ISSN: 3049-0952
www.jsetms.com

An Artificial Intelligence Framework for Detecting Deepfake Videos to Enhance Digital Forensic Investigations

[1]Mulaparthi Anu
[1]M.Tech, Department of Computer Science and Engineering
DNR College of Engineering & Technology (Autonomous)
Bhimavaram – 534202, Andhra Pradesh, India
anumulaparthi779@gmail.com

[2]Dr. A. Ramamurthy.
[2]Professor, Department of Computer Science and Engineering
DNR College of Engineering & Technology (Autonomous)
Bhimavaram – 534202, Andhra Pradesh, India
ram111.sai@gmail.com

Abstract— The rapid development of artificial intelligence and deep-learning technologies has enabled the generation of highly realistic DeepFake videos, creating significant challenges for media authenticity, cybersecurity, public trust, and digital forensic investigations. DeepFake manipulation can alter facial appearance, expressions, movements, and audio-visual characteristics, making conventional verification approaches increasingly ineffective. This survey reviews existing artificial intelligence and deep-learning techniques for DeepFake video detection, including convolutional neural networks, recurrent neural networks, spatial-temporal analysis, multimodal learning, attention mechanisms, artifact-based analysis, and transformer-based approaches. Existing methods are examined with respect to their detection capabilities, robustness, generalization, computational requirements, and limitations in real-world forensic scenarios. Based on the surveyed literature, an integrated AI-based DeepFake detection framework is presented that combines video acquisition, preprocessing, face detection, spatial feature extraction, temporal inconsistency analysis, multimodal forensic evidence assessment, and explainable detection. The framework further considers confidence estimation, robustness against unseen manipulation techniques, and human-assisted forensic verification. The survey identifies important research challenges related to dataset diversity, cross-domain generalization, adversarial manipulation, computational efficiency, interpretability, and reliable forensic evidence. The proposed framework provides a structured foundation for developing intelligent and trustworthy DeepFake detection systems capable of supporting digital forensic investigations and improving the reliability of digital video evidence.

Keywords— DeepFake Detection, Artificial Intelligence, Deep Learning, Video Forgery Detection, Digital Forensics, Convolutional Neural Networks, Spatial-Temporal Analysis, Multimodal Learning, Facial Forgery, Transformer, Explainable AI, Video Authentication, Forensic Evidence.

I. INTRODUCTION

DeepFake technology has emerged as a significant application of artificial intelligence and deep learning, enabling the generation of highly realistic manipulated images, audio, and videos. DeepFake videos can modify facial appearance, expressions, lip movements, identity, and other visual characteristics while maintaining a realistic appearance. The increasing accessibility of generative artificial intelligence has made such manipulated content easier to produce and distribute across digital platforms. As a result, DeepFake videos pose serious challenges to information authenticity, cybersecurity, public trust, and the reliability of digital evidence. Recent advances in artificial intelligence have therefore created opportunities for automated DeepFake detection and intelligent digital forensic analysis [1]–[4].

Conventional video verification generally depends on manual inspection, source verification, metadata examination, and visual analysis by investigators. Although these approaches provide useful evidence, they become difficult when manipulated videos contain highly realistic facial characteristics or when large volumes of digital content must be examined. Deep-learning techniques can automatically learn discriminative visual patterns from video frames and identify subtle manipulation artifacts [8], [9], [12]. Convolutional neural networks have therefore been investigated for extracting spatial facial features, while advanced approaches extend detection toward temporal and multimodal characteristics [4], [7], [11].

DeepFake videos are inherently temporal because manipulation artifacts can vary across consecutive frames. Spatial-temporal approaches can capture relationships between facial information across video sequences [1], [6], [14]–[16]. Gaze inconsistency provides an additional clue for identifying manipulated facial videos [1], while frame-level detection can improve video-level decisions [6]. Multimodal approaches further examine relationships between audio and visual information for detecting manipulated segments [7], [11].

The increasing sophistication of DeepFake generation creates challenges for detection generalization. Detectors may experience reduced performance on unseen manipulation methods, compressed videos, social-media transformations, noisy labels, and adversarial modifications [2], [3], [5]. These challenges are particularly important in digital forensic investigations, where detection results should be reliable, interpretable, and supported by meaningful evidence.

DeepFake detection alone cannot provide complete support for digital forensic investigation. A practical artificial-intelligence system should integrate video acquisition, preprocessing, face detection, frame extraction, spatial feature analysis, temporal inconsistency analysis, DeepFake classification, confidence estimation, and forensic evidence interpretation. This survey therefore aims to:

1. Review artificial-intelligence and deep-learning approaches for DeepFake video detection.
2. Examine spatial, temporal, multimodal, artifact-based, and attention-based detection techniques.
3. Analyze the capabilities and limitations of existing DeepFake detection methods.
4. Compare the robustness, generalization, and computational requirements of existing approaches.
5. Examine facial artifacts, spatial-temporal inconsistencies, and multimodal forensic indicators.
6. Formulate an integrated AI-based DeepFake detection framework for digital forensic investigations.
7. Identify evaluation requirements, research challenges, and future directions.

The remainder of this paper discusses existing detection approaches, the proposed comprehensive framework, forensic representation, spatial-temporal analysis, explainable decision support, evaluation requirements, research challenges, future directions, and concluding observations.

II. EXISTING METHODS FOR AI-BASED DEEPFAKE VIDEO DETECTION

Existing DeepFake video detection approaches can be broadly classified into manual and traditional verification, artifact-based detection, convolutional-neural-network-based detection, temporal and recurrent learning, multimodal detection, attention and transformer-based approaches, and robust or generalizable detection frameworks.

A. Manual and Traditional Video Verification

Traditional DeepFake verification depends on human visual inspection, video-source examination, metadata analysis, and identification of inconsistencies in facial appearance or video content. Investigators may examine facial boundaries, unnatural expressions, lighting variations, abnormal movements, and other visible characteristics to determine whether a video has been manipulated. Such approaches can provide useful initial evidence, but their effectiveness depends strongly on investigator expertise and the visibility of manipulation artifacts.

Highly realistic DeepFake videos can contain subtle modifications that are difficult to identify through visual inspection alone. Manual examination also becomes time-consuming when large numbers of videos must be investigated. Therefore, automated artificial-intelligence-based detection has become increasingly important for supporting digital forensic analysis.

B. Artifact-Based DeepFake Detection

Artifact-based approaches identify visual abnormalities introduced during the DeepFake generation process. These artifacts may include unnatural facial boundaries, blending errors, texture inconsistencies, lighting irregularities, and other manipulation traces. Artifact-based DeepFake detection attempts to learn these characteristics and distinguish manipulated content from authentic video [9].

Artifact analysis provides an important foundation for automated detection because manipulated videos can contain subtle visual traces even when their overall appearance appears realistic. However, the effectiveness of artifact-based detection can decrease when generation techniques improve or when post-processing operations such as compression and resizing modify the original manipulation traces.

C. Convolutional Neural Network-Based Detection

Convolutional neural networks have been extensively investigated for DeepFake detection because of their ability to automatically learn discriminative spatial features from facial images and video frames. CNN-based approaches can identify differences in facial texture, local regions, boundaries, and other visual characteristics associated with manipulated content [10], [12].

Deep-learning-based detection reduces dependence on manually designed features and supports automated classification of real and manipulated content. However, frame-level CNN approaches primarily concentrate on spatial information and may not fully capture inconsistencies occurring across consecutive frames. Their performance can also depend on the training dataset and manipulation techniques represented during training.

D. Temporal and Recurrent DeepFake Detection

DeepFake manipulation occurs within video sequences, making temporal information important for reliable detection. Temporal approaches analyze relationships between consecutive frames to identify inconsistencies in facial movement, expression changes, gaze behavior, and other sequence-level characteristics [1], [6], [14], [16].

Recurrent architectures can process sequential feature representations and identify temporal patterns associated with manipulated videos. Frame-level detection can also be improved by assigning greater importance to informative frames. Robust frame-level detection with lightweight inference weighting provides a mechanism for combining frame-level evidence for video-level classification [6]. Although temporal approaches provide additional information beyond individual-frame analysis, they may require greater computational resources and can be sensitive to video length and frame-sampling strategies.

E. Multimodal DeepFake Detection

DeepFake manipulation can affect multiple information channels, including facial appearance, speech, lip movements, and audio-visual synchronization. Multimodal detection methods therefore examine complementary information from audio and visual streams to improve detection reliability.

Content-driven audio-visual DeepFake analysis has been investigated for detecting manipulated content and locating forged segments within videos [7]. Multimodal approaches can provide stronger forensic evidence when visual and audio characteristics exhibit inconsistencies. However, multimodal detection requires processing multiple information streams and can increase computational complexity. Missing audio, background noise, poor recording quality, and compression can further affect detection reliability.

F. Attention and Transformer-Based Detection

Attention-based architectures provide mechanisms for identifying important spatial and temporal regions within video sequences. Instead of treating every frame or facial region equally, attention mechanisms can emphasize informative characteristics that contribute to DeepFake classification.

Recent approaches have investigated spatial-temporal attention and transformer-based video representations for DeepFake detection [15]. These approaches can model relationships across video sequences and provide an alternative to conventional sequential architectures. However, transformer-based models can require substantial computational resources and sufficiently diverse training data.

G. Robust and Generalizable DeepFake Detection

A major limitation of existing DeepFake detectors is their ability to generalize to unseen manipulation methods. A detector trained using one dataset or generation technique may perform well on familiar samples but experience reduced performance when exposed to different manipulation techniques. Generalizability has therefore become an important research objective [2].

DeepFake detection systems can also be affected by noisy training labels, adversarial perturbations, and real-world social-network transformations. Research has investigated detection under noisy-label attacks [3] and adversarial protection in real-world social-network scenarios [5]. These studies demonstrate that reliable DeepFake detection requires robustness in addition to conventional classification performance.

The literature therefore indicates a transition from simple artifact and frame-based detection toward spatial-temporal, multimodal, attention-based, robust, and generalizable DeepFake analysis. However, many existing approaches primarily emphasize classification rather than integration with a complete digital forensic workflow. A comprehensive architecture should therefore connect video acquisition, preprocessing, facial analysis, spatial feature extraction, temporal analysis, classification, confidence estimation, forensic evidence interpretation, and investigator-assisted decision support.

III. PROPOSED AI-POWERED DEEPFAKE DETECTION AND DIGITAL FORENSIC INVESTIGATION SYSTEM

The proposed system considers DeepFake detection as an integrated sequence of interconnected processing stages:

Video Input + Digital Evidence
↓
Video Acquisition
↓
Data Preprocessing
↓
Face Detection and Facial Region Extraction
↓
Frame Sampling and Sequence Formation
↓
ResNeXt50-Based Spatial Feature Extraction
↓
LSTM-Based Temporal Feature Analysis
↓
DeepFake Classification
↓
Confidence Estimation and Suspicious-Frame Identification
↓
Forensic Evidence Analysis
↓
Investigator Monitoring and Verification

The architecture consists of five major functional layers: video acquisition and preprocessing, facial and spatial analysis, temporal sequence analysis, DeepFake classification and confidence estimation, and forensic decision support. The first layer receives digital video evidence and prepares it for automated analysis. The preprocessing stage performs frame extraction, resizing, normalization, and other operations required to provide consistent input to the detection model.

The facial analysis layer identifies and extracts facial regions from selected video frames to reduce irrelevant background information. The spatial feature extraction layer uses a pre-trained ResNeXt50 network to learn discriminative representations from the extracted facial frames. These representations can capture subtle texture variations, facial artifacts, boundary inconsistencies, and other visual characteristics associated with manipulated content.

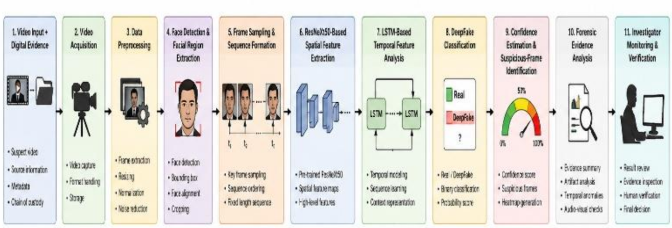

Fig. 1. Overall Architecture of the Proposed AI-Based DeepFake Detection and Digital Forensic System

The temporal analysis layer processes the sequence of spatial features using an LSTM network. This stage identifies relationships between consecutive frames and detects temporal inconsistencies that may not be apparent from individual frames. Facial movement, expression transitions, frame-to-frame feature variations, and other sequential characteristics can contribute to the final detection decision [1], [6], [14], [16]. The classification layer combines the learned spatial and temporal representations to determine whether the analyzed video is real or manipulated. A confidence value can accompany the classification result to indicate prediction reliability. Suspicious frames or facial regions can additionally be retained as supporting evidence for forensic examination.

Finally, the digital forensic decision-support layer presents the detection result, confidence information, suspicious frames, and relevant forensic observations through an accessible interface. The system can be implemented as a Flask-based web application that enables users or investigators to upload videos and obtain automated detection results. The proposed system is intended to assist forensic investigation rather than completely replace expert judgement.

IV. DEEPFAKE VIDEO DATA AND FORENSIC REPRESENTATION

A. DeepFake Video Acquisition and Processing

The proposed system considers digital videos as the primary input for DeepFake analysis. Videos may originate from online platforms, digital communication systems, repositories, surveillance sources, or other digital evidence collections. The input video is converted into a sequence of frames suitable for computational analysis.

Preprocessing handles variations in video resolution, frame dimensions, compression, and other input characteristics. Selected frames are sampled and prepared for facial analysis. Face detection is then applied to identify relevant facial regions. Cropping these regions reduces unnecessary background information and enables the model to concentrate on manipulation-related characteristics.

This stage is important because online videos may undergo compression, resizing, format conversion, and other transformations that can modify manipulation artifacts. Robust detection should therefore consider practical video conditions [5], [6].

B. Facial Region Representation

The facial region provides the primary visual representation for many DeepFake detection techniques. The proposed framework extracts facial regions from selected frames and provides them to the spatial feature extraction module.

The facial representation contains information related to skin texture, facial boundaries, eyes, mouth, expressions, lighting, and local visual artifacts. These characteristics can provide useful evidence for identifying manipulated content. Artifact-based research demonstrates the importance of analyzing visual abnormalities introduced during DeepFake generation [9].

Facial Action Unit-based analysis provides another perspective by examining facial movements and expression-related characteristics [13]. Gaze-related analysis can similarly identify inconsistencies in facial behavior across video sequences [1].

C. Spatial Feature Representation

After facial-region extraction, the proposed framework applies a pre-trained ResNeXt50 network to extract spatial features from individual frames. The network generates high-level representations of facial visual information that can support real-versus-fake classification.

Spatial features can capture differences in facial texture, blending regions, local distortions, illumination patterns, and other characteristics that distinguish manipulated frames from

ISSN: 3049-0952
www.jsetms.com

genuine frames. Deep-learning-based approaches demonstrate the effectiveness of learned visual representations for DeepFake detection [10], [12].

D. Temporal Sequence Representation

Individual frames provide only partial information about a video. DeepFake manipulation can produce inconsistencies that become visible only when consecutive frames are examined together. The proposed framework therefore processes extracted spatial features as an ordered temporal sequence.

The LSTM network learns relationships between successive frame representations and captures changes in facial expressions, movements, feature transitions, and other sequence-level patterns. Spatial-temporal DeepFake research demonstrates the importance of combining spatial and temporal information for video analysis [1], [6], [14], [16].

E. Forensic Evidence Representation

The proposed framework extends DeepFake classification toward digital forensic evidence analysis. Instead of providing only a real-or-fake decision, the system can retain the analyzed video, extracted facial frames, suspicious frames, confidence information, and relevant spatial-temporal observations.

This representation assists investigators in reviewing the evidence contributing to the automated decision. Multimodal analysis can further provide audio-visual evidence when both streams are available [7], [11], while source-anomaly analysis can provide additional forensic information [8].

V. SPATIAL-TEMPORAL DEEPFAKE DETECTION AND FORENSIC ANALYSIS

A central component of the proposed system is the combination of spatial and temporal DeepFake analysis. Conventional frame-based detection can identify visual artifacts within individual images, whereas temporal analysis examines how facial characteristics evolve across consecutive frames.

The spatial analysis stage uses ResNeXt50 to extract high-level facial representations from sampled frames. The temporal analysis stage subsequently uses LSTM to examine the sequence of extracted features and identify inconsistencies occurring over time.

The combined architecture enables the system to consider both the appearance of individual facial regions and their behavior across a video sequence. This is important because a manipulated video may contain visually convincing individual frames while still producing abnormal transitions between frames. Spatial-temporal gaze analysis, robust frame-level detection, CNN-RNN analysis, and temporal-spatial approaches demonstrate the importance of sequence-level information [1], [6], [14], [16].

Multimodal analysis can further extend this process by examining relationships between audio and visual information. A mismatch between speech, lip movement, facial behavior, or other audio-visual characteristics can provide additional evidence of manipulation [7], [11].

The proposed forensic analysis therefore considers multiple sources of evidence rather than relying exclusively on a single visual characteristic. The final detection output can include the predicted class, confidence level, suspicious frames, and relevant observations for investigator review.

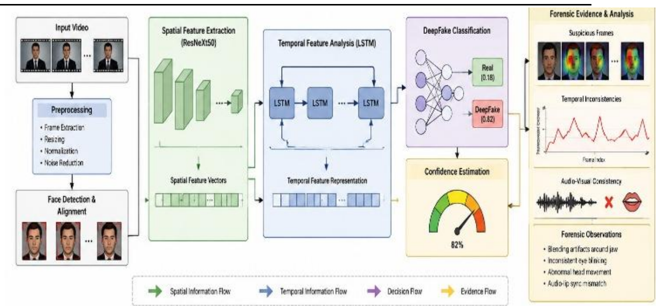

Fig. 2. Spatial-Temporal DeepFake Detection and Forensic Evidence Analysis

The framework is designed to support both automated classification and forensic interpretation. Automated classification provides rapid screening of potentially manipulated videos, while retained evidence enables investigators to examine suspicious content in greater detail. This integrated approach provides a foundation for connecting DeepFake detection with practical digital forensic investigation.

VI. EXPLAINABLE AI AND DIGITAL FORENSIC DECISION SUPPORT

AI-based DeepFake detection systems should provide more than a binary prediction. Digital forensic investigators require understandable evidence that can support interpretation of an automated result. Explainability can therefore improve transparency and assist experts in evaluating whether the prediction is consistent with observable characteristics.

For each analyzed video, the proposed system can provide the DeepFake classification, detection confidence, important suspicious frames, detected facial regions, spatial visual irregularities, temporal inconsistencies, audio-visual inconsistencies when audio is available, major characteristics contributing to the prediction, and forensic observations for investigator review.

The system can identify frames receiving strong evidence of manipulation and present them to investigators. This frame-level information can reduce the effort required to manually inspect an entire video. Gaze inconsistencies, multimodal face-forgery clues, and artifact-based visual abnormalities can provide additional interpretable forensic evidence [1], [4], [9].

The proposed system treats explainability as a decision-support component rather than a replacement for detection. A useful forensic report should distinguish between detection performance and forensic interpretability. A highly accurate model without meaningful evidence may be difficult to validate in practical investigations. The final forensic decision should therefore remain with the investigator, while AI provides automated analysis and evidence organization.

VII. ADAPTIVE DEEPFAKE INTELLIGENCE AND HUMAN-IN-THE-LOOP FORENSIC VERIFICATION

DeepFake generation methods continuously evolve through improved facial synthesis, audio manipulation, post-processing, compression, and adversarial strategies. Consequently, detectors trained on older datasets may experience reduced performance when exposed to unseen manipulation techniques. Generalizability is therefore an important research challenge [2].

The proposed framework should support periodic model evaluation and controlled updating. Newly observed DeepFake samples can be incorporated into future training and validation after appropriate verification. Model updates should be evaluated using both existing datasets and emerging manipulation techniques before deployment.

Robustness against noisy labels and adversarial manipulation should also be considered. Noisy-label attacks can affect detector learning, while adversarial perturbation research highlights the importance of real-world social-network scenarios [3], [5].

The human-in-the-loop workflow can therefore be organized as:

Video Acquisition → AI Detection → Suspicious Content Identification → Evidence Explanation → Investigator Verification → Forensic Interpretation → Final Investigation Decision

The investigator remains responsible for the final interpretation of digital evidence. AI-generated results can be combined with metadata, source information, chain-of-custody information, contextual evidence, and other forensic observations.

VIII. EVALUATION FRAMEWORK

A comprehensive evaluation framework is required to assess the proposed system across classification performance, temporal detection, robustness, generalization, explainability, efficiency, and forensic usefulness. Multiple complementary evaluation dimensions should be considered rather than relying on a single metric.

A. Classification Performance

The primary detection component can be evaluated using accuracy, precision, recall, F1-score, and AUC. These metrics measure overall correctness, detection reliability, identification capability, and balanced classification performance. Both false positives and false negatives are important in digital forensic applications [1]–[16].

B. Frame-Level and Video-Level Detection

Evaluation should be performed at both frame and video levels. Frame-level evaluation determines whether suspicious individual frames can be identified, while video-level evaluation determines whether evidence from multiple frames can produce a reliable final classification [6].

C. Spatial-Temporal Detection

The spatial-temporal component should be evaluated by comparing spatial-only detection with systems incorporating temporal information. Different video lengths, frame-sampling strategies, facial movements, and expression changes should be considered [14], [16].

D. Multimodal Detection

When audio is available, visual-only and audio-visual configurations can be compared. This evaluation determines whether complementary audio-visual information improves detection reliability and temporal localization [7], [11].

E. Generalization and Robustness

Generalization should be evaluated through cross-dataset testing, unseen manipulation methods, compressed videos, resized videos, social-media transformations, and different video qualities. Robustness should additionally consider noisy labels and adversarial perturbations [2], [3], [5].

F. Explainability and Forensic Usefulness

Explainability should be assessed according to the relevance, consistency, clarity, and usefulness of generated forensic evidence. Human evaluation can determine whether investigators or trained reviewers can understand and verify the automated results.

G. Computational Efficiency

Practical deployment requires evaluation of video-processing time, frame-processing rate, detection latency, memory consumption, and scalability. These measurements are important when large collections of digital videos must be screened.

TABLE II: EVALUATION DIMENSIONS OF THE PROPOSED DEEPFAKE FORENSIC FRAMEWORK

Dimension	Evaluation Metrics / Criteria	Purpose
Classification	Accuracy, Precision, Recall, F1-Score, AUC	Evaluate DeepFake identification
Frame-Level Detection	Frame accuracy, suspicious-frame identification	Identify manipulated portions
Video-Level Detection	Video accuracy, precision, recall, F1-score	Evaluate complete-video classification
Spatial-Temporal Analysis	Spatial-only vs. spatial-temporal comparison	Assess temporal feature contribution
Multimodal Detection	Visual-only vs. audio-visual comparison	Assess complementary evidence
Generalization	Cross-dataset and unseen-manipulation testing	Evaluate detector generalization
Robustness	Compression, resizing, noise, adversarial conditions	Assess reliability
Explainability	Relevance, consistency, clarity, human assessment	Evaluate forensic interpretability
Efficiency	Processing time, latency, memory, throughput	Assess computational practicality
Forensic Utility	Evidence identification, investigator assessment	Determine investigation usefulness

IX. RESEARCH CHALLENGES AND FUTURE ROADMAP

Although AI-powered DeepFake detection can improve identification of manipulated video content, several challenges remain in practical digital forensic investigations.

A. Dataset Availability and Diversity

DeepFake datasets may contain limited manipulation methods, controlled recording conditions, and restricted diversity. Future datasets should include diverse identities, environments, recording devices, compression levels, manipulation techniques, and video qualities.

B. Generalization to Unseen DeepFakes

DeepFake generation techniques continuously evolve. Detectors trained on known manipulation methods may fail against previously unseen techniques. Improving cross-dataset and cross-manipulation generalization is therefore a major research direction [2]. Future systems should focus on manipulation-independent forensic characteristics.

C. Adversarial Manipulation

Attackers may intentionally modify manipulated videos to avoid detection. Adversarial perturbations, post-processing, and compression can weaken forensic traces. Future systems should incorporate adversarially resistant representations and robustness testing [5].

D. Noisy and Unreliable Training Data

Incorrect labels, ambiguous samples, and variations in manipulation quality can negatively influence model learning. Reliable dataset construction, sample verification, noise-resistant learning, and confidence-aware training are therefore important [3].

E. Temporal and Multimodal Complexity

Reliable DeepFake detection requires analysis beyond individual frames. Temporal inconsistencies and audio-visual relationships provide valuable evidence, but processing multiple frames and modalities increases computational requirements [7], [14]–[16].

F. Explainability and Forensic Reliability

AI predictions should be accompanied by understandable evidence when used in forensic investigations. Future research should improve suspicious-frame localization, facial-region analysis, interpretable evidence generation, and investigator-oriented explanations.

G. Social-Media Compression and Real-World Conditions

Videos shared through social networks may undergo compression, resizing, frame-rate changes, and format conversion. These transformations can modify artifacts used by detection models. Future systems should therefore evaluate multiple levels of compression and post-processing [5].

H. Human–AI Forensic Interaction

AI should assist investigators rather than completely replace expert judgement. Interactive interfaces should allow investigators to inspect suspicious frames, review confidence information, compare evidence, and validate automated findings.

I. Privacy and Security

DeepFake investigations can involve sensitive digital videos. Secure storage, controlled access, evidence integrity, appropriate anonymization, and secure processing are therefore important requirements for practical deployment.

X. CONCLUSION

This paper presented an artificial-intelligence framework for DeepFake video detection and digital forensic investigation. The framework integrates video preprocessing, face detection, spatial feature extraction, temporal sequence analysis, DeepFake classification, confidence estimation, forensic evidence analysis, explainability, and human-in-the-loop decision support. The combination of ResNeXt50-based spatial representation and LSTM-based temporal analysis enables examination of both facial characteristics and temporal inconsistencies within manipulated videos. The survey further highlights multimodal analysis, robustness, generalization, and interpretable forensic evidence as important requirements for practical deployment. Future research should focus on unseen DeepFake generation methods, adversarial robustness, diverse datasets, efficient spatial-temporal architectures, explainable detection, privacy-preserving processing, and reliable AI-assisted digital forensic investigation.

REFERENCES

[1] C. Peng, Z. Miao, D. Liu, N. Wang, R. Hu, and X. Gao, “Where Deepfakes Gaze at? Spatial-Temporal Gaze Inconsistency Analysis for Video Face Forgery Detection,” *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 4507–4517, 2024, doi: 10.1109/TIFS.2024.3381823.

[2] K. Zhang, Z. Hou, Z. Hua, Y. Zheng, and L. Y. Zhang, “Boosting Deepfake Detection Generalizability via Expansive Learning and Confidence Judgement,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 1, pp. 953–966, 2025, doi: 10.1109/TCSVT.2024.3462985.

[3] T. Qiao, S. Xie, Y. Chen, F. Retraint, R. Shi, and X. Luo, “Deepfake Detection Fighting Against Noisy Label Attack,” *IEEE Transactions on Multimedia*, vol. 26, pp. 9047–9059, 2024, doi: 10.1109/TMM.2024.3385286.

[4] D. Liu, Z. Zheng, C. Peng, Y. Wang, N. Wang, and X. Gao, “Hierarchical Forgery Classifier on Multi-Modality Face Forgery Clues,” *IEEE Transactions on Multimedia*, vol. 26, pp. 2894–2905, 2024, doi: 10.1109/TMM.2023.3304913.

[5] Z. Qu, Z. Xi, W. Lu, X. Luo, Q. Wang, and B. Li, “DF-RAP: A Robust Adversarial Perturbation for Defending Against Deepfakes in Real-World Social Network Scenarios,” *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 3943–3957, 2024.

[6] L. Zhou, C. Ma, Z. Wang, Y. Zhang, X. Shi, and L. Wu, “Robust Frame-Level Detection for Deepfake Videos With Lightweight Bayesian Inference Weighting,” *IEEE Internet of Things Journal*, vol. 11, no. 7, pp. 13018–13028, 2024, doi: 10.1109/JIOT.2023.3337128.

[7] Z. Cai, K. Stefanov, A. Dhall, and M. Hayat, “Do You Really Mean That? Content Driven Audio-Visual Deepfake Dataset and Multimodal Method for Temporal Forgery Localization,” in *Proc. IEEE Int. Conf. Digital Image Computing: Techniques and Applications (DICTA)*, 2022, doi: 10.1109/DICTA56598.2022.10034605.

[8] Y. Wang and G. Liao, “Deepfake Video Detection Based on Image Source Anomaly,” in *Proc. IEEE 2nd Int. Conf. Image Processing and Computer Applications (ICIPCA)*, 2024, doi: 10.1109/ICIPCA61593.2024.10709022.

[9] P. Preeti and S. Bansal, “Artifact Based Deepfake Detection Methods,” in *Proc. 2023 2nd Int. Conf. Informatics (ICI)*, 2023, doi: 10.1109/ICI60088.2023.10421629.

[10] “DeepFake Video Detection,” in *Proc. IEEE Conference Publication*, 2022, doi: 10.1109/ICCCNT54827.2022.9984438.

[11] “A Deepfake Video Detection Method Based on Multi-Modal Deep Learning Method,” in *Proc. IEEE Int. Conf. Artificial Intelligence and Computer Applications (ICAICA)*, 2022, doi: 10.1109/ICAICA54496.2022.9742162.

[12] “Deep Learning Based DeepFake Video Detection,” in *Proc. IEEE 7th Int. Conf. Information Technology and Digital Applications (ICITDA)*, 2023, doi: 10.1109/ICITDA60866.2023.10087584.

[13] Q. Jaleel and I. Hadi, “Facial Action Unit-Based Deepfake Video Detection Using Deep Learning,” in *Proc. 2022 4th Int. Conf. Current Research in Engineering and Science Applications (ICCRESA)*, 2022, doi: 10.1109/ICCRESA57091.2022.10352085.

[14] E. M. S. Reddy, A. P. Kumar, and P. Swetha, “Deepfake Video Detection Using CNN and RNN With Optical Flow Features,” in *Proc. 2024 IEEE Int. Students' Conf. Electrical, Electronics and Computer Science (SCEECS)*, 2024, doi: 10.1109/SCEECS61402.2024.10482344.

[15] G. Sun and Z. Lian, “Deepfake Video Detection Based on the Decomposition of Spatial-Temporal Attention Mechanism in ViViT,” in *Proc. IEEE Int. Symp. Parallel and Distributed Processing with Applications (ISPA)*, 2024, pp. 1629–1634.

[16] F. Zafar, T. A. Khan, S. Akbar, M. T. Ubaid, S. Javaid, and K. A. Kadir, “A Hybrid Deep Learning Framework for Deepfake Detection Using Temporal and Spatial Features,” *IEEE Access*, vol. 13, pp. 79560–79570, 2025, doi: 10.1109/ACCESS.2025.3566008.