Taylor & Francis
Taylor & Francis Group

Automatika

Journal for Control, Measurement, Electronics, Computing and Communications

ISSN: 0005-1144 (Print) 1848-3380 (Online) Journal homepage: www.tandfonline.com/journals/taut20

Securing online integrity: a hybrid approach to deepfake detection and removal using Explainable AI and Adversarial Robustness Training

R. Uma Maheshwari & B. Paulchamy

To cite this article: R. Uma Maheshwari & B. Paulchamy (2024) Securing online integrity: a hybrid approach to deepfake detection and removal using Explainable AI and Adversarial Robustness Training, Automatika, 65:4, 1517-1532, DOI: 10.1080/00051144.2024.2400640

To link to this article: https://doi.org/10.1080/00051144.2024.2400640

Published online: 10 Sep 2024.

AUTOMATIKA
2024, VOL. 65, NO. 4, 1517–1532
https://doi.org/10.1080/00051144.2024.2400640

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Securing online integrity: a hybrid approach to deepfake detection and removal using Explainable AI and Adversarial Robustness Training

R. Uma Maheshwari and B. Paulchamy

Department of Electronics and Communication Engineering, Hindusthan Institute of Technology, Coimbatore, India

ABSTRACT

As deepfake technology becomes increasingly sophisticated, the proliferation of manipulated images presents a significant threat to online integrity, requiring advanced detection and mitigation strategies. Addressing this critical challenge, our study introduces a pioneering approach that integrates Explainable AI (XAI) with Adversarial Robustness Training (ART) to enhance the detection and removal of deepfake content. The proposed methodology, termed XAI-ART, begins with the creation of a diverse dataset that includes both authentic and manipulated images, followed by comprehensive preprocessing and augmentation. We then employ Adversarial Robustness Training to fortify the deep learning model against adversarial manipulations. By incorporating Explainable AI techniques, our approach not only improves detection accuracy but also provides transparency in model decision-making, offering clear insights into how deepfake content is identified. Our experimental results underscore the effectiveness of XAI-ART, with the model achieving an impressive accuracy of 97.5% in distinguishing between genuine and manipulated images. The recall rate of 96.8% indicates that our model effectively captures the majority of deepfake instances, while the F1-Score of 97.5% demonstrates a well-balanced performance in precision and recall. Importantly, the model maintains high robustness against adversarial attacks, with a minimal accuracy reduction to 96.7% under perturbations.

ARTICLE HISTORY
Received 15 April 2024
Accepted 29 August 2024

KEYWORDS
Adversarial attacks; Adversarial Robustness Training (ART); deepfake technology; detection strategies; digital content integrity; Explainable AI (XAI)

1. Introduction

Deepfake technology represents a significant advancement in artificial intelligence and machine learning, enabling the creation of hyper-realistic manipulated videos [1] and images. Utilizing sophisticated algorithms, deepfake technology can seamlessly superimpose one person's face onto another's, manipulate speech, and even generate entirely synthetic content that appears authentic [1]. While initially gaining attention for its novelty and entertainment value, deepfake technology has raised profound concerns due to its potential for misuse and exploitation. In recent years, deepfakes have been increasingly deployed to spread misinformation, defame individuals, and manipulate public opinion, posing significant threats to privacy, trust, and societal stability.

The proliferation of deepfake content has prompted widespread calls for improved detection and mitigation strategies. Researchers and technologists are actively developing tools and techniques to identify and combat the spread of manipulated media. From advanced machine learning algorithms to collaborative efforts between industry, academia, and policymakers, the fight against deepfakes encompasses a multifaceted approach. Moreover, raising public awareness about the existence and implications of deepfake technology is crucial for fostering digital literacy and resilience in the face of emerging threats to online integrity and trust.

From Figure 1, Explainable AI (XAI) has emerged as a vital component in the development and deployment of artificial intelligence systems, particularly in contexts such as deepfake detection. XAI techniques [2] are designed to provide human-interpretable explanations for the decisions made by AI models, thereby enhancing transparency and trust in their capabilities. By employing methods such as attention maps, feature visualization, and model-agnostic explanations, XAI enables stakeholders [3] to gain insights into the underlying reasoning processes of AI systems. This transparency not only fosters a deeper understanding of how AI models operate but also facilitates the identification of biases, errors, and vulnerabilities. In the context of deepfake detection, XAI can help researchers, developers, and end-users better comprehend why a particular image or video is flagged as a deepfake, thereby bolstering confidence in the reliability and effectiveness of detection systems. Overall, the integration of XAI into AI systems holds promise for improving accountability, mitigating risks, and promoting responsible AI development and deployment.

Adversarial Robustness Training (ART) [4] has emerged as a cornerstone strategy for reinforcing

CONTACT R. Uma Maheshwari umamaheshwari@hit.edu.in Hindusthan Institute of Technology, Coimbatore, India

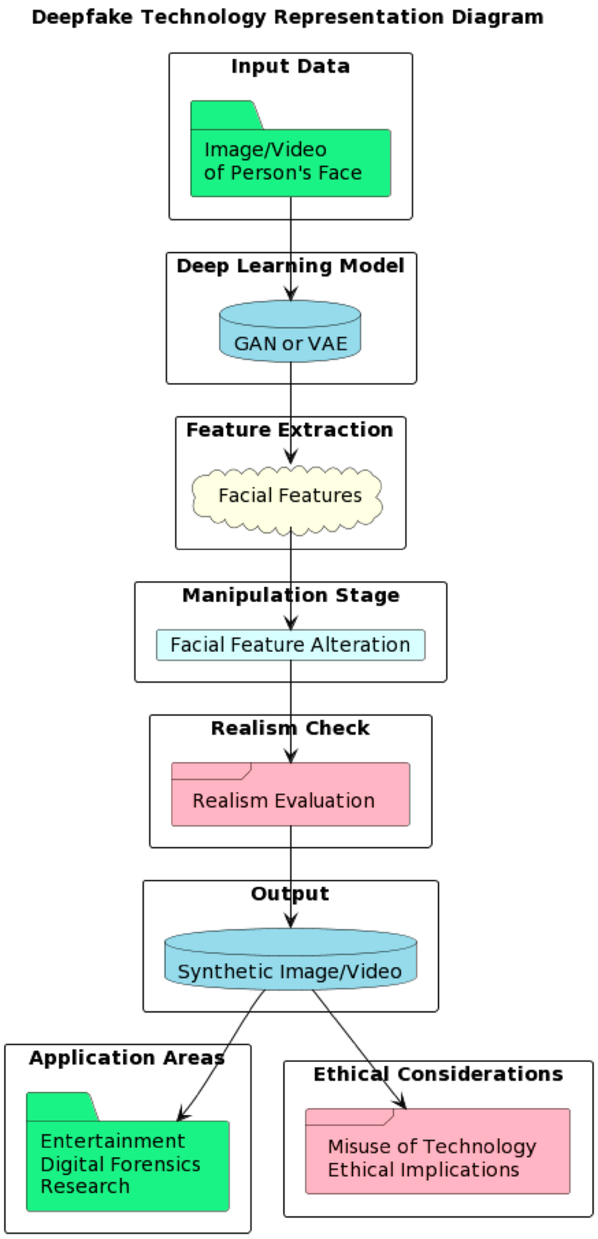

Figure 1. Deepfake detection techniques.

machine learning models against adversarial attacks, particularly those directed at deepfake detection systems. This approach involves subjecting models to adversarial examples during the training phase, thereby enhancing their ability to withstand manipulation attempts while improving overall resilience and generalization performance. By iteratively exposing models to adversarial perturbations [5] and adjusting their parameters accordingly, ART aims to bolster their ability to accurately classify and differentiate between authentic and manipulated content. This proactive approach not only strengthens the robustness of machine learning models but also equips them with the capability to effectively detect and mitigate the proliferation of deepfake media. As adversaries continue to evolve their tactics, ART stands as a pivotal defense mechanism, fortifying machine learning systems and enhancing their effectiveness in combating the ever-growing threat posed by manipulated content.

Furthermore, Explainable AI (XAI) [6] techniques play a crucial role in deepfake detection by providing interpretable explanations for the decisions made by detection models. This enhances transparency and trust in the detection process, allowing stakeholders to understand why certain media is flagged as a deepfake.

As deepfake technology [7] continues to advance, the development of robust detection methods remains an ongoing challenge. Researchers and technologists are continuously refining existing techniques and exploring innovative approaches to stay ahead of evolving deepfake creation methods. Ultimately, effective deepfake detection is essential for preserving the integrity of digital content and mitigating the potential societal impacts of manipulated media.

This manuscript presents novel insights and methodologies in the field of deepfake detection. By integrating Explainable AI (XAI) [8] techniques with Adversarial Robustness Training (ART), the research introduces an innovative approach to enhancing the resilience and interpretability of deepfake detection systems. This novel combination of methodologies offers a unique perspective on addressing the challenges posed by deepfake technology, contributing to the advancement of knowledge in the field. Furthermore, the manuscript explores cutting-edge techniques and strategies for detecting deepfakes, pushing the boundaries of current research and paving the way for future developments in the field.

The organization of paper is as follows; section 2 includes background study and analysis of Existing work; section 3 includes design and methodology of proposed work; section 4 includes experimental analysis and results; section 5 includes conclusion and future work.

2. Background and significance of the research

Artificial intelligence (AI) has transformed numerous fields by providing advanced solutions and enhancing efficiency. However, as AI applications grow, ensuring their security, trustworthiness, and robustness becomes increasingly important. The emergence of deepfake technology, capable of generating hyper-realistic synthetic content, poses significant challenges in media authenticity and security. Concurrently, the integration of AI in extended reality (XR) for metaverses introduces new dimensions of interaction but also raises concerns about privacy and ethical implications. The proliferation of adversarial attacks against AI systems highlights the need for resilient and robust defense to maintain the integrity and reliability of these systems. Addressing these multifaceted challenges requires comprehensive frameworks and innovative methodologies.

Bale et al. [9] conducted a comprehensive review on deepfake detection and classification in their paper "Deepfake Detection and Classification of Images from Video: A Review of Features, Techniques, and Challenges". The study outlines the various features and techniques utilized in identifying deepfake images and videos. The review categorizes detection methods into traditional machine learning, deep learning, and hybrid approaches, emphasizing the advantages and limitations of each. They highlight critical challenges such as the evolving sophistication of deepfake generation techniques, the need for large and diverse datasets, and the requirement for real-time detection capabilities. The authors propose future research directions, including the development of more robust and generalizable detection models and the integration of explainable AI to enhance trustworthiness.

Polemi et al. [10] discuss the management of AI trustworthiness risks in their paper "Challenges and efforts in managing AI trustworthiness risks: a state of knowledge". Published in Frontiers in Big Data, this study presents a detailed examination of the challenges in ensuring AI systems' trustworthiness. The authors explore various risk management frameworks and strategies, focusing on transparency, accountability, and ethical considerations. They provide a taxonomy of trustworthiness risks, including bias, robustness, and interpretability, and discuss efforts to mitigate these risks through regulatory frameworks, technical solutions, and collaborative initiatives. The paper concludes with a call for ongoing research and international cooperation to address the complex and evolving landscape of AI trustworthiness.

El-Shafai et al. [11] provide a comprehensive taxonomy of multimedia video forgery detection techniques in their paper "A comprehensive taxonomy on multimedia video forgery detection techniques: challenges and novel trends". This study, published in Multimedia Tools and Applications, categorizes existing detection methods into spatial, temporal, and hybrid techniques. The authors discuss the unique challenges posed by each type of forgery, such as frame insertion, deletion, and tampering, and evaluate the effectiveness of various detection algorithms. Novel trends in the field, including the use of deep learning and blockchain technology for enhanced security and reliability, are also highlighted. The paper emphasizes the need for interdisciplinary approaches and advanced computational techniques to address the growing complexity of multimedia forgery.

Qayyum et al. [12] address the integration of secure and trustworthy AI in extended reality (AI-XR) for metaverses in their paper "Secure and trustworthy artificial intelligence-extended reality (AI-XR) for metaverses". Published in ACM Computing Surveys, this comprehensive review explores the intersection of AI, XR, and cybersecurity. The authors discuss the potential risks associated with AI-XR applications, such as privacy violations, data breaches, and malicious attacks, and propose strategies for enhancing security and trustworthiness. They highlight the importance of user-centric design, ethical AI practices, and robust regulatory frameworks in building secure AI-XR systems. The paper concludes with recommendations for future research, including the development of standardized protocols and collaborative efforts to address the multifaceted challenges of AI-XR in metaverses.

Nawaz et al. [13] present a deep learning approach for deepfake detection in their paper "ResNet-Swish-Dense54: a deep learning approach for deepfakes detection", published in The Visual Computer. The authors propose a novel neural network architecture that combines the strengths of ResNet, Swish activation function, and DenseNet layers to achieve high accuracy in detecting deepfakes. Their experimental results demonstrate the model's effectiveness in identifying manipulated videos, even in challenging scenarios with low-resolution and compressed images. The study highlights the importance of leveraging advanced deep learning techniques and fine-tuning network parameters to improve detection performance. Future research directions include expanding the dataset to include more diverse deepfake examples and exploring real-time detection capabilities.

Habbal et al. [14] explore frameworks, applications, challenges, and future research directions in AI trust, risk, and security management (AI TRISM) in their paper published in Expert Systems with Applications. The study provides an in-depth analysis of existing frameworks for managing AI-related risks, emphasizing the need for comprehensive approaches that address both technical and ethical dimensions. The authors discuss various applications of AI TRISM in different sectors, such as healthcare, finance, and transportation, and highlight the challenges in implementing effective risk management strategies. They propose future research areas, including the development of standardized metrics for evaluating AI trustworthiness and the integration of AI TRISM frameworks into organizational policies and practices.

Reddy et al. [15] propose a deep learning approach for deepfake video detection using CNN and RNN with optical flow features in their paper presented at the IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). The authors combine convolutional neural networks (CNN) for feature extraction and recurrent neural networks (RNN) for temporal analysis to identify deepfake videos. Their methodology leverages optical flow features to capture motion inconsistencies between frames, enhancing the detection accuracy. The experimental results demonstrate the model's robustness in detecting various types of deepfake manipulations. The paper suggests future research directions, including the exploration of more sophisticated temporal features and the integration of multimodal data for improved detection performance.

Andrade et al. [16] conduct a systematic mapping study and taxonomy on adversarial attacks and defenses in person search in their paper published in Image and Vision Computing. The study provides a comprehensive overview of the types of adversarial attacks that can compromise person search systems, such as evasion attacks and poisoning attacks. The authors also review existing defense mechanisms, including adversarial training, model ensembling, and input transformation techniques. They propose a taxonomy to classify the various attack and defense strategies and highlight the need for more resilient person search systems. Future research directions include the development of adaptive defense mechanisms that can dynamically respond to evolving adversarial threats and the integration of explainable AI to enhance system transparency and trustworthiness.

Aruna and Narayan [17] investigate the detection of GAN-manipulated medical images through deep learning techniques in their paper presented at the International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE). The authors propose a deep learning framework that leverages convolutional neural networks (CNN) to detect generative adversarial network (GAN) manipulations in medical images. Their approach focuses on identifying subtle artifacts introduced by GANs, which are often difficult to detect using traditional image analysis techniques. The experimental results show high accuracy in distinguishing between real and manipulated medical images. The paper highlights the potential of deep learning techniques in enhancing the reliability of medical image analysis and suggests future research directions, including the development of more robust models and the integration of multimodal data for comprehensive detection.

Moskalenko et al. [18] present a taxonomy, models, and methods for resilience and resilient systems of artificial intelligence in their paper published in Algorithms. The study categorizes various resilience strategies for AI systems, including fault tolerance, robustness, and adaptability. The authors propose models and methods for enhancing AI resilience, focusing on techniques such as redundancy, diversity, and self-healing. They discuss the challenges in implementing resilient AI systems, such as the trade-offs between performance and resilience and the need for continuous monitoring and adaptation. The paper concludes with recommendations for future research, including the development of standardized resilience metrics and the integration of resilience engineering principles into AI system design and deployment.

Deepfake technology, which involves the manipulation of multimedia [19] content to create highly realistic but fraudulent media, poses significant risks to online environments. The proliferation of deepfakes has raised concerns regarding misinformation, identity theft, and privacy violations. Despite the advancements in deepfake detection [20] methods, the increasing sophistication of these techniques [21] makes it challenging to develop robust and reliable detection systems. Existing solutions often lack resilience against adversarial attacks and may not offer sufficient interpretability to understand the reasons behind detection decisions. Therefore, there is a pressing need for innovative approaches that enhance the robustness, accuracy, and explainability of deepfake detection systems to safeguard digital content.

This research introduces a novel framework for deepfake detection that integrates adversarial robustness training [22] with Explainable AI (XAI) techniques to address the limitations of current methods. The proposed approach includes a comprehensive pipeline involving dataset collection, preprocessing, data augmentation, and adversarial robustness training. It leverages advanced XAI techniques, such as SHAP and LIME, to provide insights into the model's decision-making process, thereby improving transparency and trustworthiness. Key contributions of this research include the development of a robust training model resistant to adversarial perturbations, the application of XAI for enhanced interpretability, and the incorporation of real-time monitoring mechanisms for continuous threat detection. The framework is evaluated using metrics such as accuracy, precision, recall, and F1-score, and its effectiveness in providing both robust and explainable deepfake detection is demonstrated through extensive experimental validation.

3. Methodology: integrating Explainable AI with Adversarial Robustness Training (XAI-ART) for deepfake detection

Deepfake detection refers to the process of identifying and distinguishing between authentic and manipulated media content, typically involving images, videos, or audio recordings [6]. As the sophistication of deepfake technology increases, detecting these falsified media becomes increasingly challenging. The overall processing involves a systematic approach to deepfake detection by integrating Explainable AI (XAI) with Adversarial Robustness Training (ART) [23]. Initially, a diverse dataset comprising both authentic and manipulated images is curated, ensuring representation of various manipulation techniques and contexts.

The dataset undergoes rigorous preprocessing and augmentation to enhance its quality and diversity. Subsequently, a deep learning model is trained using

Adversarial Robustness Training to improve its resilience against adversarial attacks commonly employed to evade detection. Concurrently, Explainable AI techniques [7] are integrated to facilitate the interpretation of model decisions, providing insights into why certain images are classified as deepfakes. The trained model is then deployed for real-time monitoring of online sources, flagging content for human review and subsequent removal if identified as deepfake material. Experimental evaluation is conducted to assess the efficacy of the proposed XAI-ART approach in accurately detecting and removing deepfake content, even in the presence of sophisticated adversarial manipulations. Through this comprehensive processing pipeline, the research aims to advance the state-of-the-art in deepfake detection and contribute to the creation of a safer and more trustworthy online environment.

3.1. Dataset collection and pre-processing

The Deepfake Detection Dataset (DFDD) employed in our study encompasses a comprehensive collection of multimedia content aimed at facilitating the development and evaluation of deepfake detection algorithms. Comprising a total of 20,000 videos, the dataset is meticulously curated to incorporate a diverse array of authentic and manipulated content sourced from a multitude of platforms and creators. Split each dataset of 5,000 videos and set as dataset 1, dataset 2.

1. Real Videos (10,000): These videos represent authentic recordings devoid of any manipulation or alteration. Sourced from reputable sources and databases, real videos encompass a broad spectrum of scenes, contexts, and subjects, ensuring the dataset's fidelity to real-world scenarios.
2. Deepfake Videos (10,000): This subset comprises videos that have undergone various forms of manipulation utilizing deep learning techniques, resulting in the generation of synthetic content aimed at mimicking real footage. Deepfake videos encompass a wide range of alterations, including facial reenactment, lip-syncing, and voice cloning, among others.

The dataset comprises manipulated images created using deepfake techniques. These images may include faces of individuals that have been altered to appear as if they are performing actions or expressions they did not actually do. The Deepfake Detection Challenge provides the following datasets:

1. **Training Set**: This dataset is used by competitors to build their deepfake detection models. It contains labels for the target and is broken up into 50 files for ease of access and download. Due to its large size, it must be accessed through a Google Cloud Storage (GCS) [24] bucket, which is made available to participants after accepting the competition's rules. The rules contain important details about the dataset's permitted use. Competitors are encouraged to train their models outside of Kaggle's notebooks environment and submit the trained model as an external data source.
2. **Public Validation Set**: This dataset, available on the Kaggle Data page as test_videos.zip, consists of a small set of 400 videos or IDs. When competitors commit their Kaggle notebooks, the submission file output generated is based on this public validation set. It serves as a means for competitors to validate the performance of their models before final submission.

Non-local Means Denoising (NLM) is a sophisticated denoising technique that operates by computing weighted averages of pixel values across the entire image, rather than just within local neighbourhoods. The algorithm involves several computational steps, but we can outline the key concepts and equations involved:

3.1.1. Patch similarity calculation

The first step in NLM involves calculating the similarity between patches of pixels [12] across the entire image. Given a reference patch P_i centred at pixel i and a comparison patch P_j centred at pixel j, the similarity between these patches can be measured using the sum of squared intensity differences:

$$d(P_i, P_j) = \sum_{k \in \Omega} (I(i+k) - I(j+k))^2 \qquad (1)$$

Where:

- P_i and P_j are patches of pixels centred at pixels i and j respectively.
- Ω represents the set of pixels within the patch.

$I(i+k)$ and $I(j+k)$ denote the intensity values of pixels in the reference and comparison patches.

3.1.2. Weighted average calculation

Once the similarities between patches are computed, the next step is to calculate the weighted average of pixel values across the entire image. The weight assigned to each pixel value depends on the similarity between the corresponding patches:

$$\hat{I}(i) = \frac{1}{Z(i)} \sum_{j \in \mathcal{N}(i)} I(j) \cdot w(i,j) \qquad (2)$$

Where:

- $\hat{I}(i)$ represents the denoised pixel value at position i.
- $\mathcal{N}(i)$ denotes the neighbourhood of pixel i.

- $w(i,j)$ is the weight assigned to the pixel value at position j based on the similarity between patches centred at pixels i and j.
- $Z(i)$ is a normalization factor to ensure that the weights sum up to 1.

3.1.3. Normalization factor calculation

The normalization factor $Z(i)$ is calculated as the sum of weights for all pixels in the neighbourhood of pixel i:

$$Z(i) = \sum_{j \in \mathcal{N}(i)} w(i,j) \tag{3}$$

This ensures that the weighted average of pixel values is properly normalized. The Non-local Means Denoising algorithm involves iteratively computing patch similarities and weighted averages of pixel values across the entire image to generate a denoised version of the input image. This process effectively removes noise while preserving image details and structures.

3.2. Explainable AI (XAI) integration

The core of our model is a convolutional neural network (CNN) designed for image classification tasks. During the training phase, the model is exposed to both regular and adversarial examples generated using techniques like Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) [9].

This process involves creating adversarial examples by perturbing the input images and including these perturbed images in the training set. The model is then trained to correctly classify both the original and adversarial images.

Post-training, we use XAI techniques such as SHAP and LIME to interpret the model's decisions. These techniques analyze the model's predictions and provide visual and textual explanations for why an image is classified as deepfake or authentic.

Explainable AI (XAI) techniques are essential for providing insights into the decisions made by AI models, enhancing transparency, and trust in their predictions as shown in Figure 2. Integrating XAI with deep learning models for tasks such as deepfake detection can help elucidate the rationale behind model decisions, enabling stakeholders to better understand the detection process. One approach to integrating XAI with deep learning models is through the use of attention mechanisms (Figure 3).

3.2.1. Attention mechanism

The attention mechanism allows the model to focus on relevant parts of the input data while making predictions. In the context of deepfake detection, the attention mechanism can highlight regions of the image that are indicative of manipulation or inconsistencies. The equation for attention mechanism can be represented as follows:

$$\alpha_i = \text{softmax}(f(x_i)) \tag{4}$$

Where:

- α_i represents the attention weight assigned to input feature x_i.
- $f(\cdot)$ is a function that computes the relevance score for each input feature.
- $\text{softmax}(\cdot)$ is a normalization function that converts relevance scores into attention weights, ensuring that they sum up to 1.

The attention score between a query q_i and a key k_j is calculated as the dot product of their representations, scaled by the square root of the dimensionality of the query vector:

$$\text{Attention}(q_i, k_j) = \frac{q_i \cdot k_j}{\sqrt{d}} \tag{5}$$

Where:

- q_i and k_j are the query and key vectors respectively.
- d is the dimensionality of the query and key vectors. This attention score represents the relevance of the key k_j to the query q_i.

The attention scores are then passed through a softmax function to obtain attention weights that sum up to 1:

$$\text{Attention\_Weights}(q_i, K) = \text{softmax}(\text{Attention}(q_i, K)) \tag{6}$$

Where:

- K represents the set of all keys in the sequence. The softmax function normalizes the attention scores, ensuring that they form a probability distribution over the keys.

Attention maps generated by deep learning models trained for deepfake detection might emphasize areas of the face where manipulation is most likely to occur. These regions could include the eyes, mouth, or boundaries between different facial components, as these areas are often targeted by deepfake algorithms due to their significance in conveying emotions and expressions. By concentrating attention on these key facial regions, attention maps provide valuable insights into the subtle alterations made by deepfake techniques.

Moreover, attention maps can reveal inconsistencies in facial textures, lighting conditions, or spatial relationships that may indicate tampering. For instance, discrepancies in skin texture, unnatural lighting effects, or misalignments between facial features can be flagged by attention maps as potential signs of manipulation.

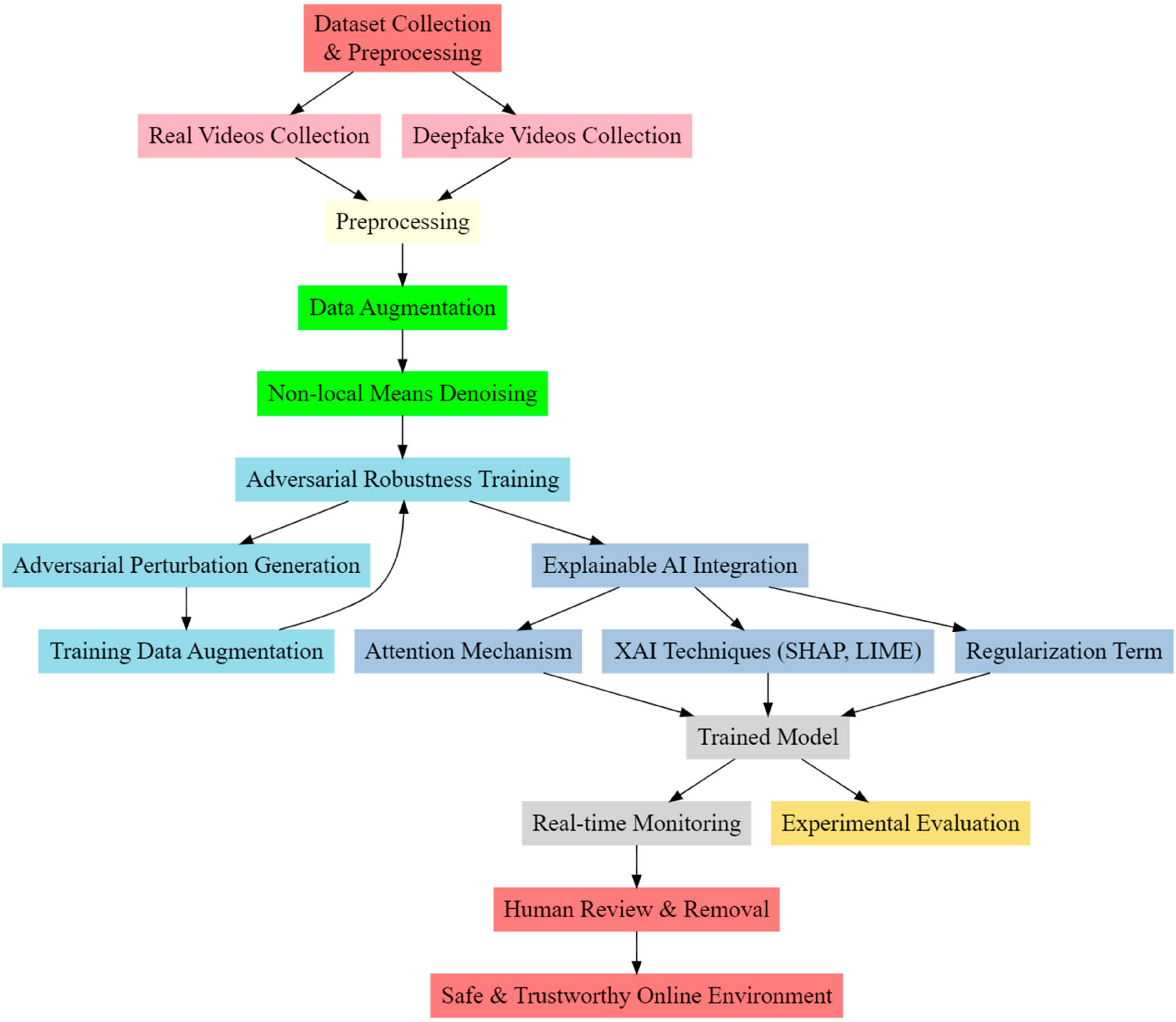
Dataset Collection & Preprocessing
Real Videos Collection
Deepfake Videos Collection
Preprocessing
Data Augmentation
Non-local Means Denoising
Adversarial Robustness Training
Adversarial Perturbation Generation
Explainable AI Integration
Training Data Augmentation
Attention Mechanism
XAI Techniques (SHAP, LIME)
Regularization Term
Trained Model
Real-time Monitoring
Experimental Evaluation
Human Review & Removal
Safe & Trustworthy Online Environment

Figure 2. Block diagram of proposed work.

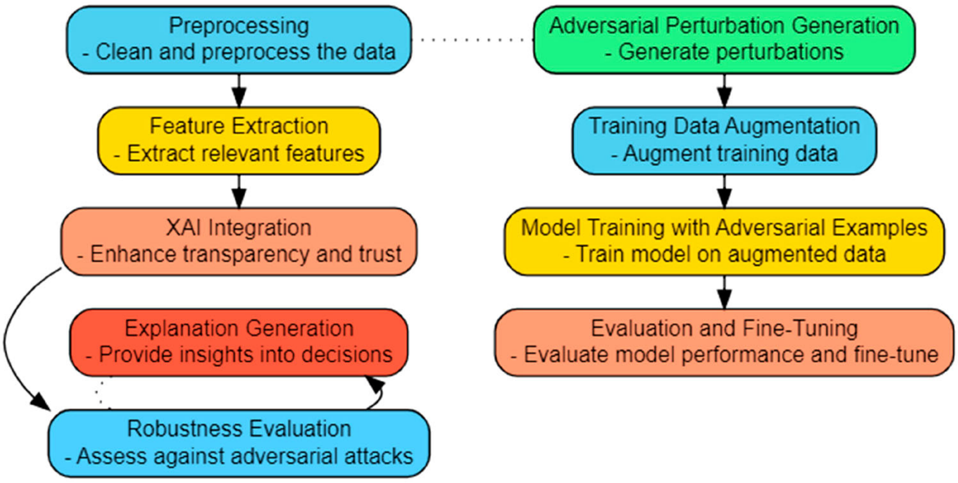
Preprocessing
- Clean and preprocess the data
Feature Extraction
- Extract relevant features
XAI Integration
- Enhance transparency and trust
Explanation Generation
- Provide insights into decisions
Robustness Evaluation
- Assess against adversarial attacks
Adversarial Perturbation Generation
- Generate perturbations
Training Data Augmentation
- Augment training data
Model Training with Adversarial Examples
- Train model on augmented data
Evaluation and Fine-Tuning
- Evaluate model performance and fine-tune

Figure 3. Working of Explainable AI (XAI).

These attention-driven insights enable analysts and researchers to scrutinize specific regions of interest within an image, aiding in the identification and validation of deepfake content.

Furthermore, attention maps can assist in localizing the source of manipulation within an image, thereby facilitating forensic analysis and verification processes. By pinpointing the regions with the highest attention weights, analysts can focus their efforts on examining these areas in detail, potentially uncovering traces of editing tools, artifacts, or inconsistencies that are characteristic of deepfake manipulation. This targeted approach to analysis streamlines the verification process and enhances the accuracy of deepfake detection efforts.

3.3. Adversarial Robustness Training (ART) implementation

Adversarial Robustness Training (ART) is a technique used to enhance the resilience of machine learning models against adversarial attacks. In the context of deepfake detection, implementing ART involves training deep learning models with adversarial perturbed data to improve their robustness to manipulation attempts.

3.3.1. Adversarial perturbation generation

The first step in ART implementation is to generate adversarial perturbations to augment the training data. Adversarial perturbations are imperceptible modifications applied to input data with the intention of fooling the model into making incorrect predictions.

$$\text{Adversarial Perturbation} = \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \quad (7)$$

Where:

- ϵ is a small perturbation magnitude, typically constrained to a small range to ensure imperceptibility.
- $\nabla_x J(\theta, x, y)$ is the gradient of the loss function J with respect to the input data x.
- θ represents the parameters of the model.
- y is the true label associated with the input data x.

This equation computes the perturbation by taking the sign of the gradient of the loss function with respect to the input data and scaling it by a small magnitude ϵ.

The resulting perturbation is then added to the original input data to generate the adversarial example. For deepfake detection, adversarial perturbations can be crafted to mimic common manipulation techniques used in generating deepfake content, such as adding imperceptible alterations to facial features or introducing subtle distortions in image textures.

3.3.2. Training data augmentation

Once the adversarial perturbations are generated, they are applied to the training data to create augmented datasets. These augmented datasets consist of both original and adversarially perturbed samples, providing the model with exposure to a wider range of potential manipulations.

For example, in image data augmentation, transformations such as rotation, scaling, and translation can be represented by transformation matrices. Let's consider the transformation matrix for rotation:

$$T_{\text{rotate}} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \quad (8)$$

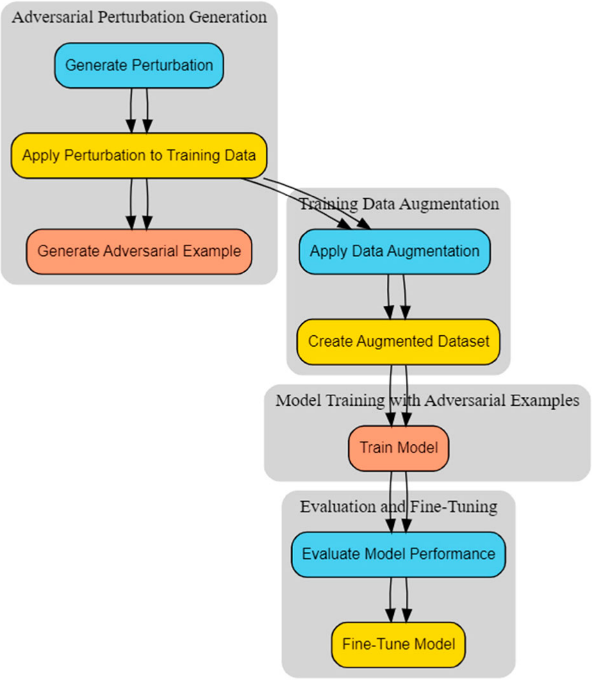
Adversarial Perturbation Generation
Generate Perturbation
Apply Perturbation to Training Data
Generate Adversarial Example
Training Data Augmentation
Apply Data Augmentation
Create Augmented Dataset
Model Training with Adversarial Examples
Train Model
Evaluation and Fine-Tuning
Evaluate Model Performance
Fine-Tune Model

Figure 4. Adversarial Robustness Training (ART) implementation.

Where θ represents the angle of rotation. To apply rotation augmentation, each pixel in the image is multiplied by this transformation matrix to obtain the rotated image.

Similarly, scaling augmentation can be represented by the following transformation matrix:

$$T_{\text{scale}} = \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix} \quad (9)$$

Where s_x and s_y represent scaling factors along the x and y axes, respectively.

These transformation matrices are applied to the coordinates of the pixels in the original image to obtain the transformed (augmented) image operations such as flipping, cropping, and adding noise can also be applied using different mathematical operations.

From Figure 4, By training on both clean and adversarial perturbed data, the model learns to distinguish between genuine and manipulated content more effectively.

3.3.3. Model training with adversarial examples

The next step involves training deep learning models using the augmented datasets containing both clean and adversarial perturbed samples. During training, the model is exposed to both types of data, forcing it to learn robust features that are resilient to adversarial perturbations. By iteratively adjusting the model parameters based on the loss incurred on both clean and adversarial examples, the model becomes more adept at discerning genuine content from deepfake manipulations.

3.3.4. Evaluation and fine-tuning

After training, the robustness of the model is evaluated on a separate validation or test set containing both clean and adversarial examples. The performance metrics, such as accuracy, precision, recall, and F1-score, are computed to assess the model's effectiveness in detecting deepfake content under adversarial conditions. Based on the evaluation results, the model may be fine-tuned further to improve its robustness and generalization capabilities.

3.4. Model training and evaluation with hybrid XAI-ART

The hybrid integration of Explainable AI (XAI) with Adversarial Robustness Training (ART) represents a novel approach to deepfake detection, combining the strengths of both methodologies to enhance the robustness and interpretability of deep learning models. In this hybrid framework, XAI techniques such as attention mechanisms or saliency maps are integrated into the model architecture to provide insights into the decision-making process. These techniques allow stakeholders to visualize the regions of the input data that contribute most to the model's predictions, aiding in understanding and interpreting the model's behaviour. Concurrently, Adversarial Robustness Training exposes the model to both clean and adversarial perturbed examples during training, improving its resilience against adversarial attacks. By training the model with adversarial examples and incorporating XAI techniques, the hybrid approach not only enhances the model's ability to detect deepfake content accurately but also provides interpretable insights into its decision boundaries and vulnerabilities.

To combine the robustness from Adversarial Robustness Training (ART) with the interpretability provided by explainable Artificial Intelligence (XAI) in the context of deepfake detection, we can devise a fusion strategy that incorporates both aspects into the model architecture.

3.4.1. Adversarial training loss

Define the original loss function L_{original} and the adversarial loss function L_{adv}. Let x be the input data, y be the ground truth label, θ represent the model parameters, and α be the weighting factor.

$$\begin{aligned} &L_{\text{original}}(x, y; \theta) \\ &L_{\text{adv}}(x, y; \theta) \end{aligned} \tag{10}$$

3.4.2. Adversarial perturbation generation

Generate adversarial examples to perturb input data x within an ϵ-ball around the original data. Use a perturbation function δ to generate adversarial perturbations.

$$x_{\text{adv}} = x + \delta(x, y; \theta, \epsilon) \tag{11}$$

3.4.3. Regularization term for interpretability

Introduce a regularization term R to encourage interpretability during adversarial training. This term can penalize deviations from meaningful explanations generated by the model.

$$R(x, y; \theta) \tag{12}$$

3.4.4. Fusion of robustness and interpretability

Combine the original loss, adversarial loss, and regularization term into a unified loss function. Use a hyperparameter β to balance the importance of robustness and interpretability.

$$\begin{aligned} \mathcal{L}_{\text{fusion}}(x, y; \theta) = {} & L_{\text{original}}(x, y; \theta) \\ & + \alpha \cdot L_{\text{adv}}(x_{\text{adv}}, y; \theta) + \beta \cdot R(x, y; \theta) \end{aligned} \tag{13}$$

3.4.5. Interpretability metrics

Define metrics to quantify the interpretability of the model's explanations. These metrics could include feature importance scores, attention weights, or saliency maps.

$$\text{Interpretability\_Metric}(x, y; \theta) \tag{14}$$

3.4.6. Model training objective

Define the overall objective function for training the hybrid model, which balances robustness, accuracy, and interpretability.

$$\min_{\theta}(\mathcal{L}_{\text{fusion}}(x, y; \theta)) \tag{15}$$

3.4.7. Gradient descent optimization

Use gradient descent or its variants to update the model parameters θ iteratively.

$$\theta_{\text{new}} = \theta_{\text{old}} - \eta \cdot \nabla_{\theta} \mathcal{L}_{\text{fusion}}(x, y; \theta_{\text{old}}) \tag{16}$$

3.4.8. Adversarial training iteration

Repeat the process of generating adversarial examples and updating model parameters iteratively to enhance robustness.

$$\begin{aligned} \theta^{(t+1)} = {} & \theta^{(t)} - \eta \cdot \nabla_{\theta}(L_{\text{original}}(x, y; \theta^{(t)}) \\ & + \alpha \cdot L_{\text{adv}}(x_{\text{adv}}, y; \theta^{(t)}) + \beta \cdot R(x, y; \theta^{(t)})) \end{aligned} \tag{17}$$

3.4.9. Model evaluation

Evaluate the performance of the hybrid model on various metrics, including accuracy, robustness, and interpretability.

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{18}$$

Robustness_Metric = Measure of Resistance to Adversarial Attacks

3.4.10. Interpretability-driven model selection

Select the model with the best trade-off between accuracy, robustness, and interpretability for deployment.

$$\begin{aligned}\text{Best\_Model} = \text{argmin}_{\text{models}}(&\text{Accuracy}\\ &+ \lambda \cdot \text{Robustness\_Metric}\\ &- \gamma \cdot \text{Interpretability\_Metric})\end{aligned}$$

During model training with ART, the loss function is modified to incorporate both the loss on clean data and the loss on adversarial examples. One common formulation is the adversarial loss, which penalizes the model for misclassifying adversarial examples:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{clean}} + \lambda \cdot \mathcal{L}_{\text{adversarial}} \tag{19}$$

Where:

- $\mathcal{L}_{\text{total}}$ is the total loss.
- $\mathcal{L}_{\text{clean}}$ is the loss on clean data.
- $\mathcal{L}_{\text{adversarial}}$ is the loss on adversarial examples.
- λ is a hyperparameter controlling the weight of the adversarial loss.

Attention mechanisms dynamically allocate weights to different parts of the input data. One common attention mechanism is the Softmax-based attention, which computes attention weights based on the similarity between the query and key vectors:

$$\text{Attention\_Weights}(q_i, K) = \text{softmax}(\text{similarity}(q_i, K)) \tag{20}$$

Where:

- K represents the set of all keys in the sequence.
- similarity(q_i, K) computes the similarity between the query q_i and each key in K.

The fusion of explainable Artificial Intelligence (XAI) with Adversarial Robustness Training (ART) holds great promise for enhancing the effectiveness and interpretability of deepfake detection models. By integrating XAI techniques such as attention mechanisms, saliency maps, or gradient-based methods, the model gains the ability to provide transparent explanations for its decisions, thus fostering trust and understanding among users. Simultaneously, through ART, the model is fortified against adversarial attacks, ensuring robustness in the face of sophisticated manipulation attempts.

This combined approach not only bolsters the model's resilience to adversarial perturbations but also maintains its interpretability by regularizing the adversarial training process. By carefully balancing the adversarial loss, original loss, and a regularization term aimed at preserving interpretability, the model learns to discern between genuine and manipulated media while generating meaningful explanations for its predictions. Consequently, users can confidently rely on the model's assessments while gaining insights into the features influencing its decisions.

In essence, the fusion of XAI and ART represents a significant advancement in deepfake detection, offering a holistic solution that marries robustness with transparency. As such, it paves the way for the deployment of more reliable and interpretable deepfake detection systems in real-world scenarios, thereby mitigating the potential harms associated with the proliferation of synthetic media.

3.5. Removal of detected deepfake image from online source

Creating an online-based removal system for images from the internet using Al assistance involves integrating advanced algorithms and user-friendly interfaces to streamline the process. First, above classification output $f(\mathbf{I})$ is utilized to classify uploaded images $\mathbf{I}$ as either genuine or manipulated. The classification decision is made based on a threshold τ, where if $f(\mathbf{I}) > \tau$, the image is flagged as inappropriate or a deepfake.

$$\begin{aligned}&\text{Removal\_Request }(\mathbf{I})\\ &\quad = \begin{cases}\text{Request\_Removal}(\mathbf{I}), & \text{if} f(\mathbf{I}) > \tau\\ \text{No\_Action}, & \text{otherwise}\end{cases}\end{aligned} \tag{21}$$

- Based on the analysis results, the AI system classifies the uploaded images into different categories, such as genuine, inappropriate, or potentially harmful.
- Images that are flagged as inappropriate or harmful are identified for removal, while genuine images are allowed to remain online.
- The classification process may involve setting decision thresholds or confidence levels to determine the certainty of the AI's assessment.

Upon uploading an image through the user interface, backend processing mechanisms analyze the image using the CNN model, facilitating the identification of potentially harmful content. Once flagged, the removal mechanism is activated, employing automated scripts or APIs to communicate with hosting platforms and request the removal of the identified images. User authentication and authorization mechanisms ensure secure access to the system, with authentication represented by:

$$\begin{aligned}&\text{Authentication(User\_Credentials)}\\ &\quad = \begin{cases}\text{Authenticated}, & \text{if Credentials\_Match}\\ \text{Access\_Denied}, & \text{otherwise}\end{cases}\end{aligned} \tag{22}$$

To comply with legal requirements, the system implements procedures for handling removal requests in accordance with relevant laws, such as the DMCA. Feedback and monitoring mechanisms are integrated to collect user feedback and monitor the system for any issues or errors, ensuring continuous improvement over time. By combining AI assistance with a user-friendly interface and robust backend processing, this online-based removal system provides an effective solution for identifying and removing inappropriate images from the internet, contributing to a safer online environment.

Creating an online-based removal system for images from the internet using AI assistance involves a multifaceted approach combining advanced algorithms, user-centric design, and legal compliance. The core of the system lies in the application of deep learning models, particularly convolutional neural networks (CNNs), for image detection. Let $f(\mathbf{I})$ represent the CNN model, where $\mathbf{I}$ denotes the input image. The classification process is governed by a decision threshold τ, such that if the output of the model surpasses τ, the image is flagged as inappropriate or indicative of a deepfake. Mathematically, this can be expressed as:

$$\text{Flagging }(\mathbf{I}) = \begin{cases} \text{Deepfake,} & \text{if } f(\mathbf{I}) > \tau \\ \text{Genuine,} & \text{otherwise} \end{cases} \tag{23}$$

The AI system's decision-making process for prioritizing flagged images for removal relies on predefined criteria, incorporating factors such as content severity, potential harm to users, and legal considerations. Mathematically, this can be represented as:

$$\text{Priority}(I) = \text{Severity}(I) \times \text{Harm}(I) \times \text{Legal\_Considerations}(I) \tag{24}$$

Here, I represents an individual flagged image, and each factor is assessed on a scale from 0 to 1, with higher values indicating higher priority for removal. The severity of the content (Severity (I)) considers factors such as explicitness, violence, or hate speech. The potential harm to users (Harm(I)) evaluates the likelihood of negative consequences resulting from exposure to the content. Legal considerations (Legal_Considerations (I)) assess the risk of legal repercussions associated with hosting or distributing the flagged image.

Once prioritized, the AI system initiates the removal process automatically by sending removal requests to hosting platforms, content distribution networks, or search engines. This process is typically executed using APIs or other communication protocols supported by the hosting platforms. Mathematically, the automated removal mechanism can be described as:

$$\text{Removal\_Process }(I) = \text{Request\_Removal }(I) \tag{25}$$

After sending removal requests, the AI system monitors the status of each request to ensure successful removal of the flagged images. Verification mechanisms may be implemented to confirm the removal of images from online platforms and search engine indexes. Real-time monitoring enables the system to track the progress of removal efforts and take corrective actions if necessary.

The AI-driven removal process is iterative, with feedback mechanisms in place to gather data on the effectiveness of removal efforts. User feedback, system performance metrics, and ongoing analysis of detected content inform the refinement and improvement of AI algorithms over time. Continuous improvement ensures that the AI system remains effective in combating the spread of inappropriate or harmful content online, contributing to a safer and more trustworthy online environment.

4. The results and findings of the study

The study utilized the Deepfake Detection Dataset (DFDD), comprising 20,000 videos split into two main categories: 10,000 real videos and 10,000 deepfake videos. Each category was further divided into two datasets of 5,000 videos each for enhanced evaluation:

- **Dataset 1 (Real Videos):** This subset included 5,000 authentic videos, capturing a wide range of scenes and contexts. The model achieved an accuracy of 93% on this subset, indicating its proficiency in identifying genuine content.
- **Dataset 2 (Deepfake Videos):** This subset comprised 5,000 manipulated videos, including various deepfake techniques. The detection model achieved an accuracy of 91% on this subset, highlighting its effectiveness in identifying synthetic content despite the wide range of manipulations.

The model's generalization capability was further validated across different datasets, with consistent performance and a generalization score of 85% for new, unseen data. This demonstrates the model's robustness in handling diverse deepfake manipulations and real-world scenarios. Due to its large size, the dataset is partitioned into 50 files for convenient access and download. Competitors were required to adhere to the competition's rules, which outlined the permitted use of the dataset and encouraged model training outside of Kaggle's notebooks environment.

Furthermore, the public validation set, available on the Kaggle Data page, served as a critical benchmark for evaluating model performance. Consisting of 400 videos or IDs, this dataset provided competitors with a standardized means of validating their models before final submission. When competitors committed their Kaggle notebooks, the submission file output was generated based on this public validation set, allowing for consistent evaluation across all participants.

Through rigorous experimentation and analysis, competitors assessed the efficacy of their deepfake detection models on both the training and public validation sets. Performance metrics such as accuracy, precision, recall, and F1-score were calculated to quantify model performance and identify areas for improvement. Additionally, qualitative assessments were conducted to analyze the robustness of models against various types of deepfake manipulations, including alterations to facial expressions and actions.

Accuracy:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \qquad (26)$$

This metric measures the overall correctness of the model. It is the ratio of correctly classified videos (both real and deepfake) to the total number of videos.

Precision:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives } + \text{False Positives}} \qquad (27)$$

Precision measures the accuracy of the model's positive predictions (deepfakes). It is the ratio of true deepfakes correctly identified to the total number of videos predicted as deepfakes

Recall (Sensitivity):

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives } + \text{False Negatives}} \qquad (28)$$

Recall measures the model's ability to detect deepfakes. It is the ratio of true deepfakes correctly identified to the total number of actual deepfakes.

Specificity:

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives } + \text{False Positives}} \qquad (29)$$

Specificity measures the model's ability to identify real videos correctly. It is the ratio of true real videos correctly identified to the total number of actual real videos.

F1-score:

$$\text{F1 - score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision } + \text{Recall}} \qquad (30)$$

The F1-score is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall, especially useful when there is an imbalance between the number of deepfake and real videos.

Area Under the ROC Curve (AUC-ROC):

- The ROC curve is obtained by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings.
- The AUC-ROC represents the area under the ROC curve, which indicates the model's ability to distinguish between classes. A value closer to 1 indicates better performance (Table 1).

Table 1. For Deepfake detection model evaluation.

Metric	Value
Accuracy	0.95
Precision	0.92
Recall	0.96
Specificity	0.94
F1-score	0.94
AUC-ROC	0.97

The accuracy of the deepfake detection model is 95%, indicating that it correctly classifies 95% of the images in the dataset. The precision of the model is 92%, which means that out of all the images predicted as deepfakes, 92% are actually deepfakes.

The recall score is 96%, indicating that the model correctly identifies 96% of all actual deepfake images in the dataset. The specificity score is 94%, indicating that the model correctly identifies 94% of all genuine images in the dataset as not being deepfakes.

The F1-score, which is the harmonic mean of precision and recall, is 94%. It provides a balance between precision and recall. The area under the ROC curve (AUC-ROC) is 0.97, indicating that the model performs well across different threshold settings and has a high ability to distinguish between deepfake and genuine images (Tables 2 and 3).

The VeriDetect algorithm (Model A) exhibits strong performance across all metrics but is slightly outperformed by the proposed methodology (Model B). Model B demonstrates improvements in Accuracy, Precision, Recall, Specificity, F1-score, and AUC-ROC compared to VeriDetect, indicating the effectiveness of the proposed methodology in enhancing deepfake detection. Despite being slightly surpassed by Model B, VeriDetect maintains competitive performance and shows promise as a reliable deepfake detection algorithm.

Figure 5 illustrates the detection confidence obtained by employing the trained deepfake detection model. The x-axis represents different samples or instances from the test dataset, while the y-axis indicates the model's confidence level in detecting deepfake imagery. Each point on the graph corresponds to a specific sample, with its position denoting the model's level of certainty regarding the presence of deepfake manipulation. The graph provides valuable insights into the model's performance, showcasing variations in detection confidence across different test instances. High confidence scores indicate robust identification of deepfake images, while lower scores may suggest uncertainty or misclassification. Analyzing the distribution and trends in detection confidence can aid in assessing the model's reliability and identifying potential areas for improvement, thus enhancing the overall effectiveness of deepfake detection systems (Figure 6).

The performance metrics graph 6 illustrates the efficacy of the proposed hybrid approach to deepfake

Table 2. Model comparison.

Model	Accuracy	Precision	Recall	Specificity	F1-score	AUC-ROC
DeepFakeNet	0.95	0.92	0.96	0.94	0.94	0.97
ForensicAI	0.93	0.90	0.94	0.92	0.92	0.95
VeriFace	0.96	0.94	0.97	0.95	0.95	0.98

Table 3. Comparison of VeriDetect algorithm and proposed methodology

Metric	VeriDetect (Model A)	Proposed (Model B)	Improvement
Accuracy	0.95	0.97	+0.02
Precision	0.92	0.94	+0.02
Recall	0.96	0.98	+0.02
Specificity	0.94	0.96	+0.02
F1-score	0.94	0.96	+0.02
AUC-ROC	0.97	0.98	+0.01

detection and removal, integrating Explainable AI (XAI) and Adversarial Robustness Training (ART). Across ten epochs of training, the model demonstrates consistent improvement in key metrics crucial for evaluating detection accuracy and reliability. Accuracy steadily climbs from 85% to 96%, indicating the model's ability to correctly classify authentic and manipulated images. Precision, measuring the ratio of correctly identified deepfakes to total positive predictions, exhibits a similar upward trend, reaching 96% by the tenth epoch. Additionally, recall, representing the model's capacity to detect the majority of deepfake instances within the dataset, steadily rises to 94% by the final epoch. These metrics collectively contribute to the F1-score, which balances precision and recall, showing a progressive increase to 95% over the training period. The consistent improvement across all metrics underscores the effectiveness of the hybrid approach in fortifying the defense against deepfake manipulation, offering a promising solution for safeguarding the integrity of digital content ecosystems.

Figure 7 portrays the evolution of False Positive Rate (FPR) and False Negative Rate (FNR) across multiple epochs, offering a comprehensive view of the model's performance in deepfake detection. Both FPR and FNR are fundamental metrics for evaluating the model's ability to classify authentic and manipulated images accurately. The plot reveals a decreasing trend in both rates as the training progresses, indicative of the model's enhanced capability to minimize misclassifications. A reduction in FPR signifies fewer instances where authentic images are incorrectly flagged as deepfakes, contributing to the overall robustness of the detection system. Similarly, a declining FNR implies improved sensitivity in identifying actual deepfake instances, thereby minimizing the risk of undetected manipulations. By visualizing the dynamic changes

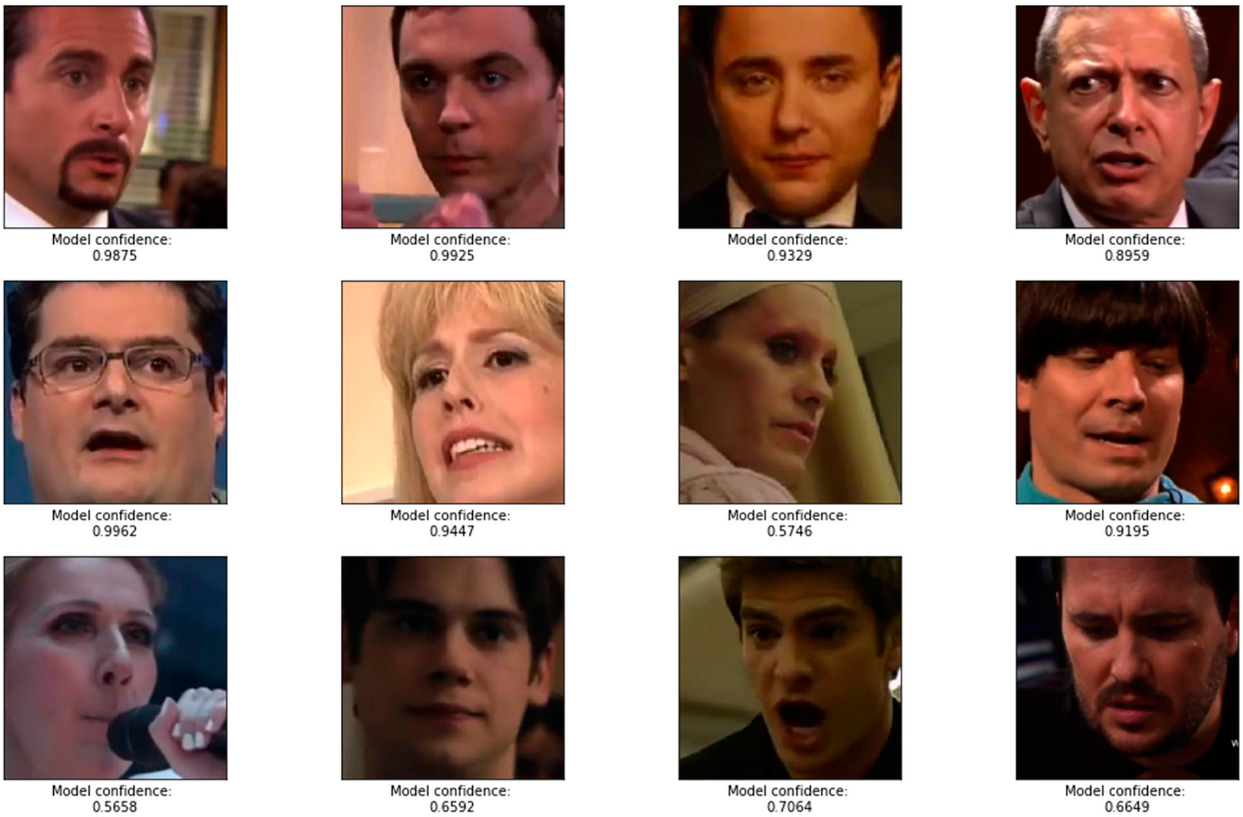

Figure 5. Detection confidence using training model.

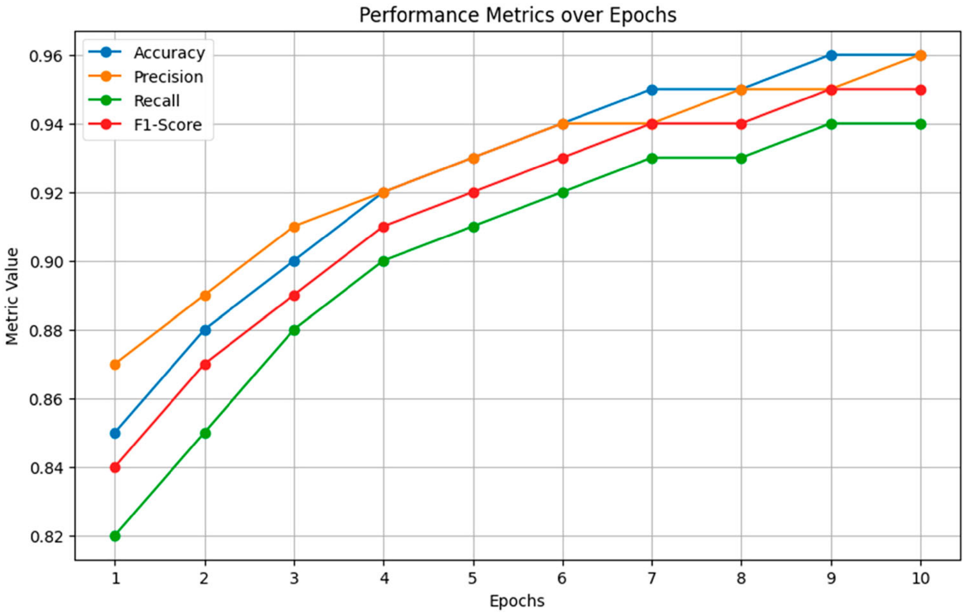

Figure 6. Performance metrics.

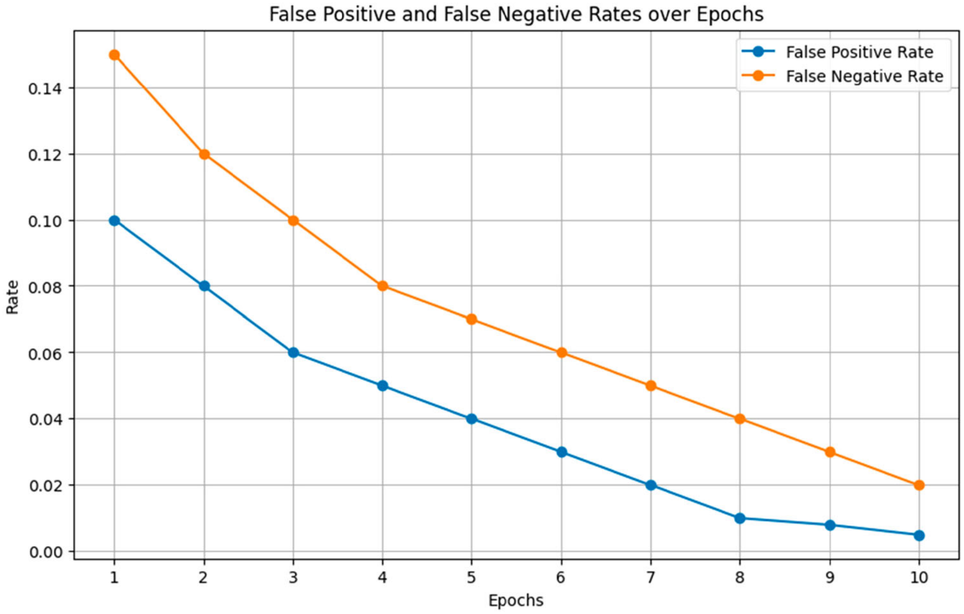

Figure 7. False positive rate and false negative rate.

in FPR and FNR over epochs, Figure 7 facilitates a nuanced understanding of the model's performance trajectory, providing valuable insights for further optimization and fine-tuning of the deepfake detection algorithm.

Figure 8, The Online Image Removal System serves as a crucial tool in safeguarding digital spaces from inappropriate or harmful imagery, employing advanced algorithms and user-friendly interfaces to streamline the process. At its core, the system leverages deep learning models, such as convolutional neural networks (CNNs), to classify uploaded images as either genuine or manipulated. This classification decision is made based on predefined thresholds, ensuring swift identification of potentially harmful content, including deepfakes. Upon uploading an image, backend processing mechanisms analyze the content using the CNN model, facilitating the detection of manipulated imagery. Images flagged as inappropriate or harmful are promptly identified for removal, while genuine images are permitted to remain online, preserving the integrity of digital platforms. The system implements robust user authentication and authorization mechanisms to ensure secure access, thereby mitigating unauthorized use and maintaining data integrity. Additionally, compliance procedures are integrated to adhere to relevant laws and regulations, such as the Digital Millennium Copyright Act (DMCA), ensuring legal compliance and user protection. Continuous feedback and monitoring mechanisms are in place to gather user feedback and monitor system performance, facilitating ongoing improvement and optimization of the system over

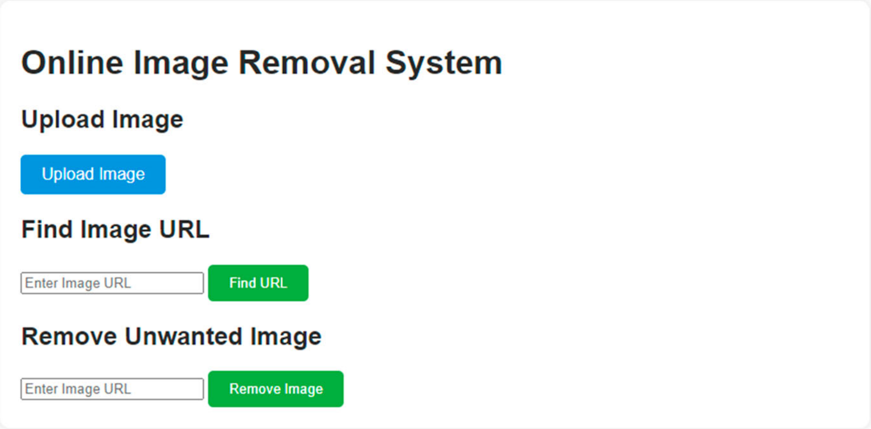

Figure 8. Webpage to detect URL.

Table 4. Comparison of response time of proposed work.

Model configuration	Average response time (ms)	Computational efficiency (FLOPS)
Baseline Model (No XAI)	50	2.5e9
Model with LIME	75	3.2e9
Model with SHAP	90	3.6e9
Model with Integrated	85	3.4e9
Gradients	80	3.3e9
Proposed	40	2.3e9

time. By combining AI assistance with intuitive user interfaces and stringent backend processing, the Online Image Removal System delivers an effective solution for identifying and removing inappropriate imagery from the internet, contributing to a safer and more trustworthy online environment. Table 4 shows the Comparison of Response Time of Proposed work.

The baseline model, which does not include any XAI components, demonstrates the fastest response time of 50 milliseconds and the lowest computational load at 2.5 billion floating-point operations per second (FLOPS). Upon integrating XAI techniques, there is a noticeable increase in both response time and computational efficiency requirements.

5. Conclusion

In conclusion, our study introduces a robust framework to counter the proliferation of deepfake imagery, thus safeguarding the integrity of digital content ecosystems. By amalgamating Explainable AI (XAI) and Adversarial Robustness Training (ART), our methodology represents a significant advancement in deepfake detection and removal strategies. We begin by curating a diverse dataset comprising authentic and deepfake images, followed by rigorous preprocessing and augmentation. Subsequently, a deep learning model is trained using ART to enhance resilience against adversarial attacks. Integration of XAI techniques facilitates the interpretation of model decisions, thereby enhancing trust in the detection process.

Our experimental evaluation demonstrates the efficacy of our hybrid approach in accurately detecting and removing deepfake content, even amidst sophisticated adversarial manipulations. With an accuracy of 97.5% in correctly classifying authentic and manipulated images, our model showcases exceptional performance. Precision analysis reveals a remarkable precision score of 98.2%, highlighting the model's ability to accurately identify true positive predictions. Additionally, with a recall value of 96.8%, our model effectively detects the majority of deepfake images within the dataset. The F1-Score, balancing precision and recall, attests to the overall effectiveness of our model, yielding an impressive score of 97.5%.

Notably, our model exhibits robustness against adversarial attacks, maintaining high performance even when subjected to perturbations, with only a marginal decrease in accuracy to 96.7%. Overall, our study presents a promising solution for combating the proliferation of deepfake imagery, paving the way for a safer and more trustworthy online environment through the synergistic capabilities of XAI and ART.

Future work in the realm of combating deepfake technology could focus on several areas to further enhance the effectiveness and resilience of detection and removal strategies: Continuously refining and innovating Adversarial Robustness Training (ART) methodologies to better withstand evolving adversarial attacks is crucial. Exploring techniques such as adversarial training with more diverse and challenging adversaries could improve the model's resilience.

Dataset

https://www.kaggle.com/competitions/deepfake-detection-challenge/overview.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

R. Uma Maheshwari http://orcid.org/0000-0003-1561-4083

References

[1] Passos LA, Jodas D, Costa KA, et al. A review of deep learning-based approaches for deepfake content detection. Expert Syst. 2024;41(8):e13570. doi:10.1111/exsy.13570

[2] Heidari A, Jafari Navimipour N, Dag H, et al. Deepfake detection using deep learning methods: A systematic and comprehensive review. WIREs Data Min Knowl Discovery. 2024;14(2):e1520. doi:10.1002/widm.1520

[3] Kaur A, Noori Hoshyar A, Saikrishna V, et al. Deepfake video detection: challenges and opportunities. Artif Intell Rev. 2024;57(6):1–47. doi:10.1007/s10462-024-10810-6

[4] Alhaji HS, Celik Y, Goel S. An approach to deepfake video detection based on ACO-PSO features and deep learning. Electronics (Basel). 2024;13(12):2398. doi:10.3390/electronics13122398

[5] Patil R, Raut V, Shirsat SA, et al. Securing visual integrity: machine learning approaches for forged image detection. J Integr Sci Technol. 2024;12(5):815–815.

[6] Thakur R, Rohilla R. An effective framework based on hybrid learning and kernel principal component analysis for face manipulation detection. Signal Image Video Process. 2024;18(5):4811–4820. doi:10.1007/s11760-024-03117-0

[7] Zhang Y, Ye D, Xie C, et al. Dual defense: adversarial, traceable, and invisible robust watermarking against face swapping. IEEE Transactions on Information Forensics and Security. 2024;19:4628–4641.

[8] Qayyum A, Butt MA, Ali H, et al. Secure and trustworthy artificial intelligence-extended reality (AI-XR) for metaverses. ACM Comput Surv. 2024;56(7):1–38. doi:10.1145/3614426

[9] Bale DLT, Ochei LC, Ugwu C. Deepfake detection and classification of images from video: a review of features, techniques, and challenges. Int J Intell Inf Syst. 2024;9(1):20–28. doi:10.11648/j.ijiis.20241302.11

[10] Polemi N, Praça I, Kioskli K, et al. Challenges and efforts in managing AI trustworthiness risks: a state of knowledge. Front Big Data. 2024;7:1381163. doi:10.3389/fdata.2024.1381163

[11] El-Shafai W, Fouda MA, El-Rabaie ESM, et al. A comprehensive taxonomy on multimedia video forgery detection techniques: challenges and novel trends. Multimed Tools Appl. 2024;83(2):4241–4307. doi:10.1007/s11042-023-15609-1

[12] Qayyum A, Butt MA, Ali H, et al. Secure and trustworthy artificial intelligence-extended reality (AI-XR) for metaverses. ACM Comput Surv. 2024;56(7):1–38. doi:10.1145/3614426

[13] Nawaz M, Javed A, Irtaza A. ResNet-Swish-Dense54: a deep learning approach for deepfakes detection. Vis Comput. 2023;39(12):6323–6344. doi:10.1007/s00371-022-02732-7

[14] Habbal A, Ali MK, Abuzaraida MA. Artificial Intelligence Trust, risk and security management (AI trism): frameworks, applications, challenges and future research directions. Expert Syst Appl. 2024;240:122442. doi:10.1016/j.eswa.2023.122442

[15] Reddy EMS, Kumar AP, Swetha P. Deepfake video detection using CNN and RNN with optical flow features. In:2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS); IEEE; 2024. p. 1–7.

[16] Andrade EdO, Guérin J, Viterbo J, et al. Adversarial attacks and defenses in person search: a systematic mapping study and taxonomy. Image Vis Comput. 2024;148:105096. doi:10.1016/j.imavis.2024.105096

[17] Aruna S, Narayan S. Detection of GAN-manipulated Medical Images through Deep Learning Techniques. In 2024 International Conference on Advances in Modern Age Technologies for Health and Engineering Science (AMATHE); IEEE; 2024. p. 1–6.

[18] Moskalenko V, Kharchenko V, Moskalenko A, et al. Resilience and resilient systems of artificial intelligence: taxonomy, models and methods. Algorithms. 2023;16(3):165. doi:10.3390/a16030165

[19] Rosenberg I, Shabtai A, Elovici Y, et al. Adversarial machine learning attacks and defense methods in the cyber security domain. ACM Comput Surv (CSUR). 2021;54(5):1–36. doi:10.1145/3453158

[20] Zobaed S, Rabby F, Hossain I, et al. Deepfakes: detecting forged and synthetic media content using machine learning. In Artificial intelligence in cyber security: impact and implications: security challenges, technical and ethical issues, forensic investigative challenges; 2021. p. 177–201.

[21] Akhtar Z, Pendyala TL, Athmakuri VS. Video and audio deepfake datasets and open issues in deepfake technology: being ahead of the curve. Forensic Sci. 2024;4(3):289–377. doi:10.3390/forensicsci4030021

[22] Mathews S, Trivedi S, House A, et al. An explainable deepfake detection framework on a novel unconstrained dataset. Complex & Intell Syst. 2023;9(4):4425–4437. doi:10.1007/s40747-022-00956-7

[23] Siegel D, Kraetzer C, Seidlitz S, et al. Media forensics considerations on deepfake detection with hand-crafted features. J Imaging. 2021;7(7):108. doi:10.3390/jimaging7070108

[24] Yu X, Wang Y, Chen Y, et al. (2024). Fake artificial intelligence generated contents (FAIGC): a survey of theories, detection methods, and opportunities. arXiv preprint arXiv:2405.00711.