International Journal of Emerging Engineering and Technology (IJEET)
ISSN (e): 2958-3764
Volume: 5, Number: 1, Pages: 46 - 54, Year: 2026

A Critical Analysis of Deepfake Detection Software: Emerging Challenges and Quality Assurance

Imran Nawaz Rana* and Irshad Ahmed Sumra

School of Science and Technology, University of Management and Technology, Lahore, 54770, Pakistan
*Corresponding author: Imran Nawaz Rana (Email: s2025108007@umt.edu.pk)

Received: 12/02/2026, Revised: 22/05/2026, Accepted: 15/06/2026

***Abstract*— Deepfakes have proven their identity in the field of digital forensics. The article explains how Deepfakes create highly realistic but synthetic videos, images, and audio clips, which create doubt about whether they are real or fake. This article identifies some important tools and technologies. In every field of life, such as entertainment, education, and content writing, deepfakes help impose a sense of authenticity. Digital forensics investigation agencies, law enforcement officials, and offices are also concerned about day-to-day social engineering attacks, identity theft, and misinformation. Software Quality Assurance (SQA) and testing features of the deepfake forensic systems, including accuracy, reliability, security, robustness, scalability, and explainability. The discussion covers testing challenges, including dataset validation, performance evaluation, adversarial testing, and real-world scenario testing, to enhance the reliability of AI-driven forensic solutions.**

***Index Terms*—Digital Forensics, Deepfake, Artificial Intelligence, Digital Evidence, Cybercrime, Cyber Security, Multimedia Forensics, Software Quality Assurance.**

I. Introduction

Deepfakes are ready to rule the world of digital media by creating highly realistic synthetic media that will compromise the authenticity, integrity, and trustworthiness of digital evidence. The study aims to improve explainable AI techniques and methods of forensic reliability and newly introduced threat detection in AI learning [1-4]. The Custom CNN model achieves better results in deepfake audio tests, even with larger datasets, than other models such as VGG-16, ResNet, and FG-RCNN, using MFCC, Mel-spectrogram, Spectrogram, and Chronogram. Where VGG-16 performs very well for the detection of forensics [5-8]. This study empowers cybercriminals and strengthens techniques for fake news, revenge, pornography, and multimedia manipulation. And weakens trust in digital data evidence [9-13]. A detailed study from 2021 to 2024 shows the importance and trust in digital forensic evidence. The authors found that CNNs and GANs are critical for deepfake generation and detection, and stronger detection frameworks, interdisciplinary collaboration, and ethical guidelines are necessary to maintain forensic integrity [14-19]. Section II identifies the importance and challenges of deepfake and types of digital forensics. Section III classifies deepfake detection techniques. Section IV: The Influence of Deepfakes and AI-Generated Evidence on Digital Forensics. Section V discusses the importance of digital forensics in the presence of deepfakes. Legal and ethical issues, deepfake datasets, and benchmarking metrics are essential in deepfake detection technologies. Section VI examines the comparative analysis of existing deepfake detection methods. Section VII: Detection of deepfake tampering is a challenge for digital forensics. Collaboration of researchers, industry participants, and regulatory bodies sets the future direction. Section under SQA. Section VIII: Comparative Analysis of Existing Deepfake Detection Methods. Section IX: Facing the challenges during deepfake detection without considering SQA. Section X provides the results of this survey paper.

II. Elementary thought of Deefake and Types

The majority of deepfake systems rely on deep neural networks, in particular Generative Adversarial Networks (GANs) and autoencoders. GANs consist of two rival neural networks: the generator produces synthetic content, and the discriminator determines whether the generated content is authentic. These models are trained continuously to produce highly convincing media and can replicate a person's facial expressions, voice, and mannerisms. As a result, deepfakes have been used in entertainment, education, virtual assistants, and content creation [20-25]. In the digital age, information is disseminated and consumed at an unprecedented rate, and the authenticity and integrity of information have become ever more important. Deepfake technology has become one of the biggest challenges to the authenticity of information in recent years. Deepfakes are synthetic media produced using artificial intelligence (AI) techniques that substitute a person in an existing image or video with someone else's likeness. The term deepfake is a portmanteau of the words ''deep learning'' and ''fake'' [26-30]. The application of Artificial Intelligence (AI) through deepfake expertise does not make manipulating digital media, such as videos, images, and audio recordings, difficult. "Deepfake" combines "deep learning" and "fake," signifying the technology's capability to generate highly realistic synthetic content that closely mimics genuine media. Deepfake generation tools have become much more powerful and readily available in the recent past, thanks to advances in machine learning, and are now widely available for both legitimate and

This work is licensed under a Creative Commons AttributionShareAlike4.0 International License, which permits Unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited

nefarious purposes.

Digital forensics is the science of gathering, preserving, analyzing, and presenting electronic evidence from a digital device or multimedia source to support the investigation and prosecution of a crime. This is critical in digital content verification and integrity, especially in identifying manipulated media, including deepfakes. With the rise of AI-generated content, digital forensics has evolved to embrace more advancedmethods for detection, attribution, and authentication to confirm digital evidence [10]. Digital forensics is the basis of cybercrime investigations; it ensures the authenticity, integrity, and reliability of digital evidence in the increasingly complex digital landscape [31].

A. Computer Forensics

The identification, gathering, preservation, and analysis of digital evidence kept on computers and other storage devices are all part of computer forensics, a subfield of digital forensics. To reconstruct events related to cyber incidents, it examines hard drives, file systems, operating systems, logs, application data, and other file systems. In the field of deepfakes, computer forensics can help investigators identify the origin, history of alterations, and veracity of altered photos, videos, and audio recordings. Sophisticated forensic tools can retrieve erased files, examine metadata, and find evidence of manipulation. Computer forensics is crucial for investigating cybercrimes, intellectual property theft, and legal actions. Forensic analysts now face additional difficulties due to the increasing use of artificial intelligence in media manipulation [19].

B. Mobile Forensics

Mobile forensics is the extraction and analysis of digital evidence from cell phones, tablets, and other portable devices. Investigators examine call logs, messages, application data, multimedia files, GPS data, and internet activity to find evidence related to cybercrime. With mobile devices serving as the primary platforms for the creation and sharing digital content, their association with the distribution of deepfakes and manipulated media has grown. Mobile forensic techniques are used to identify the origin of suspicious content and communication patterns related to fraudulent activity. This process requires specialized tools that can collect data without compromising the integrity of the evidence. Challenges include device encryption, cloud sync, and frequent OS updates. Thus, mobile forensics is important in modern digital investigations involving AI-generated media.

C. Network Forensics

To look into cyberattacks and unauthorised behaviour, network forensics monitors, records, and analyses network traffic. Investigators can follow the spread of harmful files, identify suspicious conversations, and identify the origin of cyberthreats thanks to it. Network forensics can also be utilised in deepfake investigations to monitor the propagation of altered content over communication networks and internet platforms. Investigators can detect compromised systems and piece together attack timing by examining packet, log, and traffic patterns. Network forensic methods are frequently used in incident response, fraud investigation, and intrusion detection. The need for real-time monitoring and forensic analysis has increased along with the volume of AI-generated content on digital networks. Network forensics is a key component in reducing [18].

D. Cloud Forensics

Cloud forensics is the identification, collection, and analysis of digital evidence within a cloud computing environment or that is processed. It manages forensic investigations of virtual machines, cloud storage services, and distributed infrastructures. The growing use of cloud-based artificial intelligence systems has increased the importance of cloud forensics in deepfake investigations. Investigators will often have to analyze datasets, machine learning models, and user activities related to media manipulation hosted in the cloud. However, cloud environments present challenges such as multi-tenancy, data volatility, and restricted access to the underlying hardware. In cloud forensic investigations, it is important to consider the integrity of evidence and compliance with legal requirements. Consequently, cloud forensics has become an essential discipline for investigating cybercrimes using AI-based applications and cloud-stored digital evidence [12].

III. BASIC CONCEPT OF DEEPFAKE AND ITS DETECTION TECHNIQUES

The growth of deepfake technologies has primarily transformed the digital environment. Deepfakes can credibly reproduce human faces, voices, gestures, and phrases.

A. Audio and Audio-Visual Detection Techniques

The techniques are based on the detection of manipulated speech and synthetic voices by analyzing acoustic features (Fig. 1), emotional cues, and synchronization between audio and video streams [21].

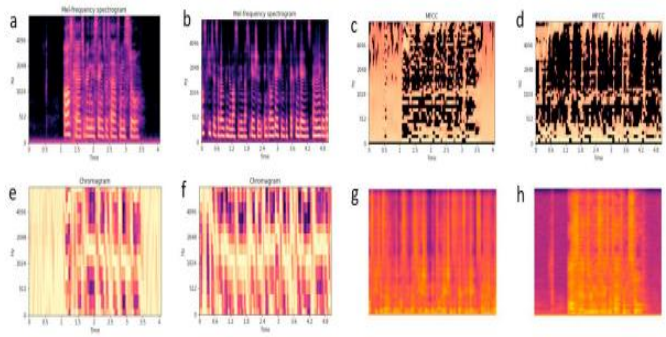

Fig. 1: Deepfake Audio Detection [21].

B. Visual Artifact-Based Detection Technique

Deepfake content creation generally results in subtle anomalies in the visual data, commonly referred to as artifacts, e.g., irregular pixel arrangements, distortions at the edges, or anomalies in the spectral domain. These anomalies are exploited by visual artefact detection methods as clues of synthetic manipulation. Features for deepfake detection, for example. The full face (a) has visual artifacts such as unnatural

pixel formations around the facial features (e.g., glasses and skin edges) (Fig. 2). They indicate irregularities induced by the generative algorithm or compression distortions. (b) The zoomed region around the mouth shows possible texture inconsistencies, such as unnatural blending of lip textures and a lack of smooth transitions, which are typical of natural skin and lip patterns [2].

Fig. 2: Deepfake detection framework [2]

C. Multimedia Manipulation Detection Technique

It is very interesting to broadcast videos and photos in digital format with half-truths. Three well-known primary types of manipulation that can be utilized on a multimedia file are copy–move, splicing, and deepfake. Even though the outcome is similar overall, primarily involving the alteration of objects, faces, or voices in multimedia content, the techniques used to create and improve these manipulations, along with the machine learning methods applied to identify them, are markedly different and can pose significant difficulties for a completely automated detection system. Two strong examples illustrate how copy–move manipulation can effectively disseminate false information and remove the original photo from its context. Almost ten years ago, Iran faced allegations of tampering with a photograph of its missile tests. The image shown in Fig.3. Posted on the official website of Iran's Revolutionary Guards, which asserted that four missiles were firing upward at the same time, although only three missiles were actually launched [30].

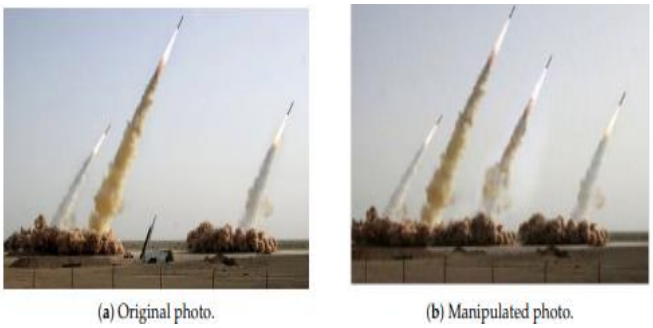

Fig. 3: Multimedia Manipulation Techniques [30].

D. 3D Deepfake Generation and Detection (3DDGD) dataset technique

This method employs 3D facial meshes to enhance biometric security against deepfakes. It employs a multi-step approach that involves reconstructing faces from 2D images, extracting geometric features, and using custom-trained deep learning models to identify real faces from fake ones. The 3DDGD dataset is developed as a public 3D deepfake dataset that helps enhance the security of 3D facial recognition and remote identity verification systems against 3D deepfake attacks. Furthermore, it offers guidelines on evaluating the effectiveness of the 3D deepfake detection system. The dataset contains different ages and genders, mostly Asian, and 20 different facial expressions. In contrast to current deepfake detection datasets that mainly concentrate on 2D deepfakes, the 3DDGD dataset is designed explicitly for 3D deepfake situations. The process of developing the dataset is described in the following sections as illustrated in Fig. 4 [13].

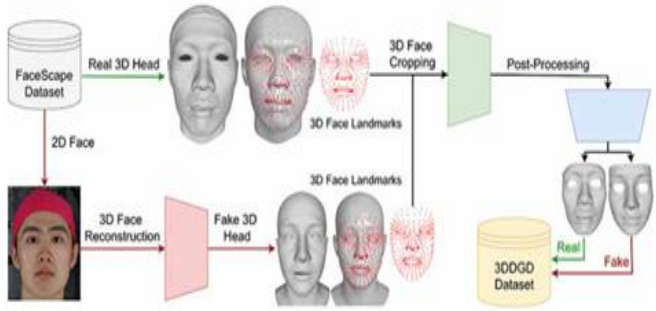

Fig 4. 3D Deepfake Generation and Detection Using 3d Face Meshes. Source [13].

IV. Impact of Deepfakes and AI-Generated Evidence on Digital Forensics

The rise of deepfakes and AI-generated evidence is revolutionizing the field of digital forensics, raising doubts about the authenticity, integrity, and reliability of digital evidence. Advanced artificial intelligence can create highly convincing fake images, videos, and audio content that is increasingly difficult to distinguish from real content, making traditional forensic approaches less effective. As a result, digital forensic investigators are increasingly turning to deep learning, explainable AI (XAI), multimodal analysis, forensic attribution, and media authentication tools to detect alterations and ensure confidence in digital evidence. Deepfakes have been rapidly evolving and are continually being used in media production, while at the same time forensic tools and methods are continuously being enhanced to prevent their use [33].

Evidence Tampering: The alteration of images, deepfakes, and tampering with evidence are threats. It can overcome major obstacles to digital forensic validation by facilitating the generation of extremely convincing but deceptive digital material. Networks (GANs) can also be used to generate images. Another approach to generating images with deep learning is Generative Adversarial Networks (GANs). Networks (GANs) and diffusion models can change or create images, videos, and audio recordings, which makes it harder to read. To distinguish between true and manipulated media. The integrity, admissibility, and reliability of advancements are at stake. Digital evidence is applied in criminal investigations, legal Processes, and cybersecurity incident responses. So, digital the existing forensic authentication relies on complicated Detection techniques such as forensic artifact analysis and multimodal, In addition, AI systems must be verifiable, explainable AI (XAI), and source-attributable. Watermarking

and tracking of origin to verify authenticity, raising awareness of digital content and maintaining trust in forensic investigations [18]. Similarly, audio recordings created by AI can closely mimic an individual's voice, complicating the process of authenticating recorded dialogues and declarations [21].

The swift progress of with the advent of artificial intelligence, the media has become much more complicated. Manipulation methods. Deepfake technologies can create realistic videos of people. Deepfakes can make realistic videos of people. Face mimicking, voice replicas, and synthetic videos that closely mimic faces. Be able to replicate authentic content, which can be difficult for forensic analysis [25]. Offenders could exploit these technologies. These technologies could be abused by offenders. To produce deceptive documentation, spread deceptive information, or pretend to be someone else to cover up or conceal illegal behavior of others. Consequently, the conventional trust that you have in digital photos, videos, and recordings of sound. Depicting actual events is no longer universally valid. Techniques for Cropping, rotating, inserting objects into an image, etc., are all types of image manipulation. Color, font, and shading of the background, and editing the content are possible. Context and significance of visual evidence. These changes could impact decisions of the court, lie to detectives, or compromise results of legal cases [30]. The increasing ease of deepfake creation tools increases the chances of evidence tampering on a variety of digital channels [32].

V. Relationship of Deepfakes with digital forensics

With the advent of deepfakes, the world has changed. The digital environment, new challenges for investigators, law enforcement, and the judicial systems. Deepfakes can convincingly attempt to copy human faces, voices, gestures, and phrases, causing manipulated content that appears authentic to people and machine systems [2]. Digital evidence has become a cornerstone of evidence. An element of contemporary investigations, assisting in crime, the following activities are prohibited: prosecutions, cybersecurity incident management, and civil. Lawsuits. However, with the advent of “deepfakes,” the rise in these causes is undermining the whole concept. The authenticity of, or the extent of their use in, documents and artifacts. Doubt regarding documents and artifacts and the extent of their use of digital media. Authenticity and trustworthiness of sources [10]. Investigators can no longer simply accept that photographs, video surveillance, audio recordings, or any other type of documentation are just a part of the process. True-to-life emails, records, or social media posts are not necessarily an accurate reflection of real events. As a result, forensic professionals need to implement sophisticated Methods of validation before accepting digital evidence as evidence.

Modern digital forensic processes use artificial intelligence and machine learning algorithms to identify anomalies that may indicate that the media has been tampered with. The systems are used to analyze the faces. Many discrepancies, many lighting anomalies, many compression artifacts, Frequency domain properties, and behavioral traits make it difficult to duplicate exactly in artificial content [20]. The advent of sophisticated AI tools has made it easier for bad actors to exploit. With the introduction of advanced AI tools, malicious actors have had an easier time exploiting. To produce ‘misleading’ material with minimal effort and without technical expertise, increasing the likelihood of misinformation, identity theft, fraud, and the fabrication of evidence [32].

Deepfakes are becoming more and more widespread. Growing in popularity due to advances in artificial intelligence. Generative models make it very easy to create really simply. Fake images, videos, and audio recordings that look real. These altered Media are one of the major challenges to the credibility and reliability of digital evidence used in investigations and court proceedings. Reliability of digital evidence used in investigations and judicial proceedings. Processes. Typical forensic techniques often struggle to identify it. Challenging to identify advanced deepfakes that closely resemble authentic material. Thus, digital forensics has evolved to add advanced detection technologies based on deep learning, multimodal analysis, as well as an investigation of forensic artifacts. Explainable Artificial Intelligence (XAI) is being used by researchers.

To improve the clarity and reliability of detection systems. Methods Similar to source attribution, watermarking, and media provenance. The world of digital media offers numerous methods for tracking the origin and legality of digital content. There are many ways to validate the source and authenticity of digital media. Content. The existing deepfake detection systems analyze images. Errors, inconsistencies, temporal changes, and connections between sounds and sights. Yet, despite all these developments, a continuous arms competition has continued. The battle between creating deepfakes and forensic analysis continues. Technologies. The growing complexity of AI-produced content requires continuous improvement of forensic tools and techniques. Therefore, Digital Forensics has a crucial role in preserving. The integrity, credibility, and admissibility of electronic evidence in the face of deepfakes [33]. Multimodal detection methods enhance this by simultaneously measuring visual and auditory data and using situational data to identify discrepancies between media types [34].

A. Legal and Ethical Issues

This occurrence, called the "liar's dividend," is allowed. People denying that valid evidence is created by Artificial Intelligence (AI) has been identified as an area of focus by researchers. The necessity to strengthen regulatory systems and monitoring of the media, to methods, transparent AI techniques, and global collaboration to maintain public confidence and ensure responsibility in digital. Media and forensic investigations [2]. The swift progress of deepfake technology has brought considerable legal and ethical issues that affect people, groups, and law. A major Identity theft and impersonation is the main concern, as AI-generated images, videos, or audio of individuals are used to impersonate them. Without authorization, it leads to fraud and damage to reputation [12]. Also, there are ethical concerns arising from the collection and use of personal information. Data for training deepfake generation models without obtaining informed consent. Further, biases in AI datasets can also have the effect of causing errors. Result in unfair or biased outcomes when creating and using them. Identification of deepfakes [18]. Deepfakes can be used to advance the promotion of the impacting on

dissemination of misinformation and disinformation, public awareness, politics, and peace among the people. Invented media content [24]. A significant concern is the non-consensual and harmful content, like revenge porn, harassment and slander that violates Any consideration of individual privacy and human rights [25] Legally, deepfakes pose problems related to the admissibility and authenticity of digital Evidence in court, as media can be convincingly altered and this can erode trust in the evidence, in photographs, videos, and audio recordings utilized in Investigations, legal processes [33].

B. Deepfake Datasets and Benchmarking Metrics

Deepfake datasets and benchmarking systems play a crucial role in the development and evaluation of deepfake detection systems. They provide real and fake images, videos, and audio examples that can be used for training, validation, and testing the accuracy of forensic models. Benchmarking helps to assess the efficacy, accuracy, flexibility, and reliability of detection methods in the practical context. As the technology for creating deepfakes continues to evolve, the need to continuously feed large datasets into the detection system to ensure its continued effectiveness with the new wave of manipulation techniques is increasing. As deepfake technology advances, the need to constantly flood the system with massive amounts of data to keep it effective against new deepfakes is growing. Recent studies have revealed the need for diverse datasets with different age, gender, ethnicity, facial expression, lighting conditions, and compression levels for better generalization of the model [20].

C. FaceForensics++

One of the most widely used benchmark sets for deepfake detection research is FaceForensics++. It includes thousands of altered videos created with different facial manipulation methods, such as Face2Face, FaceSwap, and DeepFakes. Data is available in several compression levels, which can be used to evaluate detection models in realistic scenarios [25].

D. Celeb-DF

To address the limitations of existing datasets, Celebr- DF aims to provide more realistic and higher-quality deepfake videos. It includes videos with the faces of celebrities turned into visual artifacts, making identification much more difficult. As a result, Celeb-DF is becoming an important benchmark to evaluate the generalization capabilities of deepfake detection systems [20].

E. DeepFake Detection Challenge (DFDC)

The purpose of the DFDC dataset is to share and motivate the development of successful deepfake detection methods. It contains over 100,000 videos of various identities, backgrounds, and manipulation techniques. To assess deep learning models in real-life and large-scale scenarios, DFDC is frequently used [28].

F. ASVspoof Dataset

The ASVspoof dataset focuses on speaker verification and audio deepfake detection, addressing spoofing threats. It includes real and fake speech samples generated using voice conversion and speech-to-text technologies. The dataset is used as a benchmark for the evaluation of audio forensic techniques and anti-spoofing techniques [29].

G. 3DDGD Dataset

The 3D Deepfake Generation and Detection (3DDGD) dataset was specially created for detecting 3D facial deepfakes. It uses 3D facial meshes, geometric features, and facial expressions to evaluate the strength of biometric authentication systems against a 3D deepfake attack. In contrast to conventional datasets that emphasize 2D manipulations, 3DDGD tackles new challenges related to 3D face reconstruction and identity verification systems [13].

Researchers commonly use the following metrics to compare deepfake detection models:

Accuracy: Overall percentage of correctly classified samples [25].

Precision: Ability to correctly identify manipulated media [28].

Recall (Sensitivity): Ability to detect all deepfake instances [20]

F1-Score: Harmonic mean of precision and recall [11].

Area Under the Curve (AUC): Measures classifier discrimination capability [7].

Equal Error Rate (EER): Frequently used in audio and biometric spoofing detection [29].

Cross-Dataset Generalization: Evaluates model performance on unseen datasets [28].

VI. Software Testing Challenges in Deepfake Detection Software

A. Dataset Multiplicity and Real-World Testing Challenges

Deepfake detection software is hard to develop since models are regularly tested against small datasets, which don't replicate real-world scenarios. Deepfake generation methods are continually growing, and new methods of manipulation are being established that may not be detected by current systems [9].

Few accessible high-quality benchmark datasets. Testing methods are challenging to use against hidden deepfake generation procedures. How the video performs at various quality, environment, and device levels. Performance ranges from one video quality to another, one environment to another, one device to another. Difficulty implementing standardized testing procedures. Problems with standardized testing procedures.

B. Generalization and Robustness Testing Challenges

One of the main software quality challenges is ensuring a deepfake detection system operates efficiently in fluctuating environments. Deepfake detectors can be incorrectly trained to recognize unwanted patterns, rather than manipulation features. Poor generalization to new deepfake algorithms is a drawback. Learn to deal with overfitting to certain datasets. The ability to detect compression in images and video. Sensitivity to image/video compression. Failure due to noise, changes in light, and resolution differences.

C. Assessment Metrics and Reliability Testing Challenges

Measuring accuracy isn't enough to evaluate deepfake detection tools. When it comes to forensic applications, results must be reliable and explainable [28]. Forensic reliability is not measurable by accuracy alone. Struggle to assess false positives and false negatives. Inadequate explanation of detection

decisions made by AI. Requirements for testing detection speed and computational efficiency.

VII. Quality Assurance Requirement for Deepfake Detection Software

A. Accuracy of Detection and the Performance Evaluation

One of the most important quality assurance needs of deepfake detection software is high detection accuracy. The system should correctly classify manipulated media but with minimized false classifications [3], [9], [25]. Deepfake detection systems must be judged not only in terms of accuracy, since good accuracy in small data sets does not necessarily translate to good accuracy in real forensic conditions [9], 25], [28].

B. Measure uncertainty and determine reliability

Detection results in digital forensic investigations need to be reliable. A system should give a level of confidence and give reasons for the classification of content as fake [1], [7].

Requirements:

Assurance score generation.
Uncertainty assessment.
The same results that are observed in several tests.
Able to make decisions when information is limited.

Critical Analysis:

Reliability testing is critical as forensic investigators need evidence that is not only truthful, but is also backed by a level of assurance and uncertainty that can be measured [7].

C. Explainability and Transparency

Typically, AI-powered deepfake detection tools operate as a "black box." Quality assurance requires understandable outputs [1]. [3].

Requirements:

- Explainable AI (XAI) support
- Visualization of manipulated regions
- Justify the decisions made regarding detection.
- Transparent forensic reporting

Critical Analysis:

Trust in deepfake detection software is enriched by explainability, enabling forensic experts to understand and confirm the system's decisions [1, 3].

D. Robustness and Generalization Testing are identical

A good deepfake detector must be able to detect attacks that haven't been seen before and are performed through various manipulation techniques [8], [14], [28].

Requirements:

- Cross-dataset testing
- Evaluating deepfake detection techniques for new generation techniques
- The property of resisting compression, noise, and changes in quality.
- Ability to adjust to changing threats.

Critical Analysis:

Generalization testing is also a critical requirement of QA for deepfakes, as there is a nonstop evolution in deepfakes and detection systems need to be able to work alongside unknown manipulation techniques [8], [14].

E. The Quality and Validation of Datasets

The accuracy of training and test sets is directly related to the reliability of detection [9], [33].

Requirements:

- Diverse datasets
- Balanced real/fake samples
- There are many types of deepfake attacks
- Regular dataset updates

Critical Analysis:

Poor data diversity may result in less efficiency of DeepFakes detection systems and poor Forensic results [25], [33].

F. Examination of Efficiency and Scalability of Computational Approaches

For hands-on forensic investigations, detecting Deepfakes should be swift to investigate [4], [26].

Requirements:

- Quicker processing time
- Effective AI models
- Low computational cost
- Real-time detection ability

Critical Analysis:

In addition to detection performance, computational efficiency should be a 6th aspect of quality assurance, given the need for swift media investigation in forensic inquiries [4] [28].

VIII. Comparative Analysis of Existing Deepfake Detection Methods

Various improvements in model trustworthiness, clarity, and generalizability across datasets have been applied, such as contrastive learning, uncertainty-aware models, and Explainable AI (XAI) frameworks. A set of deep learning models, particularly Convolutional Neural Networks (CNNs), has demonstrated remarkable accuracy in detecting altered media, capable of autonomously learning distinguishing features from images and videos. The evolution of deepfake detection has moved beyond traditional forensic artifact analysis to more advanced AI-powered methods such as CNNs, Transformers, and multimodal learning, significantly improving detection accuracy. Despite this, challenges like adversarial attacks and poor generalization across datasets, as well as rapidly evolving deepfake generation methods, continue to hinder reliable use in real-world settings [15]. In recent years, architectures utilizing Transformers and attention mechanisms have also improved detection accuracy by comprehending complex spatial and temporal relationships in the multimedia information [17]. Despite these advances, the existing methods still face challenges of unknown manipulations, adversarial attacks, and practical implementation, highlighting the need for more adaptable and effective forensic solutions [28]. The methods used to detect deepfakes have evolved from traditional forensic tools to advanced AI-based systems. Early efforts focused on detection of visual artefacts, inconsistencies in metadata, and compression artefacts, but their effectiveness has waned as deepfake production has become more complex [30]. Multimodal strategies that incorporate visual, auditory, and behavioral signals are more resistant to advanced deepfakes

[31].

IX. Challenges in Deepfake Detection and Digital Forensics under Software Quality Assurance and Testing

This means that there is a growing need for detection mechanisms to be explainable, adaptive, and reliable, which is a challenge due to the rise of deepfake technologies [1]. For technically effective and legally compliant solutions, there is a need for ongoing cooperation between AI researchers, the cybersecurity community, forensic experts, and the legal profession [10]. Additionally, it is of great significance to develop a standard evaluation framework that can evaluate different deepfake detection models in terms of accuracy, robustness, explainability, computational complexity, and real-world performance [11]. In conclusion, future studies will need to challenge ethical and privacy issues that involve the use of AI-powered forensic systems, including the creation of principles of accountable use and transparency in decision-making. As deepfake technology advances swiftly, digital forensics systems need to be adaptive and incorporate mechanisms for threat intelligence and continuous learning to effectively beat new forms of manipulation. Future justifiable, adaptive, explainable, secure, and multimodal forensic frameworks will become critical for the credibility, reliability, and admissibility of digital evidence [18].

X. Conclusion

In the age of digital evidence, Deepfake detection software is an important tool for certifying the trustworthiness of digital evidence and enhancing forensic investigations. But as the technology of deepfakes continues to advance, so do the challenges for detection procedures, confirming accuracy and reliability over time. To ensure that these systems can manage numerous kinds of manipulated media and give trustworthy results, there is a requirement for effective testing and quality assurance. The analysis indicates that deepfake detection tools should not just prioritize accuracy of detection but include other characteristics of the tools as well, such as robustness, explainability, security, and performance. However, challenges like growing deepfake approaches, small sample sizes for testing, and challenges in validation in real-world settings highlight the necessity for continuing innovation and sophisticated testing approaches. SQA and effective Explainable and Adaptive AI techniques will help create more reliable deepfake detection systems in the future. These enhancements will further encourage the competency of digital forensic investigations and endorse confidence in the application of AI-powered detection capabilities.

FUNDING STATEMENT

The author(s) received no specific funding for this study.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest to report regarding the present study.

AUTHOR CONTRIBUTIONS

Conceptualization, D.K. and P.K.; methodology, D.K.; software, D.K. and P.K.; validation, D.K.; writing—original draft preparation, D.K.; writing—review and editing, P.K.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

Data is available on reasonable request.

References

[1] S. Aoudi et al., "Beyond Accuracy: Auditing Image Forgery Localization via Scene-Disjoint Evaluation and XAI Faithfulness," IEEE Access, vol. 14, pp. 53101–53115, 2026.

[2] I. Amerini et al., "Deepfake Media Forensics: Status and Future Challenges," Journal of Imaging, vol. 11, no. 3, p. 73, 2025.

[3] M. Al-Imran et al., "DeFaX: A Cross-Attention Fusion Framework for Robust and Explainable Deepfake Detection," IEEE Access, vol. 13, pp. 213962–213979, 2025.

[4] A. Al Redhaei, S. Fraihat, and M. A. Al-Betar, "A Self-Supervised BEiT Model with a Novel Hierarchical PatchReducer for Efficient Facial Deepfake Detection," Artificial Intelligence Review, vol. 58, no. 9, p. 278, 2025.

[5] M. Zou et al., "Semantics-Oriented Multitask Learning for DeepFake Detection: A Joint Embedding Approach," IEEE Transactions on Circuits and Systems for Video Technology, 2025.

[6] T. D. Mohottalalage, D. Saha, and M. Schmidt, "A Comprehensive Review of Deepfake Detection Methods and Challenges in Digital Forensics," preprint, 2025.

[7] X. Jin et al., "Towards Reliable Deepfake Detection from Uncertainty Calibration Perspective," Visual Intelligence, vol. 3, no. 1, p. 28, 2025.

[8] O. Pontorno, L. Guarnera, and S. Battiato, "DeepFeatureX-SN: Generalization of Deepfake Detection via Contrastive Learning," Multimedia Tools and Applications, pp. 1–20, 2025.

[9] M. Alrashoud, "Deepfake Video Detection Methods, Approaches, and Challenges," Alexandria Engineering Journal, vol. 125, pp. 265–277, 2025.

[10] R. Raman et al., "Opposing Agents Evolve the Research: A Decade of Digital Forensics," Multimedia Tools and Applications, vol. 84, no. 14, pp. 13485–13513, 2025.

[11] A. A. Alrawahneh et al., "Video Authentication Detection Using Deep Learning: A Systematic Literature Review," Applied Intelligence, vol. 55, no. 4, p. 239, 2025.

[12] C. J. Zhang et al., "AI-Based Identity Fraud Detection: A Systematic Review," arXiv preprint arXiv:2501.09239, 2025.

[13] H. Felouat et al., "3DDGD: 3D Deepfake Generation and Detection Using 3D Face Meshes," IEEE Access, 2025.

[14] N. Bai et al., "Towards Generalizable Face Forgery Detection via Mitigating Spurious Correlation," Neural Networks, vol. 182, p. 106909, 2025.

[15] D. Javiya et al., "Enhancing Deepfake Detection with Adaptive-DCGAN and Lite-CNN: A Novel Approach to Image Classification," Discover Applied Sciences, vol. 7, no. 11, p. 1229, 2025.

[16] X. Jin et al., "Towards Reliable Deepfake Detection from Uncertainty Calibration Perspective," Visual Intelligence, vol. 3, no. 1, p. 28, 2025.

[17] G. Chen et al., "A Deepfake Image Detection Method Based on a Multi-Graph Attention Network," Electronics, vol. 14, no. 3, p. 482, 2025.

[18] M. Wazid et al., "A Secure Deepfake Mitigation Framework: Architecture, Issues, Challenges, and Societal Impact," Cyber Security and Applications, vol. 2, p. 100040, 2024.

[19] M. Tampubolon, "Digital Face Forgery and the Role of Digital Forensics," International Journal for the Semiotics of Law, vol. 37, no. 3, pp. 753–767, 2024.

[20] A. Kaur et al., “Deepfake Video Detection: Challenges and Opportunities,” Artificial Intelligence Review, vol. 57, no. 6, p. 159, 2024.
[21] M. Mcuba et al., “The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation,” Procedia Computer Science, vol. 219, pp. 211–219, 2023.
[22] Y. Cao et al., “Three-Classification Face Manipulation Detection Using Attention-Based Feature Decomposition,” Computers & Security, vol. 125, p. 103024, 2023.
[23] Z. Guo et al., “A Data Augmentation Framework by Mining Structured Features for Fake Face Image Detection,” Computer Vision and Image Understanding, vol. 226, p. 103587, 2023.
[24] M. T. Raj et al., “A Review on the Detection of Deep Fake and Propaganda Videos and Images-Based Voice and Facial Manipulation Using AI Techniques,” in Proc. 2nd Int. Conf. on Automation, Computing and Renewable Systems (ICACRS), 2023, pp. 1–6.
[25] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, “Deepfakes Generation and Detection: State-of-the-Art, Open Challenges, Countermeasures, and Way Forward,” Applied Intelligence, vol. 53, no. 4, pp. 3974–4026, 2023.
[26] A. Alharbi et al., “Novel 59-Layer Dense Inception Network for Robust Deepfake Identification,” Scientific Reports, vol. 15, no. 1, p. 24159, 2025.
[27] R. Yao et al., “Deepfake Detection in Image Sequences: A Temporal Approach for Anomaly Detection,” International Journal of Intelligent Systems, vol. 2025, no. 1, Art. no. 8566328, 2025.
[28] N. U. R. Ahmed et al., “Visual Deepfake Detection: Review of Techniques, Tools, Limitations, and Future Prospects,” IEEE Access, vol. 13, pp. 1923–1961, 2024.
[29] H. Delgado et al., “ASVspoof 2021: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan,” arXiv preprint arXiv:2109.00535, 2021.
[30] S. Ferreira *et al*., “Exposing Manipulated Photos and Videos in Digital Forensics Analysis,” *Journal of Imaging*, vol. 7, no. 7, p. 102, 2021.
[31] T. Mittal et al., “Emotions Don’t Lie: An Audio-Visual Deepfake Detection Method Using Affective Cues,” in Proc. 28th ACM International Conference on Multimedia, 2020, pp. 2823–2832.
[32] J. Kietzmann, L. W. Lee, I. P. McCarthy, and T. C. Kietzmann, “Deepfakes: Trick or Treat?,” Business Horizons, vol. 63, no. 2, pp. 135–146, 2020.
[33] Loovens, Jenifer, and Hasan Tinmaz. "A Systematic Literature Review of Deepfakes in Forensic Science." Forensic Imaging (2025): 200647.
[34] Hao, Hanxiang, et al. "Deepfake detection using multiple data modalities." Handbook of digital face manipulation and detection: from deepfakes to morphing attacks. Cham: Springer International Publishing, 2022. 235-254.

TABLE 1: Comparative Analysis of various methods

Author Name / Year	Proposed Method	Working of Model	Limitation
Aoudi et al. (2026)	**DF2023-XAI Framework using Scene-Disjoint Evaluation and XAI Faithfulness Auditing**. The authors compare a CNN-based U-Net and a Transformer-based SegFormer-B2 for image forgery localization and introduce **Saliency Total Variation (TV)** as an Explainable AI (XAI) metric to evaluate forensic reliability and attribution faithfulness.	With a strict Scene-Disjoint evaluation protocol, the framework localizes manipulated regions in images based on U-Net and SegFormer architectures. Saliency TV is used to quantify the model's attention to the semantically relevant manipulated areas rather than boundary artifacts in the images by analyzing the Grad-CAM++ attribution maps The anomalies detected by SegFormer are semantic object-level ("Blob" behavior), and the anomalies detected by U-Net are mainly edge detection ("Wireframe" behavior).	Performance Drops: SegFormer struggles with enhancement-based manipulations like sharpening, blurring, and contrast adjustments. High Overhead: Transformer architectures require more computation and memory than CNNs. Domain Limitations: The study is restricted to RGB representations, omitting frequency-domain or noise-residual features that could detect subtle manipulations
I. Amerini et al . (2025).	Deepfake Detection Forensic Attribution and Recognition Passive Authentication Active Authentication Detection in Real-World Scenarios	The framework examines image, video, or audio materials to determine whether they are real or created by artificial intelligence. It then verifies the authenticity by source attribution and authentication techniques before reaching a final trust assessment.	The framework struggles with increasingly sophisticated deepfakes and real-world conditions, while its authentication relies on watermarking and provenance technologies that are not yet universally adopted.
Loovens, J., & Tinmaz, H. (2025).	The study systematically collects, evaluates, and categorizes existing deepfake forensic research. It compares available detection approaches and identifies gaps that need further investigation.	It classifies and compares deepfake detection technologies to uncover strengths and weaknesses, and research gaps in forensic science.	Loovens et al. (2025) provide a systematic review of deepfake research in forensic science, highlighting existing detection methods, challenges, limitations, and future research opportunities.
Ahmed et al. (2024)	Comprehensive review of visual deepfake detection techniques, datasets, tools, and forensic approaches. The study analyzes traditional forensic methods, deep learning models, CNNs, Transformers, and multimodal detection frameworks.	The paper provides a comprehensive overview of current research on visual deepfake detection techniques and classifies them according to their techniques. It contrasts datasets, assessment criteria, good points, bad points, and applications to practical uses of various detection methods. The study highlights the existing problems and future research opportunities to enhance the deepfake detection system.	Does not offer a new detection framework or algorithm. Findings are based on the quality and extent of literature reviewed. Some reviewed techniques might rapidly become obsolete due to the swift progress in deepfake creation.
M. Mcuba et al. (2023).	Audio files are converted into visual feature representations (MFCC, Mel-Spectrogram, Spectrogram, and Chromagram) and fed into deep learning models. The CNN architectures learn distinguishing patterns between genuine and synthetic voices and classify the audio as real or fake.	The model transforms the audio file into feature representations like MFCC and spectrograms and then uses CNN-based networks to process these representations. model converts audio recordings into feature representations such as MFCCs and spectrograms, which are processed by CNN-based networks. CNN detects patterns in the audio and determines whether it is authentic or a deepfake.	The number of training or testing data was not large enough. Results might not be broadly applicable to newer methods of voice cloning. The choice of the audio feature representation sets a key constraint for model performance. The authors pointed out that more tuning of the parameters and more datasets are required to increase the accuracy.