33rd European Signal Processing Conference (EUSIPCO), IEEE, 2025, pp. 1357–1361.

[6] G. Dorai, P. Rad, F. Breitinger, R. Bardhan, V. Ramasamy, Mapping the research landscape-an exploratory analysis of ai applications in digital forensics, in: International Conference on Availability, Reliability and Security, Springer, 2025, pp. 113–130.

[7] V. Abrol, S. Kumari, A. Ghosh, V. Diwakar, V. Jha, D. Suhag, A systematic review of evidence extraction and analysis in digital disk forensics, in: International Conference on Data Analytics & Management, Springer, 2025, pp. 510–521.

[8] S. K. Punjabi, S. Chaure, Forensic intelligence- combining artificial intelligence with digital forensics, in: 2022 2nd International Conference on Intelligent Technologies (CONIT), 2022, pp. 1–5. doi:10.1109/CONIT55038.2022.9848406.

[9] S. L. Sanna, L. Regano, D. Maiorca, G. Giacinto, Improving cybercrime detection and digital forensics investigations with artificial intelligence, arXiv preprint arXiv:2510.14638 (2025).
[10] A. A. Solanke, M. A. Biasiotti, Digital forensics ai: evaluating, standardizing and optimizing digital evidence mining techniques, KI-Künstliche Intelligenz 36 (2022) 143–161.
[11] J. Faehndrich, W. Honekamp, R. Povalej, H. Rittelmeier, S. Berner, D. Labudde, Digital forensics and strong ai: A structured literature review, Forensic Science International: Digital Investigation 46 (2023) 301617.
[12] M. Thomson, S. McKeown, R. Macfarlane, P. Leimich, Exploring dataset diversity for genai image inpainting localisation in digital forensics, in: Proceedings of the Digital Forensics Doctoral Symposium, 2025, pp. 1–7.
[13] C. Grajeda, F. Breitinger, I. Baggili, Availability of datasets for digital forensics–and what is missing, Digital Investigation 22 (2017) S94–S105.
[14] H. Studiawan, F. Breitinger, M. Scanlon, Towards a standardized methodology and dataset for evaluating llm-based digital forensic timeline analysis, arXiv preprint arXiv:2505.03100 (2025).
[15] J.-N. Hilgert, M. Lambertz, M. Rybalka, R. Schell, Syntactical carving of pngs and automated generation of reproducible datasets, Digital Investigation 29 (2019) S22–S30.
[16] L. L. Voigt, F. Freiling, C. J. Hargreaves, Re-imagen: Generating coherent background activity in synthetic scenario-based forensic datasets using large language models, Forensic Science International: Digital Investigation 50 (2024) 301805.
[17] N. Saba, W. K. Balwan, Artificial intelligence in forensic science: Can it be a revolution or else?, Scholars Academic Journal of Biosciences 13 (2025) 335–339. doi:10.36347/sajb.2025.v13i03.005.
[18] R. Brighi, Informatica forense, algoritmi e garanzie processuali, Ars interpretandi 1 (2021) 153–164. URL: https://www.rivisteweb.it/doi/10.7382/100798. doi:10.7382/100798.
[19] K.-J. Kim, C.-H. Lee, S.-E. Bae, J.-H. Choi, W. Kang, Digital forensics in law enforcement: A case study of llm-driven evidence analysis, Forensic Science International: Digital Investigation 54 (2025) 301939. doi:10.1016/j.fsidi.2025.301939.
[20] D. Vyas, M. Shah, A. Kothari, J. Golakia, V. Parikh, Enhancing digital forensics: Machine learning techniques for social media investigation 258 (2025-01-01) 2290–2301. URL: https://www.sciencedirect.com/science/article/pii/S1877050925015868. doi:10.1016/j.procs.2025.04.483.
[21] J. Hendrix, D. Morozoff, Media forensics in the age of disinformation, in: H. T. Sencar, L. Verdoliva, N. Memon (Eds.), Multimedia Forensics, Springer Singapore, 2022, pp. 7–40. URL: https://link.springer.com/10.1007/978-981-16-7621-5_2. doi:10.1007/978-981-16-7621-5_2, series Title: Advances in Computer Vision and Pattern Recognition.
[22] H. Vaghela, N. Varshney, R. Jain, Leveraging AI and ML to innovate forensic frameworks for the identification of illicit operations and extraction of digital artifacts within deep web and dark web environments 2 (2025-05-16) 20–35. URL: https://www.digitalsecurityforensics.org/digisecforensics/article/view/43. doi:10.29121/digisecforensics.v2.i1.2025.43, number: 1.
[23] S. Rizvi, M. Scanlon, J. Mcgibney, J. Sheppard, Application of artificial intelligence to network forensics: Survey, challenges and future directions 10 (2022) 110362–110384. URL: https://ieeexplore.ieee.org/abstract/document/9919162. doi:10.1109/ACCESS.2022.3214506.
[24] S. Costantini, G. De Gasperis, R. Olivieri, Digital forensics and investigations meet artificial intelligence 86 (2019-07) 193–229. URL: http://link.springer.com/10.1007/s10472-019-09632-y. doi:10.1007/s10472-019-09632-y.
[25] R. S. A. Faqir, Digital criminal investigations in the era of artificial intelligence: A comprehensive overview 17 (2023) 77–94. URL: https://cybercrimejournal.com/menuscript/index.php/cybercrimejournal/article/view/189.
[26] R. Liepiņa, F. Lagioia, M. Lippi, P. Pałka, H.-W. Micklitz, G. Sartor, Automating legal tasks: LLMs, legal documents, and the AI act, in: M. Zou, C. Poncibò, M. Ebers, R. Calo (Eds.), The Cambridge Handbook of Generative AI and the Law, 1 ed., Cambridge University Press, 2025-08-07, pp. 407–424.

URL: https://www.cambridge.org/core/product/identifier/9781009492553%23c23/type/book_part. doi:10.1017/9781009492553.029.
[27] L. Floridi, P. Illari, The Philosophy of Information Quality, Synthese library, Springer, 2014.
[28] F. Costantini, F. Galvan, M. A. De Stefani, S. Battiato, Assessing "information quality" in iot forensics: Theoretical framework and model implementation, Journal of Applied Logics – IfCoLog Journal of Logics and their Applications 8 (2021) 2373–2406.
[29] M. Scanlon, F. Breitinger, C. Hargreaves, J.-N. Hilgert, J. Sheppard, ChatGPT for digital forensic investigation: The good, the bad, and the unknown 46 (2023) 301609. URL: https://www.sciencedirect.com/science/article/pii/S266628172300121X. doi:10.1016/j.fsidi.2023.301609.
[30] B. Sharma, J. Ghawaly, K. McCleary, A. M. Webb, I. Baggili, ForensicLLM: A local large language model for digital forensics 52 (2025-03-01) 301872. URL: https://www.sciencedirect.com/science/article/pii/S2666281725000113. doi:10.1016/j.fsidi.2025.301872.
[31] European Union Agency for Law Enforcement Cooperation., Europol (2023), Online fraud schemes: a web of deceit, Technical Report, Europol, 2023.
[32] European Union Agency for Law Enforcement Cooperation., IOCTA, internet organised crime threat assessment 2024., Publications Office, 2024. URL: https://data.europa.eu/doi/10.2813/442713.
[33] N. Patki, R. Wedge, K. Veeramachaneni, The synthetic data vault, in: 2016 IEEE international conference on data science and advanced analytics (DSAA), IEEE, 2016, pp. 399–410.
[34] B. Settles, R. J. Brachman, Active learning. morgan claypool, Synthesis Lectures on AI and ML (2012).
[35] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information processing & management 45 (2009) 427–437.
[36] J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, H. Yu, Accountable algorithms, University of Pennsylvania Law Review 165 (2017) 633.