Hybrid Learning Techniques for Image Forensics and Privacy Protection in the Face of Deep Fake Threats

Arif Ullah[1], Sidra Pervez[2], Muhammad Wajidullah Khan[3], Aina Hassan[4]and Muazam Ali[2]
[1]Faculty of Computing and Artificial Intelligence, Air University, Islamabad, Pakistan
[2]Department of Management Sciences, HITEC University, Taxila, Pakistan
[3]Faculty of Engineering and Technology UTHM, Malaysia
[4]Othman Yeop Abdullah Graduate School of Business, UUM, Malaysia

arifullahms88@gmail.com
sidra.pervez@hitecuni.edu.pk
hn230030@student.uthm.edu.my
alinashah597@gmail.com
muazam.ali@hitecuni.edu.pk

Abstract: This study investigates the detection of deepfake images and videos on social media platforms such as Instagram for forensic analysis using hybrid-learning approaches. It highlights the critical importance of safeguarding privacy and authenticity in digital media. The background draws attention to the growing threat posed by deepfakes, which pose significant challenges across multiple domains, such as politics and entertainment. Existing methods often depend on visual features specific to a dataset and struggle to generalize across different manipulation techniques. Moreover, most approaches focus exclusively on either temporal or spatial features, which limits their capacity to identify complex anomalies involving fused facial features like the mouth, nose and eyes. Important solutions to these challenges include Convolutional Neural Network (CNN), Recurrent Neural Networks (RNN) and hybrid architectures that simultaneously capture spatial and temporal information in deepfake content, such as Convolutional Neural Network - Long Short-Term Memory (CNN-LSTM), Gated Recurrent Unit (GRU) and Vision Transformers (ViT). Additionally, this paper introduces a novel combination of artifact inspection and facial landmark recognition to enhance detection accuracy and employs Gated Recurrent Units (GRUs) and Vision Transformers (ViT) for data augmentation thereby improving model robustness. The effectiveness of the proposed approach is validated through experiments demonstrating substantially improved deduction accuracy, with improvement exceeding 1.5% across multiple datasets. However, several challenges remain, including limited robustness to noise, difficulty in detecting deepfakes in compressed video formats, and dataset imbalances issues. The proposed enhanced hybrid model exhibits superior detection performance while maintaining adaptability across multiple datasets. Future research will focus strengthening model generalization to effectively counter emerging deepfake generation techniques.

Keyword: Deepfakes, Image forensics, Privacy protection, Hybrid model, Web technology

1. Introduction

Information is now widely accessible due to the development of web technology. Every day, a great deal of information is transmitted through print and online media, yet it can be challenging to assess its veracity [1]. The deliberate attempt to damage or enhance the reputation of a company, institution, or individual in order to gain financial or political advantages is known as the dissemination of false information [2]. A wide range of fabricated stories, falsified news articles, and digitally manipulated images circulate on social media. The consumption and distribution of information, as well as various modes of communication including blogging and texting, have been completely transformed by digital platforms [3]. Deepfake technology benefits the fashion industry by allowing customers to make fashion decisions more quickly. Furthermore, this technology has the potential to help individuals with Alzheimer's disease communicate by enabling them to engage with a digitally modified version of their younger selves [4]. Beyond detection, forensic analysis in the deepfake arena covers serval important study areas. These include techniques designed to function effectively in authentic, real-world situations, as well as deepfake attribution, recognition, and passive and active authentication. Deepfake attribution and recognition seek to track the origins of synthetic content by identifying the precise models utilized in its production. This entails examining the traces left by generative models in order to attribute content to its source. Without the need for additional embedded data, passive authentication methods concentrate on assessing the legitimacy of media using its natural characteristics, such as statistical anomalies. For retrospective analysis, these techniques work especially well [5]. Conversely, active authentication methods incorporate verifiable data, such as digital watermarks or cryptographic signatures, into media during the creation process. Deepfake technology is developing very quickly. Apart from the conventional ways of modifying particular regions of an image or video, computer graphics and deep learning (DL) have introduced in a number of new approaches [6]. It has been demonstrated that auto-encoders and Generative Adversarial Networks (GANs) aid in producing high-quality face synthesis with a high degree of photorealism. Additionally, a segmentation map

can be used to create fake videos or image. A written representation or a drawing can be used to create an image. Similarly, a series of auditory stimuli can change a person's facial features. Often, the modification involves altering pre-existing photos or videos [7].

Deepfake Generation Process: A class of machine learning techniques known as generative models is designed to discover the underlying distribution of a dataset and produce new samples that closely resemble the original data. Generative models seek to capture the data distribution $pdata$ and create new samples with comparable features, in contrast to discriminative models, which concentrate on distinguishing between classes or generating predictions [8]. These models have gained significant attention in a variety of fields, such as data augmentation, text generation, music composition, and image synthesis. Generative Adversarial Networks and Diffusion Models are two most widely used generative modeling techniques. Although the goal of both approaches is to create high-quality synthetic data, their underlying principles and learning paradigms are very different. Generating a deepfake is a multi-step pipeline that uses cutting-edge machine learning algorithms to create incredibly lifelike media [9].

Deepfake Detection: As deepfake, technology develops and creates more realistic synthetic media, ensuring the authenticity of digital information has become a crucial challenge in multimedia forensics. In addition to providing the foundation for forensic analysis, attribution, and authentication, detecting modified content is essential for mitigating the risks of disinformation, identity fraud, and threat to media integrity concerns [10]. This section offers a structured summary of detection methodologies, ranging from traditional forensic methods to contemporary data-driven strategies. The primary goal of early forensic techniques was to detect obvious spatial anomalies, such as compression artifacts, pixel-level irregularities, and inconsistent illumination [11]. Although these techniques remain relevant, they frequently fail to account for highly sophisticated adjustments. To overcome these constraints, deep learning algorithms utilize data-driven patterns to identify subtle discrepancies that are invisible to the naked eye [12].

2. Related Work

Table 1 below presents a summary of the related work, showing the authors' names, research areas, publication years, and findings.

Table 1: Summary of related work

Ref	Technique	Findings
[13]	Closest Neighbor Classification using pretrained vision-language model	Conventional DNN classifiers are unable to identify fake images from cutting-edge generative models. The closest neighbor classification approach increased accuracy by 25.90% and mAP by 15.07 m on unobserved diffusion and autoregressive models.
[14]	Statistical Consistency Assault (StatAttack) and MStatAttack	Deepfake detecting statistical discrepancies are decreased using MStatAttack. The updated MStatAttack added layers of deterioration and adjusted combination weights. It may be able to get around sophisticated Deepfake detection systems because it operated in both white-box and black-box scenarios spanning detectors and datasets.
[15]	Hybrid technique combining Residual Network and KK-NN methods	Used a methodical strategy to classify Deepfake pictures with an accuracy of 89.5%. The hybrid approach was strong and successful in identifying Deepfake, indicating that it may lessen the dangers of propaganda and defamation on social media like Instagram.
[16]	ID-unaware Deepfake Detection Model	Enhanced generalization of binary classifiers because of implicit identity leakage. To address this issue, the ID-unaware model was proposed, and it performed better than previous methods over a wide range of datasets.
[17]	Deepfake Predictor (DFP) using VGG16 and CNN architecture	Created the DFP framework, which had a 94% accuracy and 95% precision rate in identifying Deepfake. The technique proved more effective at identifying Deepfake media and resolving ethical issues than transfer learning and other contemporary techniques.

3. Proposed Model

The proposed methodology addresses deepfake detection and privacy protection on social media platforms such as Instagram in image forensics by combining Gated Recurrent Units (GRUs) with Vision Transformers (ViTs). Sequential dependencies and temporal artifacts found in image and video sequences are captured by GRUs. High-dimensional global spatial characteristics and structural patterns are extracted from image patches using ViTs. The model improves sensitivity to both spatial and temporal inconsistencies typical of deepfakes by fusing

the ViT's proficiency in contextual image representation with the GRU's capacity to model temporal changes. The framework uses GRU to identify minute temporal irregularities in motion coherence and facial expressions. ViT meanwhile assist by identifying frequency domain distortions, visual artifacts, and inconsistencies unique to GAN-generated content. Strong feature fusion from both sequential and patch-based perspectives is ensured by this hybridization. Figure 1 presents the proposed model structure along with its different components.

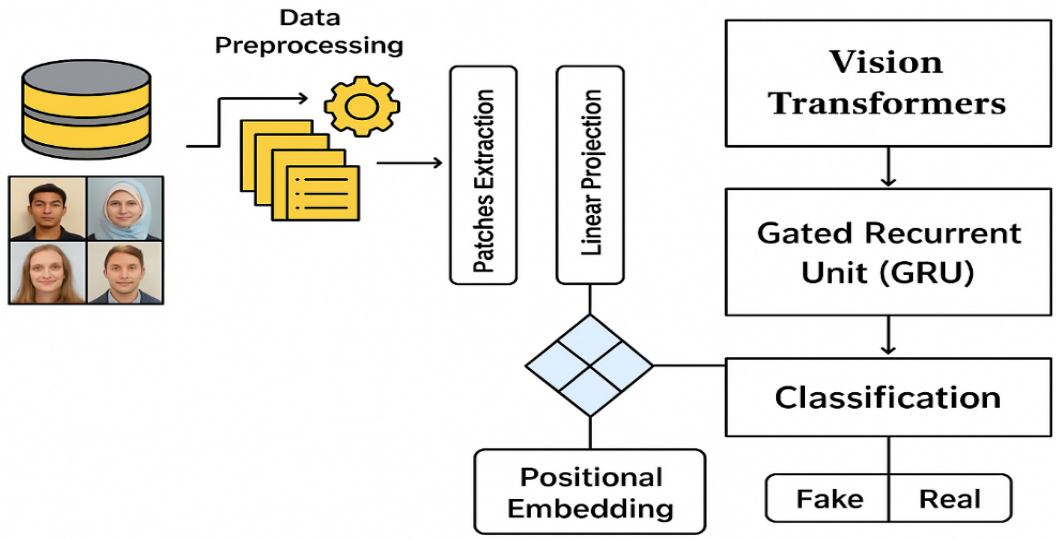

Figure.1: Proposed model for deepfakes detection in social media like Instagram

An outline of our deepfake detection procedure is presented in Figure 1, with particular attention paid to steps such as data preprocessing, model training, and binary classification topology categorization. When paired with behavioral or biological signal cues, this integration facilitates multi-modal forensic analysis. By limiting the availability of raw data while preserving high detection accuracy, the approach prioritizes privacy. Overall, this GRU–ViT hybrid framework provides a robust, scalable, and explainable defense against deepfake threats in digital forensics. A brief mathematical explanation of the proposed hybrid model, which combines privacy-aware training (federated learning with differential privacy) and image forensics (deepfake detection and localization) using a Vision Transformer (ViT) backbone and a Gated Recurrent Unit (GRU) module, is provided below. This section introduces the notation, privacy-aware federated update equations, model heads, loss functions, core equations (1), and the feature fusion process described in Equation (1).

$$\ell - 1 = LN(zi\ell - 1)MHA(\ell - 1)\sum = 1head(\ell - 1)MHA(z{\sim}\ell - 1)i = m = 1\sum Hheadm(z{\sim}\ell - 1)i\ell = \ell - 1 + MHA(\ell - 1) + MLP(LN(\ell - 1))zi\ell = zi\ell - 1 + MHA(z{\sim}\ell - 1)i + MLP(LN(zi\ell - 1)) \quad \textbf{(1)}$$

The Transformer model's layer operations are described in the four lines that follow. Prior to being sent into a Multi-Head Attention (MHA) mechanism, where attention is computed over several heads and their results are averaged, the input from the preceding layer, l − 1 l−1, is first normalized using Layer Normalization (LN). Updated contextualized representations of the input are thus produced. The original input zl−1 z l−1 is then added to the attention output by a residual connection, and the features are further refined by a Layer Normalization and Multi-Layer Perceptron (MLP). Ultimately, the residual input, the MHA output, and the MLP transformation are combined to create the output zlz l at layer ll, guaranteeing stability, richer representation learning, and improved gradient flow across the deep network.

$$t = Er(xt) = (+h - 1+)zt = \sigma(Wzrt + Uzht - 1 + bz)'(+h - 1+)rt' = \sigma(Wrrt + Urht - 1 + br)h{\sim} = tanh\,(h + h(' \odot h - 1) + h)h{\sim}t = tanh(Whrt + Uh(rt' \odot ht - 1) + bh)h = (1-) \odot h - 1 + \odot h{\sim}ht = (1 - zt) \odot ht - 1 + zt \odot h{\sim}t \quad \textbf{(2)}$$

The updating procedure in a Gated Recurrent Unit (GRU) is described by in equations (2) Initially, the update gate z t and reset the gate rtr t are calculated using sigmoid functions, which determine the amount of historical data to retain or discard. Next, a potential concealed state h ~ t h ~ t is produced with the reset gate's assistance, and the ultimate hidden state ht h t is calculated as the weighted sum of the prior state ht−1 h t−1 and the candidate state, which the update gate regulates.

4. Datasets

A sizable dataset called WildFake was created to facilitate the deduction of AI-generated images. It includes a wide range of synthetic photos from various generative models and open communities, ensuring a wide range of styles and content. Images in the dataset are categorized hierarchically by generative model types, architectures, weights, versions, and release dates in a hierarchical manner [17]. Both authentic and synthetic facial photos are included in the sample. We employed a validation set of 76,161 photos for our assessment, of which 38,080 were authentic and 38,081 were fraudulent. The evaluation metrics used include accuracy, precision, recall, and the F1-score.

5. Experiment Results and Discussion

This study evaluates both standard GU models and the ViT-based model for deepfake detection.

The results demonstrate a notable performance advantage for the ViT, which is attributed to its ability to integrate both local and global features through self-attention mechanisms. A thorough presentation of the results is provided in the following section and illustrated in Figure 2.

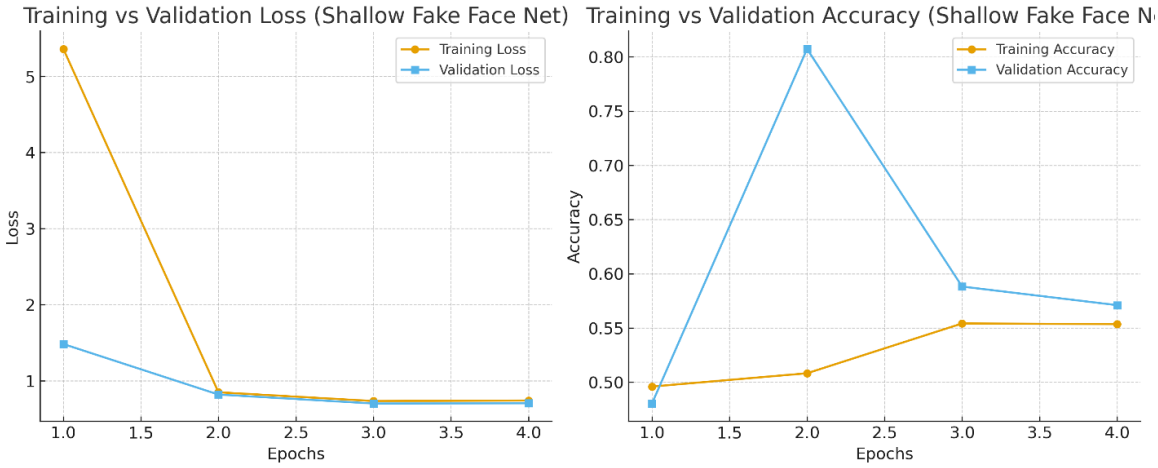

Figure 2: Accuracy and Loss values over training dataset

Figure 2 displays the accuracy and loss of curves for both training and validation. The Y-axis in training and validation accuracy displays accuracy, which ranges from 0 to 1 and indicates how well the model predicts. Complete runs of the training dataset, or epochs, are the units on the X-axis. The training accuracy is displayed by the yellow line (R) graph, which begins approximately 0.4962 and gradually increases to almost 0.5538 in the fifth epoch. The upward trend indicates that the model is getting better at what it does by learning from the training data. The validation accuracy is indicated by the orange line, which starts at 0.4803, experiences a slight decline after the second epoch, and subsequently increases to 0.5710 by the fifth epoch. Two high performing models for identifying GAN-generated fused artifacts in facial features TCN and CNN-LSTM with GAN. For strong generalization, we achieve excellent alignment between the accuracy and loss curves of both models. CNN and CNN-GRU both provide dependable stability, however CNN-GRU does not achieve the same accuracy levels. CNN performs reasonably well, but it could use some further enhancements to handle the intricacy of fused GAN-generated features. Our findings build on the research of [18], where the authors employed CNN models like VGG16, VGG19, and the InceptionV3 model to detect deepfakes with an accuracy of up to 90%. The effectiveness of attention-based architectures in deepfake identification is demonstrated by this work, which uses PViT to provide a 9% performance boost with 99% accuracy. As indicated in Figure 3, the outcomes of both investigations are comparable because they use the same Kaggle dataset. The fact that ViT performs so well demonstrates the benefit of processing image patches in their holistic form, which makes it possible to identify changes across local and global inputs. This approach addresses several limitations of CNN architectures and reinforces the importance of transformer models in multimedia forensics.

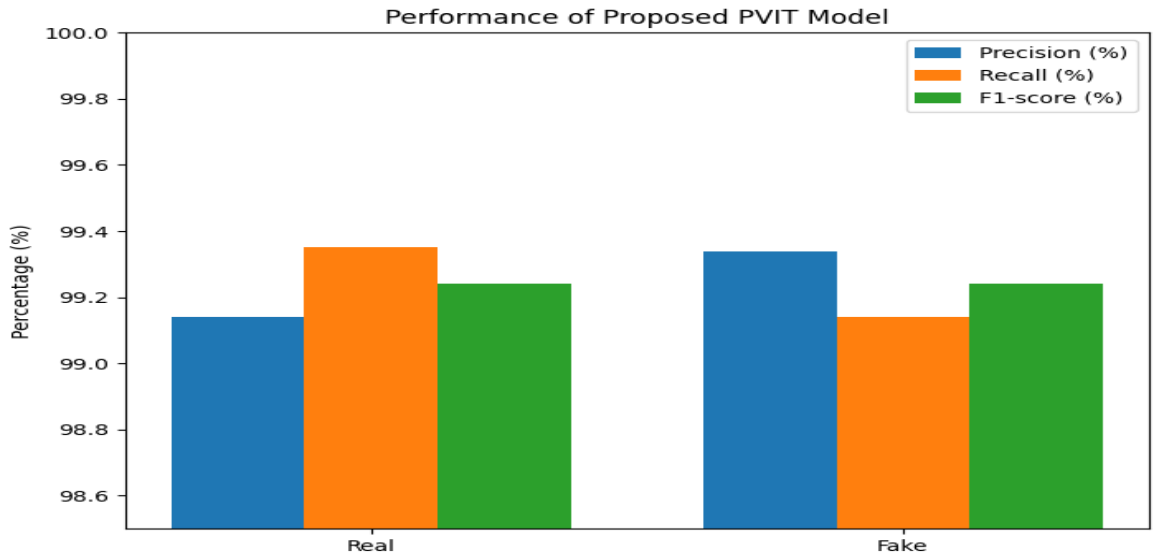

Figure 3: Performance of Proposed PVIT Model

According to these findings, the ACC-GAN model has improved the mean values of performance metrics like 95.7% PRE, 99.3% SEN, 90.0% SPE, and 98.40% ACC. Additionally, the suggested ACC-GAN model attains a ρ-value of less than 5% (1.6%), indicating that the outcomes obtained by ACC-GAN are noteworthy. With 1.4% accuracy, 3.3% sensitivity, 1.6% specificity, 3.8% precision, and an 8.4% ρ-value, the suggested model outperforms GPT-4o. The suggested model achieves minimum SD values of 0.7% accuracy, 0.3% sensitivity, 2.6% specificity, 2.8% precision, and 0.8% ρ-value when taking into account the SD value of the evaluation metrics.

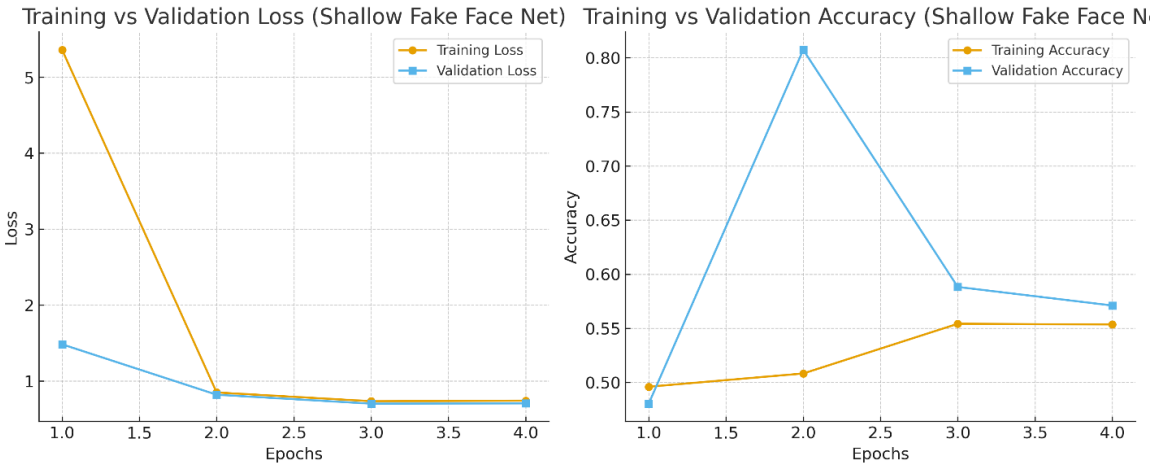

Figure 4: Accuracy and Loss values of CNN-PCA approach

Figure 4 illustrates the training and validation curves for the CNN-Principal Component Analysis approach, focusing on accuracy and loss. Accuracy, a number between 0 and 1, is displayed on the Y-axis in training and confirmation accuracy, indicating how well the model predicted. The X-axis displays the epochs, which indicate the number of full runs of the training dataset. The training accuracy is displayed by the yellow line. It begins at roughly 0.5138 and increases to roughly 0.7526 by the fifth phase, demonstrating how the model continues to get better as it absorbs information from the training set. The validation accuracy is displayed by the orange line. After the first epoch, it begins at roughly 0.5908, falls to roughly 0.5123, and then increases to roughly 0.5833. Based on these results, the proposed model achieves a 1.5% performance improvement across multiple datasets using various parameters.

6. Conclusion

Numerous scholars have been working extensively to identify misleading information using a variety of computational approaches. This review study demonstrates existing methods for the identification and classification of deepfake images. Because deepfake images are so realistic, it might be difficult for conventional

image analysis techniques struggle to emphasizes on the difficulty in determining the authenticity of the images and videos. There are various techniques for creating deepfakes, from deep neural networks to GANs. The model needs to be resistant to different methods of producing deepfakes. To ensure its efficacy over time, the solution should be designed to generalize and adapt effectively to emerging deepfake generation technologies. Adversarial attacks can be used by malicious actors to covertly alter deepfake content in order to fool detection systems. The legitimacy and security of digital content are seriously threatened by the growing realism and accessibility of modified images, particularly those produced by sophisticated AI methods like GANs. Even though they are effective and interpretable, traditional picture forensic techniques are unable to combat the sophistication of contemporary visual forgeries. Convolutional Neural Networks (CNNs), in particular, are deep learning-based models that provide high detection accuracy but have limited transparency and are susceptible to hostile attacks. This study addresses these constraints by proposing a hybrid strategy that combines deep learning methods with traditional forensic analysis. The proposed model achieves high accuracy, better generalization, and increased interpretability by fusing CNNs' feature-learning capabilities with Photo-Response Non-Uniformity analysis, compression artifacts, and other manually created features. Experimental results confirm the effectiveness of this hybrid framework, demonstrating that it outperforms standalone methods in identifying both traditional and AI-generated manipulations. Additionally, the model is more resilient to adversarial environments and offers clearer justifications for its conclusions, which makes it more appropriate for practical application in high-stakes fields like digital security, legal investigations, and journalism. Future research should examine additional fusion mechanism optimization, expand the system to manage video-based alterations involving videos, and enhance real-time detection capabilities. Such reliable and comprehensible detection techniques are essential for preserving confidence in visual media in the digital age, where disinformation is only becoming more widespread and complicated.

Ethics Declaration: This research did not require ethical approval as it did not involve any secret human participants, personal data collection, or sensitive information. All the sources and examples cited in this paper are publicly available and properly referenced/cited.

AI Declaration: AI tools, such as Google Gemmini and DeepSeek, were used to assist in refining academic language, and organizing the structure of the manuscript. All content generated by AI was critically reviewed, edited, and verified by the author(s) to ensure accuracy, originality, and adherence to scholarly standards. However, this manuscript similarity index was checked using plagiarism detection software and found to be within the acceptable range, indicating proper academic integrity and originality.

References

Camacho, I. C., et al. (2021). A Comprehensive Review of Deep-Learning-Based Image Forensics Techniques. *PMC*.

Gandhi, A., & Jain, S. (2020). Adversarial Perturbations Fool Deepfake Detectors.

Gandhi, A., & Jain, S. (2020). Adversarial Perturbations Fool Deepfake Detectors.

Hong, Y., Feng, J., Chen, H., Lan, J., Zhu, H., Wang, W., & Zhang, J. (2025, April). Wildfake: A large-scale and hierarchical dataset for ai-generated images detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 39, No. 4, pp. 3500-3508).

Killekar, A., et al. (2020). Two-Branch Recurrent Network for Isolating Deepfakes in Videos.

Kim, M., et al. (2021). FReTAL: Generalizing Deepfake Detection using Knowledge Distillation and Representation Learning.

Kim, M., et al. (2021). FReTAL: Generalizing Deepfake Detection using Knowledge Distillation and Representation Learning.

Kumar, B. A. (2025). Hybrid CMNV2: DeepFake Faces Classification and Detection. *ScienceDirect*.

Kumar, S., & Narang, R. (2025). Combating Digitally Altered Images: Deepfake Detection Using Vision Transformer.

Qureshi, S. M., et al. (2024). Deepfake Forensics: A Survey of Digital Forensic Methods for Deepfake Detection. *PMC*.

Singh, S. (2025). Unmasking Digital Deceptions: An Integrative Review of Deepfake Generation and Detection. *ScienceDirect*.

Sohail, S., Sajjad, S. M., Zafar, A., Iqbal, Z., Muhammad, Z., & Kazim, M. (2025). Deepfake Image Forensics for Privacy Protection and Authenticity Using Deep Learning. *Information*, 16(4), 270.

Verdoliva, L. (2021). Multimedia Forensics: Challenges and Opportunities in the Age of Deepfakes. *IEEE*.

Verdoliva, L. (2021). Multimedia Forensics: Challenges and Opportunities in the Age of Deepfakes. *IEEE*.

Wang, R., et al. (2020). FakeTagger: Robust Safeguards against DeepFake Dissemination via Provenance Tracking. Shan, S., et al. (2020). Fawkes: Protecting Privacy against Unauthorized Deep Learning Models.

Wang, R., et al. (2020). FakeTagger: Robust Safeguards against DeepFake Dissemination via Provenance Tracking.

Yadav, S. (2025). Enhancing Fake Image Detection with a Hybrid Approach. *IJSAT*.

Zhang, Y., et al. (2025). A Review of Generated Images and Deepfake Detection Techniques. *Springer*.