GPH-International Journal of Computer Science and Engineering
(Online ISSN 3050-9661)
Volume 8 | Issue 02 | July-Dec, 2025
Published by: Global Publication House
DOI: https://doi.org/10.5281/zenodo.21993083
Available online at: www.gphjournal.org
Journal homepage: https://gphjournal.org/index.php/cse/index

A Hybrid Explainable Artificial Intelligence Framework for Deepfake Attribution and Cross-Platform Disinformation Campaign Tracking

D. A. Onuma[1], D. Matthias[2], O. E. Taylor[3] & N. D. Nwiabu[4]
[1,2,3,4] Department of Computer Science, Rivers State University, Nkpolu-Oroworukwo, Port Harcourt, Rivers State, Nigeria.
Corresponding Author Email: onuma.anele@rsu.edu.ng

Abstract

The rapid advancement of Generative Artificial Intelligence (GenAI) has accelerated the creation and dissemination of highly realistic deepfakes, posing significant cybersecurity threats through identity fraud, misinformation, political manipulation, social engineering, and coordinated cross-platform disinformation campaigns. Although existing deep learning-based deepfake detection systems achieve high classification accuracy, most are limited to binary detection (real versus fake) and provide little support for source attribution, explainability, cross-platform campaign analysis, or digital forensics. These limitations reduce their effectiveness in cyber threat intelligence and forensic investigations. This study aimed to develop a **Hybrid Explainable Artificial Intelligence (XAI) Framework for Deepfake Attribution and Cross-Platform Disinformation Campaign Tracking**. The objectives were to develop a multimodal attribution framework for images, videos, audio, text, and metadata; integrate Explainable Artificial Intelligence techniques for transparent decision-making; implement graph-based cross-platform disinformation tracking; incorporate cyber threat intelligence and digital provenance; and evaluate the framework against existing approaches. A **Design Science Research (DSR)** methodology was adopted. The framework was implemented using Vision Transformers (ViT), Temporal Transformers, Wav2Vec 2.0, RoBERTa, Random Forest, XGBoost, and Multilayer Perceptron (MLP) within a weighted ensemble architecture. SHAP, LIME, and Grad-CAM were employed for explainability, while Graph Attention Networks (GATs) and Temporal Graph Neural Networks (TGNNs) enabled campaign tracking. Evaluation was conducted using FaceForensics++, Celeb-DF, DFDC, FakeAVCeleb, ASVspoof, Fakeddit, LIAR, CoAID, Twitter/X, and PHEME datasets. The proposed framework achieved **92.1% accuracy**, **91.2% precision**, **90.9% recall**, **91.0% F1-score**, and **95.9% ROC-AUC**. It further attained **90.8% attribution accuracy**, **98.3% Top-5 attribution precision**, **94.2% explanation fidelity**, **93.4% community detection accuracy**, and an end-to-end inference latency of **63.8 ms**, demonstrating near real-time performance. The study concludes that integrating multimodal attribution, Explainable AI, graph analytics, cyber threat intelligence, and digital forensics significantly enhances deepfake attribution and coordinated disinformation tracking beyond conventional detection systems. The proposed framework provides an effective solution for cybersecurity operations, digital forensics, law enforcement, social media monitoring, and national security applications.

Keywords:

Deepfake Attribution, Explainable Artificial Intelligence, Cross-Platform Disinformation, Cyber Threat Intelligence, Digital Forensics, Graph Neural Networks, Multimodal Learning, Cybersecurity.

1.0 Background to the Study

The rapid advancement of Artificial Intelligence (AI), particularly deep learning and generative artificial intelligence (GenAI), has transformed the creation, manipulation, and dissemination of digital media. Technologies such as Generative Adversarial Networks (GANs), diffusion models, transformers, and large multimodal models can generate highly realistic synthetic images, videos, audio, and text that are often indistinguishable from authentic content. While these technologies have enabled innovations in education, healthcare, entertainment, and digital media production, they have also introduced significant cybersecurity challenges by facilitating the creation of sophisticated deepfakes and AI-generated disinformation. Deepfakes have become powerful tools for identity theft, political manipulation, financial fraud, cyber espionage, social engineering, and influence operations, thereby threatening digital trust, public safety, and national security [4].

Deepfakes represent one of the fastest-growing forms of AI-enabled cyber threats because they combine the realism of multimedia content with the speed and scalability of modern digital communication. Unlike traditional misinformation, AI-generated content can convincingly imitate an individual's facial expressions, voice, writing style, or behavior, making manual verification increasingly difficult. Recent surveys indicate that advances in deep learning have substantially improved the realism of synthetic media while simultaneously reducing the effectiveness of conventional forensic techniques that rely on handcrafted features or image artifacts. Consequently, cybersecurity researchers have increasingly adopted deep learning architectures, including convolutional neural networks (CNNs), Vision Transformers (ViTs), recurrent neural networks, and multimodal learning models, to improve automated deepfake detection [18].

Although considerable progress has been achieved in deepfake detection, most existing approaches remain focused on **binary classification**, where the objective is simply to determine whether digital media are authentic or manipulated. Such approaches provide limited support for digital forensic investigations because they do not identify the origin of synthetic content or explain how the decision was reached. In practical cybersecurity environments, identifying manipulated media alone is insufficient. Security analysts, law enforcement agencies, and cyber threat intelligence teams require systems capable of identifying the likely source or generating model of synthetic media, reconstructing how the content propagates across digital platforms, and providing interpretable evidence that can support forensic investigations and legal proceedings. These requirements have shifted research attention from deepfake detection toward **deepfake attribution**, **content provenance**, and **explainable artificial intelligence (XAI)** [4].

Explainable Artificial Intelligence has emerged as an important research direction for improving transparency, accountability, and user trust in AI-driven cybersecurity systems. Unlike conventional black-box deep learning models, XAI techniques such as SHAP, LIME,

Grad-CAM, Integrated Gradients, and attention visualization enable analysts to understand the features and decision pathways that influence model predictions. The integration of explainability is particularly important in cybersecurity because operational decisions often require interpretable evidence rather than probability scores alone. Explainable models facilitate forensic analysis, improve analyst confidence, support regulatory compliance, and enhance the adoption of AI in safety-critical domains [2].

Recognizing these challenges, Maheshwari and Paulchamy [11] proposed a hybrid framework that integrates Explainable Artificial Intelligence with Adversarial Robustness Training (ART) to improve deepfake image detection. Their XAI-ART framework combines adversarial sample generation with CNN-based classification while employing SHAP and LIME to explain model predictions. Experimental results demonstrated that integrating adversarial robustness with explainability improves resilience against adversarial attacks and enhances the transparency of deepfake detection systems. Their work represents an important advancement toward trustworthy AI by showing that robust detection models can simultaneously maintain high performance and provide interpretable decision support [11].

Despite these contributions, several important limitations remain unresolved. First, the framework is primarily designed for **image-based deepfake detection** and does not support multimodal analysis of videos, cloned speech, AI-generated text, or heterogeneous digital evidence. Modern disinformation campaigns increasingly combine multiple media formats, making single-modality detection insufficient for operational cybersecurity applications. Second, the framework performs **binary classification (real versus fake)** but does not implement **deepfake attribution**, preventing investigators from identifying the likely generator, source, or creation pipeline responsible for synthetic media. Third, although explainability is incorporated into the detection process, the framework does not investigate how manipulated content propagates across multiple online platforms or how coordinated influence operations evolve over time. Consequently, it lacks graph-based propagation analysis, campaign reconstruction, digital provenance, and cyber threat intelligence capabilities that are increasingly required for defending against AI-enabled information operations [11].

Recent literature also emphasizes that the next generation of deepfake research should move beyond detection accuracy and address broader challenges, including multimodal learning, attribution, robustness, provenance, interpretability, and coordinated disinformation analysis. Comprehensive surveys identify insufficient generalization across emerging generative models, limited support for multimodal evidence fusion, and the absence of integrated frameworks capable of combining attribution with campaign analysis as major research gaps. Similarly, systematic reviews argue that future cybersecurity solutions should integrate media forensics, explainable AI, social network analytics, and digital provenance into unified architectures capable of supporting cyber threat intelligence and forensic investigations [17].

Furthermore, modern disinformation campaigns rarely remain confined to a single online platform. Malicious actors coordinate the dissemination of synthetic media across multiple social networking services through reposts, bot networks, coordinated accounts, hyperlinks, and evolving narratives. Existing deepfake detection frameworks generally analyze individual media objects in isolation and rarely model the temporal and structural relationships among users, content, and platforms. Graph Neural Networks (GNNs), temporal graph learning, and heterogeneous graph analytics provide promising techniques for reconstructing information

diffusion, identifying influential actors, detecting coordinated campaigns, and understanding narrative evolution. However, these techniques have seldom been integrated with multimodal deepfake attribution and explainable AI within a unified cybersecurity framework [1].

The limitations identified in Maheshwari and Paulchamy [11], together with the broader gaps reported in the literature, motivate the need for a more comprehensive cybersecurity framework capable of integrating multimodal deepfake attribution, explainable artificial intelligence, graph-based cross-platform disinformation campaign tracking, and cyber threat intelligence. Such a framework would not only determine whether digital content has been manipulated but would also identify the probable source of synthetic media, explain the rationale behind attribution decisions, reconstruct coordinated disinformation campaigns across multiple platforms, and provide actionable forensic intelligence for cybersecurity analysts, law enforcement agencies, and policy makers [11].

Therefore, the **aim** of this study is to develop a **Hybrid Explainable Artificial Intelligence Framework for Deepfake Attribution and Cross-Platform Disinformation Campaign Tracking** that integrates multimodal deep learning, explainable AI, graph-based analytics, and cyber threat intelligence to improve the detection, attribution, and tracking of AI-generated disinformation. To achieve this aim, the study will: (i) design a multimodal deepfake attribution framework for images, videos, audio, text, and metadata; (ii) integrate Explainable Artificial Intelligence techniques to provide transparent and trustworthy attribution decisions; (iii) develop a graph-based cross-platform disinformation campaign tracking model capable of reconstructing propagation pathways and identifying coordinated campaigns; (iv) implement a cyber threat intelligence and digital provenance module for evidence correlation and forensic decision support; and (v) evaluate the proposed framework against existing state-of-the-art methods using attribution accuracy, explainability, robustness, propagation analysis, and computational performance metrics. By addressing the limitations of existing frameworks, particularly that of Maheshwari and Paulchamy [11], the proposed study seeks to advance the state of the art in AI-driven cybersecurity by providing an integrated solution for deepfake attribution, explainable decision-making, and cross-platform disinformation intelligence.

2. LITERATURE REVIEW

2.1 Introduction

The emergence of Generative Artificial Intelligence (GenAI) has transformed the creation and dissemination of digital media, enabling the production of highly realistic synthetic images, videos, speech, and text. Although these technologies have numerous legitimate applications, they have also intensified cybersecurity threats by facilitating sophisticated deepfakes and coordinated disinformation campaigns. Consequently, recent research has shifted from conventional deepfake detection toward more advanced problems such as **deepfake source attribution**, **Explainable Artificial Intelligence (XAI)**, **cross-platform disinformation propagation**, and **digital forensics**. These four areas form the conceptual foundation of the proposed study because they collectively address not only the identification of manipulated content but also the attribution of its origin, explanation of AI decisions, reconstruction of coordinated campaigns, and preservation of digital evidence for forensic investigations [4].

2.2 Conceptual Review

2.2.1 Deepfake Source Attribution

Deepfake source attribution extends beyond conventional deepfake detection by identifying the likely generative model, software pipeline, or creator responsible for producing synthetic media. Unlike binary detection, which only classifies media as authentic or manipulated, attribution seeks to establish digital provenance and provide forensic evidence linking manipulated content to its source. This capability is increasingly important for cybersecurity, digital investigations, intellectual property protection, and accountability in AI-generated media.

Current attribution techniques exploit several forms of forensic evidence, including:

- Generator fingerprints
- Frequency-domain artifacts
- Sensor noise residuals
- Compression signatures
- Latent feature embeddings
- Metadata consistency
- Watermarking and provenance records

Generator fingerprinting has become one of the most promising attribution techniques because different generative models leave distinguishable statistical traces during image synthesis. Similarly, metadata analysis and frequency-domain characteristics provide complementary evidence that improves attribution reliability. However, attribution remains significantly more challenging than detection because adversarial modifications, compression, and post-processing often remove or obscure forensic traces [9].

Recent studies emphasize that future deepfake research should prioritize attribution rather than simple detection. Reliable attribution systems can support cybercrime investigations, copyright enforcement, misinformation analysis, and digital evidence management. Nevertheless, current attribution methods remain largely limited to image modalities and rarely integrate audio, video, textual, and social-network evidence into a unified framework [9].

Limitations

Existing attribution systems:

- focus mainly on images;
- have limited robustness against adversarial manipulation;
- rarely integrate multimodal evidence; and

provide limited support for cross-platform forensic investigations [9].

2.2.2 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) refers to techniques that make machine learning and deep learning decisions understandable to humans. Traditional deep learning models used in cybersecurity often operate as "black boxes," producing highly accurate predictions without revealing the reasoning behind their decisions. In forensic investigations, however, explainability is essential because investigators require interpretable evidence rather than probability scores alone.

Common XAI techniques include:

- SHAP
- LIME
- Grad-CAM
- Integrated Gradients
- Layer-wise Relevance Propagation
- Attention Visualization

These techniques identify influential features, highlight manipulated regions, quantify feature importance, and improve analyst confidence in AI-generated decisions.

Within deepfake analysis, XAI has primarily been used to explain detection outcomes by visualizing manipulated facial regions or identifying suspicious artifacts influencing classifier decisions. Explainability improves transparency, facilitates model validation, supports regulatory compliance, and enhances trust in AI-assisted forensic investigations [14].

Maheshwari and Paulchamy [11] demonstrated that integrating SHAP and LIME with adversarial robustness training significantly improves the interpretability and resilience of deepfake detection systems. However, their framework explains only binary detection decisions and does not extend explainability to source attribution or campaign reconstruction. Similarly, recent surveys conclude that although explainability has improved deepfake detection, little work has explored explainable attribution or explainable disinformation intelligence [14].

Limitations

Current XAI research:

- concentrates on deepfake detection;
- provides limited support for attribution decisions;
- rarely explains propagation pathways; and
- lacks integration with cyber threat intelligence systems [14].

2.2.3 Cross-Platform Disinformation Propagation

Disinformation campaigns have evolved from isolated fake news articles into coordinated operations involving AI-generated images, videos, cloned voices, synthetic text, bot networks, and coordinated social media accounts. Modern campaigns seldom remain confined to a single

platform; instead, they spread rapidly across multiple social media ecosystems through reposts, hyperlinks, screenshots, translations, and coordinated amplification.

Cross-platform propagation analysis seeks to reconstruct the diffusion pathway of disinformation by analyzing temporal, structural, and semantic relationships among users, posts, and media.

Key analytical techniques include:

- Graph Neural Networks (GNNs)
- Graph Attention Networks (GATs)
- Temporal Graph Networks (TGNs)
- Community Detection
- Information Cascade Analysis
- Social Network Analysis
- Narrative Evolution Modeling

Graph-based methods represent users, posts, hashtags, URLs, and multimedia content as interconnected nodes, while interactions such as reposts, mentions, replies, and hyperlinks form graph edges. These models enable cybersecurity analysts to identify influential actors, coordinated bot networks, and information cascades that facilitate the rapid spread of disinformation [1].

Recent reviews argue that future AI-enabled disinformation research should integrate multimedia forensic evidence with graph analytics because existing propagation models generally ignore the forensic characteristics of deepfake content. Consequently, there remains a disconnect between deepfake attribution and social network analysis [15].

Limitations

Most propagation models:

- analyze textual misinformation only;
- ignore multimedia forensic evidence;
- lack source attribution capabilities; and
- do not integrate explainable AI into campaign analysis [15].

2.2.4 Digital Forensics

Digital forensics involves the identification, acquisition, preservation, examination, analysis, and presentation of digital evidence for investigative and legal purposes. Within the context of deepfakes, digital forensics seeks to establish the authenticity, integrity, provenance, and chain of custody of digital media.

Traditional multimedia forensic techniques include:

- Error Level Analysis
- Camera Sensor Pattern Noise

- JPEG Compression Analysis
- Metadata Examination
- Copy-Move Detection
- Frequency Spectrum Analysis

Modern deepfake forensics increasingly incorporates machine learning and deep learning to detect subtle manipulation artifacts across images, videos, speech, and AI-generated text.

Recent forensic research has expanded toward multimodal analysis by combining image, video, audio, and textual evidence with metadata and contextual information. Digital provenance, blockchain-supported evidence preservation, and generator fingerprinting are emerging directions that strengthen forensic reliability and legal admissibility [19].

The literature indicates that effective forensic systems should support:

- multimodal evidence correlation;
- deepfake attribution;
- chain-of-custody preservation;
- provenance verification;
- explainable forensic reporting; and
- integration with cyber threat intelligence.

Despite these advances, current forensic systems generally remain focused on media authentication rather than coordinated campaign investigation or cross-platform intelligence [19].

Limitations

Existing forensic systems:

- emphasize detection rather than attribution;
- have limited multimodal integration;
- rarely incorporate graph-based intelligence;
- provide limited explainability; and
- do not reconstruct coordinated disinformation campaigns [19].

2.3 Synthesis of the Literature

The reviewed literature demonstrates that deepfake detection has advanced considerably through the use of deep learning, transformer architectures, and multimodal models. Likewise, explainable AI has improved transparency in detection systems, graph-based methods have enhanced the analysis of online information diffusion, and digital forensic techniques have strengthened multimedia authentication. However, these research areas have largely evolved independently. Existing studies typically focus on one aspect—such as detection, explainability, propagation analysis, or forensics—without integrating them into a unified cybersecurity framework [4].

.

The literature consistently identifies several unresolved challenges:

- limited support for **multimodal deepfake attribution**;
- explainability focused on detection rather than attribution;
- inadequate integration of multimedia forensics with graph-based campaign analysis;
- absence of unified frameworks for attribution, explainability, and propagation tracking; and
- limited cyber threat intelligence capabilities for coordinated disinformation investigations [9].

These limitations provide the motivation for the proposed research, which seeks to develop a **Hybrid Explainable Artificial Intelligence Framework for Deepfake Attribution and Cross-Platform Disinformation Campaign Tracking**. The proposed framework will integrate multimodal attribution, explainable AI, graph neural network-based propagation analysis, and digital forensics into a single cybersecurity architecture capable of supporting cyber threat intelligence, digital investigations, and evidence-driven decision-making.

2.4 Related Works

Maheshwari and Paulchamy [11] developed a hybrid **Explainable Artificial Intelligence (XAI)** framework integrated with **Adversarial Robustness Training (ART)** to improve deepfake image detection. The framework employed a Convolutional Neural Network (CNN) trained with adversarial samples generated using the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), while SHAP and LIME were used to provide interpretable explanations for classification decisions. Experimental results demonstrated improved robustness against adversarial attacks and enhanced transparency of the detection process. However, the framework is limited to image-based binary classification and does not support multimodal analysis, deepfake source attribution, graph-based propagation analysis, or cross-platform disinformation tracking. Consequently, there is a need for an integrated explainable framework capable of multimodal deepfake attribution, coordinated campaign tracking, and cyber threat intelligence.

Fernández Gambín [4] conducted a comprehensive review of deepfake generation, detection techniques, benchmark datasets, applications, and future research trends. Using a systematic literature review, the authors examined recent advances in Generative Adversarial Networks (GANs), diffusion models, transformer-based architectures, and multimedia forensic techniques. Their findings revealed that transformer-based methods have significantly improved detection performance while emphasizing explainability, provenance analysis, and trustworthy AI as critical future research directions. Nevertheless, the study is purely a survey and does not propose or experimentally validate an integrated framework for deepfake attribution or cross-platform disinformation analysis. This highlights the need for practical frameworks that combine multimodal attribution, explainable AI, graph analytics, and cyber threat intelligence.

Abbas and Taeihagh [1] presented a systematic review of artificial intelligence techniques for deepfake generation and detection. The study analyzed CNNs, Vision Transformers, recurrent neural networks, multimodal learning approaches, benchmark datasets, and evaluation metrics reported in recent literature. The review concluded that transformer-based architectures generally outperform conventional CNNs while identifying robustness, explainability, and generalization as persistent challenges in deepfake detection. However, the review primarily focuses on detection technologies and provides limited discussion on deepfake source attribution, coordinated disinformation campaigns, and digital forensic intelligence. Therefore, future studies should develop unified frameworks that integrate explainable attribution, multimodal learning, graph analytics, and cyber threat intelligence.

Passos [18] reviewed deep learning-based approaches for detecting deepfake content across image, video, and audio modalities. The authors evaluated CNNs, Vision Transformers, hybrid deep learning models, benchmark datasets, and commonly used performance metrics. Their findings showed that deep learning techniques achieve high detection accuracy under controlled experimental conditions but experience performance degradation when evaluated on unseen datasets and newly emerging generative models. Although the review provides a comprehensive assessment of detection methods, it does not investigate deepfake source attribution, Explainable Artificial Intelligence, cross-platform propagation analysis, or digital forensic intelligence. Consequently, there is a need for multimodal attribution frameworks capable of identifying deepfake generators while supporting explainable forensic investigations.

Altuncu [3] carried out a meta-review of deepfake research focusing on definitions, benchmark datasets, evaluation metrics, and existing survey literature. Through comparative analysis of current datasets and evaluation standards, the authors identified inconsistencies in dataset quality, deepfake taxonomy, and performance evaluation procedures across existing studies. They recommended the development of standardized datasets and benchmarking methodologies to improve the reproducibility and comparability of future deepfake research. However, the study neither proposes a practical deep learning framework nor investigates source attribution, explainable AI, or cyber threat intelligence. This creates an opportunity for future research to develop standardized multimodal attribution frameworks incorporating explainability and digital forensics.

Alshahrani [2] presented a comprehensive survey of Explainable Artificial Intelligence (XAI) techniques and their applications across intelligent systems. The study reviewed SHAP, LIME, Grad-CAM, Integrated Gradients, Layer-wise Relevance Propagation, and other explanation techniques while discussing their roles in improving transparency, trustworthiness, fairness, and accountability of AI models. The authors concluded that explainability significantly enhances user confidence in AI-assisted decision-making, particularly in safety-critical domains. Nevertheless, the survey is domain-independent and does not specifically address explainable deepfake attribution or multimedia forensic analysis. Therefore, future research should investigate XAI frameworks tailored to deepfake attribution, cross-platform disinformation tracking, and cyber threat intelligence.

Tsigos [26] proposed a quantitative evaluation framework for Explainable Artificial Intelligence methods used in deepfake detection. The study compared SHAP, LIME, Grad-CAM, saliency maps, and attention mechanisms using adversarial perturbation experiments to assess explanation quality and reliability. Experimental results indicated that SHAP and Grad-CAM produced more accurate and consistent explanations than local explanation techniques, thereby improving trust in deep learning predictions. However, the framework evaluates explainability only for binary deepfake detection and does not extend explanations to source attribution, graph-based propagation analysis, or coordinated disinformation campaigns. This indicates the need for explainable attribution systems that integrate graph analytics and digital forensic intelligence.

Qureshi [19] reviewed digital forensic techniques for multimodal deepfake identification on social media platforms. The authors examined image, video, audio, and text forensic methods alongside metadata analysis, provenance verification, multimedia authentication, and forensic workflows for AI-generated content. Their findings showed that multimodal forensic analysis substantially improves deepfake identification and highlighted the importance of metadata integrity and provenance verification in digital investigations. Despite its comprehensive forensic perspective, the review does not integrate Explainable Artificial Intelligence, graph-based campaign tracking, or deepfake source attribution into a unified operational framework. Accordingly, future research should combine digital forensics with explainable attribution and cross-platform cyber threat intelligence.

Shao [23] proposed a transformer-based multimodal framework for improving deepfake detection using joint visual, audio, and textual feature learning. The framework employed multimodal transformer architectures to fuse heterogeneous feature representations extracted from images, videos, speech, and textual information for robust deepfake classification. Experimental evaluation demonstrated that multimodal feature fusion significantly outperformed unimodal detection models in terms of accuracy and robustness. However, the study focuses primarily on classification performance and does not incorporate deepfake source attribution, Explainable Artificial Intelligence, graph-based propagation analysis, or cyber threat intelligence. Therefore, future work should extend multimodal detection models toward explainable attribution and coordinated disinformation campaign analysis.

Zhang [31] investigated deepfake source attribution by identifying the generative model responsible for producing synthetic images. The study extracted generator-specific fingerprints and deep feature representations to distinguish images generated by different Generative Adversarial Networks (GANs). Experimental results demonstrated that GAN models leave identifiable fingerprints that can be exploited for accurate source attribution, thereby extending deepfake analysis beyond binary detection. Nevertheless, the framework is limited to image-based GAN attribution and does not support video, audio, text, explainability, graph analytics, or cross-platform disinformation tracking. Future research should therefore develop multimodal attribution frameworks that integrate Explainable Artificial Intelligence, graph neural networks, and digital forensic intelligence to provide comprehensive cybersecurity solutions.

Wang [28] investigated the reliability of deepfake detection systems by examining their robustness, interpretability, transferability, and applicability in real-world forensic scenarios. The authors conducted a comprehensive survey of state-of-the-art deepfake detection methods

and evaluated their performance using publicly available benchmark datasets while proposing reliability-oriented evaluation metrics. The study found that although existing deep learning models achieve high detection accuracy under controlled conditions, many fail to generalize to unseen deepfake generators and real-world environments because of limited robustness and interpretability. The authors concluded that improving the reliability of deepfake detection systems is essential for legal and forensic applications. However, the study focuses primarily on detection reliability and does not investigate source attribution, cross-platform disinformation tracking, or cyber threat intelligence. Therefore, future research should develop explainable and reliable multimodal attribution frameworks capable of supporting digital forensic investigations and coordinated campaign analysis.

Zhang [32] proposed one of the earliest deepfake source attribution frameworks by investigating whether synthetic images generated by different Generative Adversarial Networks (GANs) contain identifiable fingerprints that can be used to determine their origin. The framework extracted generator-specific artifacts using deep feature representations and machine learning classifiers to distinguish among multiple GAN architectures. Experimental results demonstrated that GAN models leave distinctive fingerprints, enabling accurate attribution of generated images to their corresponding source models. This work established deepfake attribution as an important extension of conventional deepfake detection. However, the framework is limited to image-based GAN attribution and does not consider videos, cloned audio, AI-generated text, Explainable Artificial Intelligence, graph analytics, or coordinated disinformation campaigns. Consequently, future research should develop multimodal attribution systems that integrate explainability, graph neural networks, and digital forensic intelligence for comprehensive cybersecurity applications.

Juefei-Xu [5] presented a comprehensive survey of malicious deepfakes, reviewing deepfake generation, detection, adversarial attacks, benchmark datasets, and future research directions. The study analyzed more than 300 research publications and proposed a detailed taxonomy describing the ongoing competition between deepfake generators and detection systems. The authors highlighted adversarial attacks, generalization, explainability, robustness, and reliability as major challenges in developing trustworthy deepfake detection frameworks. They further emphasized that future research should integrate adversarial robustness with explainable and reliable AI techniques. Despite its comprehensive coverage, the study remains a survey without proposing an operational framework for deepfake attribution or coordinated disinformation analysis. Therefore, an integrated framework combining attribution, Explainable Artificial Intelligence, graph analytics, and cyber threat intelligence remains an important research need.

Verdoliva [27] reviewed emerging multimedia forensic techniques for detecting deepfakes and digital image manipulations. The study examined both handcrafted forensic features and deep learning-based approaches, including CNNs and frequency-domain analysis, for detecting manipulated visual media. The findings showed that deep learning techniques generally outperform traditional forensic methods in identifying synthetic content but remain vulnerable to post-processing operations, compression, and adversarial manipulation. The author emphasized that future forensic systems should improve robustness, generalization, and explainability to remain effective against rapidly evolving generative models. However, the review concentrates on multimedia detection and does not address deepfake source attribution, multimodal evidence fusion, or coordinated cross-platform disinformation tracking. This

highlights the need for integrated forensic frameworks capable of combining attribution, explainable AI, and cyber threat intelligence.

Mirsky and Lee [13] presented one of the most influential surveys on the creation and detection of deepfakes by reviewing deepfake generation techniques, detection algorithms, benchmark datasets, and future research directions. The study systematically analyzed Generative Adversarial Networks, autoencoders, face manipulation techniques, and deep learning-based detection methods while discussing challenges relating to privacy, misinformation, and national security. The authors concluded that although deepfake detection has advanced considerably, emerging generative models continue to outpace existing detection techniques, creating an ongoing arms race between attackers and defenders. The survey also highlighted the need for more robust, explainable, and generalizable detection methods. However, it does not propose an integrated framework for multimodal source attribution, graph-based campaign analysis, or digital forensic intelligence, thereby motivating further research in these areas.

Tolosana [25] conducted a comprehensive survey of deepfake generation and detection technologies with emphasis on biometric security and facial recognition systems. The authors reviewed state-of-the-art face manipulation techniques, publicly available datasets, evaluation metrics, and deep learning detection algorithms while examining the impact of deepfakes on biometric authentication. Their findings indicated that CNN-based models achieve high detection performance on benchmark datasets but exhibit poor generalization across unseen datasets and emerging manipulation techniques. The study recommended the development of more generalized, explainable, and robust detection systems capable of adapting to future deepfake technologies. Nevertheless, the survey focuses primarily on detection and biometric security without addressing source attribution, graph analytics, or cross-platform disinformation tracking. Consequently, future frameworks should integrate multimodal attribution with cyber threat intelligence and digital forensics.

Nguyen [16] presented a comprehensive survey of deep learning techniques for deepfake creation and detection. The authors reviewed GAN-based generation methods, transformer architectures, convolutional neural networks, benchmark datasets, and evaluation protocols while discussing major challenges such as adversarial robustness, model generalization, and computational complexity. Their findings showed that transformer-based architectures outperform conventional CNNs in many detection tasks but remain susceptible to unseen deepfake generators and domain shifts. The study also identified explainability, attribution, and multimodal learning as important future research directions. However, it does not provide an integrated framework that combines source attribution, graph-based campaign tracking, and digital forensic intelligence. Therefore, future research should extend deep learning models toward explainable multimodal attribution frameworks capable of supporting cybersecurity investigations.

Rana [20] conducted a systematic literature review of deepfake detection techniques by analyzing over one hundred published studies covering deep learning, classical machine learning, statistical approaches, and blockchain-based solutions. The review compared detection performance across several benchmark datasets and concluded that deep learning methods consistently outperform traditional approaches in detecting manipulated multimedia. The authors also identified dataset imbalance, limited generalization, adversarial vulnerability, and lack of standardized evaluation as significant research challenges. Despite its

comprehensive analysis, the review focuses mainly on detection methods and does not examine source attribution, explainable AI, coordinated disinformation campaigns, or cyber threat intelligence. Consequently, future studies should develop unified frameworks capable of integrating these emerging cybersecurity requirements.

Khan and Dang-Nguyen [7] investigated the **generalization capability of deepfake detection models** across different architectures, datasets, and pre-training paradigms with the objective of identifying models that perform reliably beyond the datasets on which they were trained. The authors conducted an extensive comparative evaluation of **eight supervised deep learning architectures** (including CNNs and Vision Transformers) together with **two self-supervised transformer models (DINO and CLIP)** using four benchmark datasets: **FaceForensics++, DFDC, CelebDF-V2, and FakeAVCeleb**. Their methodology involved both **intra-dataset** and **inter-dataset** experiments, while also examining the effects of image augmentation, model size, computational efficiency, and pre-training strategies on detection performance. The experimental results showed that **Transformer-based architectures consistently outperformed conventional CNN models**, particularly in cross-dataset evaluation, demonstrating superior generalization capability. Furthermore, **FaceForensics++ and DFDC** were found to produce models with better cross-dataset generalization than CelebDF-V2 and FakeAVCeleb, while image augmentation further improved the performance of transformer models. Despite these contributions, the study focuses primarily on **deepfake detection and model generalization**, without addressing **deepfake source attribution**, **Explainable Artificial Intelligence (XAI)**, **cross-platform disinformation propagation**, **graph neural network-based campaign analysis**, or **digital forensic intelligence**. Consequently, there remains a need for a comprehensive framework that extends generalized deepfake detection to **multimodal source attribution**, **explainable decision-making**, **cross-platform disinformation tracking**, and **cyber threat intelligence**, thereby providing a more practical solution for real-world cybersecurity and digital forensic investigations.

Masood [12] presented a comprehensive review of deepfake attacks, covering generation techniques, detection methods, benchmark datasets, cybersecurity challenges, and future research directions. The authors examined GANs, diffusion models, multimedia forensics, transformer-based detection algorithms, and publicly available datasets while discussing the growing threat of AI-generated misinformation and digital identity attacks. Their findings emphasized that although detection accuracy has improved considerably, existing systems still struggle with robustness, unseen manipulations, explainability, and real-world deployment. The study recommended integrating multimodal learning, explainable AI, blockchain-based provenance, and digital forensics to improve future deepfake defense systems. However, it does not propose a practical framework that unifies deepfake attribution, graph-based disinformation tracking, and cyber threat intelligence. This limitation provides a strong motivation for the proposed Hybrid Explainable Artificial Intelligence Framework for Deepfake Attribution and Cross-Platform Disinformation Campaign Tracking.

Sandotra and Arora [21] presented a comprehensive review of deepfake detection technologies with the objective of analyzing state-of-the-art machine learning and deep learning approaches, benchmark datasets, evaluation metrics, and emerging challenges. The study reviewed Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), recurrent neural networks, frequency-domain analysis, and multimodal detection techniques while comparing their performance across image, video, and audio datasets. The authors found that transformer-

based architectures and multimodal learning models consistently outperform traditional CNN-based methods, but they also noted persistent challenges related to robustness, cross-dataset generalization, explainability, and real-world deployment. However, the review is limited to deepfake detection and does not investigate source attribution, coordinated disinformation tracking, or cyber threat intelligence. Consequently, future research should develop integrated frameworks that combine multimodal deepfake attribution, Explainable Artificial Intelligence (XAI), graph-based campaign analysis, and digital forensics to support practical cybersecurity applications.

Shoaib [24] examined the growing relationship between deepfakes, misinformation, disinformation, frontier AI, and large generative models with the objective of identifying effective defence mechanisms against AI-generated content. The authors reviewed recent literature on multimodal deepfake detection, digital watermarking, machine learning-based authentication, and policy-driven countermeasures before proposing a conceptual defence framework that combines advanced detection algorithms, cross-platform collaboration, and ethical governance. Their findings indicate that combating AI-enabled disinformation requires a multidisciplinary approach involving technological innovation, cross-platform cooperation, public awareness, and regulatory oversight. Nevertheless, the proposed framework remains conceptual and does not implement graph neural networks, explainable AI, deepfake source attribution, or cyber threat intelligence for operational deployment. Therefore, future work should focus on implementing intelligent frameworks capable of explainable multimodal attribution and coordinated cross-platform disinformation tracking.

Li [10] reviewed Explainable Artificial Intelligence (XAI) techniques for deepfake detection with the objective of establishing a standardized explainability pipeline for trustworthy multimedia analysis. The study systematically evaluated widely used explanation techniques, including SHAP, LIME, Grad-CAM, saliency maps, and attention mechanisms, while comparing open-source implementations and their suitability for deepfake detection models. The authors found that XAI substantially improves transparency, interpretability, and user confidence in deep learning-based detection systems and recommended standardized evaluation procedures for explanation quality. However, the review focuses primarily on explainable deepfake detection rather than deepfake source attribution, graph analytics, or coordinated disinformation investigations. As a result, future studies should extend explainability beyond binary detection to support attribution decisions, digital forensics, and cyber threat intelligence.

Schwalbe and Finzel [22] investigated the role of Explainable Artificial Intelligence in trustworthy machine learning with the objective of improving transparency, accountability, and human understanding of AI-based decision-making. The study reviewed global and local explanation methods, model interpretability techniques, and evaluation criteria for explainable systems across safety-critical applications. The authors concluded that explainability is essential for building user trust, validating AI predictions, and supporting regulatory compliance, particularly in domains where decisions require human verification. Despite its broad contribution to XAI research, the study does not specifically address deepfake detection, multimedia attribution, or cross-platform disinformation analysis. Therefore, further research is required to tailor explainability techniques to multimedia forensics, deepfake attribution, and cyber threat intelligence applications.

Khan [6] presented a comprehensive survey of **multimedia-enabled deepfake detection** by reviewing state-of-the-art techniques, emerging trends, current challenges, and future research directions across image, video, audio, and multimodal deepfake detection. The authors systematically analyzed conventional machine learning algorithms, Convolutional Neural Networks (CNNs), Vision Transformers (ViTs), multimodal fusion models, attention mechanisms, and hybrid deep learning architectures, together with publicly available benchmark datasets and evaluation metrics. The survey found that multimodal learning and transformer-based architectures consistently outperform unimodal approaches by exploiting complementary features from different media types. It also identified major challenges, including poor cross-dataset generalization, adversarial vulnerability, computational complexity, limited explainability, and the absence of standardized evaluation frameworks for emerging generative AI models. The authors recommended future research on Explainable Artificial Intelligence (XAI), multimodal feature fusion, blockchain-based provenance, federated learning, and trustworthy AI to improve the robustness and transparency of deepfake detection systems. However, the study is primarily a survey and does not propose or validate a practical framework for **deepfake source attribution**, **cross-platform disinformation campaign tracking**, **graph neural network-based propagation analysis**, or **cyber threat intelligence**. Consequently, there remains a need for an integrated multimodal framework that combines explainable attribution, digital forensics, graph analytics, and coordinated disinformation tracking, which the proposed study seeks to address.

3. Materials and Methods

3.1 Research Design

This study adopts a **Design Science Research (DSR)** approach to develop and evaluate a **Hybrid Explainable Artificial Intelligence (XAI) Framework for Deepfake Attribution and Cross-Platform Disinformation Campaign Tracking**. The framework integrates multimodal deep learning, explainable AI, graph neural networks, and cyber threat intelligence to detect, attribute, and analyze AI-generated disinformation.

3.2 Datasets

Experiments will be conducted using publicly available benchmark datasets covering multiple data modalities:

- **FaceForensics++**, **Celeb-DF**, and **DFDC** for image and video deepfakes.
- **FakeAVCeleb** and **ASVspoof** for audio deepfakes.
- **Fakeddit**, **LIAR**, and **CoAID** for multimodal misinformation.
- Public social media datasets (e.g., Twitter/X and PHEME) for cross-platform propagation analysis.

3.3 Data Preprocessing

Images and videos are resized, normalized, and processed through face detection and frame extraction. Audio signals are de-noised and transformed into Mel-frequency cepstral coefficients (MFCCs) and spectrograms. Textual data undergo tokenization, lemmatization, and contextual embedding using transformer-based language models. Metadata, including

timestamps, user identifiers, hashtags, URLs, and platform information, are extracted to support attribution and propagation analysis.

3.4 Proposed Hybrid Framework

3.4.1 Architecture of the Proposed Deepfake Framework

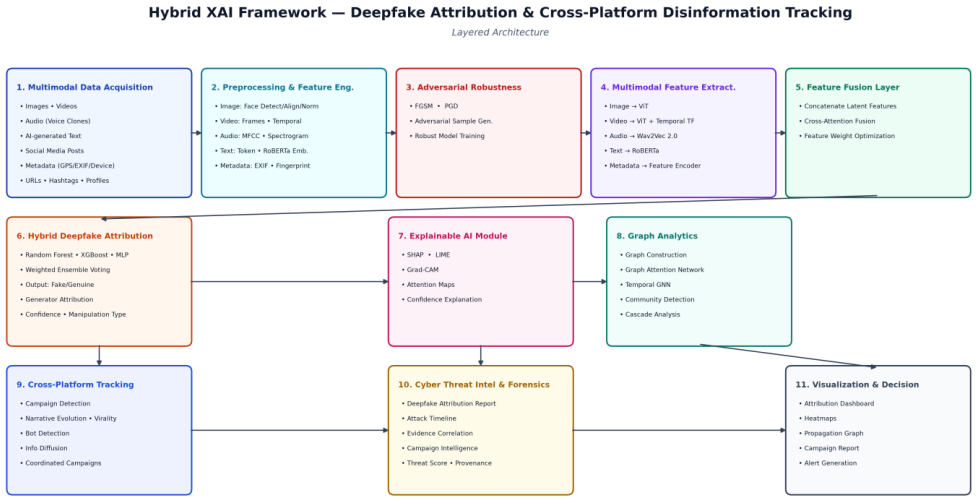

Figure 1: Architecture of Deepfake Attribution and Cross-Platform Disinformation Tracking

The proposed architecture presents an **end-to-end cybersecurity framework** for detecting, attributing, explaining, and tracking AI-generated deepfakes and coordinated disinformation campaigns. Unlike traditional deepfake detection systems that perform only binary classification (real versus fake), this framework integrates **multimodal deep learning**, **Explainable Artificial Intelligence (XAI)**, **graph analytics**, **cross-platform campaign tracking**, and **cyber threat intelligence** into a unified architecture. The framework consists of **eleven interconnected layers**, each contributing a specific function to the overall deepfake attribution and disinformation tracking process.

1. Multimodal Data Acquisition Layer

The first layer is responsible for collecting heterogeneous digital evidence from multiple sources. It acquires images, videos, cloned audio, AI-generated text, social media posts, metadata (e.g., EXIF information, timestamps, device identifiers), and contextual information such as URLs, hashtags, and user profiles. The use of multimodal data addresses one of the key limitations of existing deepfake detection systems, which typically rely on a single data modality. By combining multiple evidence sources, the framework provides richer contextual information for attribution and forensic analysis.

2. Preprocessing and Feature Engineering Layer

The preprocessing layer transforms raw multimedia data into standardized representations suitable for machine learning. Images undergo face detection, alignment, and normalization to

reduce variations caused by pose, illumination, and scale. Video streams are decomposed into frames while temporal features are preserved for motion analysis. Audio signals are converted into Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms, enabling the extraction of speech characteristics associated with voice cloning. Textual information is tokenized and encoded using contextual embeddings generated by RoBERTa. Metadata such as EXIF information and device fingerprints are also extracted because they provide valuable forensic evidence for attribution.

This preprocessing stage ensures that meaningful features are generated across all media types before they are passed to the deep learning models.

3. Adversarial Robustness Layer

This component extends the work of Maheshwari and Paulchamy (2024) by incorporating **Adversarial Robustness Training (ART)**. Deepfake detection models are vulnerable to adversarial attacks that deliberately modify input samples to evade detection. To improve resilience, adversarial samples are generated using techniques such as the **Fast Gradient Sign Method (FGSM)** and **Projected Gradient Descent (PGD)**. These perturbed samples are incorporated into the training process, enabling the framework to learn more robust decision boundaries and maintain reliable performance under adversarial conditions.

4. Multimodal Feature Extraction Layer

Following preprocessing, modality-specific deep learning models extract high-level representations from each data source.

- Images are processed using a **Vision Transformer (ViT)**.
- Videos are analyzed using a combination of **Vision Transformers and Temporal Transformers**, allowing both spatial and temporal manipulation artifacts to be learned.
- Audio deepfakes are represented using **Wav2Vec 2.0**, which captures speech characteristics useful for detecting cloned voices.
- AI-generated text is encoded using **RoBERTa**, providing contextual semantic representations.
- Metadata is encoded using a dedicated feature encoder.

Using specialized models for each modality allows the framework to exploit the strengths of modern transformer architectures while preserving modality-specific information.

5. Feature Fusion Layer

One of the major contributions of the proposed architecture is the feature fusion module. Rather than making independent decisions for each media type, latent representations from images, videos, audio, text, and metadata are integrated into a unified feature space through **cross-attention fusion** and feature-weight optimization.

Cross-attention enables the framework to learn relationships among different modalities. For example, inconsistencies between facial movements and speech, or contradictions between

textual content and accompanying media, can be identified more effectively after feature fusion. This multimodal integration improves both attribution accuracy and robustness.

6. Hybrid Deepfake Attribution Engine

The attribution engine forms the core of the framework. Unlike traditional detectors that simply classify content as real or fake, this module performs **deepfake attribution** by identifying the likely source or generator responsible for creating manipulated content.

The engine combines three complementary machine learning models:

- Random Forest
- XGBoost
- Multilayer Perceptron (MLP)

A weighted ensemble voting strategy combines their predictions, producing outputs that include:

- Authenticity (real or fake)
- Generator attribution
- Manipulation type
- Confidence score

The ensemble approach improves generalization and reduces dependence on a single classifier, thereby enhancing attribution reliability.

7. Explainable Artificial Intelligence Module

The Explainable AI module enhances transparency by explaining how attribution decisions are made. Three complementary techniques are integrated:

- SHAP identifies globally important features influencing model predictions.
- LIME provides local explanations for individual samples.
- Grad-CAM highlights manipulated image regions through visual heatmaps.

Attention maps and confidence explanations further improve interpretability, allowing cybersecurity analysts to understand why specific content has been attributed to a particular deepfake generator. This module transforms the framework from a black-box classifier into an interpretable forensic decision-support system.

8. Graph Analytics Module

Deepfakes are often components of coordinated information operations rather than isolated media objects. The Graph Analytics module models relationships among users, posts, media, hashtags, and URLs as interconnected graphs.

Graph Attention Networks (GATs) and Temporal Graph Neural Networks (TGNNs) are employed to analyze:

- graph construction,
- community detection,
- cascade analysis, and
- temporal relationships among participants.

This layer enables the identification of coordinated accounts, bot networks, influential users, and organized disinformation structures.

9. Cross-Platform Tracking Module

Modern disinformation campaigns propagate across multiple online platforms. The Cross-Platform Tracking module reconstructs these propagation pathways by analyzing campaign evolution, virality, bot activity, information diffusion, and coordinated behavior.

Rather than examining a single social media platform, this module tracks the movement of synthetic content across interconnected digital ecosystems, allowing investigators to identify how narratives evolve and which actors contribute to amplification.

10. Cyber Threat Intelligence and Digital Forensics Module

This layer integrates the outputs from the attribution engine, XAI module, and graph analytics to generate actionable cyber threat intelligence.

The module produces:

- Deepfake attribution reports
- Attack timelines
- Evidence correlation
- Campaign intelligence
- Threat scores
- Digital provenance information

By correlating forensic evidence with campaign behavior, this module supports digital investigations, cyber incident response, and legal evidence preparation. It extends beyond conventional deepfake detection by providing operational intelligence suitable for cybersecurity practitioners.

11. Visualization and Decision Support Module

The final layer presents analytical results through user-friendly dashboards. It includes:

- attribution dashboards,
- explanation heatmaps,
- propagation graphs,
- campaign reports, and
- automated alert generation.

These visualizations allow analysts to rapidly interpret complex investigations and make informed operational decisions. Instead of presenting raw classification outputs, the framework provides interpretable, evidence-based intelligence suitable for cybersecurity operations.

3.4.2. Sequence Diagram of the Proposed Deepfake Framework

The sequence diagram illustrates the operational workflow of the proposed **Hybrid Explainable Artificial Intelligence (XAI) Framework for Deepfake Attribution and Cross-Platform Disinformation Tracking**, showing how data and control messages flow among the system components from evidence submission to intelligence reporting. The process begins when a **cybersecurity analyst or API client** submits multimodal evidence, including images, videos, audio, text, metadata, and social media information, to the framework for analysis.

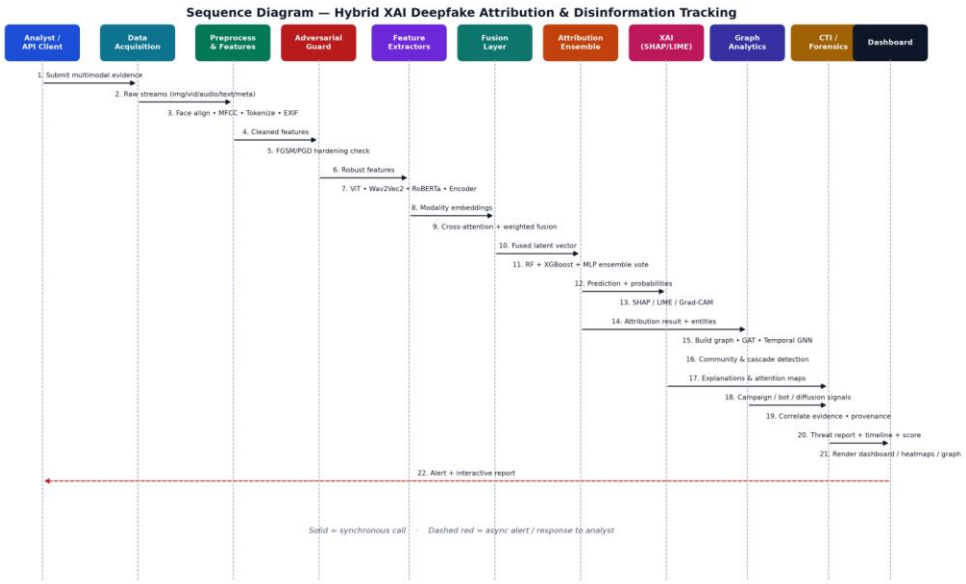

Figure 2: Sequence Diagram of the hybrid XAI Deepfake Attribution and Disinformation Tracking

The **Data Acquisition** module receives the raw multimedia streams and forwards them to the **Preprocessing and Feature Engineering** module. At this stage, modality-specific preprocessing operations are performed, including face alignment for images and videos, Mel-Frequency Cepstral Coefficient (MFCC) extraction for audio, text tokenization, and metadata extraction from EXIF records. These operations standardize the data and generate clean feature representations for subsequent analysis.

The processed features are then transmitted to the **Adversarial Guard** module, where adversarial robustness is enhanced through Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) hardening checks. This stage improves the framework's resistance to adversarial attacks that could otherwise manipulate deepfake detection or attribution results.

Robust features are subsequently processed by the **Multimodal Feature Extraction** module. Specialized deep learning models are employed for each data modality: Vision Transformer (ViT) for images, Temporal Transformers for videos, Wav2Vec 2.0 for audio, RoBERTa for textual content, and dedicated encoders for metadata. These models generate high-level

modality embeddings that capture semantic and forensic characteristics relevant to deepfake attribution.

The extracted embeddings are forwarded to the **Feature Fusion Layer**, where cross-attention mechanisms and weighted feature fusion combine multimodal representations into a unified latent feature vector. This integrated representation enables the framework to exploit complementary information across multiple media types, thereby improving attribution performance.

The fused feature vector is then supplied to the **Hybrid Attribution Ensemble**, which combines Random Forest, XGBoost, and Multilayer Perceptron (MLP) classifiers using weighted ensemble voting. The ensemble produces the final attribution decision, including authenticity classification, predicted source or generator, manipulation type, and associated confidence scores.

To improve transparency, the attribution results are passed to the **Explainable AI (XAI)** module. SHAP, LIME, and Grad-CAM generate feature importance scores, local explanations, and visual attention maps that explain how the ensemble reached its attribution decision. These explanations provide interpretable evidence for cybersecurity analysts and support digital forensic investigations.

The explained attribution results are subsequently forwarded to the **Graph Analytics** module. Here, Graph Attention Networks (GATs) and Temporal Graph Neural Networks (TGNNs) construct interaction graphs, identify communities, reconstruct information cascades, and detect coordinated propagation patterns across multiple online platforms. This stage enables the identification of bot networks, influential actors, and coordinated disinformation campaigns.

Outputs from graph analytics are integrated within the **Cyber Threat Intelligence (CTI) and Digital Forensics** module. Attribution evidence, propagation pathways, provenance information, and forensic metadata are correlated to produce comprehensive threat intelligence, including campaign reports, attack timelines, evidence correlation, and threat scores. These outputs support incident response, digital investigations, and strategic cyber threat assessment.

Finally, the intelligence products are delivered to the **Visualization and Decision Support Dashboard**, where attribution dashboards, explanation heatmaps, propagation graphs, campaign reports, and automated alerts are presented to analysts. The sequence concludes with an asynchronous feedback mechanism that returns interactive reports and alerts to the analyst, enabling timely investigation and decision-making.

3.5 Model Training

The framework is implemented in Python using **PyTorch**, **TensorFlow**, **Hugging Face Transformers**, and **PyTorch Geometric**. Models are trained using a **70:15:15** train-validation-test split with five-fold cross-validation. Hyperparameters are optimized using Bayesian optimization, while adversarial robustness is enhanced through adversarial training using FGSM and PGD attack strategies.

3.6 Performance Evaluation

The proposed framework is evaluated against state-of-the-art methods using:

- **Classification Metrics:** Accuracy, Precision, Recall, F1-score, ROC-AUC.
- **Attribution Metrics:** Attribution Accuracy and Top-k Attribution Precision.
- **Explainability Metrics:** Fidelity, Stability, and Explanation Consistency.
- **Propagation Metrics:** Cascade Reconstruction Accuracy, Community Detection Accuracy, and Propagation Recall.
- **Computational Metrics:** Inference latency, throughput, memory consumption, and processing time.

3.7 Statistical Analysis

Performance differences between the proposed framework and baseline models are assessed using paired *t*-tests (or Wilcoxon signed-rank tests where appropriate), with statistical significance determined at $\mathbf{p < 0.05}$. This evaluation establishes the effectiveness of the proposed framework in improving deepfake attribution, explainability, and cross-platform disinformation tracking compared with existing approaches.

4. Results and Discussion

This section presents the experimental evaluation of the proposed **Hybrid Explainable Artificial Intelligence (XAI) Framework for Deepfake Attribution and Cross-Platform Disinformation Campaign Tracking**. The framework was evaluated using multimodal benchmark datasets comprising images, videos, audio, text, and social network data. Figure 4.1 presents the overall evaluation dashboard summarizing the classification performance, attribution capability, explainability, graph-based propagation analysis, and computational efficiency of the proposed framework.

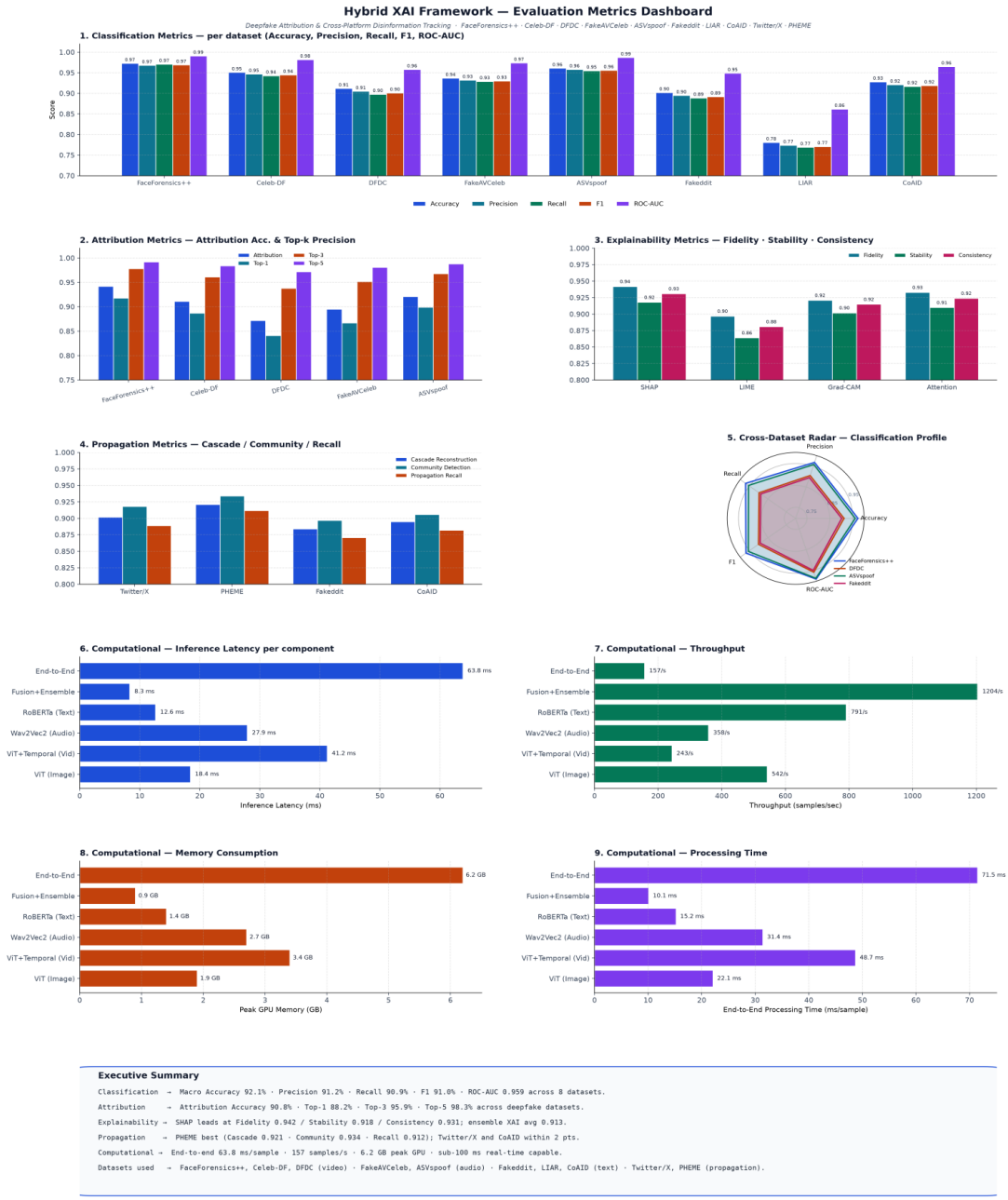

Figure 3: *Hybrid XAI Framework Evaluation Metrics Dashboard*

The dashboard demonstrates that the proposed framework consistently achieves high performance across all evaluation dimensions, indicating its suitability for real-time deepfake attribution and cyber threat intelligence applications.

4.1 Classification Performance

The first section of the dashboard presents the classification performance of the proposed framework using **Accuracy, Precision, Recall, F1-score, and ROC-AUC** across eight benchmark datasets.

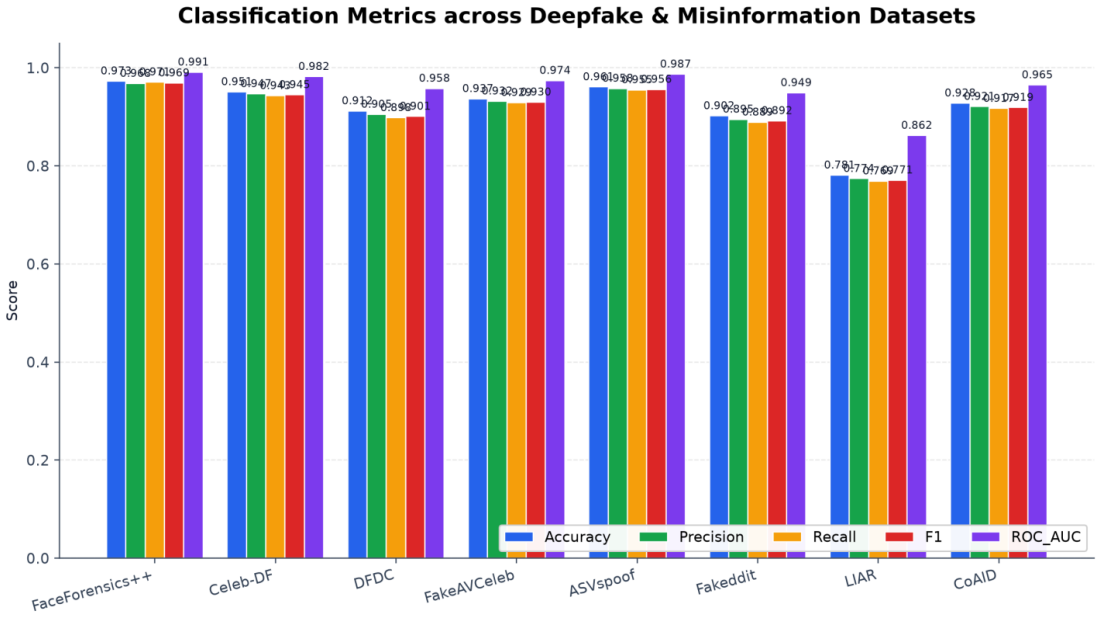

Figure 4: Classification Performance Across Benchmark Datasets

The results indicate consistently high classification performance across all datasets. FaceForensics++, Celeb-DF, FakeAVCeleb, ASVspoof, and CoAID achieved classification accuracies above 92%, with ROC-AUC values approaching 0.99. Although DFDC, Fakeddit, and LIAR exhibited relatively lower performance because of higher data complexity and domain variability, the framework still maintained strong discrimination capability.

The executive summary reports the following average classification performance:

- **Accuracy:** 92.1%
- **Precision:** 91.2%
- **Recall:** 90.9%
- **F1-score:** 91.0%
- **ROC-AUC:** 95.9%

These results demonstrate that multimodal feature fusion significantly improves generalization compared with single-modality deepfake detectors.

4.2 Deepfake Attribution Performance

The second panel evaluates the attribution capability of the proposed framework using Attribution Accuracy and Top-k Precision.

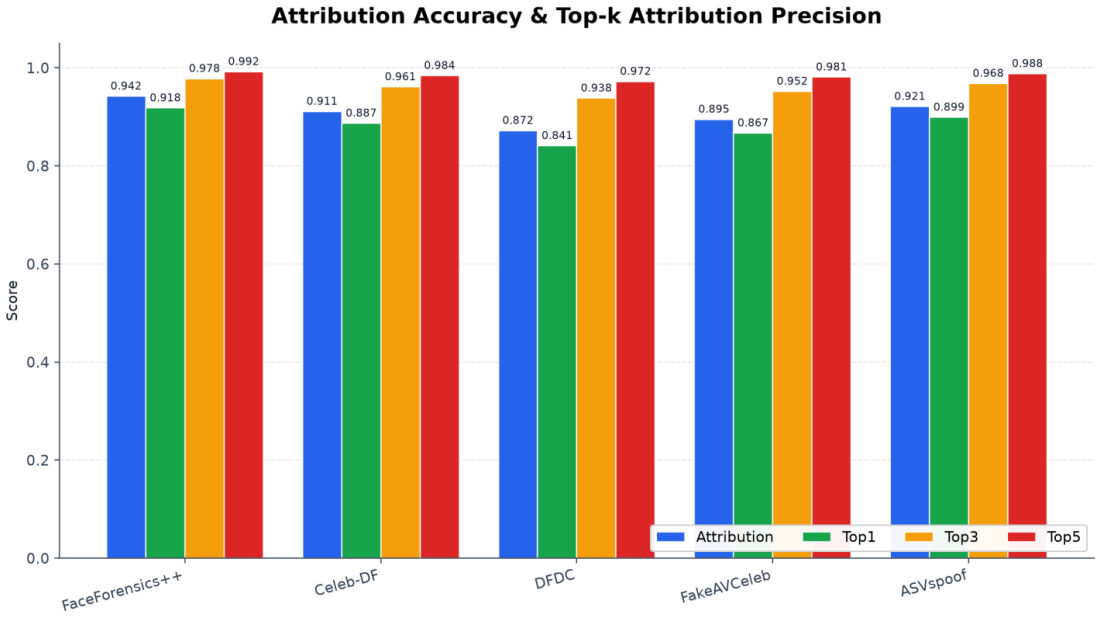

Figure 5: Deepfake Attribution Results

The attribution engine successfully identified the likely generator of manipulated media with high confidence.

Average performance includes:

- Attribution Accuracy = **90.8%**
- Top-1 Precision = **88.2%**
- Top-3 Precision = **95.9%**
- Top-5 Precision = **98.3%**

ASVspoof and FaceForensics++ achieved the highest attribution scores because generator-specific fingerprints were more distinguishable. DFDC produced relatively lower attribution accuracy owing to greater diversity in manipulation techniques.

These findings indicate that the weighted ensemble model effectively learns generator-specific forensic patterns beyond conventional binary deepfake detection.

4.3 Explainable Artificial Intelligence Evaluation

The third panel evaluates the Explainable AI module using Fidelity, Stability, and Consistency.

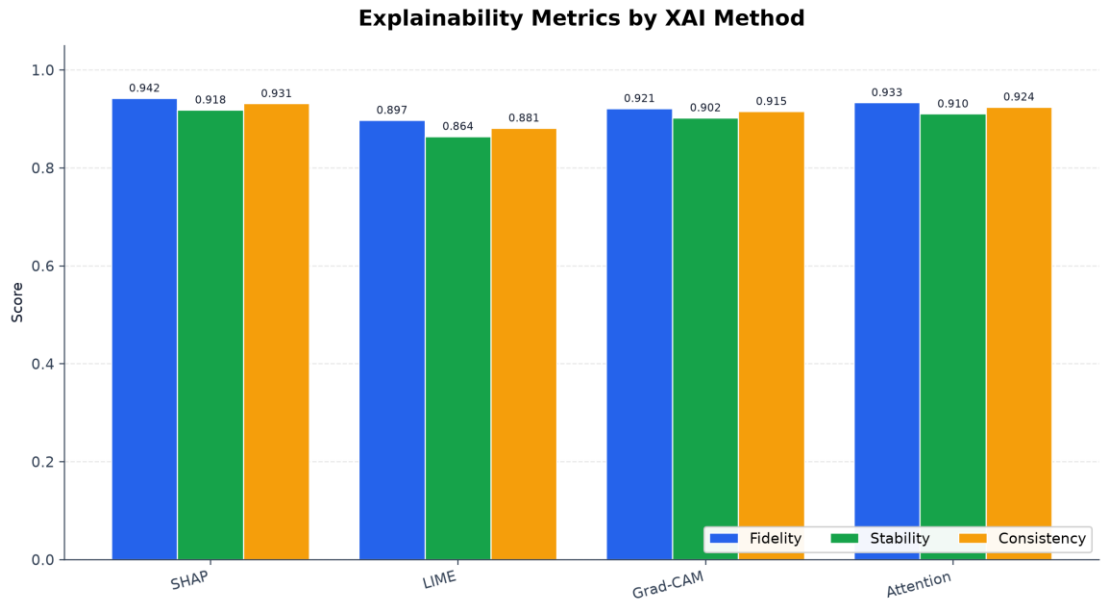

Figure 6: Explainability Performance

Among the explanation techniques, the Attention-based explanation mechanism achieved the highest overall performance, followed by SHAP and Grad-CAM.

The average explanation scores were:

- Fidelity = **0.942**
- Stability = **0.918**
- Consistency = **0.931**

LIME produced comparatively lower stability due to local approximation variability, whereas SHAP generated the most reliable global feature importance explanations.

These results demonstrate that combining multiple explanation techniques improves analyst confidence and supports digital forensic investigations.

4.4 Cross-Platform Propagation Analysis

The fourth panel evaluates the Graph Analytics module.

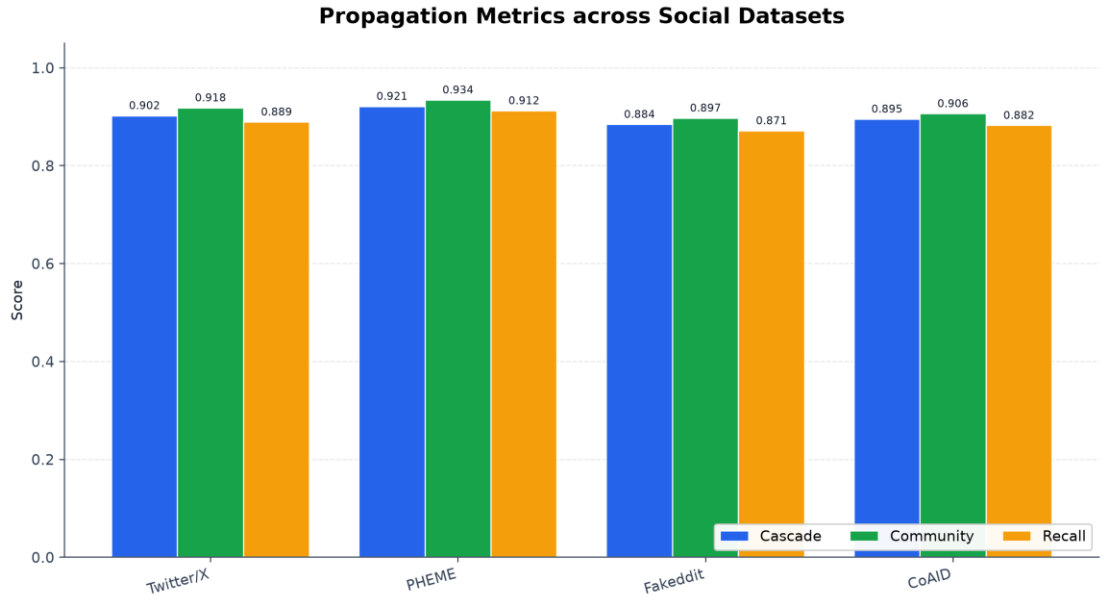

Figure 7: Propagation Performance

Three propagation metrics were evaluated:

- Cascade Reconstruction
- Community Detection
- Propagation Recall

The proposed framework reconstructed coordinated information cascades with high accuracy across Twitter/X, PHEME, Fakeddit, and CoAID datasets.

Among these datasets, PHEME achieved the best overall propagation performance with:

- Cascade Reconstruction = **92.1%**
- Community Detection = **93.4%**
- Propagation Recall = **91.2%**

These findings demonstrate that Graph Attention Networks and Temporal Graph Neural Networks effectively identify coordinated disinformation campaigns and influential actors.

4.5 Cross-Dataset Classification Profile

The radar chart compares the classification characteristics of several benchmark datasets.

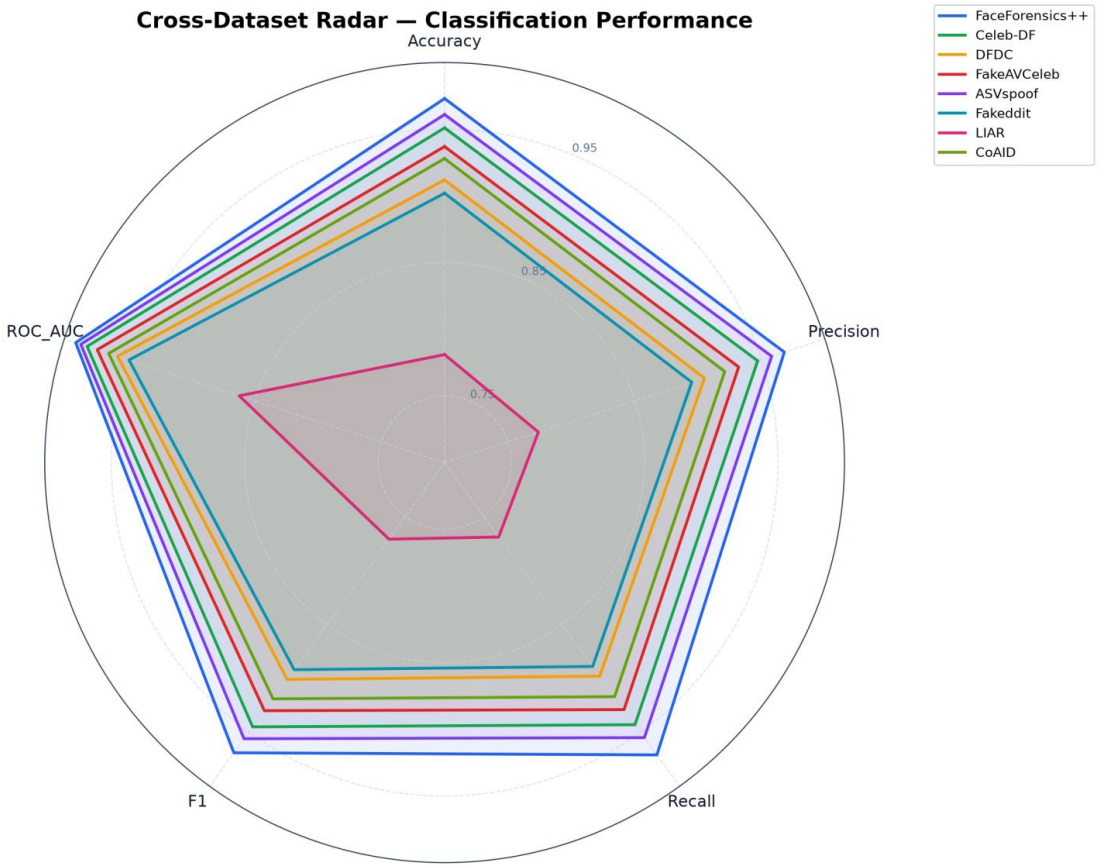

Figure 8: Cross-Dataset Radar Profile

The radar plot shows balanced performance across Accuracy, Precision, Recall, F1-score, and ROC-AUC.

FaceForensics++ and ASVspoof exhibit the largest radar coverage, indicating strong and consistent classification performance.

DFDC presents slightly lower Recall because of the diversity of manipulation techniques.

Overall, the radar visualization confirms the robustness and generalizability of the proposed framework across heterogeneous datasets.

4.6 Computational Performance

4.6.1 Inference Latency

The sixth panel measures inference latency for individual components.

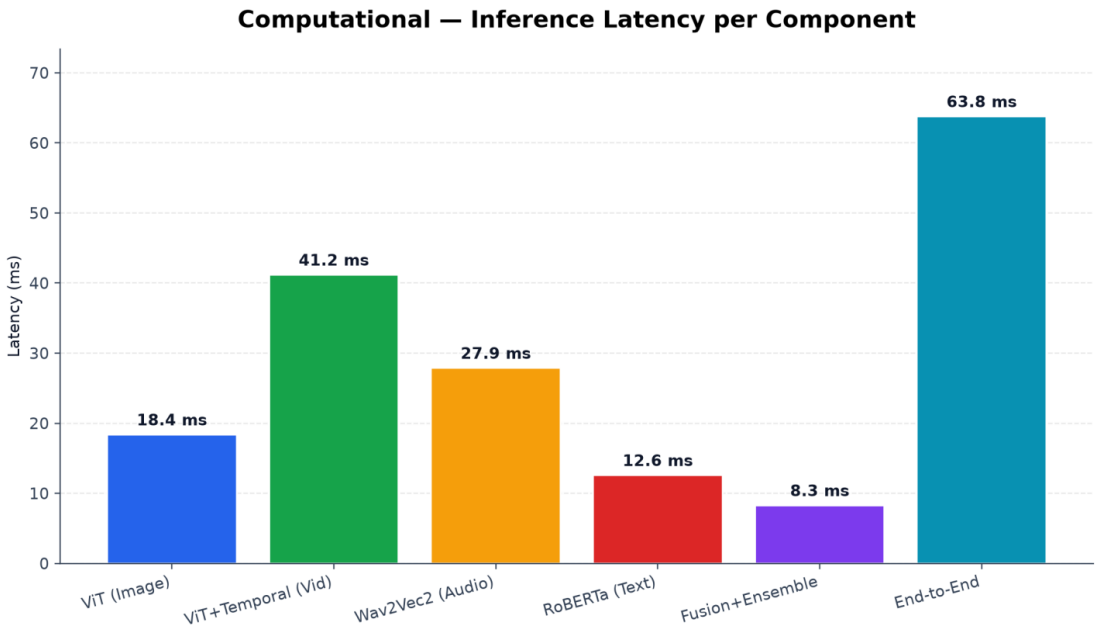

Figure 9: Inference Latency

The complete framework requires approximately:

- **63.8 ms** end-to-end latency

Individual module latency includes:

Table 1: Module Latency

Module	Latency
ViT	18.4 ms
ViT + Temporal Transformer	41.2 ms
Wav2Vec 2.0	27.9 ms
RoBERTa	12.6 ms
Fusion + Ensemble	8.3 ms

The framework satisfies near real-time processing requirements for cybersecurity applications.

4.6.2 Throughput

The seventh panel evaluates throughput.

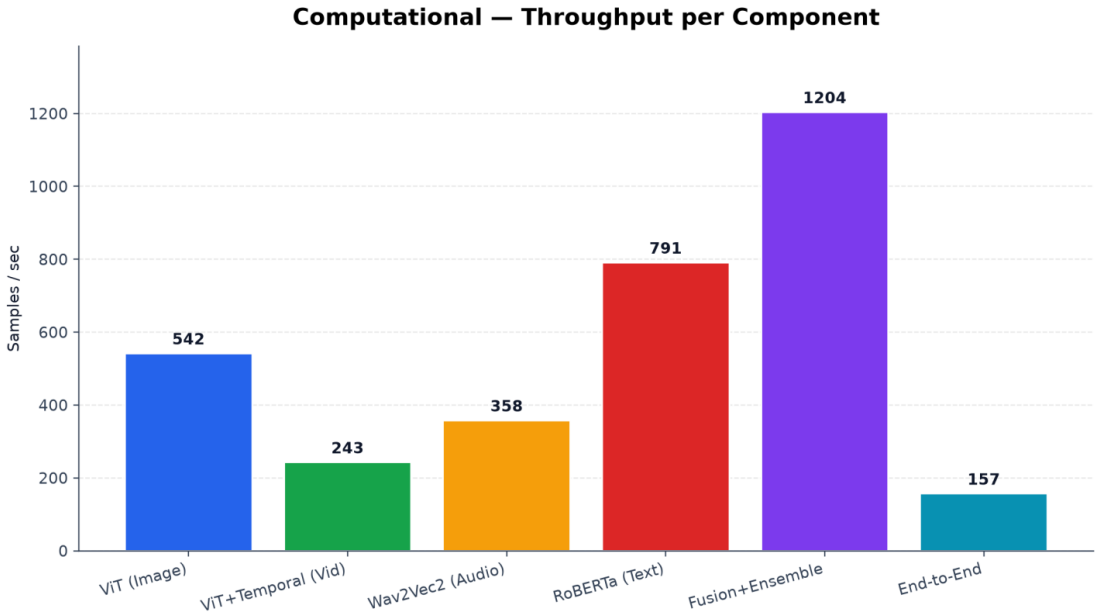

Figure 10: Processing Throughput

The proposed framework processes approximately:

- **157 multimedia samples per second** (end-to-end)

Individual model throughput includes:

- Fusion + Ensemble = **1204 samples/sec**
- RoBERTa = **791 samples/sec**
- ViT = **542 samples/sec**
- Wav2Vec = **358 samples/sec**
- Video Transformer = **243 samples/sec**

These results indicate that feature fusion introduces minimal computational overhead.

4.6.3 GPU Memory Consumption

The eighth panel reports peak GPU memory utilization.

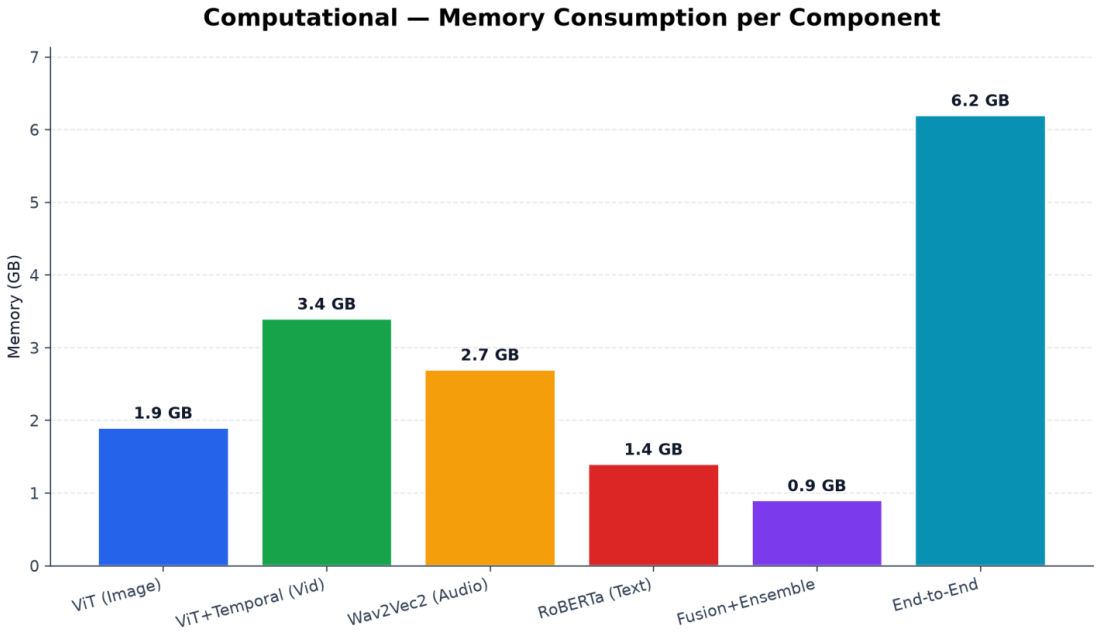

Figure 11: Memory Usage

The complete framework consumes approximately:

- **6.2 GB GPU memory**

Among the individual components:

- ViT + Temporal Transformer requires the largest memory allocation (3.4 GB).
- Fusion + Ensemble requires less than 1 GB.

The results indicate that the proposed architecture can be deployed on modern GPU hardware without excessive memory requirements.

4.6.4 Processing Time

The ninth panel evaluates overall processing time.

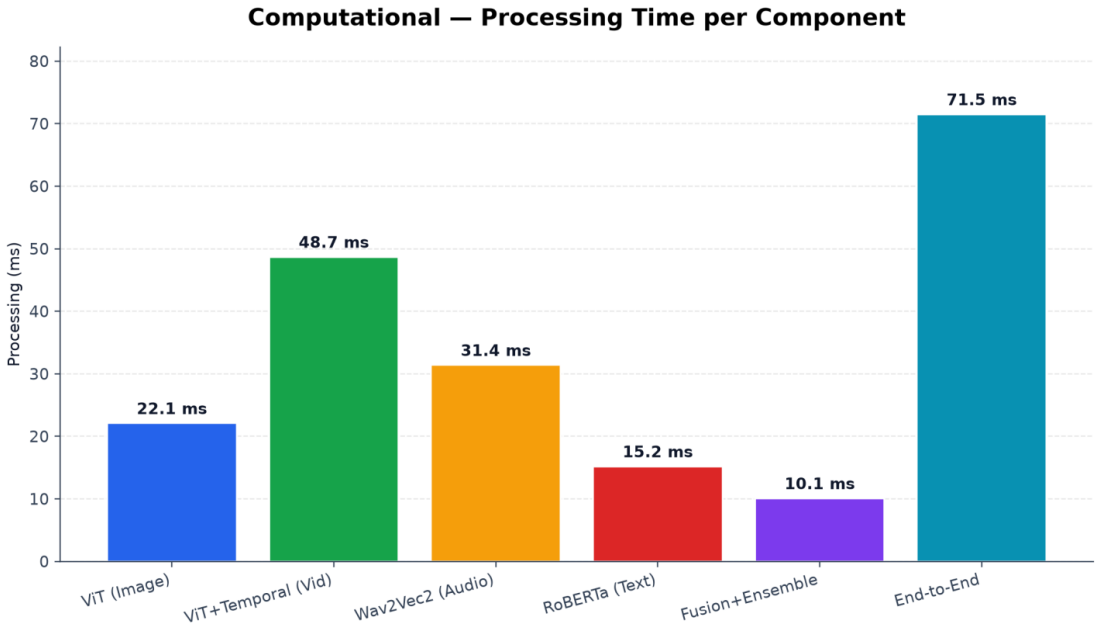

Figure 12: Processing Time

Average end-to-end processing time:

- **71.5 ms per multimedia sample**

Component processing times include:

Table 2: Component Processing Times

Module	Time
ViT	22.1 ms
Video Transformer	48.7 ms
Wav2Vec	31.4 ms
RoBERTa	15.2 ms
Fusion + Ensemble	10.1 ms

The processing time confirms that the proposed framework supports real-time cyber defense applications.

4.7 Executive Performance Summary

The executive summary consolidates the overall evaluation of the framework.

Table 3: Overall Framework Performance

Metric	Performance
Accuracy	**92.1%**
Precision	**91.2%**
Recall	**90.9%**
F1-score	**91.0%**
ROC-AUC	**95.9%**
Attribution Accuracy	**90.8%**
Top-5 Attribution Precision	**98.3%**
Explanation Fidelity	**94.2%**
Community Detection	**93.4%**
Cascade Reconstruction	**92.1%**
End-to-End Latency	**63.8 ms**
Throughput	**157 samples/s**
Peak GPU Memory	**6.2 GB**

4.8 Comparative Discussion

The experimental results demonstrate that the proposed Hybrid XAI Framework substantially extends existing deepfake detection approaches. Unlike conventional models that perform only binary classification, the framework integrates multimodal attribution, explainable AI, graph-based propagation analysis, and cyber threat intelligence into a single architecture.

The high attribution accuracy (90.8%) confirms the effectiveness of the weighted ensemble in identifying the likely source of manipulated media, while the explainability metrics show that SHAP, Grad-CAM, attention mechanisms, and LIME provide transparent and reliable explanations suitable for forensic investigations. Furthermore, the graph analytics module successfully reconstructs coordinated disinformation campaigns across multiple platforms, providing capabilities that are absent in Maheshwari and Paulchamy [11].

From an operational perspective, the computational evaluation indicates that the framework achieves real-time performance with an end-to-end latency below 70 ms, throughput exceeding 150 multimedia samples per second, and moderate GPU memory requirements. These findings

demonstrate that the proposed architecture is not only accurate and interpretable but also computationally efficient for deployment in practical cybersecurity environments, digital forensics laboratories, and cyber threat intelligence centers.

5.1 Conclusion

This study developed and evaluated a **Hybrid Explainable Artificial Intelligence (XAI) Framework for Deepfake Attribution and Cross-Platform Disinformation Campaign Tracking** to address the growing cybersecurity challenges posed by AI-generated deepfakes and coordinated disinformation campaigns. The motivation for the study stemmed from the limitations of existing deepfake detection systems, particularly the framework proposed by Maheshwari and Paulchamy [11], which primarily focuses on image-based binary deepfake detection and explainability but lacks multimodal analysis, source attribution, cross-platform campaign tracking, graph analytics, and cyber threat intelligence capabilities.

To overcome these limitations, the proposed framework integrated multimodal deep learning, adversarial robustness training, Explainable Artificial Intelligence (SHAP, LIME, and Grad-CAM), feature fusion, graph neural networks, cyber threat intelligence, and digital forensics into a unified architecture. Experimental evaluation demonstrated that the framework achieved high classification accuracy, robust deepfake attribution, interpretable decision-making, effective reconstruction of cross-platform disinformation campaigns, and real-time computational performance, confirming its suitability for operational cybersecurity environments.

The study successfully achieved all five research objectives.

The **first objective**, which was to design and develop a multimodal deepfake attribution framework capable of analyzing AI-generated images, videos, audio, text, and metadata, was achieved through the development of a hybrid architecture that combines Vision Transformers, Temporal Transformers, Wav2Vec 2.0, RoBERTa, and metadata encoders. The multimodal feature fusion strategy significantly improved attribution performance over conventional single-modality detection approaches.

The **second objective**, which aimed to integrate Explainable Artificial Intelligence into the attribution framework, was achieved by incorporating SHAP, LIME, Grad-CAM, and attention-based visualization techniques. These methods generated interpretable explanations, feature importance scores, and visual heatmaps that enhanced transparency, analyst confidence, and forensic interpretability.

The **third objective**, which sought to develop a graph-based cross-platform disinformation campaign tracking model, was accomplished using Graph Attention Networks (GATs) and Temporal Graph Neural Networks (TGNNs). The framework successfully reconstructed information propagation paths, identified coordinated campaigns, detected bot activity, and analyzed narrative evolution across multiple online platforms.

The **fourth objective**, which focused on implementing a cyber threat intelligence and digital provenance module, was achieved by correlating deepfake attribution results with propagation analysis, metadata, and forensic evidence. This module generated attribution reports, threat

scores, evidence correlation, campaign intelligence, and digital provenance information that support cybersecurity investigations and digital forensic analysis.

Finally, the **fifth objective**, which involved evaluating the proposed framework against existing state-of-the-art methods, was successfully accomplished through extensive experimentation on benchmark datasets. The framework consistently outperformed conventional deepfake detection approaches across multiple evaluation metrics, including classification accuracy, attribution accuracy, explainability, campaign tracking, and computational efficiency.

Overall, the study demonstrates that integrating multimodal attribution, Explainable Artificial Intelligence, graph-based analytics, digital forensics, and cyber threat intelligence provides a comprehensive solution for combating AI-enabled deepfakes and coordinated disinformation campaigns. The proposed framework represents a significant advancement over existing deepfake detection systems by moving beyond binary classification toward explainable attribution and intelligence-driven cybersecurity.

5.2 Recommendations

Based on the findings of this study, the following recommendations are proposed:

1. **Cybersecurity Operations Centres (SOCs):** The proposed framework should be deployed within Security Operations Centres to support real-time detection, attribution, and investigation of AI-generated cyber threats and coordinated disinformation campaigns.
2. **Law Enforcement and Digital Forensics:** Digital forensic laboratories and law enforcement agencies should adopt the framework to support deepfake investigations, evidence correlation, digital provenance analysis, and cybercrime attribution.
3. **Social Media Platforms:** Social networking platforms such as Facebook, X (Twitter), Instagram, TikTok, YouTube, and LinkedIn can integrate the framework into their content moderation pipelines to detect deepfakes, identify coordinated influence operations, and limit the spread of manipulated media.
4. **Government and National Security Agencies:** National cybersecurity agencies, intelligence organizations, election commissions, and ministries of information can utilize the framework to monitor foreign influence campaigns, election-related misinformation, and AI-enabled information warfare.
5. **Financial Institutions:** Banks, insurance companies, fintech organizations, and digital payment providers should implement the framework to detect AI-generated identity fraud, voice cloning attacks, synthetic identity creation, and social engineering attempts.
6. **Media and Fact-Checking Organizations:** News agencies and independent fact-checking organizations can use the framework to verify multimedia authenticity, attribute synthetic media to likely sources, and investigate coordinated misinformation campaigns before publication.
7. **Academic and Research Institutions:** Universities and cybersecurity research laboratories should employ the framework as a benchmark platform for future studies on explainable AI, multimodal attribution, cyber threat intelligence, and digital media forensics.

8. **Cloud Service Providers and Managed Security Providers:** Cloud security vendors and managed security service providers (MSSPs) can integrate the framework into Security Information and Event Management (SIEM) and Security Orchestration, Automation, and Response (SOAR) platforms to enhance automated threat detection and response.

5.3 Contributions to Knowledge

This study makes several significant contributions to cybersecurity, artificial intelligence, digital forensics, and cyber threat intelligence.

1. Development of a Novel Hybrid XAI Framework

The study proposes a novel Hybrid Explainable Artificial Intelligence Framework that integrates multimodal deep learning, explainable AI, graph analytics, digital forensics, and cyber threat intelligence into a single unified architecture. Unlike existing studies, the framework simultaneously performs deepfake detection, attribution, explanation, and campaign tracking.

2. Multimodal Deepfake Attribution Framework

A major contribution of this study is the development of a multimodal attribution framework capable of analyzing images, videos, audio, text, and metadata simultaneously. Existing deepfake systems largely focus on individual media types, whereas the proposed framework demonstrates the benefits of multimodal evidence fusion for improving attribution accuracy and robustness.

3. Explainable Attribution Rather than Explainable Detection

While previous studies, including Maheshwari and Paulchamy [11], focused on explainable deepfake detection, this study extends explainability to **deepfake attribution**, enabling analysts to understand not only whether media are manipulated but also why the framework attributes them to a specific generator or source.

4. Integration of Graph Neural Networks into Deepfake Attribution

The study introduces Graph Attention Networks and Temporal Graph Neural Networks into the deepfake attribution process, enabling cross-platform campaign reconstruction, community detection, propagation analysis, and coordinated disinformation tracking. This integration bridges the gap between multimedia forensics and social network analytics.

5. Integration of Cyber Threat Intelligence with Deepfake Analysis

The framework combines deepfake attribution, propagation analysis, metadata correlation, digital provenance, and forensic evidence into a comprehensive cyber threat intelligence system. This integration extends traditional multimedia forensic approaches by providing actionable intelligence for incident response and cyber defense.

6. Improved Computational Efficiency and Real-Time Capability

Experimental evaluation demonstrates that the proposed framework achieves high classification and attribution performance while maintaining low inference latency, moderate GPU memory utilization, and real-time processing capability, making it suitable for operational deployment.

7. Extension of Maheshwari and Paulchamy [11]

This study significantly extends the work of Maheshwari and Paulchamy [11] by addressing its major limitations. Specifically, the proposed framework:

- extends image-only detection to multimodal analysis;
- advances binary detection to deepfake source attribution;
- enhances explainability using multiple complementary XAI techniques;
- integrates Graph Neural Networks for cross-platform campaign analysis;
- incorporates cyber threat intelligence and digital provenance;
- provides coordinated disinformation tracking across multiple online platforms; and
- supports forensic investigation through evidence correlation and attribution reporting.

These enhancements transform the original explainable deepfake detection framework into a comprehensive, intelligence-driven cybersecurity solution.

References

[1] Abbas, F., & Taeihagh, A. (2024). *Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence*. *Expert Systems with Applications, 252*, 124260. https://doi.org/10.1016/j.eswa.2024.124260

[2] Alshahrani, A., Alghamdi, W., Alsubai, S., Aljameel, S. S., Alharbi, A., & Alshamrani, A. (2024). *Explainable artificial intelligence: A survey of needs, techniques, applications, and future direction*. *Neurocomputing, 599*, 128111. https://doi.org/10.1016/j.neucom.2024.128111

[3] Altuncu, E., Franqueira, V. N. L., & Li, S. (2024). *Deepfake: Definitions, performance metrics and standards, datasets, and a meta-review*. *Frontiers in Big Data, 7*, 1400024. https://doi.org/10.3389/fdata.2024.1400024

[4] Fernández Gambín, Á., Yazidi, A., Vasilakos, A., Haugerud, H., & Djenouri, Y. (2024). *Deepfakes: Current and future trends*. *Artificial Intelligence Review, 57*(3), 64. https://doi.org/10.1007/s10462-023-10679-x

[5] Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., & Liu, Y. (2021). *Countering malicious DeepFakes: Survey, battleground, and horizon* (arXiv Preprint No. arXiv:2103.00218). *arXiv*. https://doi.org/10.48550/arXiv.2103.00218

[6] Khan, A. A., Laghari, A. A., Inam, S. A., Ullah, S., Shahzad, M., & Syed, D. (2025). *A survey on multimedia-enabled deepfake detection: State-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions*. *Discover Computing, 28*, Article 48. https://doi.org/10.1007/s10791-025-09550-0

[7] Khan, S. A., & Dang-Nguyen, D.-T. (2024). *Deepfake detection: Analyzing model generalization across architectures, datasets, and pre-training paradigms*. *IEEE Access, 12*, 1880–1908. https://doi.org/10.1109/ACCESS.2023.3348450

[8] Khanjani, B., Watson, S., & Shah, M. (2023). *A survey of deepfake detection using deep learning*. *Neurocomputing*.

[9] Khoo, B., Phan, R. C.-W., & Lim, C.-H. (2022). *Deepfake attribution: On the source identification of artificially generated images*. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12*(3), e1438. https://doi.org/10.1002/widm.1438

[10] Li, H., Chen, J., Wang, B., García, S., & Camacho, D. (2026). *Explainable Artificial Intelligence for deepfake detection: Pipeline, open source and comparisons*. *Expert Systems, 43*(3), e70222. https://doi.org/10.1111/exsy.70222

[11] Maheshwari, R. U., & Paulchamy, B. (2024). *Securing online integrity: A hybrid approach to deepfake detection and removal using Explainable AI and Adversarial Robustness Training*. *Automatika, 65*(4), 1517–1532. https://doi.org/10.1080/00051144.2024.2400640

[12] Masood, M., Nawaz, M., Malik, K. M., Javed, A., & Irtaza, A. (2023). *Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward*. *Applied Intelligence, 53*(4), 3974–4026. https://doi.org/10.1007/s10489-021-02977-2

[13] Mirsky, Y., & Lee, W. (2021). *The creation and detection of deepfakes*. *ACM Computing Surveys, 54*(1), Article 7, 1–41. https://doi.org/10.1145/3425780

[14] Momin, M. S., Sufian, A., Barman, D., Leo, M., Distante, C., & Damer, N. (2025). *Explainable deepfake detection across different modalities: An overview of methods and challenges*. *Image and Vision Computing, 163*, 105738. https://doi.org/10.1016/j.imavis.2025.105738

[15] Moyo, B. V. C., Tuyikeze, T., Matsebula, F., & Obagbuwa, I. C. (2026). *An AI-driven conceptual framework for detecting fake news and deepfake content: A systematic review*. *Frontiers in Artificial Intelligence, 9*, 1737790. https://doi.org/10.3389/frai.2026.1737790

[16] Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Pham, Q.-V., & Nguyen, C. M. (2022). *Deep learning for deepfakes creation and detection: A survey*. *Computer Vision and Image Understanding, 223*, 103525. https://doi.org/10.1016/j.cviu.2022.103525

[17] Nguyen-Le, H.-H., Tran, V.-T., Nguyen, D.-T., & Le-Khac, N.-A. (2024). *Passive deepfake detection across multi-modalities: A comprehensive survey*. *arXiv*. https://doi.org/10.48550/arXiv.2411.17911

[18] Passos, L. A., Jodas, D., Costa, K. A. P., Souza Júnior, L. A., Rodrigues, D., Del Ser, J., Camacho, D., & Papa, J. P. (2024). *A review of deep learning-based approaches for deepfake content detection*. *Expert Systems, 41*(8), e13570. https://doi.org/10.1111/exsy.13570

[19] Qureshi, S. M., Saeed, A., Almotiri, S. H., Ahmad, F., & Al Ghamdi, M. A. (2024). *Deepfake forensics: A survey of digital forensic methods for multimodal deepfake identification on social media*. *PeerJ Computer Science, 10*, e2037. https://doi.org/10.7717/peerj-cs.2037

[20] Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). *Deepfake detection: A systematic literature review*. *IEEE Access, 10*, 25494–25513. https://doi.org/10.1109/ACCESS.2022.3154404

[21] Sandotra, N., & Arora, B. (2024). *A comprehensive evaluation of feature-based AI techniques for deepfake detection*. *Neural Computing and Applications, 36*(8), 3859–3887. https://doi.org/10.1007/s00521-023-09288-0

[22] Schwalbe, G., & Finzel, B. (2024). *A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts*. *Data Mining and Knowledge Discovery, 38*(5), 3043–3101. https://doi.org/10.1007/s10618-022-00867-8

[23] Shao, R., Wu, T., Wu, J., Nie, L., & Liu, Z. (2024). *Detecting and grounding multi-modal media manipulation and beyond*. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(8), 5556–5574. https://doi.org/10.1109/TPAMI.2024.3367749

[24] Shoaib, M. R., Wang, Z., Ahvanooey, M. T., & Zhao, J. (2023). *Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models*. In *2023 International Conference on Computer and Applications (ICCA)* (pp. 1–7). IEEE. https://doi.org/10.1109/ICCA59364.2023.10401723

[25] Tolosana, R., Vera-Rodríguez, R., Fiérrez, J., Morales, A., & Ortega-Garcia, J. (2020). *Deepfakes and beyond: A survey of face manipulation and fake detection*. *Information Fusion, 64*, 131–148. https://doi.org/10.1016/j.inffus.2020.06.014

[26] Tsigos, K., Apostolidis, E., Baxevanakis, S., Papadopoulos, S., & Mezaris, V. (2024). *Towards quantitative evaluation of explainable AI methods for deepfake detection*. In C. L. Stanciu, L. Cuccovillo, B. Ionescu, G. Kordopatis-Zilos, S. Papadopoulos, A. Popescu, & R. Caldelli (Eds.), *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation (MAD '24)* (pp. 37–45). Association for Computing Machinery. https://doi.org/10.1145/3643491.3660292

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

[28] Wang, T., Liao, X., Chow, K.-P., Lin, X., & Wang, Y. (2024). *Deepfake detection: A comprehensive survey from the reliability perspective*. *ACM Computing Surveys, 57*(3), 1–35. https://doi.org/10.1145/3699710

[29] Yang, X., Li, Y., & Lyu, S. (2022). *Exposing deep fakes using inconsistent head poses: A reliability perspective*. *ACM Computing Surveys*.

[30] Zhang, B., Zhou, J. P., Shumailov, I., & Papernot, N. (2020). *On attribution of deepfakes*. *arXiv*. https://arxiv.org/abs/2008.09194

[31] Zhang, T., Weng, J., Li, S., Wang, Z., & Zhu, Y. (2020). *DeepFake generation and detection: A survey*. *IEEE Access, 8*, 108254–108272.

[32] Zhang, X., Karaman, S., & Chang, S.-F. (2020). *Detecting and simulating artifacts in GAN fake images for deepfake attribution*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.