Understanding and Empowering Intelligence Analysts: User-Centered Design for Deepfake Detection Tools

Y. KELLY WU, Rochester Institute of Technology, USA
SANIAT JAVID SOHRAWARDI, Rochester Institute of Technology, NY
CANDICE R. GERSTNER, National Security Agency, USA
MATTHEW WRIGHT, Rochester Institute of Technology, USA

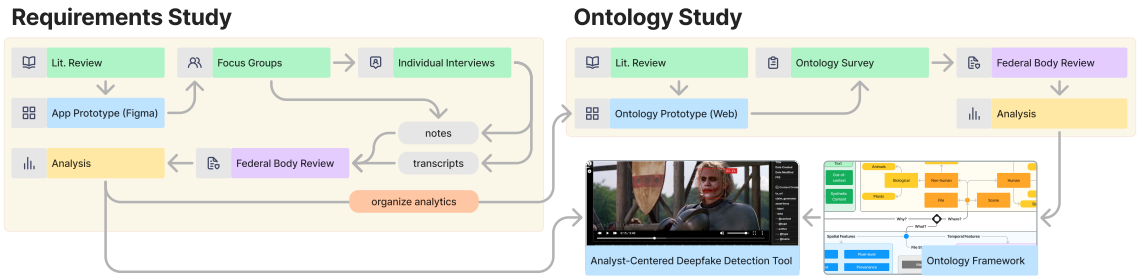


Fig. 1. Study flow showing the two phases of user studies: the *Requirements Study*, followed by the *Ontology Study* and their sub-steps.

Intelligence analysts must quickly and accurately examine and report on information in multiple modalities, including video, audio, and images. With the rise of Generative AI and deepfakes, analysts face unprecedented challenges, and require effective, reliable, and explainable media detection and analysis tools. This work explores analysts' requirements for deepfake detection tools and explainability features. From a study of 30 practitioners from the United States Intelligence Community, we identified the need for a comprehensive and explainable solution that incorporates a wide variety of methods and supports the production of intelligence reports. In response, we propose a design for an analyst-centered tool, and introduce a digital media forensics ontology to support analysts' interactions with the tool and understanding of its results. We conducted a study grounded in work-related tasks as an initial evaluation of this approach, and report on its potential to assist analysts and areas for improvement in future work.

CCS Concepts: • **Human-centered computing** → **User studies**; • **Security and privacy** → *Human and societal aspects of security and privacy*; • **Applied computing** → *Investigation techniques.*

Additional Key Words and Phrases: Deepfake, Intelligence Community, Qualitative Studies, Ontology, Explainability

ACM Reference Format:
Y. Kelly Wu, Saniat Javid Sohrawardi, Candice R. Gerstner, and Matthew Wright. 2025. Understanding and Empowering Intelligence Analysts: User-Centered Design for Deepfake Detection Tools. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 36 pages. https://doi.org/10.1145/3706598.3713711

Authors' Contact Information: Y. Kelly Wu, kellywu@mail.rit.edu, Rochester Institute of Technology, Rochester, NY, USA; Saniat Javid Sohrawardi, saniat.s@mail.rit.edu, Rochester Institute of Technology, Rochester, NY; Candice R. Gerstner, crgerst@uwe.nsa.gov, National Security Agency, Fort George G. Meade, MD, USA; Matthew Wright, matthew.wright@rit.edu, Rochester Institute of Technology, Rochester, NY, USA.

Manuscript submitted to ACM

1 Introduction

Intelligence analysts working for governments around the world are responsible for collecting, analyzing, and reporting on information that assists decision-makers in national security and foreign policy. A critical part of their workflow involves *content triage*, where they must process vast amounts of raw data from multiple sources to distill valuable insights [6]. And discerning whether the data is authentic or not is critical to developing accurate analyses. The emergence of deepfakes—media synthesized or manipulated using deep-learning technology—has significantly complicated analysts' ability to discern real content from fake [71].

Deepfake detection tools offer a potential solution to this challenge by running media through algorithms for analysis—referred to as *analytics* in the Intelligence Community—and reporting the findings. Traditional analytics often rely on detecting artifacts or inconsistencies that can be manually identified or detected through well-defined statistical models [84]. As deepfakes and other media manipulation techniques have become more sophisticated, however, these traditional methods are proving less reliable. Recent deepfake detection tools leverage deep learning [7, 18, 19, 76, 83], which shows promise but introduces new challenges. Deep learning models often function as "black boxes," providing limited transparency into how decisions are made [49, 84]. This lack of explainability is a significant drawback for intelligence analysts who must provide evidence-based conclusions and justify their findings in high-stakes scenarios [27]. Furthermore, these models frequently struggle with real-world data that may diverge from their training datasets [74], raising concerns about their reliability in operational environments.

Despite the availability of deepfake detection technologies, there is limited understanding of how analysts perceive these tools and what features they require for effective integration into their workflows. This study aims to address this gap by exploring the specific needs of intelligence analysts when working with deepfake detection tools.

The first phase of our research is guided by several key questions:

- **RQ1**: *How do analysts perceive the growing prevalence of deepfakes and the need for advanced detection tools?* This question is crucial in understanding analysts' concerns and how tools can best support their work in this domain.
- **RQ2**: *What specific features analysts would find most valuable in a tool designed for their use?* Understanding these preferences will help inform the design of more effective and user-friendly detection systems.
- **RQ3**: *What kinds of explainability features do analysts prefer when using deepfake detection tools?* Explainability is a crucial factor in digital media forensics, because it ensures that tools not only detect manipulated media but also provide interpretable insights that can be effectively communicated in reports.

To answer these questions, as shown in Figure 1, we conducted a *Requirements Study* through semi-structured interviews with 30 practitioners from the Intelligence Community (see details in Section 3). As part of the study, we presented participants with a prototype deepfake detection tool designed for journalists [76], which we updated based on our preliminary knowledge about analysts' needs. This prototype served as a baseline for gathering detailed feedback on design choices.

Key takeaways from this qualitative study include:

- The growing sophistication and prevalence of deepfakes is an immediate concern for intelligence analysts, underscoring the need for advanced tools that can effectively detect these manipulations and explain their results (**RQ1**). Analysts emphasized the importance of explainability, as it fosters trust and enables them to understand and communicate the outputs of deepfake detection tools (**RQ2**).

- The fragmented nature of existing tools poses significant challenges for analysts, who often struggle to manage and consolidate results from different analyses. Participants expressed a strong preference for a comprehensive solution that integrates various analytics relevant to evolving manipulation techniques into a unified user interface, reducing the complexity of managing multiple tools and formats (**RQ2**).
- Analysts expressed a desire for more intuitive and transparent explanations to accompany deepfake detection results, as current explanation methods (e.g., heatmaps) were deemed insufficient and difficult to interpret without additional guidance (**RQ3**).

Based on these findings, we entered the second phase of our study, in which we sought to take steps to address the needs of analysts. We recognized that creating a comprehensive tool for deepfake detection would present significant challenges. As the number of available analytics increases, analysts face the burden of selecting the most appropriate analytic for their task, which can be overwhelming, especially given that new analytics may require extra effort to understand and interpret the results. This challenge forms the basis of our fourth research question:

- **RQ4**: *How can a large list of analytics be presented to users in an interpretable way?*

To address this issue, we conducted a literature review and proposed a *Digital Media Forensics Ontology* (hereafter referred to as the *ontology*) as a potential solution. The ontology is designed around an innovative *why, where, what* structure that systematically organizes analytics based on their capabilities to help users navigate through the many options. This structure, detailed in Section 5, makes tools more intuitive to use while also providing greater transparency about the purpose and function of each analytic.

Having proposed this ontology, we wanted to understand whether it could help analysts. In particular:

- **RQ5**: *How well does the ontology help analysts to find desired analytics more easily and improve their ability to interpret results for use in report writing?*

To answer this research question, we designed and carried out an *Ontology Study*, which involved a survey grounded in work-based tasks with 11 analysts. The ontology was adapted into a *sentence-forming interface* to facilitate interaction. We found that ontology-based search has the potential to enhance analysts' confidence in their analytic choices and improve both the quality and clarity of their report writing. Analysts reported that the structured approach provided by the ontology made it easier to navigate through complex sets of analytics and select the appropriate tools for their tasks. Furthermore, they noted that the ontology helped them better understand the underlying purpose and methods of each analytic, which in turn facilitated clearer communication of results in their reports.

Finally, based on the findings from the two studies, we propose a modular design for an analyst-centered deepfake detection tool that incorporates the ontology framework. Such a tool, detailed in Section 7, could streamline analytic selection and ensure the results could be interpreted within the specific context of each analysis. Future work will focus on testing the usability and effectiveness of this ontology-based strategy. These efforts lay the foundation for tools that empower analysts to navigate the complexities of deepfake detection with confidence and precision.

2 Related Work

2.1 Deepfake Detection

The urgency of the deepfake threat has spurred the development of various deepfake detection methods. Some of these methods target specific artifacts introduced during the generation process, such as mismatched eye color or reflections [28, 32], inconsistencies in the mouth region across frames [17], or inconsistencies in other parts of the

face and body [4, 5]. Other methods leverage deep neural networks to automatically learn discriminative features that can distinguish between real and fake media. Techniques like convolutional neural networks (CNNs) [2, 89], recurrent neural networks (RNNs) [11, 12, 33], and more recently transformers [86, 93] have shown promising results in detecting various types of deepfakes.

Unfortunately, deepfake detection remains far from solved. The diversity of deepfake generation techniques and the difficulty of obtaining representative training data hinder the generalization capabilities of many detectors, making them less effective against new or unseen types of deepfakes [42, 89]. Additionally, deepfake detectors are vulnerable to adversarial attacks, where subtle, imperceptible perturbations can be added to the media to evade detection [36, 72]. These issues mean that detection tools can make mistakes, such that analysts cannot rely uncritically on their outputs.

Due to the risk of errors and the black-box nature of deep learning models, which makes their decisions difficult to interpret [84], it is critical to build explainability into deepfake detection tools. Explainability helps users to understand the reasoning behind a model's decision, which then builds trust and ensures responsible use of the tool. Heatmap visualizations have been employed to offer post-hoc explanations, their effectiveness in deepfake detection remains limited [49, 66, 87, 88]. Highlighted regions often fail to reveal discernible artifacts or may not align with human interpretable reasoning, hindering their practical utility for analysts and decision-makers. Detection methods focused on specific artifacts – analyzing features like eye color or inconsistencies in the mouth – offer some inherent explainability, but their efficacy may diminish as deepfake generation techniques become more sophisticated [85].

Addressing these challenges while also catering to the specific needs of various user groups has become a priority. Recent research has focused on studying detection tools with a focus on journalists [40, 76, 77] and forensic analysts [84]. Furthermore, the Defense Advanced Research Projects Agency (DARPA) has spearheaded initiatives like MediFor [15] and SemaFor [16] to advance media integrity verification and semantic analysis. MediFor focused on developing technologies to automatically assess the integrity of photos and videos, while SemaFor aims to analyze the semantic content of media, identifying inconsistencies or manipulations that alter the intended meaning. These efforts, though extensive, have not included a significant focus on usability of the resulting software tools. Our work seeks to bridge this gap by first understanding the needs of intelligence analysts in the context of deepfake detection, and then taking initial steps to develop a broader comprehensive, usable, and explainable system.

2.2 Human-Machine Teaming in Intelligence Analysis

Intelligence analysis is a complex process involving various disciplines such as human intelligence (HUMINT), imagery intelligence (IMINT), and open-source intelligence (OSINT) [61]. Analysts follow a structured five-step cycle—direction, collection, processing, analysis, and dissemination [24]—to transform raw data into actionable intelligence. Automated tools have the potential to assist analysts in this process [82], but their adoption has been met with resistance due to the black-box nature of AI systems [30]. Explainability significantly influences analysts' trust in AI systems, as they require transparency into how the system generates its outputs [20].

Human-machine teaming (HMT) addresses these concerns by fostering collaboration between analysts and AI systems, where machine-driven data processing complements human interpretive skills. Effective HMT requires transparency to build trust and improve decision-making, which are critical in high-stakes environments like intelligence analysis. Lyons et al. [47] emphasize that transparency through explainability enables analysts to understand AI decisions, allowing them to intervene when necessary, which enhances both trust and overall team performance. Traditionally, analytics allowed for a balanced interaction where machines supported analysts without overshadowing their expertise. Current deepfake detection tools, however, often rely heavily on automated verdicts, risking the exclusion of analysts'

input. Our research aims to restore this balance by exploring how analysts envision collaborating with these tools in a more interactive manner, thereby enhancing both trust and decision-making.

2.3 Digital Media Forensics Frameworks

In recent years, the landscape of digital media manipulation has significantly transformed, evolving from manual forgeries to sophisticated deep-learning-based techniques, such as deepfakes. This shift has required corresponding advancements in the field of digital media forensics, including the development of new analytics tailored to address these emerging challenges. To better understand and navigate the increasingly complex media forensics landscape, researchers have created taxonomies to systematically categorize manipulation types and detection analytics.

Existing taxonomies within the field of digital media forensics show considerable variation in their structure and design. Many of the taxonomies are separated by data modalities like images [8, 50, 53], videos [22, 38, 73], and audio [91], or by underlying generation methods like manual forgeries [25, 73] or AI-based techniques [48, 51, 79]. A more recent work by Lin et al. [45] offers a comprehensive view of the current landscape of generation and detection of synthetic media in the age of generative AI models. While these taxonomies provide valuable insights, it is difficult to find coherence across taxonomies. For instance, many taxonomies isolate elements like data modalities, manipulation types, and detection techniques, making it difficult to represent analytics that have broader applicability across multiple domains. This fragmented approach limits the utility of these taxonomies for serving as comprehensive guides for forensics tasks.

While much of the existing literature focuses on the construction of taxonomies, we propose shifting towards an ontology framework to address these challenges, as ontologies can capture the dynamic relationships between various domain concepts, and their definitions and properties [44]. In the context of cybercrime investigations, ontologies have proven beneficial by enabling automated reasoning, facilitating anomaly detection and supporting the chain of custody [75]. They also help clarify technical digital forensic terminologies encountered during investigations, potentially reducing analysis time [39]. Our proposed framework emphasizes the interconnections among various components of digital media forensics, including manipulation types, entities available for analysis, and potential artifacts for scrutiny, thereby fostering greater clarity, usability, and explainability within the realm of digital media forensics.

3 Requirements Study Design

In this section, we describe the semi-structured interviews we conducted with intelligence analysts to understand their needs for a deepfake detection tool. Our findings are presented in Section 4.

3.1 Sample and Consent

Given the unique nature of their work, recruiting U.S.-based intelligence workers as participants required a specialized approach. We recruited 30 practitioners from our professional network within the Intelligence Community, including analysts, researchers, and technical support specialists. Their daily responsibilities ranged from verifying digital media provenance, identities of people, and entities depicted in the media, and writing reports to inform additional downstream exploration work. To preserve participant confidentiality and adhere to institutional policies, no authors were directly involved in the recruitment phase; instead, collaborators within the community handled initial participant outreach.

We conducted two types of semi-structured interview sessions to accommodate participants' availability and operational constraints. Five participants were interviewed individually using Microsoft Teams, and the remaining 25

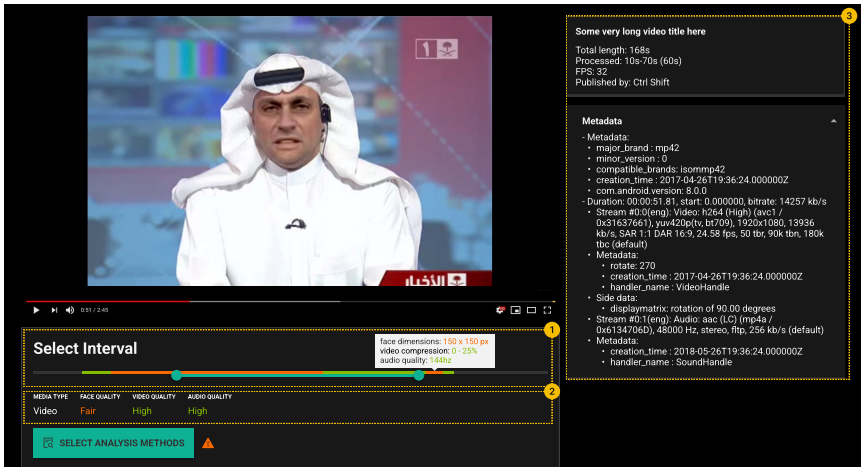


Fig. 2. Prototype of the screen prior to processing the video with (1) color-coded content quality warnings over a timeline and the ability to select snippets to process, (2) summarized content quality warnings, and (3) initial metadata information. Video frame from FaceForensics++ [68] with the design based on DeFake [76].

participants were interviewed in person either individually or in small groups ranging from two to seven participants. The in-person sessions varied in size due to our use of a drop-in system, where participants could attend interview sessions during pre-designated time slots based on their availability. This made it easier to recruit participants in a situation where we had limited access. In each interview session, one researcher served as the primary interviewer, while a second researcher took notes. All participants were notified about the nature and the scope of the study and received an informed consent document prior to the interview. Consent was obtained through a checkmark on the consent statement indicating their willingness to participate in the study; for online interviews, participants were also asked if they would consent to be recorded and had the right to turn on or off their cameras.

The study protocol received approval from the Institutional Review Boards (IRB) of the participating institutions. Additionally, the protocol was endorsed by a federal Human Research Protection Office (HRPO), ensuring that all ethical considerations for research involving Intelligence Community professionals were fully addressed.

3.2 Ethical Statement

The primary objective of this study was to inform the development of an analyst-centered tool for deepfake detection that meets their needs for reliability and explainability. The tool itself is intended to benefit intelligence analysts in their important work.

Given their background working with classified materials, the participants were well-equipped to evaluate the potential risks of their involvement in the study and to respond to the questions with appropriate consideration. Participants were asked to answer all the questions during the interviews at an unclassified level to ensure that no sensitive or classified information was disclosed. In addition, all the data collected during the study including interview notes and recording transcripts underwent review by a federal body before being made available to the research team for analysis. This process added another layer of protection so that participants' identities and any potentially sensitive details in their answers were thoroughly removed.

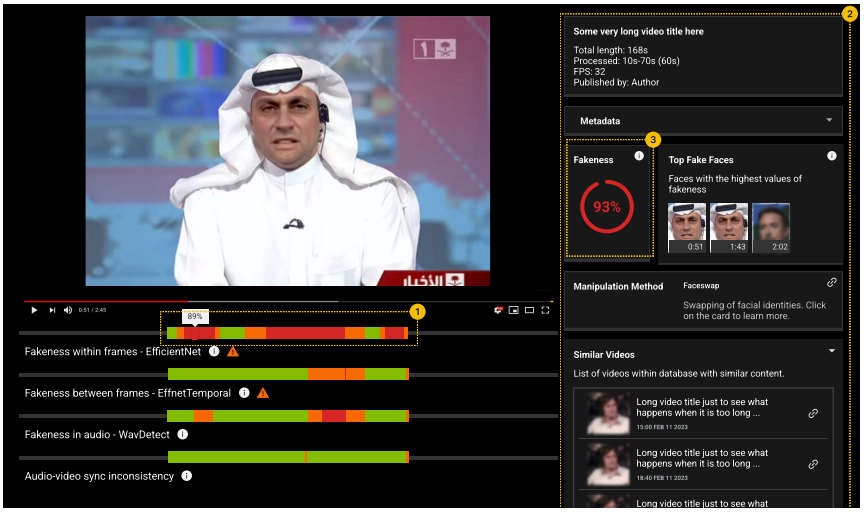


Fig. 3. Prototype of the results screen. (1) Timeline color-codes for each analytic show red for fake, orange for suspicious, and green for real; warnings (⚠) note a higher risk of potential inaccuracy. (2) A supplementary results sidebar showing metadata, top fake faces, approximate manipulation method, and similar videos. (3) Fakeness score shows the highest detection score among all the models used. Video frame from FaceForensics++ [68] with the design based on DeFake [76].

3.3 Study Development

To gain a comprehensive understanding of intelligence analysts' workflow with media verification, perceptions of deepfakes, their requirements for detection tools, and their preferences for explainability, we conducted semi-structured interviews aided by prototype screens. These interviews were structured into sections, each building upon the previous one, to facilitate a thorough exploration of the topics.

Current Workflow. The interviews began with a discussion about the participants' current ways of working. Participants were asked to describe the types of media they interacted with, their existing workflow concerns, and the primary audiences for their analyses. This section provided a broad overview of the analysts' daily tasks and challenges, helping us identify areas where our tool could be tailored to address specific needs and enhance their workflow.

Deepfake Detection. Next, we delved into the topic of deepfakes, asking about the participants' familiarity with this technology, their exposure to it, and their concerns regarding its impact. We also explored their expectations for a detection tool, including what features and capabilities they would find most useful. This part of the interview helped participants articulate their own thoughts on deepfakes and detection capabilities, setting the stage for discussing the ideal characteristics of a detection tool during prototype evaluation.

Prototype Evaluation. Following the deepfake discussion, we presented participants with prototype screens to gather feedback on a potential detection interface. The screens, such as shown in Figure 2 (preprocessing) and Figure 3 (results), were based on a tool initially designed for journalists and made available to us by its developers [76]. While intelligence analysts and journalists share a similar high-level goal in the verification of media, intelligence analysts have different training, outcomes, and methodologies. To better tailor the prototypes to the needs of analysts, we incorporated insights from a detailed document written by a current analyst, outlining what a typical day looks like in their role [27]. Analysts

often have more technical expertise in their content modality and are expected to perform an in-depth analysis of the content to communicate findings clearly to the other stakeholders [57, 59]. Thus, it would be fair to provide them with a greater saturation of information in a single interface. This additional context helped us refine the prototypes before starting the evaluation.

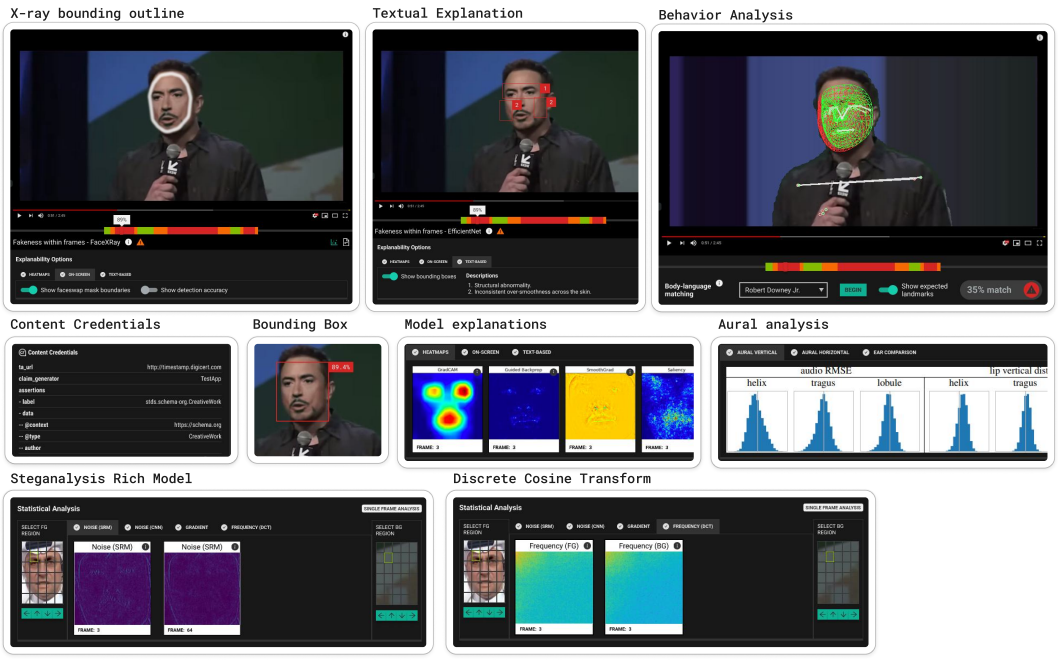


Fig. 4. Explanation screens used in the Requirements Study. While the figure shows only the most relevant segments of each screen, participants were able to interact with a full prototype. Images were sourced from the DeepFaceLab [65] tutorial video, FaceForensics++ [68], and [3] for Aural analysis.

Explainability. In the final section, we introduced various explainability formats that could potentially be integrated into the tool using prototype screens. Shown in Figure 4, they included bounding boxes, Face X-ray outlines [43], model explanation heatmaps [70], biometric-based features [3, 90], frequency maps from Discrete Cosine Transformation (DCT) [67], noise maps from Steganalysis Rich Model (SRM) filters [34], Content Credential (CR) extraction from Content Provenance and Authenticity (C2PA) [10], and textual explanations accompanying the visuals. Participants were shown each format and asked to explain their understanding of the information presented, as well as to evaluate its effectiveness. This section helped us understand what types of explainability features were most useful for analysts and how they could aid in their work.

Importance Ratings. The interviews concluded with participants rating the importance of key factors in a deepfake detection tool: speed of analysis, low false positives, low false negatives, and the explainability of results.

3.4 Data Collection and Analysis

Our data sources included detailed notes taken during the in-person interviews by the note-taker and, for online sessions, full transcripts of the recorded conversations.

Since our primary goal is to understand analysts' requirements for the tool and explainability formats, we performed a thematic analysis of the collected data [9]. Two of the authors independently reviewed and became familiarized with the data and, using an open coding approach, developed their own initial codebook of the participants' answers and sentiments, including preliminary themes. Then they convened to discuss their codebooks and use axial coding to merge conceptually similar themes, resolve any discrepancies in interpretation, and refine and name the themes to reach the final codebook [69].

AI Assistance in Thematic Analysis. Following the growing trend of using generative AI in qualitative research [26, 92], we incorporated Gemini Advanced [80] and ChatGPT 4o [63] into our open coding process. These AI tools, used in private mode to ensure data confidentiality, were provided with organized notes, transcripts, and our initial set of codes. Functioning as a third coder, the AI was prompted to generate additional codes and propose alternative themes, thereby complementing our manual analysis. The AI's role was carefully defined and contextualized within the research framework, ensuring its contributions aligned with our specific research objectives. This strategic integration of AI assistance enhanced the comprehensiveness of our analysis and may have helped to reduce researcher bias.

4 Requirements Study Results

The Requirements Study revealed valuable insights into the workflows, challenges, and needs of intelligence analysts when using deepfake detection tools. As shown in the study overview in Figure 5, several recurring themes emerged, reflecting diverse perspectives on the current state of tools and the opportunities for improvement. Analysts consistently highlighted frustrations with fragmented toolsets, desires to accommodate diverse media formats, and the need for more integrated and explainable detection solutions. These findings indicate significant room for improvement in existing workflows and underscore the importance of developing tools that not only detect deepfakes but also provide interpretable results to support decision-making.

4.1 Analyst Workflow and Challenges

Analysts often engage in content triage, where they sift through vast amounts of data in various formats, ultimately producing comprehensive reports about their analysis as the final product. They are accustomed to using various tools in their work, but with that comes problems that affect their work efficiency. Understanding these workflows and the challenges analysts face provides important context for our study, helping us identify specific areas where deepfake detection tools can be better integrated into their processes to enhance productivity and decision-making.

Diverse Media Handling. Many of our participants reported working with more than one type of media. Some participants (P5, P6, P7, P9, P14, P23, P24) regularly dealt with images, videos, audio, and text in their daily tasks. The nature of their work varied widely, including activities such as information exploration, information retrieval, language translation and analysis, and the writing and reviewing of reports. This diversity in media handling indicates a need for a versatile and adaptable tool that can support various data formats and analytical tasks.

Frustrations with Tool Fragmentation. A recurring theme in the interviews was the challenge of maintaining work efficiency due to the need to navigate through multiple tools. Analysts expressed significant frustration with the

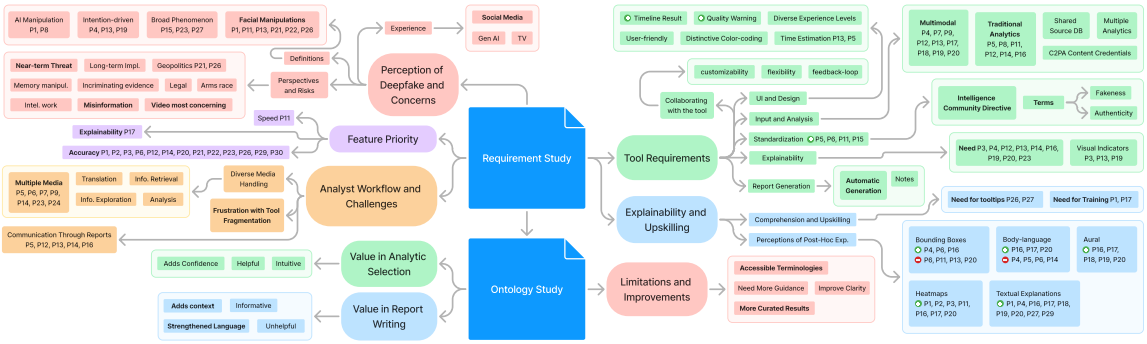


Fig. 5. The results of thematic analysis from both the Requirements Study and the Ontology Study. The first children nodes of the studies represent the overarching themes, with further nodes showing the sub-themes, and relevant codes with participants listed in significant nodes. Highlights show the most prominent codes (**bold**) and sentiments (⊕, ⊖) under that theme.

fragmented nature of existing tools, which complicates the organization and summarization of their findings. For example, P19 mentioned, *"A lot of times I'd have to use different tool[s] for a bunch of pieces of my workflow … having those not talk to each other was a big pain … hard to organize my thoughts into one."* This sentiment was echoed by many participants, highlighting the need for a seamless integration of deepfake detection capabilities into their existing toolsets. Additionally, P16 pointed out the difficulty of standardization, noting that different contents and varying agency protocols make it challenging to establish a consistent workflow. Addressing these frustrations through the development of a comprehensive tool could significantly enhance analysts' workflow efficiency and overall productivity.

Communication Through Reports. The primary output of analysts' work is often in the form of reports, which summarize their findings along with explanations and evidence. The audiences of these reports can range from fellow analysts to high-level decision-makers involved in policy-making and strategy. This can be a tedious task that requires clear reasoning for their analysis choices and explanations for the results they obtain. Since report audiences can have varying levels of technical expertise, the reports need to effectively communicate the findings in a clear and useful manner and bridge the gap between technical details and strategic insights, ensuring that the information is accessible and actionable (P5, P12, P13, P14, P16).

4.2 Perception of Deepfake and Concerns

Our participants demonstrated a good understanding of the creation and use of deepfakes, despite limitations on discussing specific experiences in their classified work. Instead, they shared their exposure to deepfakes through news, social media, and publicly available tools in their daily lives. A unanimous concern emerged during the interviews regarding the potential impact of deepfakes, both personally and on a societal level. These discussions about their perceptions and concerns directly contribute to answering **RQ1**, as they provide insight into analysts' views on the growing need for reliable deepfake detection tools.

Participants' Definition of Deepfake. In general, participants understood the synthetic and imitative nature of deepfakes, and recognized the role of AI in their creation. Some common descriptors used by the participants during interviews included *"synthetic media," "manipulated media,"* and *"creating narratives."* We noted that some participants (P1, P11,

P13, P21, P22, P26) focused primarily on facial manipulation, while others viewed deepfakes as a broader phenomenon capturing multimodal media such as images, videos, audio, and text (P15, P23, P27).

Several participants (P4, P13, P19) emphasized the intention behind deepfakes as a crucial aspect of their definition. P4 said, *"While it is important to know that an image is manipulated, if the sky color is changed from cloudy to blue that is not as important as if more people are added to an image. The purpose of the manipulation is important."* This highlights that the perceived threat level of a deepfake can be influenced by the intent and context of the manipulation. Two participants (P1, P8) observed that "deepfake" has become an overloaded term, akin to "Kleenex" in its widespread and generic usage. This suggests a potential need for more precise terminology in media forensics.

Exposure to Deepfakes. Outside of their professional roles, most participants reported limited direct interaction with deepfakes. Their exposure primarily came through news, social media, or television. They referenced widely recognized deepfakes, such as the image of the pope wearing a white puffer coat [52] and Metaphysic's performance on America's Got Talent [78]. Two participants (P26, P27) had hands-on experience with publicly available tools like Midjourney,[1] DALL-E,[2] and Stable Diffusion.[3]

Analysts' Perspectives on the Risks of Deepfakes. All participants expressed concerns about deepfakes, with many viewing them as a near-term threat while also emphasizing the importance of considering long-term implications. Analysts expressed a range of concerns, spanning both personal and societal dimensions. On a personal level, participants worried about deepfakes manipulating individual memories (P24) or creating false incriminating evidence (P26). From a societal perspective, the manipulation of public sentiment and the rapid spread of disinformation were significant concerns. P29 elaborated: *"A lie makes it three times around the world and becomes a conspiracy theory before the truth has time to put its shoes on."* This worry also extended to potential geopolitical manipulation (P21, P26).

Professionally, analysts highlighted the challenges deepfakes pose to their work. They noted that identifying subjects in a video could become more difficult. In legal contexts, where introducing fake evidence could have severe consequences, P27 commented, *"It is a big deal because the people on the stand [on trial] are facing significant sentences,"* a sentiment also shared by P17. The rapid advancement of deepfake technology raised additional concerns about keeping pace with detection methods. P3 described it as *"an arms race in getting detection methods to be able to detect manipulated media before it's too late to counter the repercussions."*

Regarding the most concerning modality, the majority of the participants picked video due to its accessibility to the general public, rapid dissemination (P17, P19), higher believability compared to audio or text (P20), and heavy consumption by society (P29).

4.3 Requirements for Analyst-centered Deepfake Detection Tool

To elicit users' requirements for a deepfake detection tool, we first asked participants to describe features they wanted and then showed them prototype screens such as shown in Figures 2 and 3 to get their feedback. This approach helped us understand their ideal tool without limitations and evaluate how existing designs could be improved. Participants envisioned a user-friendly, flexible, and comprehensive tool that delivers standardized, explainable results suitable for report integration. These insights directly address **RQ2**, revealing the desired features analysts expect from deepfake detection tools.

[1]https://www.midjourney.com
[2]https://openai.com/index/dall-e-2
[3]https://stability.ai

User Interface and Design. When presented with the prototype interfaces, most participants found the overall flow to be intuitive and user-friendly. They appreciated the logical sequence of actions and the ease of navigation. As P21 remarked, the tool *"needs to be user-friendly for all groups of people depending on user experience with interfaces."*

The color-coding system as shown in Figure 2 and 3 received mixed feedback. While some participants found it helpful to understand input quality and the level of fakeness, others suggested that there should be a clearer separation of logic between the main screen and the preprocessing screen (Figure 2) to avoid confusion. Within the preprocessing screen, participants expressed a need for additional features to enhance usability. P5 and P13 emphasized the importance of having a processing time estimation for analytics, which would allow them to select appropriate analytics based on the time available. P13 also liked the content quality warning, which helped users make better-informed decisions when selecting analytics and reduce the risk of misuse.

The presentation of the analytics over the video timeline was well-received, particularly by P1 and P11, who appreciated its utility in navigating through the media. This feature allows users to easily pinpoint and review specific segments, which is crucial for detailed analysis and reporting.

Input and Analysis. The requirements for input and analysis capabilities of a deepfake detection tool reflect the complex and diverse nature of intelligence analysts' work. A key feature emphasized by many participants is the ability to handle multimodal input without the need for tool refactoring. This requirement aligns closely with the analysts' daily workflow, which often involves processing and analyzing data across various modalities. Given the substantial volume of data analysts routinely handle, batch processing was described as a key requirement (P4, P7, P9, P12, P13, P17, P18, P19, P20). More specifically, they seek content triaging features that can identify higher-priority items, thereby improving overall efficiency.

While advanced deepfake detection methods are important, participants (P5, P8, P11, P12, P14, P16) also stressed the need to incorporate traditional analysis tools. Features such as Hex editors, metadata viewers, and camera identification tools for authentication were frequently mentioned as valuable aids to their analytical process. A specific highlight in this regard was the Content Credentials [10] component under Metadata in the prototype screens. This new initiative, aimed at offering clear and traceable information about content shared online, resonated positively with participants (P21-P25, P29, P30). Rather than relying on a single detection method, analysts expressed a preference for multiple analytics to be available within the tool.

A suggestion made by many participants was the implementation of a shared database for source comparison. This feature would potentially allow analysts to compare media under investigation with known sources. On the other hand, a challenge associated with this feature is data compartmentalization due to varying agency affiliations and access levels, which was also mentioned during the interviews.

Consistency and Standardization. A recurring theme emphasized by several participants (P5, P6, P11, P15) was the need for standardization in deepfake detection tools and processes. Central to this discussion was Intelligence Community Directive (ICD) 203 [60], which establishes analytic standards for the Intelligence Community. As P15 succinctly stated, *"[We] need standardization to communicate the findings well."* Participants stressed the importance of a common framework in the Intelligence Community, facilitating clearer communication among different stakeholders and ensuring consistency in reporting.

In discussing standardization, we specifically asked participants about their opinion regarding using the term "Fakeness" to communicate the detection results. Most participants considered "Fakeness" a fair term to use, provided that the methodology for determining it was clearly explained to users, ensuring they knew how to interpret the results.

However, one concern raised was that *"manipulation does not mean it is fake necessarily."* Some participants suggested alternative terms such as "authenticity" (P21, P22, P24). During the interviews, the SemaFor project [16], was mentioned by a few participants; its use of a 5-point score scale was also brought up as a potential way for communicating detection results.

Explainability. Our participants expressed a great need for explainability in the deepfake detection tool, as highlighted by several participants (P3, P4, P12, P13, P14, P16, P19, P20, P23). Participants emphasized that simply providing a detection result is insufficient. As P4 noted, *"an answer is not enough"*; the results must be accompanied by detailed explanations of why a certain conclusion was reached.

Visual indications of which parts of the media could be manipulated were identified as particularly useful by P3, P13, and P19. Visual aids can help communicate findings to stakeholders who may not have a technical background or an understanding of the underlying analytics. This interaction between analysts and AI systems, as emphasized in the HMT literature, can enhance both trust and decision-making by allowing analysts to engage more dynamically with the tool's outputs.

Collaborating with the Tool. Several participants expressed a desire for deeper engagement with the tool, seeking more interactivity and control over its outputs. Participant P17 emphasized the need for a feedback loop, stating they wanted *"some way to give feedback and then some way to either remove [analyzed media] from my workflow if it's nothing that I care about"*. This highlights the importance of dynamic interaction, where users can refine the tool's outputs based on their specific needs. Participants also valued flexibility in how they interact with the tool, such as being able to turn specific features on or off depending on the task at hand. P25 further reflected this desire for fine-tuning performance by asking, *"Would you be able to add your own data to use to compare?"* indicating a need for tools that allow analysts to incorporate their own references to help verify the analysis. Additionally, participants expressed interest in exploring media at various levels—whether byte-level, metadata, or pixel-level—and applying different analytics to specific segments of a media asset. This ability to interactively select and customize analytics would enable analysts to focus on areas of interest and tailor their analysis according to the context. Such a collaborative approach aligns with HMT principles by ensuring that AI systems not only automate tasks but also adapt to user preferences, enhancing both control and efficiency in the analysis process [47, 64].

Report Generation. Report writing is a significant component of analysts' work, as highlighted earlier in section 4.1. Recognizing this, many participants (P4, P5, P6, P12, P13, P14, P16, P19, P22, P26, P27) expressed a strong desire for integrated report generation capabilities within deepfake detection tools. The ability to automatically include detection results and accompanying explanations in exportable reports would help reduce the time and effort required to translate technical findings into comprehensive reports.

Additionally, a few participants emphasized the importance of note-taking capabilities within the tool. Jotting down observations and insights while analyzing media would enhance the analysts' ability to document their thought processes and maintain a detailed record of their work. One participant, P25, also suggested an aspirational feature of recording the entire analysis process taken within the tool and automatically converting this into a report, so it could be repeatable.

4.4 Explainability Formats and Upskilling

As highlighted in earlier sections, participants often emphasized the need for explainability even before being prompted to interact with prototype screens showcasing possible post-hoc explanation formats. This underscores the importance of explainability in deepfake detection tools, a key focus of **RQ3**, which explores the types of explainability features analysts prefer. While participants had varied opinions on the prototypes shown in Figure 4, there was a strong preference for detailed textual explanations that could be easily integrated into reports. Additionally, participants stressed the value of tooltips and the need for training sessions to help users better understand the tool's functionality and maximize its effectiveness.

Perceptions of Post-hoc Explanations. The study revealed diverse opinions on presented post-hoc explanation methods (Figure 4), all of which incorporated visual elements either directly on the media or in dedicated explanation panels.

Bounding boxes with fakeness scores received mixed reactions. Some participants (P4, P6, P16) found them helpful, particularly for scenes with multiple faces, while others (P6, P11, P13, P20) expressed confusion about the meaning of the associated percentages. Face X-ray, which highlighted blending artifacts on fake faces, was generally perceived as a variant of bounding boxes and received limited enthusiasm from participants.

For biometric and behavior-based explanations, including aural and body-language indicators, opinions were similarly divided. Some participants (P4, P5, P6, P14) found these explanations unclear, particularly regarding the interpretation of body-language cues and associated percentages. Others (P16, P17, P20), however, found them useful and intuitive. Notably, the ear comparison feature in the aural screen was well-received by several participants (P16, P17, P18, P19, P20), indicating potential value in focused, comparative visual explanations. The more complex correlations between ear regions with respect to mouth opening and volume received little comment.

Heatmaps, frequency maps, and noise maps were often perceived as overly technical (P1, P2, P3, P11, P16, P17, P20). Participants P11 and P14 emphasized the need for baselines or comparative data to make these visualizations more meaningful.

Textual explanations were widely favored (P1, P4, P16, P17, P18, P19, P20, P27, P29) because they were seen as *"easy to understand," "simple,"* and *"less technical."* Many participants also noted their usefulness for direct inclusion in reports.

A common challenge noted was the difficulty in interpreting the explanation results themselves. Participants often felt a need for what we might call "explanations of explanations."

Tool Comprehension and Upskilling. Beyond specific post-hoc explanations, a comprehensive understanding of tool functionality was crucial for effective use. During the interviews, participants were walked through the prototype screens again after their initial assessment. While participants generally found the overall flow of the prototype intuitive, they indicated that tooltips providing explanations to specific features, such as analytics used for analysis and fakeness scores, would be beneficial, especially for new users (P26, P27). Training and onboarding sessions were also identified as valuable for the upskilling process by P1 and P17.

4.5 Prioritization of Key Features

To better understand the priorities when building a deepfake detection system, participants were asked to rank four key characteristics: analysis speed, low false positive rate, low false negative rate, and explainability of results. This question provided crucial insights into how desired features should be balanced to meet their operational needs, contributing to addressing **RQ2**. This insight is essential for designing a tool that aligns with the practical demands of their workflows.

Accuracy was identified as a primary concern for many participants, with a particular emphasis on minimizing false positives (P1, P2, P3, P6, P12, P14, P20, P21, P22, P23, P26, P29, P30). As P29 noted, *"We don't want to call something a fake that actually wasn't one because we would lose more credibility that way than vice versa."* Explainability was frequently ranked highly, often second or third in importance. P17 prioritized it above all else because *"the more people can explain and understand how the tool works [...] the more it is trusted, and the more that people will use it."* This sentiment of having explainability was shared across the interviews as reflected in the previous sections. Analysis speed, while important, often ranked lower than accuracy and explainability. However, we did have a participant rank it the first because an *"analyst's time is most important."*

Several participants highlighted the interconnected nature of these features and that the ranking could be case-dependent. For offline media analysis, accuracy was generally more important than speed, while real-time detection scenarios prioritized rapid results. The relationship between speed and explainability was also subtle, with some suggesting that *"if the speed is significantly slow, then it matters more when compared to explainability."* Explainability was also seen as a potential mitigator for false positives and negatives because, as P19 said, *"I don't expect humans to be 99% accurate all the time. If I have explainability, how much false positives and false negatives I get initially doesn't really matter to me as long as it gets better as I keep using it."*

5 Digital Media Forensics Ontology

From our Requirements Study, it is evident that intelligence analysts require a deepfake detection tool equipped with features that seamlessly support their workflow. One of the most prominent pain points was the frustration analysts felt from managing multiple tools. This fragmentation led to a disjointed user experience and made it challenging to consolidate results and generate insights. Analysts also expressed a desire for integrating traditional analysis methods with new analytics, combining deepfake detection with the traditional media forensics pipeline, and accommodating multiple modalities of input (i.e., images and audio). These insights pointed to the need for a more cohesive and integrated solution.

Additionally, the study highlighted the need for explainability. While a few explanation prototypes had some support from participants, there was a strong preference for intuitive textual information that could be more easily included in reports. Analysts also emphasized the need for an interface that is user-friendly for analysts at all experience levels.

While an all-inclusive tool could address the need for integration and explanation by centralizing analytics with standardized results, it introduces the challenge of overwhelming users with an ever-growing list of available analytics. As new analytics are developed to address emerging technologies and manipulation techniques, analysts face increasing difficulty in selecting the appropriate analytic for their specific tasks. To alleviate this complexity, a systematic organization of digital media forensics analytics is crucial. This led us to develop a *Digital Media Forensics Ontology* as a solution for presenting a large list of analytics to analysts in an interpretable way, addressing our **RQ4**. By structuring knowledge consistently across various analytics, the ontology reduces ambiguity and streamlines decision-making, ensuring that AI outputs are more relevant and aligned with traditional workflows. This not only simplifies analytic selection but also fosters better collaboration between analysts and AI systems by creating a shared understanding of the capabilities of the analytics.

5.1 Ontology Development

Our development process began by reviewing existing taxonomies to assess their suitability for organizing digital media forensic analytics. After reviewing the literature, we identified several limitations, notably that existing taxonomies

sometimes conflate manipulation types with detection techniques (e.g. the taxonomy for classification of image forgery detection methods by Thakur and Rohilla [81]), making the taxonomies harder to navigate. Furthermore, many techniques, such as the photo response non-uniformity (PRNU) technique, apply to both deepfake detection [46] and tampering localization [41]. However, concepts regarding the traditional and deep-learning-based detection techniques are often organized separately in existing taxonomies, leading to additional fragmentation.

Given these gaps, we adopted a tailored taxonomy development approach inspired by the methodology of Nickerson et al. [56]. The development process began with identifying a *meta-characteristic*—essentially the theme describing the characteristics of interest—that would guide the structure of the taxonomy. Given that our primary goal was to create a structure that would help analysts efficiently navigate the growing landscape of digital media analytics, we focused our meta-characteristic on the capabilities of analytics.

Starting from this foundation, we created a taxonomy categorizing the technical features employed by various analytics. Following a conceptual-to-empirical approach, we identified three major dimensions for this taxonomy: file structure, spatial features, and temporal features. However, we realized that the highly technical nature of this feature-based taxonomy may create a barrier to intuitive navigation for novice analysts. To address this, we developed a second taxonomy that focused instead on the analytic capabilities for detecting different manipulation types, such as deepfakes and manual manipulation, which is more accessible.

Against the recommendation of Nickerson et al. [56], our taxonomy development did not follow strict ending conditions, as our aim was not to exhaustively capture all possible concepts within the field of digital media forensics. Rather, we aimed to propose an initial framework and evaluate its suitability through a user study. In addition, we intended for our framework to evolve organically as new technologies and analytics emerge.

As the two taxonomies evolved, we recognized that a single, comprehensive knowledge base would simplify management and enhance usability. Ontologies, with their capacity for representing complex knowledge structures in an understandable and adaptable format, offered a fitting solution [14, 31]. Rather than building an ontology from scratch, we integrated our existing taxonomies as its foundational backbone. During this process, we noted that many deepfake detection methods target specific facial regions, such as the eyes or mouth. To capture this, we added a final class to the ontology that categorizes focal areas within digital content, allowing analysts to target the precise regions that particular analytics examine.

5.2 Ontology Structure

The ontology resulting from this development process, depicted in Figure 6, consists of three primary classes:

*Analytic Capabilities space – **Why** are we analyzing this piece of media?* This space captures the underlying motivations for digital media analysis, ranging from concerns about advanced deep-learning manipulations, such as deepfakes, and traditional manual alterations achieved through image editing software. It can serve as an entry point for analysts, who have ideas about what type of manipulations they are detecting.

*Search space – **Where** does the analytic look for artifacts?* Recognizing that there may be multiple points of interest within the digital content, this space provides a structured approach to identifying potential areas of manipulation. It guides analysts in focusing their analysis on specific elements, ranging from the file's overall structure to individual entities within the scene.

*Feature space – **What** feature does the analytic use?* This space is the most technically oriented, detailing the specific features utilized by analytics for detection. This includes file structure, which covers common items used in metadata analysis and other file-related features; spatial features, which focus on frame-level visual information such as lighting inconsistencies and textural anomalies; and temporal features, which analyze time-based elements such as human behavior patterns across sequential frames and frame-to-frame jitters.

Each analytic can be tagged with concepts in these spaces, indicating what they are capable of detecting, where they look for artifacts, and what features they utilize. To ensure the ontology accurately represents knowledge and meets practical needs, we used competency questions as a validation tool [54]. For instance, the question *"What features are used in <method name> to analyze this deepfake video?"* allowed us to assess the ontology's effectiveness in providing detailed insights about the features that known analytic methods use. We created several such questions, and manually navigated the ontology to determine whether the information it contained could fully answer the questions we raised. This process allowed us to evaluate the scope of the ontology and confirmed its ability to address practical needs analysts might have in exploring analytics capabilities in their workflows.

Overall, the ontology serves as a navigational guide for analysts to hone in on the most appropriate analytics that suit their needs for the piece of media in hand. In addition, while traversing the pathways in the ontology, analysts form hypotheses for their analysis. This helps place the detection results in context, making them more interpretable.

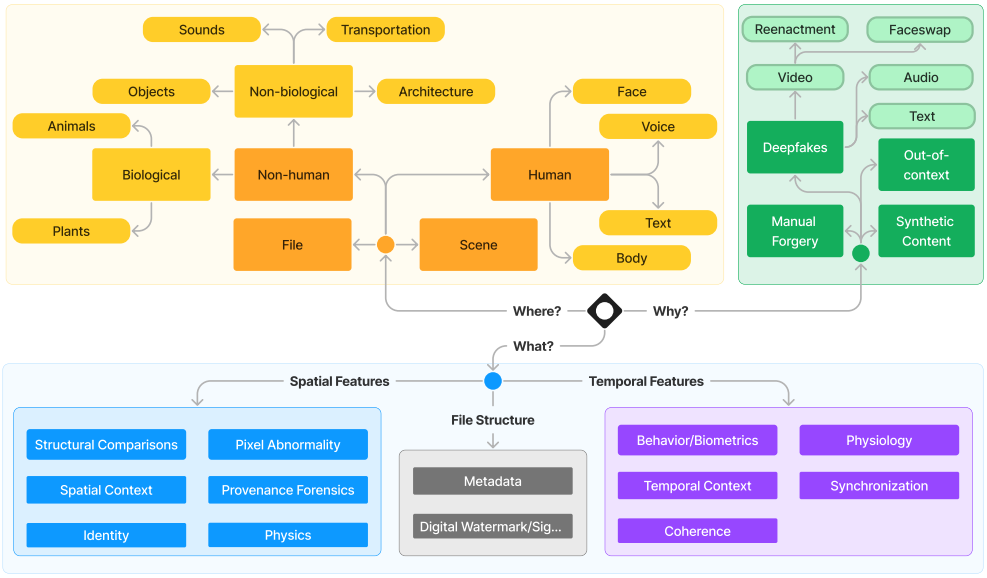


Fig. 6. Ontology graph showing initial core structure with why?, where?, and what? blocks and their corresponding nodes.

5.3 Ontology Study

To evaluate the effectiveness of this framework, especially for analytic selection and report writing, we conducted a task-based survey with analysts. The findings suggest that the proposed ontology positively impacts their experience in analytic selection and report writing by improving explainability, though there are areas that require further refinement.

5.3.1 Participants and Procedure. For this study, we recruited 11 analysts through the same collaboration network as in the Requirements Study. To maximize flexibility for participants, we designed a survey-based study. The survey materials were developed by us and distributed within the Intelligence Community by our collaborators, ensuring voluntary participation while maintaining necessary security protocols. The informed consent document was included in the survey, with participants indicating their consent by clicking to proceed. This study protocol also received approval from the IRBs of the participating institutions and was endorsed by the federally affiliated HRPO. For data analysis, we employed the same thematic analysis methodology described in section 3.4.

5.3.2 Survey Design. In our task-based survey, participants were asked to independently interact with two interfaces for choosing analytics to complete mock forensics tasks—one using the proposed ontology as its backbone (Figure 7a) and the other using fuzzy search (Figure 7b). They were also asked to write short reports based on their findings to simulate a portion of their analytical process. This approach allowed us to gather accurate feedback on their perceptions of the ontology and its implications on their work. The survey had three components, *tasks*, *interfaces*, and *questions*.

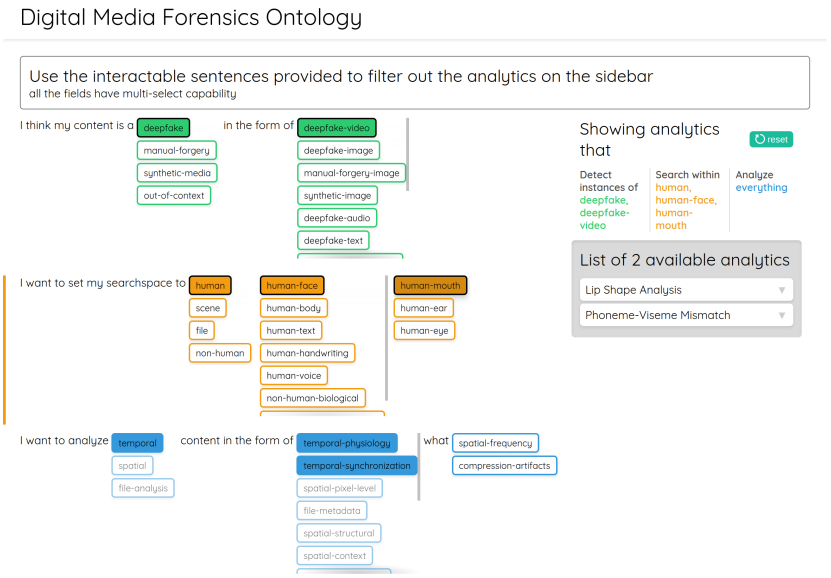
Tasks. We began the survey with two scenarios designed to mimic the types of analyses that analysts might perform. In the first scenario, participants were asked to identify the most appropriate analytic for each of two tasks: (i) detecting an image that had been manually manipulated, and (ii) examining a single frame from a deepfake video. In the second scenario, both tasks focused on report writing. Participants were provided with two deepfake videos, names of the analytics used to process the videos, and the corresponding results (see the example in Figure 7c). They were then asked to write short reports summarizing the findings.

Interfaces. To evaluate the ontology's potential in supporting analysts' workflows, we developed two distinct interfaces for the survey: an ontology-based sentence-forming interface and a fuzzy search interface for comparison.

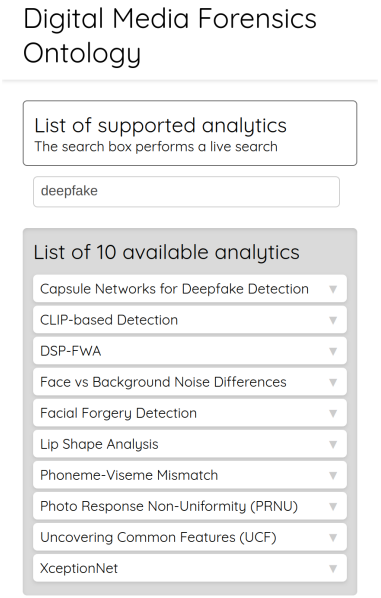
The ontology-based sentence-forming interface was conceived as a holistic guide, enabling analysts to navigate the complex landscape of analytics for digital media forensics. We introduced the concept of *pre-hoc explanations*, which are formed by combining the "why," "where," and "what" concepts associated with each analytic during the filtering stage to form hypotheses for analysis. Inspired by the preference for textual explanations observed in the Requirements Study, we implemented a sentence-forming format for filtering the analytics. For instance, to filter analytics using the concepts in the "why" space, analysts can complete the sentence "I think my content is a [deepfake] in the form of [deepfake-video]," as shown in Figure 7a. Concepts from the ontology were presented as clickable buttons, serving as filtering criteria. As participants selected concepts across the three spaces in the ontology, they constructed hypothesis sentences for verification. Recognizing the dynamic nature of the ontology framework and the varying expertise levels of analysts, the interface allowed filtering to begin from any of the three spaces. Filtered analytics were presented alphabetically, with expandable descriptions and associated ontology concepts available for each.

For comparison, we implemented a fuzzy search interface. This traditional approach presented all available analytics in a list format with a search bar, allowing participants to browse descriptions by clicking on individual items. The participants were asked to complete all the tasks first using the fuzzy search interface and then again using the ontology-based sentence-forming interface.

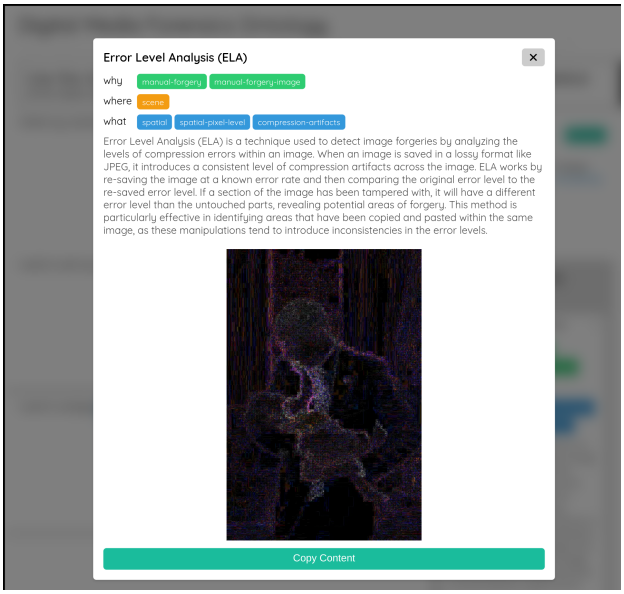
To populate these interfaces, we curated and tagged 28 digital media forensics analytics, such as Phoneme-Viseme Mismatch [4], CLIP-based Detection [13], and more (see the full list in the Appendix), assigning appropriate "why," "where," and "what" concepts based on our understanding of the corresponding research papers.



(a) Ontology text-completion filtering interface with pre-selected "deepfake-video" and "human-mouth" with their parent nodes to show the filtered results consisting of two analytics.



(b) Fuzzy search interface showing a list of analytics that appear when searching for "deepfake."



(c) The analytic result screen that the participant sees when selecting the Error Level Analysis (ELA) analytic for an image [35] provided in the study. The fuzzy search interface would not show the ontology tags.

Fig. 7. Screenshots of the web prototype used for the ontology user study.

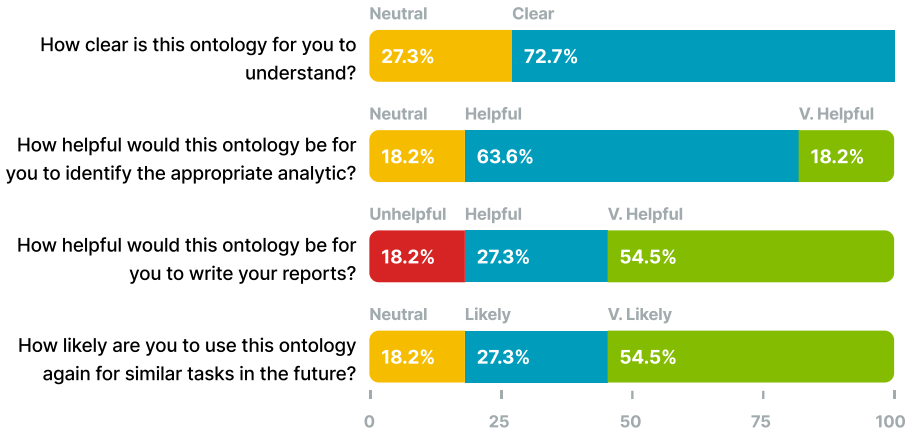


Fig. 8. Stacked bar chart showing the distribution of answers to the Likert Scale questions in the Ontology Study. We can see a mostly positive trend in the responses showing the potential usefulness of the new analytic discovery system.

Questions. Following the completion of each task, participants were asked to reflect on their experience using the two different search interfaces. This comparative feedback was crucial in understanding the usability and effectiveness of the ontology-based interface versus the traditional approach.

To obtain a quantitative overview of participants' attitudes towards the ontology, we incorporated four Likert Scale questions. These questions focused on key aspects of the ontology, including its clarity, usefulness in analytic searching and report writing, and likelihood of future use. These questions were selected based on their alignment with our study's objectives, helping us validate our hypothesis and inform areas for future development. Responses were analyzed by calculating mean scores to capture the overall impression of the ontology among the participants.

We also included three open-ended questions to capture more detailed feedback. These questions prompted participants to share their most and least favorite aspects of the ontology, identify any perceived gaps in the framework, and suggest potential improvements. These answers provided valuable insights into the strengths and limitations of the current ontology from the users' perspective.

The survey concluded with a set of demographic questions, focusing on participants' daily tasks, experience in digital media forensics analysis, and the recency of their experience. This information was collected to enable analysis of potential correlations between participants' professional backgrounds and their feedback on the ontology.

5.3.3 Survey Results. Through our analysis, we observed that the ontology had a generally positive impact on participants' experiences with both analytic selection and report writing, demonstrating its effectiveness in addressing the goals outlined in **RQ5**. When asked about the likelihood of using the ontology for similar tasks in the future (on a scale from 1, "Very Unlikely," to 5, "Very Likely"), participants gave an average rating of 4.36. Notably, 6 out of 11 participants indicated that they would be very likely to use the ontology again. These findings suggest that, with further enhancements to better align with user needs, the ontology offers a promising solution to the challenges of integration and explanation in digital media forensics. The overall themes extracted from our analysis of the survey are summarized in Figure 5, and the distribution of participants' answers for the Likert Scale questions are plotted in Figure 8.

Ontology for Analytic Selection. In the first scenario, we aimed to evaluate how the ontology influences the analytic selection process. Participants generally found the ontology-based interface to be more intuitive, as it offered a systematic and structured approach that guided their thought process while exploring available options. This filtering process allowed participants to concentrate on relevant analytics that target specific areas of interest. Participant PO2 commented, *"I felt more confident that I was being provided with the analytics that would actually help with my use case."* In addition, the ontology adds another benefit to the selection process as participant PO11 pointed out, *"it also helps in the overall explanation as you can narrow down what artifact the models are looking for to find complementary models."* When asked to rate the helpfulness of the ontology in identifying appropriate analytics on a scale from 1 ("Very Unhelpful") to 5 ("Very Helpful"), participants gave an average score of 4. While most found the ontology helpful for this task, two participants expressed a more neutral stance towards it.

Ontology for Report Writing. In scenario two, we examined how the ontology supports report writing and the communication of results. Participants generally found that the ontology added valuable context to their analysis, with the concept tags for each analytic acting as informative keywords that could be directly incorporated into reports. This not only strengthened the language but also as participant PO6 wrote *"the additional information in the second approach [ontology] was useful and lends credibility (trustworthiness?) to the results."* Additionally, participants perceived that reports written by incorporating the ontology concepts were more understandable to other stakeholders. Participants rated the usefulness of the ontology in report writing with an average score of 4.18, indicating that they found it generally helpful. However, the standard deviation was 1.11, reflecting some variability in responses. While most participants found the ontology beneficial, two participants rated it as unhelpful, with one noting that they relied more on the descriptions of the analytics rather than the ontology itself when writing their reports.

Recommendations for Enhancing the Ontology. As a preliminary framework, the ontology also received feedback regarding areas for improvement. The ontology's clarity received a mean score of 3.73 on a Likert scale of 1 to 5, indicating that participants generally found it "Clear." However, this score also suggests room for improvement. Non-expert participants especially expressed a desire for more accessible terminology and comprehensive definitions of the concepts used in the ontology, through which they can better understand the filtering options. This feedback pointed out the need to balance technical precision with user-friendly language to accommodate analysts with varying levels of expertise. Additionally, while the ontology-based search interface effectively narrowed down options, participants noted that in some cases, multiple analytics still remained after filtering. For example, many deepfake detection analytics tend to focus on human faces and use pixel-level features. In these instances, they found it challenging to make informed decisions without further guidance, as participant PO8 said *"it would be, however, nice if once you've fine-tuned your ontology, the analytics were sorted by most applicable."* As said, the system might better focus on a smaller set of default analytics and only offer analytics with overlapping utility if the user chooses a more detailed view.

6 Discussion

Definitions of Deepfake and Implications for Analysts. While most participants in our study demonstrated a solid understanding of deepfakes, we observed varied interpretations and occasional misunderstandings. This reflects a broader issue: there is no consistent or official definition of "deepfake" in either research or government that fully captures its key characteristic—the impersonation of real people. As generative AI technologies evolve, some participants conflated deepfakes with synthetic media in general, which encompasses a broader range of AI-generated content not

necessarily tied to real-world individuals. This conflation can lead to confusion, particularly for intelligence analysts who require precise terminology to ensure clarity in their reports and accuracy in their analyses.

The lack of standardized definitions poses significant challenges for analysts who must select appropriate analytics and interpret results correctly. Inconsistent terminology can undermine the reliability of their work and complicate communication among stakeholders. Standardization is crucial for ensuring that deepfake detection tools are applied consistently across different relevant contexts, helping analysts maintain the accuracy and credibility of their findings.

Standardization bodies such as the National Institute of Standards and Technology (NIST) and the Scientific Working Group on Digital Evidence (SWGDE) play a crucial role in addressing these gaps by working toward consistent definitions and standards. Our proposed ontology has the potential to contribute to this effort by integrating these definitions into analysts' workflows and identifying areas where further standardization is needed. By organizing analytics within a structured framework, the ontology helps clarify distinctions between different types of manipulated media and provides analysts with a more consistent way to interpret results. This not only enhances workflow efficiency but also facilitates clearer communication of findings. As the field evolves, such tools will be essential for maintaining precision and clarity, ultimately fostering greater trust in deepfake detection technologies.

The Multifaceted Nature of Explainability. Our participants made it clear that explainability is critical when they use deepfake detection tools which is complemented by relevant recent work [77], and current post-hoc explanations are not sufficient. While participants expressed preferences for certain formats of explanation during interviews, our findings highlight that explainability is, in fact, a multifaceted requirement that spans multiple stages of the analytic process.

Many commercial deepfake detection tools, such as Reality Defender [19], DuckDuckGoose [21], and Sensity AI [7], have incorporated explainability as a primary feature by offering heatmaps to visualize detection results. However, participants in our study found heatmaps to be overly technical and difficult to interpret without additional guidance. Even with assisted interpretation, visualizing explanations through heatmaps can be problematic. Prior research has shown that heatmaps are often inconsistent across different post-hoc explanation methods [1] and the same method could exhibit varied performance across different model backbones [55]. Studies with practitioners further highlight that these visualizations are challenging to interpret and can lead to incorrect assumptions about a model's behavior [29]. This instability raises concerns when incorporating such methods into analytic workflows that require a high level of repeatability.

Beyond merely interpreting detection results, analysts seek clarity on the overall workings of the tool and the specific capabilities of each analytic. This aligns with the standards set by documents from organizations like the Organization of Scientific Area Committees (OSAC) [58] and SWGDE [62], which provide procedural guidelines and reporting frameworks for media analysis. These standards emphasize repeatable processes and step-by-step lists that analysts can follow and refer to when explaining their analytical processes. Similarly, analysts in our study expressed a desire for a deepfake detection tool that not only produces detection results but also provides a transparent, systematic breakdown of the processes involved, akin to the documentation they rely on in manual analysis.

Our study highlights the necessity of designing an "explainable system" that is intuitive and transparent at every stage, from selecting analytics to interpreting the results. As discussed in Section 2.2, explainability is fundamental to effective human-machine teaming, fostering trust and enabling analysts to use the tool more confidently and efficiently. By organizing analytics through an ontology, our approach takes a step toward building such a system. This ontology-based

selection process provides analysts with insight into the specific capabilities of the analytics they choose, enhancing both their control over and understanding of the system's functions.

The Need for an Integrated Solution. The need for a comprehensive "one-stop shop" tool was a recurring theme among participants. Participant P4 emphasized the importance of a solution that allows users to access and manage a broad array of analytics across diverse data formats. Current tools, such as TrueMedia [83] and RealityDefender [19], employ various detection schemes and present their results in a rigid, concise format. While this may suffice for general internet users, our study indicates that intelligence analysts require more detailed information about the capabilities and specifications of these detection schemes. Such details are crucial for analysts to either justify or challenge the results of their analysis in reports. There is a strong preference from participants for tools that facilitate dynamic interaction, enabling analysts to customize functionality based on their specific analytical tasks. Such flexibility is essential for ensuring that tools not only automate repetitive tasks but also adapt to analysts' varying needs and preferences. The proposed ontology offers a structured framework for an integrated solution that balances usability with

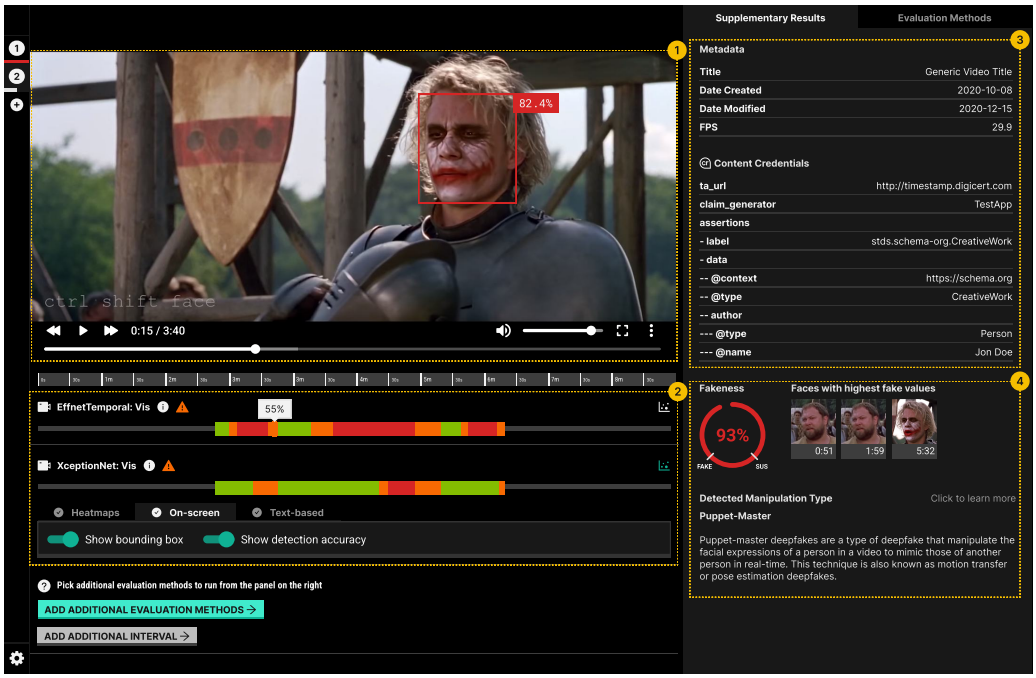


Fig. 9. The final evaluation screen showing key features: (1) video player with fine-grained video interaction capability, (2) individual analytics results, with analytic performance warnings, info bubbles, and optional post-hoc explainability tabs for each, (3) Curated EXIF metadata as well as Content Credentials [10], and (4) overall result and manipulation summary. Video frame sourced from [23].

the incorporation of diverse analytics. By systematically organizing analytics based on their capabilities, the ontology enables analysts to customize their investigative approach by navigating through different categories of analytics and choosing methods that align with their specific task requirements. Beyond streamlining this selection process, the ontology also enhances explainability by helping analysts understand how their chosen analytics contribute to their results.

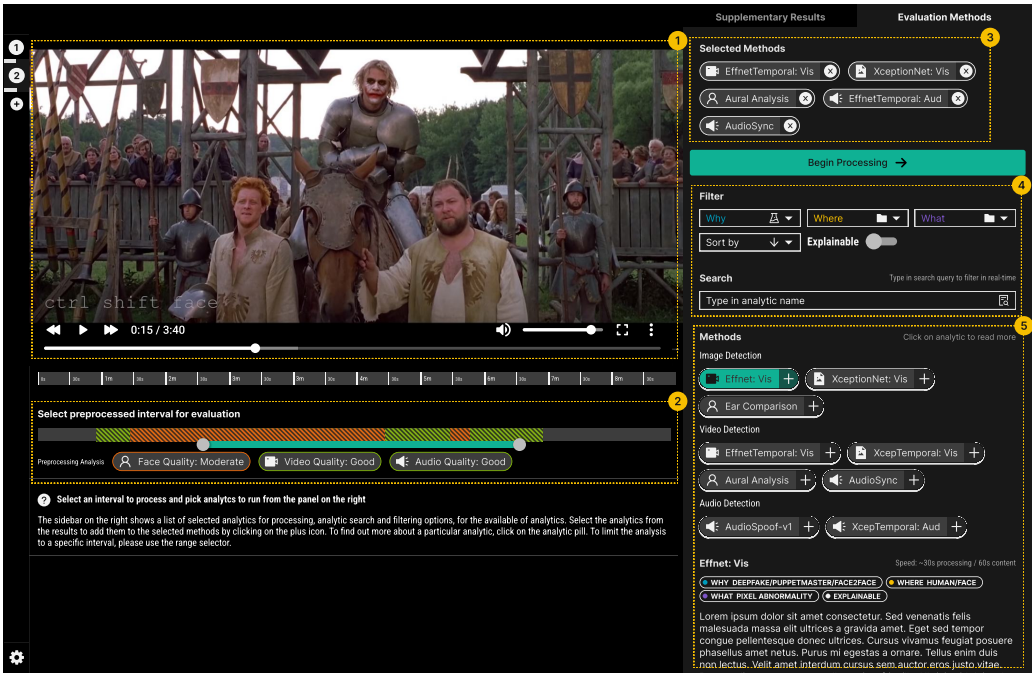


Fig. 10. Preprocessing screen showing a (1) video player, (2) color-coded content quality warning timeline, with each warning detail in a pill, (3) selected analytics for processing, (4) an ontology-guided analytic filtering system and a search functionality, (5) grouped list of analytics results with a details section showing a description, ontological tags, and expected runtime. Video frame sourced from [23].

7 Analyst-centered Deepfake Detection Tool Design

The insights gained from both of the Requirements Study and Ontology Study significantly impacted the evolution of the tool. It transitioned from a tool primarily designed for journalists to one that caters specifically to the unique needs of intelligence analysts. The redesigned interfaces shown in Figures 9 and 10 now feature a prominent **detection workbench**, serving as a central hub where analysts can access and interact with the results from various evaluation methods applied to the media content within the selected analysis interval. The layout of the workbench is divided into a larger primary results section and a multipurpose sidebar, the design of which was directly influenced by the analysts' emphasis on the need for a clear, integrated workspace that consolidates findings from multiple tools.

Primary Results Area. This section commands the most visual attention and contains all the critical interactive analytic outcomes and any associated explanations as shown in Figure 9. The media being analyzed is prominently displayed at the top within a media player featuring fine-grained video controls. Below, modular evaluation blocks present the output of the analytics in alignment with the video timeline, offering a clear and intuitive visualization. Each block also includes method-specific explainability sub-blocks, catering to users with varying levels of forensic expertise. The preference for diverse explanation formats, accommodating both novice and expert analysts, directly informed this design choice. The inclusion of informational labels indicating potential inaccuracies of the results further enhances the user experience by addressing analysts' concerns about the result interpretation and the need for contextual information.

Multipurpose Sidebar. The sidebar provides access to secondary insights and application controls. It includes blocks that display comprehensive information about the processed content, such as metadata, overall fakeness score, and identification of the individual faces with top fakeness scores, as well as an estimation of the manipulation method employed. The tool goes beyond conventional metadata extraction by incorporating the new Content Credential (CR) extraction from C2PA [10] initiative, bolstering the tool's reliability. The inclusion of CR directly addresses the analysts' positive reception of this initiative during the interviews.

The redesign of the tool prioritized usability and explainability. An alternate tab generally focused on during the preprocessing-step, and accessible at any point, offers a diverse array of analytics for analysts to apply to the content. The interface empowers the users to explore the available analytics, understand their underlying mechanisms, estimate processing times, add them to the evaluation queue, and seamlessly view the results in the primary area. To assist the analysts in navigating the potentially vast number of analytics, the tool leverages our Digital Media Forensics Ontology framework in the filter menu. This structured approach aims to alleviate the frustrations of the analysts with the tool fragmentation and their desire for a more guided and intuitive analytic selection process, as evidenced in the Ontology Study where participants reported feeling more confident in their analytic choices.

Before the application of various analytics, a preprocessing section shown in Figure 10 provides users with crucial information about potential effectiveness of different evaluation methods and allows them to select a smaller video snippet for analysis, promoting efficiency, especially considering the resource intensive nature of forensic AI tools. This feature aligns with the analysts' need for efficient content triage and the ability to focus on high-priority segments.

Addressing the challenge of fragmented toolsets, the redesigned interface introduces a tabbed structure for multiple concurrent workspaces. This allows for multiple pieces of content to be analyzed, providing a convenient overview of the progress and whether the content has been flagged within the tab interface.

The resulting user interface acts as a bridge between cutting-edge detection technology and the practical needs of the analysts. The modular design serves as a strong base for future integration of diverse manipulation methods and explainability types, which will empower analysts to navigate complexities of deepfake detection with greater efficiency, confidence, and understanding.

8 Limitations and Future Work

Tool Feature Implementation and Future Expansion. The newly designed tool, while primarily focused on empowering analysts to coherently analyze content with enhanced explainability, does not yet incorporate all the desired features identified by participants, such as note-taking, collaboration features, and batch-processing capabilities. Expanding the tool to include these features in future iterations could facilitate broader support for the analysts' workflows. Additionally, our research indicates that while many analysts encounter similar challenges and follow comparable workflows, their objectives often differ. Future work could benefit from designing with individual modalities in mind, enabling interface personalization, and maintaining ongoing collaboration with analysts to monitor and improve task efficiency with each system update.

Ontology Development and Growth. The concepts incorporated into the ontology were developed based on our current understanding of the literature. While sufficient for conducting this study and evaluating the ontology's potential to assist analysts, the definitions of these concepts lack rigorous evaluation and standardization. Future efforts should focus on collaborating with analytic developers to validate and standardize these definitions, ensuring consistency and reliability. Additionally, the current ontology is primarily focused on visual media, but its adaptable structure allows for

expansion into other modalities, such as audio and text, which are also critical in forensic analysis. This growth will not only broaden the ontology's utility but also address specific needs identified by analysts for a more versatile and multimodal approach to analysis.

Group Dynamics in Interviews. Due to logistical constraints, some in-person interviews in the Requirements Study were conducted in groups, which may introduce groupthink, potentially influencing individual responses [37]. However, our approach focused on gathering individual insights rather than facilitating group interaction, ensuring each participant had the opportunity to respond independently. Future work should consider more controlled, individual sessions, where feasible, to minimize potential bias. Also, exploring analysts' feedback in a focus group setting could provide valuable insights into shared challenges and perspectives, especially for collaborative tasks in their workflow.

Participant Expertise and Role Correlation. To maximize the participant pool in our study, given the limited access to intelligence analysts, we broadened our recruitment to include researchers, technical support specialists, and subject matter experts from the Intelligence Community. While this diversification could be viewed as a limitation, it proved advantageous in the context of our qualitative study, allowing us to gather diverse feedback from various participants who are all involved in the information analysis workflow. Although we included a question inviting participants to describe the types of tasks they often perform at an unclassified level, not all participants chose to respond. Consequently, the study was not explicitly focused on exploring the nuances of their interview answers based on their individual roles. Given that different roles offer perspectives and needs that could further inform and refine the tool's design, future work could delve deeper into role-specific requirements, enabling the development of tailored features that better align with particular tasks and expertise.

Sample Size and Participant Diversity. The Ontology Study involved a relatively small sample of 11 participants, including both experienced analysts and novice users familiar with reviewing various media formats. While this preliminary evaluation was sufficient to gain a general understanding of the ontology's reception and identify areas for improvement, future studies could benefit from a larger and more diverse study sample as the ontology grows into a mature state. Currently, our work focuses on intelligence analysts, who have specific needs for media verification and explanation. However, other potential users interested in multimedia forensics, such as journalists and law enforcement personnel, could also benefit from the solution we are proposing. Expanding future evaluations to include these groups would provide a broader perspective on the ontology's applicability across various professional contexts, enhancing its utility as an all-inclusive tool for digital media analysis.

9 Conclusion

This research contributes to the ongoing efforts to equip intelligence analysts with the tools and knowledge necessary to navigate the rapidly changing complexities of the digital media landscape, especially in the era of deepfakes. Our Requirements Study underscored the importance of tools that not only detect anomalies but also provide clear explanations for their findings. Analysts also expressed a preference for a unified, comprehensive tool capable of handling diverse media types and manipulation methods. While an all-inclusive tool that incorporates a wide range of analytics is an ideal solution, we recognize that developing such a tool would present usability challenges. To address this, we designed a Digital Media Forensics Ontology as a foundational step toward making an all-inclusive tool both functional and user-friendly.

The proposed ontology offers a significant advantage by demystifying the functions of various analytics, thereby allowing analysts to better understand the detection process and draw more informed conclusions from the results. Moreover, the adaptable nature of this ontology can incorporate new analytics as they emerge, thus remaining relevant in the face of evolving threats. Ultimately, our research shows the importance of not only focusing on tool capabilities, but also empowering analysts with the knowledge needed to trust and effectively utilize the tool in their work.

Acknowledgments

We are especially grateful to Jascha Swisher, Aaron W., Jacque J., and John L. for their advice and the extensive time they dedicated to ensuring the smooth execution of this work. We also sincerely thank all the participants who took the time to contribute to our study. This work was supported by the National Science Foundation Award no. 2040209 and 2429835, and an award from the U.S. Department of Defense (DoD).

This material is based upon work performed, in whole or in part, in coordination with the DoD. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the DoD and/or any agency or entity of the United States Government.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems (NeurIPS 2018)*, Vol. 31. Curran Associates, Inc., Montréal, Canada.

[2] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. MesoNet: a Compact Facial Video Forgery Detection Network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, Hong Kong, China, 1–7. https://doi.org/10.1109/WIFS.2018.8630761

[3] Shruti Agarwal and Hany Farid. 2021. Detecting Deep-Fake Videos from Aural and Oral Dynamics. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Nashville, TN, USA, 981–989. https://doi.org/10.1109/CVPRW53098.2021.00109

[4] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. 2020. Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Seattle, WA, USA, 2814–2822. https://doi.org/10.1109/CVPRW50498.2020.00338

[5] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. 2019. Protecting World Leaders Against Deep Fakes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, Long Beach, CA, USA, 38–45.

[6] National Security Agency. n.d.. Career Fields. https://www.intelligencecareers.gov/nsa/career-fields#intelligence-analysis. Accessed: 2024-11-13.

[7] Sensity AI. 2024. Sensity | Deepfakes detection. https://sensity.ai/deepfake-detection. Accessed: 2025-02-05.

[8] Gajanan K. Birajdar and Vijay H. Mankar. 2013. Digital image forgery detection using passive techniques: A survey. *Digit. Investig.* 10, 3 (Oct. 2013), 226–245. https://doi.org/10.1016/j.diin.2013.04.007

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[10] C2PA. 2024. Content Credentials. https://www.contentcredentials.org/.

[11] Akash Chintha, Aishwarya Rao, Saniat Sohrawardi, Kartavya Bhatt, Matthew Wright, and Raymond Ptucha. 2020. Leveraging Edges and Optical Flow on Faces for Deepfake Detection. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, Houston, USA, 1–10. https://doi.org/10.1109/IJCB48548.2020.9304936

[12] A. Chintha, B. Thai, S. J. Sohrawardi, K. M. Bhatt, A. Hickerson, M. Wright, and R. Ptucha. 2020. Recurrent Convolutional Structures for Audio Spoof and Video Deepfake Detection. *IEEE Journal of Selected Topics in Signal Processing* 14, 5 (2020), 1024–1037. https://doi.org/10.1109/JSTSP.2020.2999185

[13] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. 2024. Raising the Bar of AI-generated Image Detection with CLIP. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Seattle, WA, USA, 4356–4366. https://doi.org/10.1109/CVPRW63382.2024.00439

[14] Matteo Cristani and Roberta Cuel. 2005. A survey on ontology creation methodologies. *International Journal on Semantic Web and Information Systems (IJSWIS)* 1, 2 (2005), 49–69.

[15] DARPA. 2016. Media Forensics (MediFor). https://www.darpa.mil/program/media-forensics.

[16] DARPA. 2021. Semantic Forensics (SemaFor). https://www.darpa.mil/program/semantic-forensics.

[17] Soumyya Kanti Datta, Shan Jia, and Siwei Lyu. 2024. Exposing Lip-syncing Deepfakes from Mouth Inconsistencies. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, Niagara Falls, ON, Canada, 1–6. https://doi.org/10.1109/ICME57554.2024.10687902

[18] deepware.ai. 2025. Deepware | Scan & Detect Deepfake videos. https://deepware.ai. Accessed: 2025-02-05.

[19] Reality Defender. 2024. Enterprise-Grade Deepfake Detection. https://realitydefender.com. Accessed: 2025-02-05.

[20] Stephen L. Dorton and Samantha B. Harper. 2022. A Naturalistic Investigation of Trust, AI, and Intelligence Work. *Journal of Cognitive Engineering and Decision Making* 16, 4 (2022), 222–236. https://doi.org/10.1177/15553434221103718

[21] DuckDuckGoose. 2024. DuckDuckGoose. https://www.duckduckgoose.ai. Accessed: 2024-11-13.

[22] Walid El-Shafai, Mona A. Fouda, El-Sayed M. El-Rabaie, and Nariman Abd El-Salam. 2024. A comprehensive taxonomy on multimedia video forgery detection techniques: challenges and novel trends. *Multimedia Tools and Applications* 83, 2 (Jan. 2024), 4241–4307. https://doi.org/10.1007/s11042-023-15609-1

[23] Ctrl Shift Face. 2019. The Dark Knight's Tale [DeepFake]. https://www.youtube.com/watch?v=TgcvQA6-qBg&ab_channel=CtrlShiftFace.

[24] Federation of American scientists. n.d.. The Intelligence Cycle. https://irp.fas.org/cia/product/facttell/intcycle.htm. Accessed: 2025-02-05.

[25] William D. Ferreira, Cristiane B.R. Ferreira, Gelson da Cruz Júnior, and Fabrizzio Soares. 2020. A review of digital image forensics. *Computers & Electrical Engineering* 85 (2020). https://doi.org/10.1016/j.compeleceng.2020.106685

[26] Jessica L. Feuston and Jed R. Brubaker. 2021. Putting Tools in Their Place: The Role of Time and Perspective in Human-AI Collaboration for Qualitative Analysis. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 469 (Oct. 2021), 25 pages. https://doi.org/10.1145/3479856

[27] Laboratory for Analytic Sciences. n.d.. A Day in the Life of an NSA Intelligence Analyst. https://tae.ncsu-las.net/documents. Accessed: 2024-11-13.

[28] Candice R. Gerstner and Hany Farid. 2022. Detecting Real-Time Deep-Fake Videos Using Active Illumination. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, New Orleans, LA, USA, 53–60. https://doi.org/10.1109/CVPRW56347.2022.00015

[29] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3, 11 (Nov. 2021). https://doi.org/10.1016/S2589-7500(21)00208-9

[30] Dennis J Gleeson. 2023. Artificial Intelligence for Analysis: The Road Ahead. *Studies in Intelligence* 67, 4 (2023), 11–15.

[31] Michael Gruninger, Olivier Bodenreider, Frank Olken, Leo Obrst, and Peter Yim. 2008. Ontology Summit 2007 - Ontology, taxonomy, folksonomy: Understanding the distinctions. *Appl. Ontol.* 3, 3 (Aug. 2008), 191–200.

[32] Hui Guo, Xin Wang, and Siwei Lyu. 2023. Detection of Real-Time Deepfakes in Video Conferencing with Active Probing and Corneal Reflection. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Rhodes Island, Greece, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10094720

[33] David Güera and Edward J. Delp. 2018. Deepfake Video Detection Using Recurrent Neural Networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, Auckland, New Zealand, 1–6. https://doi.org/10.1109/AVSS.2018.8639163

[34] Bing Han, Xiaoguang Han, Hua Zhang, Jingzhi Li, and Xiaochun Cao. 2021. Fighting Fake News: Two Stream Network for Deepfake Detection via Learnable SRM. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 3, 3 (2021), 320–331. https://doi.org/10.1109/TBIOM.2021.3065735

[35] Silvan Heller, Luca Rossetto, and Heiko Schuldt. 2018. The PS-Battles Dataset - an Image Collection for Image Manipulation Detection. arXiv:1804.04866 [cs.MM]

[36] Shehzeen Hussain, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, and Julian McAuley. 2021. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Virtual, 3347–3356. https://doi.org/10.1109/WACV48630.2021.00339

[37] Irving L Janis. 1972. *Victims of groupthink: A psychological study of foreign-policy decisions and fiascoes.* Houghton Mifflin, Boston, MA.

[38] Abdul Rehman Javed, Zunera Jalil, Wisha Zehra, Thippa Reddy Gadekallu, Doug Young Suh, and Md. Jalil Piran. 2021. A comprehensive survey on digital video forensics: Taxonomy, challenges, and future directions. *Eng. Appl. Artif. Intell.* 106, C (Nov. 2021). https://doi.org/10.1016/j.engappai.2021.104456

[39] Nickson Karie and Victor Kebande. 2016. Building Ontologies for Digital Forensic Terminologies. *International Journal of Cyber-Security and Digital Forensics* 5 (04 2016), 75–82. https://doi.org/10.17781/P002032

[40] Sohail Ahmed Khan, Ghazaal Sheikhi, Andreas L. Opdahl, Fazle Rabbi, Sergej Stoppel, Christoph Trattner, and Duc-Tien Dang-Nguyen. 2023. Visual User-Generated Content Verification in Journalism: An Overview. *IEEE Access* 11 (2023), 6748–6769. https://doi.org/10.1109/ACCESS.2023.3236993

[41] Paweł Korus and Jiwu Huang. 2017. Multi-Scale Analysis Strategies in PRNU-Based Tampering Localization. *IEEE Transactions on Information Forensics and Security* 12, 4 (2017), 809–824. https://doi.org/10.1109/TIFS.2016.2636089

[42] Boquan Li, Jun Sun, Christopher M. Poskitt, and Xingmei Wang. 2024. How Generalizable are Deepfake Image Detectors? An Empirical Study. arXiv:2308.04177 [cs.CV]

[43] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2020. Face X-Ray for More General Face Forgery Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 5000–5009. https://doi.org/10.1109/CVPR42600.2020.00505

[44] Joselice Ferreira Lima, Cléia M. Gomes Amaral, and Luís Fernando R. Molinaro. 2010. Ontology: An Analysis of the Literature. In *ENTERprise Information Systems*. Springer, Berlin, Heidelberg, 426–435.

[45] Li Lin, Neeraj Gupta, Yue Zhang, Hainan Ren, Chun-Hao Liu, Feng Ding, Xin Wang, Xin Li, Luisa Verdoliva, and Shu Hu. 2024. Detecting Multimedia Generated by Large AI Models: A Survey. arXiv:2402.00045 [cs.MM]

[46] Florian Lugstein, Simon Baier, Gregor Bachinger, and Andreas Uhl. 2021. PRNU-based Deepfake Detection. In *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '21)*. ACM, Virtual Event, Belgium, 7–12. https://doi.org/10.1145/3437880.3460400

[47] Joseph B. Lyons, Kevin T. Wynne, Sean Mahoney, and Mark A. Roebke. 2019. Chapter 6 - Trust and Human-Machine Teaming: A Qualitative Study. In *Artificial Intelligence for the Internet of Everything*. Academic Press, Cambridge, MA, USA, 101–116. https://doi.org/10.1016/B978-0-12-817636-8.00006-5

[48] Asad Malik, Minoru Kuribayashi, Sani M. Abdullahi, and Ahmad Neyaz Khan. 2022. DeepFake Detection for Human Face Images and Videos: A Survey. *IEEE Access* 10 (2022), 18757–18775. https://doi.org/10.1109/ACCESS.2022.3151186

[49] Badhrinarayan Malolan, Ankit Parekh, and Faruk Kazi. 2020. Explainable Deep-Fake Detection Using Visual Interpretability Methods. In *2020 3rd International Conference on Information and Computer Technologies (ICICT)*. IEEE, San Jose, CA, USA, 289–293. https://doi.org/10.1109/ICICT50521.2020.00051

[50] Wasen Fahad Mashaan and Ismail Taha Ahmed. 2023. Manual and Automatic Feature Engineering in Digital Image Forgery Detection Algorithms: Survey. In *2023 19th IEEE International Colloquium on Signal Processing & Its Applications (CSPA)*. IEEE, Kedah, Malaysia, 81–86. https://doi.org/10.1109/CSPA57446.2023.10087398

[51] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. 2022. Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence* 53, 4 (June 2022), 3974–4026. https://doi.org/10.1007/s10489-022-03766-z

[52] Matt Novak. 2023. That Viral Image Of Pope Francis Wearing A White Puffer Coat Is Totally Fake. https://www.forbes.com/sites/mattnovak/2023/03/26/that-viral-image-of-pope-francis-wearing-a-white-puffer-coat-is-totally-fake/.

[53] Fatemeh Zare Mehrjardi, Ali Mohammad Latif, Mohsen Sardari Zarchi, and Razieh Sheikhpour. 2023. A survey on deep learning-based image forgery detection. *Pattern Recognition* 144 (2023). https://doi.org/10.1016/j.patcog.2023.109778

[54] Glaice Kelly Q. Monfardini, Jordana S. Salamon, and Monalessa P. Barcellos. 2023. Use of Competency Questions in Ontology Engineering: A Survey. In *Conceptual Modeling: 42nd International Conference, ER 2023, Lisbon, Portugal, November 6–9, 2023, Proceedings*. Springer-Verlag, Berlin, Heidelberg, 45–64. https://doi.org/10.1007/978-3-031-47262-6_3

[55] Katelyn Morrison, Ankita Mehra, and Adam Perer. 2023. Shared Interest…Sometimes: Understanding the Alignment between Human Perception, Vision Architectures, and Saliency Map Techniques. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Vancouver, BC, Canada, 3776–3781. https://doi.org/10.1109/CVPRW59228.2023.00391

[56] Robert C Nickerson, Upkar Varshney, and Jan Muntermann. 2013. A method for taxonomy development and its application in information systems. *European Journal of Information Systems* 22, 3 (2013), 336–359.

[57] Scott W. O'Connor. n.d.. What Does an Intelligence Analyst Do? https://graduate.northeastern.edu/knowledge-hub/what-does-an-intelligence-analyst-do/. Accessed: 2025-01-24.

[58] Organization of Scientific Area Committees (OSAC). n.d.. Video/Imaging Technology & Analysis Subcommittee. https://www.nist.gov/osac/subcommittees/videoimaging-technology-analysis. Accessed: 2024-11-13.

[59] Office of the Director of National Intelligence. n.d.. Intelligence Analysis. https://www.intelligence.gov/careers/explore-careers/389-intelligence-analysis. Accessed on 2025-01-27.

[60] Office of the Director of National Intelligence. 2022. Intelligence Community Directive 203 Technical Amendment. https://www.odni.gov/files/documents/ICD/ICD-203_TA_Analytic_Standards_21_Dec_2022.pdf.

[61] Office of the Director of National Intelligence. n.d.. What is Intelligence? https://www.dni.gov/index.php/what-we-do/what-is-intelligence. Accessed: 2025-02-05.

[62] Scientific Working Group on Digital Evidence(SWGDE). 2025. Published - Complete Listing. https://www.swgde.org/published-complete-listing. Accessed: 2024-11-13.

[63] OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/

[64] Rohan Paleja, Muyleng Ghuy, Nadun Ranawaka Arachchige, Reed Jensen, and Matthew Gombolay. 2021. The Utility of Explainable AI in Ad Hoc Human-Machine Teaming. In *Advances in Neural Information Processing Systems (NeurIPS 2021)*, Vol. 34. Curran Associates, Inc., Virtual, 610–623.

[65] Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, Sheng Zhang, Pingyu Wu, Bo Zhou, and Weiming Zhang. 2021. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. arXiv:2005.05535 [cs.CV]

[66] Samuele Pino, Mark James Carman, and Paolo Bestagini. 2021. What's wrong with this video? Comparing Explainers for Deepfake Detection. arXiv:2105.05902 [cs.CV]

[67] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In *Computer Vision – ECCV 2020*. Springer International Publishing, Cham, 86–103. https://doi.org/10.1007/978-3-030-58610-2_6

[68] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Seoul, Korea, 1–11. https://doi.org/10.1109/ICCV.2019.00009

[69] Johnny Saldaña. 2021. *The Coding Manual for Qualitative Researchers*. SAGE, Thousand Oaks, CA, USA. 1–440 pages.

[70] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Venice, Italy, 618–626. https://doi.org/10.1109/ICCV.2017.74

[71] National Security Agency/Central Security Service. n.d.. NSA, U.S. Federal Agencies Advise on Deepfake Threats. https://www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/3523329/nsa-us-federal-agencies-advise-on-deepfake-threats. Accessed: 2025-02-05.

[72] Shaikh Akib Shahriyar and Matthew Wright. 2022. Evaluating Robustness of Sequence-based Deepfake Detector Models by Adversarial Perturbation. In *Proceedings of the 1st Workshop on Security Implications of Deepfakes and Cheapfakes*. ACM, Nagasaki, Japan, 13–18. https://doi.org/10.1145/

3494109.3527194

[73] Nitin Arvind Shelke and Singara Singh Kasana. 2021. A comprehensive survey on passive techniques for digital video forgery detection. *Multimedia Tools Appl.* 80, 4 (Feb. 2021), 6247–6310. https://doi.org/10.1007/s11042-020-09974-4

[74] Raquel Vázquez Llorente shirin anlen. 2024. Spotting the deepfakes in this year of elections: how AI detection tools work and where they fail. https://reutersinstitute.politics.ox.ac.uk/news/spotting-deepfakes-year-elections-how-ai-detection-tools-work-and-where-they-fail.

[75] Leslie F. Sikos. 2021. AI in digital forensics: Ontology engineering for cybercrime investigations. *WIREs Forensic Science* 3, 3 (2021), 11 pages. https://doi.org/10.1002/wfs2.1394

[76] Saniat Javid Sohrawardi, Akash Chintha, Bao Thai, Sovantharith Seng, Andrea Hickerson, Raymond Ptucha, and Matthew Wright. 2020. DeFaking Deepfakes: Understanding Journalists' Needs for Deepfake Detection. In *Computation + Journalism Symposium*. C+J, Boston, MA, USA, 5 pages.

[77] Saniat Javid Sohrawardi, Y. Kelly Wu, Andrea Hickerson, and Matthew Wright. 2024. Dungeons & Deepfakes: Using scenario-based role-play to study journalists' behavior towards using AI-based verification tools for video content. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu, HI, USA, 1–17. https://doi.org/10.1145/3613904.3641973

[78] TALENTKINGHD. 2023. America's Got Talent 2022 Metaphysic Finals Full Performance & Intro. https://www.youtube.com/watch?v=nHDYpxYP6sk.

[79] Diangarti Tariang, Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2024. Synthetic Image Verification in the Era of Generative Artificial Intelligence: What Works and What Isn't There yet . *IEEE Security & Privacy* 22, 03 (May 2024), 37–49. https://doi.org/10.1109/MSEC.2024.3376637

[80] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2024. Gemini: a family of highly capable multimodal models. arXiv:2312.11805 [cs.CL]

[81] Rahul Thakur and Rajesh Rohilla. 2020. Recent advances in digital image manipulation detection techniques: A brief review. *Forensic Science International* 312 (2020). https://doi.org/10.1016/j.forsciint.2020.110311

[82] Alice Toniolo, Federico Cerutti, Timothy J. Norman, Nir Oren, John A. Allen, Mani Srivastava, and Paul Sullivan. 2023. Human-machine collaboration in intelligence analysis: An expert evaluation. *Intelligent Systems with Applications* 17 (2023). https://doi.org/10.1016/j.iswa.2022.200151

[83] TrueMedia. 2024. Identifying Political Deepfakes in Social Media using AI. https://www.truemedia.org/. https://www.truemedia.org/ Accessed: 2024-08-14.

[84] Luisa Verdoliva. 2020. Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing* 14, 5 (2020), 910–932. https://doi.org/10.1109/JSTSP.2020.3002101

[85] James Vincent. 2019. Deepfake detection algorithms will never be enough. https://www.theverge.com/2019/6/27/18715235/deepfake-detection-ai-algorithms-accuracy-will-they-ever-work

[86] Tianyi Wang, Harry Cheng, Kam Pui Chow, and Liqiang Nie. 2023. Deep convolutional pooling transformer for deepfake detection. *ACM Trans. Multimedia Comput. Commun. Appl.* 19, 6 (2023), 1–20. https://doi.org/10.1145/3588574

[87] Ying Xu, Kiran Raja, and Marius Pedersen. 2022. Supervised Contrastive Learning for Generalizable and Explainable DeepFakes Detection. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*. IEEE, Waikoloa, HI, USA, 379–389. https://doi.org/10.1109/WACVW54805.2022.00044

[88] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. 2024. Transcending Forgery Specificity with Latent Space Augmentation for Generalizable Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 8984–8994. https://doi.org/10.1109/CVPR52733.2024.00858

[89] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. 2023. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection. In *Advances in Neural Information Processing Systems*, Vol. 36. Curran Associates, Inc., New Orleans, USA, 4534–4565. https://neurips.cc/virtual/2023/poster/73502

[90] Xin Yang, Yuezun Li, and Siwei Lyu. 2019. Exposing Deep Fakes Using Inconsistent Head Poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Brighton, UK, 8261–8265. https://doi.org/10.1109/ICASSP.2019.8683164

[91] Mohammed Zakariah, Muhammad Khurram Khan, and Hafiz Malik. 2018. Digital multimedia audio forensics: past, present and future. *Multimedia Tools Appl.* 77, 1 (Jan. 2018), 1009–1040. https://doi.org/10.1007/s11042-016-4277-2

[92] He Zhang, Chuhao Wu, Jingyi Xie, Yao Lyu, Jie Cai, and John M. Carroll. 2024. Redefining Qualitative Analysis in the AI Era: Utilizing ChatGPT for Efficient Thematic Analysis. arXiv:2309.10771 [cs.HC]

[93] Cairong Zhao, Chutian Wang, Guosheng Hu, Haonan Chen, Chun Liu, and Jinhui Tang. 2023. ISTVT: Interpretable Spatial-Temporal Video Transformer for Deepfake Detection. *IEEE Transactions on Information Forensics and Security* 18 (2023), 1335–1348. https://doi.org/10.1109/TIFS.2023.3239223

A Ontology Nodes and Initial List of Analytics

Table 1. Capabilities, Search, and Feature Spaces

Block	Name	Description
Why	The Capabilities Space	Why are we analyzing this content?
Where	The Search Space	Where does the analytic focus?
What	The Feature Space	What features does the analytic use?
Why	deepfake	Media manipulated using deep learning algorithms placing identifiable entities in fictitious situations
Why	manual-forgery	Manually manipulated media
Why	out-of-context	Media used outside of the original context
Why	synthetic-media	Fully synthetic media content
Where	human	Analytic extracts human data and focuses only on that
Where	non-human	Analytic extracts non-human data and focuses on that
Where	scene	Analytic focuses on the whole scene
Where	file	Analytic focuses on the file
Why	deepfake-video	Videos manipulated using deep learning algorithms placing identifiable entities in fictitious situations
Why	deepfake-audio	Audio manipulated using deep learning algorithms placing identifiable entities in fictitious situations
Why	deepfake-image	Images manipulated using deep learning algorithms placing identifiable entities in fictitious situations
Why	synthetic-image	Fully synthetic image
Why	synthetic-video	Fully synthetic video
Why	synthetic-audio	Fully synthetic audio
Why	synthetic-text	Fully synthetic text
Why	deepfake-text	Text manipulated using deep learning algorithms to imitate an identifiable entity
Why	reenactment	Videos where an identifiable entity is manipulated using a driver content to perform actions
Why	faceswap	Videos where an identifiable entity's face is swapped
Why	manual-forgery-video	Manipulated video content created through manual editing
Why	manual-forgery-audio	Manipulated audio content created through manual editing
Why	manual-forgery-image	Manipulated image content created through manual editing
Why	manual-forgery-text	Manipulated text content created through manual editing
Where	human-face	Analytic extracts human face data and focuses only on that
Where	human-mouth	Analytic extracts human mouth data and focuses only on that
Where	human-eye	Analytic extracts human eye data and focuses only on that
Where	human-ear	Analytic extracts human ear data and focuses only on that
Where	human-voice	Analytic extracts human voice data and focuses only on that
Where	human-body	Analytic extracts human body data and focuses only on that
Where	human-handwriting	Analytic extracts handwriting data and focuses only on thatg
Where	human-text	Analytic extracts text data and focuses only on that
Where	non-human-biological	Analytic extracts non-human biological data and focuses only on that
Where	non-human-non-biological	Analytic extracts non-biological data and focuses only on that
What	spatial	Spatial distribution and arrangement of elements within media content, such as objects, shapes, and patterns.

Table 2. Capabilities, Search, and Feature Spaces

Block	Name	Description
What	spatial-structural	Focus on the structural features of the media content, such as the layout, composition, or arrangement of objects within the frame
What	spatial-pixel-level	Investigating individual pixel-level attributes of media content, such as color, intensity, and texture, to extract meaningful information
What	spatial-context	Contextual factors surrounding objects or scenes within media content, including their spatial relationships with other elements and their environmental context
What	spatial-provenance	Examine the provenance features of the media content, such as the source, origin, or chain of custody of the file based on visual signals
What	spatial-identity	Identifying and analyzing spatial features related to the identity of objects, individuals, or entities depicted in media content, such as facial recognition or object classification
What	spatial-physics	Examining spatial features influenced by physical properties and phenomena, such as lighting conditions, shadows, reflections, and perspective distortions
What	file-analysis	Investigating the structure, format, metadata, watermarks, and composition of media files to extract relevant information and detect anomalies or inconsistencies
What	file-metadata	Extracting and analyzing metadata associated with media files, including information such as creation date, authorship, location data, and device details
What	file-digital-watermark	Detecting and analyzing digital watermarks embedded within media files for purposes such as copyright protection, authentication, or tracking
What	temporal	Analyzing temporal aspects and changes in media content over time, including motion, dynamics, and temporal relationships between objects or events
What	temporal-biometrics	Extracting identity-specific biometric information from media content over time, such as facial expressions, gait analysis, or physiological signals for identity verification or behavioral analysis
What	temporal-physiology	Monitoring general species/object-specific physiological features and changes exhibited within media content over time, such as heart rate variability, pupil dilation, or facial blood flow for health monitoring or emotional analysis
What	temporal-context	Considering the temporal context and sequential relationships between events or actions depicted in media content, such as scene transitions, continuity, or narrative structure
What	temporal-synchronization	Ensuring synchronization and alignment of temporal elements within media content, such as audio-video synchronization or synchronization between different media streams
What	temporal-coherence	Assessing the coherence and consistency of temporal features within media content, such as smoothness of motion, temporal stability, or temporal consistency across frames

Table 3. Capabilities, Search, and Feature Spaces

Block	Name	Description
What	compression-artifacts	Identifying and analyzing artifacts introduced by compression algorithms during the encoding or decoding process, such as blockiness, blurring, or ringing artifacts
What	spatial-frequency	Extracting frequency-related characteristics from media content to identify patterns using operations involving spacial frequency transformation, such as DCT or FFT

Table 4. Initial List of Analytics for Detecting Media Manipulation (Part 1)

Name	Why	Where	What	Paper Title
MesoNet	deepfake, deepfake-image, deepfake-video, reenactment, faceswap	human, human-face	spatial, spatial-pixel-level	MesoNet: a Compact Facial Video Forgery Detection Network
Lip Shape Analysis	deepfake, deepfake-video, reenactment	human, human-face, human-mouth	temporal, temporal-physiology	Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection
Blood Flow Analysis	deepfake, deepfake-video, reenactment, faceswap	human, human-face	temporal, temporal-physiology	DeepRhythm: Exposing DeepFakes with Attentional Visual Rhythms
Eye Blinking Analysis	deepfake, deepfake-video, faceswap	human, human-face, human-eye	temporal, temporal-physiology	In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking
Face X-ray	deepfake, deepfake-image, reenactment, faceswap	human, human-face	spatial, spatial-pixel-level, blurring-artifacts	Face X-Ray for More General Face Forgery Detection
Error Level Analysis (ELA)	manual-forgery, manual-forgery-image	scene	spatial, spatial-pixel-level, compression-artifacts	
Head Pose Inconsistency	deepfake, deepfake-video, faceswap	human, human-face	temporal, temporal-physiology	Exposing Deep Fakes Using Inconsistent Head Poses
Cornea Reflection Analysis	synthetic-media, synthetic-image	human, human-face	spatial, spatial-structural	Exposing GAN-generated Faces Using Inconsistent Corneal Specular Highlights
F3Net	deepfake, deepfake-image, deepfake-video, reenactment, faceswap	human, human-face	spatial, spatial-pixel-level, spatial-frequency	Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues
Steganalysis Rich Model (SRM)	deepfake, deepfake-image, deepfake-video, manual-forgery, manual-forgery-image	human, human-face, scene	spatial, spatial-pixel-level, noise-artifacts	Learning Rich Features for Image Manipulation Detection

Table 5. Initial List of Analytics for Detecting Media Manipulation (Part 2)

Name	Why	Where	What	Paper Title
Spatial-Phase Shallow Learning	deepfake, deepfake-image, deepfake-video, reenactment, faceswap	human, human-face	spatial, spatial-pixel-level, spatial-frequency	Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain
Facial Forgery Detection	deepfake, deepfake-image, deepfake-video, reenactment, faceswap, synthetic-image	human, human-face	spatial, spatial-pixel-level	On the Detection of Digital Face Manipulation
Uncovering Common Features (UCF)	deepfake, deepfake-image, deepfake-video, reenactment, faceswap	human, human-face	spatial, spatial-pixel-level	UCF: Uncovering Common Features for Generalizable Deepfake Detection
Capsule-forensics	deepfake, deepfake-image, deepfake-video, reenactment, faceswap	human, human-face	spatial, spatial-pixel-level	Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos
DSP-FWA	deepfake, deepfake-image, deepfake-video, faceswap	human, human-face	spatial, spatial-pixel-level	Exposing DeepFake Videos By Detecting Face Warping Artifacts
CORE	deepfake, deepfake-image, deepfake-video, faceswap	human, human-face	spatial, spatial-pixel-level	CORE: Consistent Representation Learning for Face Forgery Detection
RECCE	deepfake, deepfake-image, deepfake-video, reenactment, faceswap	human, human-face	spatial, spatial-pixel-level	End-to-End Reconstruction-Classification Learning for Face Forgery Detection
Phoneme-Viseme Mismatch	deepfake, deepfake-video, reenactment	human, human-face, human-mouth	temporal, temporal-physiology, temporal-synchronization	Detecting Deep-Fake Videos From Phoneme-Viseme Mismatches

Table 6. Initial List of Analytics for Detecting Media Manipulation (Part 3)

Name	Why	Where	What	Paper Title
Face vs Background Noise Differences	deepfake, deepfake-image, deepfake-video, reenactment, faceswap	human, human-face, scene	spatial, spatial-pixel-level, noise-artifacts	Deepfake noise investigation and detection
Edge Region Feature Extraction	deepfake, deepfake-image, deepfake-video, reenactment, faceswap	human, human-face, scene	spatial, spatial-pixel-level	Generalized Facial Manipulation Detection With Edge Region Feature Extraction
ManTra-Net	manual-forgery, manual-forgery-image	scene	spatial, spatial-pixel-level	ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features
Model Fingerprints	synthetic-media, synthetic-image	scene	spatial, spatial-pixel-level, spatial-frequency	On the detection of synthetic images generated by diffusion models
Photo Response Non-Uniformity (PRNU)	deepfake, deepfake-image, deepfake-video, reenactment, faceswap, manual-forgery, manual-forgery-image	human, human-face, scene	spatial, spatial-pixel-level, noise-artifacts	PRNU-based Deepfake Detection
Metadata Analysis	deepfake, manual-forgery, synthetic-media	file	file-analysis, file-metadata	
Behavior Analysis	deepfake, deepfake-video, reenactment, faceswap	human, human-face	temporal, temporal-behavior	Detecting Deep-Fake Videos from Appearance and Behavior
CLIP-based Detection	synthetic-media, synthetic-image	scene	spatial, spatial-pixel-level	Raising the Bar of AI-generated Image Detection with CLIP
Aural & Oral Dynamic	deepfake, deepfake-video	human, human-face, human-ear	temporal, temporal-physiology	Detecting Deep-Fake Videos from Aural and Oral Dynamics
XceptionNet	deepfake, deepfake-video, deepfake-image	human, human-face	spatial, spatial-pixel-level	FaceForensics++: Learning to Detect Manipulated Facial Images