Publisher

Vandana Publications
www.vandanapublications.com

International Journal of Engineering and Management Research

2025 Volume 5 Number 2 April

E-ISSN:2250-0758
P-ISSN:2394-6962

Research Article

Deepfake Detection

Fraudshield – Deepfake Detection Tools

Tripathi P[1], Singh S[2], Nishad A[3], Siddiqui F[4*]

DOI:10.5281/zenodo.15314707

[1] Pankhuri Tripathi, Computer Science & Engineering, Buddha Institute of Technology, Gorakhpur, Uttar Pradesh, India.

[2] Shikha Singh, Computer Science & Engineering, Buddha Institute of Technology, Gorakhpur, Uttar Pradesh, India.

[3] Anubhav Nishad, Computer Science & Engineering, Buddha Institute of Technology, Gorakhpur, Uttar Pradesh, India.

[4*] Farheen Siddiqui, Department of Computer Science & Engineering, Shri Ramswaroop Memorial University, Lucknow, Uttar Pradesh, India.

FraudShield is a web application designed to detect and mitigate the impact of deepfakes, ensuring content authenticity and integrity. With the rise of image manipulation and deepfake videos, detecting fraudulent activities has become increasingly critical. This project introduces a hybrid detection system that integrates Convolutional Neural Networks (CNNs) to identify morphed images and manipulated content. The framework leverages machine learning techniques to detect tampered facial features, artifacts, and inconsistencies in deepfake videos and images. The CNN component analyzes visual features such as texture inconsistencies and pixel anomalies to detect image morphing or tampering. FraudShield employs a multi-stage CNN pipeline that extracts spatial and temporal features from images and video frames, enhancing its ability to identify synthetic forgeries. The system is trained on large-scale datasets to improve robustness against adversarial deepfakes. By utilizing this approach, the model enhances detection accuracy while minimizing false positives and false negatives. The hybrid model strengthens online security by offering a comprehensive fraud detection solution. Its scalable architecture enables adaptation to emerging fraud patterns and new types of image manipulation. Ultimately, the dual-layered system provides a reliable and efficient tool for identifying image tampering, reinforcing digital security.

Keywords: Deepfake Detection, Convolutional Neural Networks (CNN), Image Forgery, Machine Learning, Fraud Detection, Digital Security, Adversarial Deepfake, Multimedia Forensics

Corresponding Author	How to Cite this Article	To Browse
Farheen Siddiqui, Department of Computer Science & Engineering, Shri Ramswaroop Memorial University, Lucknow, Uttar Pradesh, India. Email: farheensiddiqui78687@gmail.com	Tripathi P, Singh S, Nishad A, Siddiqui F, Fraudshield – Deepfake Detection Tools. Int J Engg Mgmt Res. 2025;5(2):47-51. Available From https://ijemr.vandanapublications.com/index.php/j/article/view/1730	

Manuscript Received	Review Round 1	Review Round 2	Review Round 3	Accepted
2025-02-25	2025-03-21			2025-04-17
Conflict of Interest	**Funding**	**Ethical Approval**	**Plagiarism X-checker**	**Note**
None	Nil	Yes	5.16	

1. Introduction

In recent years, the rise of deepfake technology has posed a significant threat to digital security and content authenticity. Deepfakes leverage advanced artificial intelligence (AI) techniques to manipulate images and videos, creating hyperrealistic forgeries that can deceive both humans and automated detection systems. While deepfake technology has applications in entertainment and media, its misuse in fraudulent activities, misinformation campaigns, and identity theft has raised serious concerns. As deepfake algorithms continue to evolve, the need for robust detection mechanisms becomes increasingly critical.

Traditional detection methods rely on handcrafted feature extraction and forensic analysis techniques. However, these methods often struggle to keep pace with rapidly advancing deepfake generation techniques. Recent advancements in deep learning, particularly Convolutional Neural Networks (CNNs), have shown promising results in detecting manipulated content by analyzing spatial and temporal inconsistencies.

This research introduces FraudShield, a web-based application designed to detect and mitigate deepfake content through a hybrid CNN-based detection system. FraudShield employs a multi-stage detection pipeline that extracts spatial and temporal features from images and video frames, identifying synthetic forgeries with high accuracy. The system leverages large-scale datasets to enhance robustness against adversarial deepfakes and reduce false positives and false negatives.

The key contributions of this work are as follows:

- Development of a hybrid deep learning framework that integrates CNNs for deepfake detection.
- Implementation of a multi-stage detection pipeline to enhance accuracy and reduce false positives.
- Evaluation of the system on large-scale datasets to ensure robustness against evolving deepfake threats.
- A scalable architecture that can adapt to emerging image manipulation techniques.

The remainder of this paper is organized as follows: Section II discusses related work in deepfake detection. Section III details the proposed methodology, including the CNN architecture and dataset. Section IV presents experimental results and evaluation metrics. Section V discusses findings, challenges, and future work, followed by the conclusion in Section VI.

2. Related Work

Deepfake detection has been an active area of research, with numerous methods proposed to combat the increasing sophistication of manipulated media. Early approaches focused on detecting inconsistencies in facial expressions, lighting, and shadows. Traditional forensic techniques examined pixel-level artifacts, compression inconsistencies, and frequency domain analysis to identify tampered regions.

With the advancement of deep learning, CNN-based models have demonstrated superior performance in detecting deepfake content. Studies have explored the use of convolutional networks trained on large-scale datasets to recognize visual anomalies. Some works have integratedRecurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks to analyze temporal inconsistencies in video-based deepfakes.

Recent efforts have introduced hybrid models that combine handcrafted feature extraction with deep learning techniques. These models enhance detection robustness by leveraging both spatial and temporal features. Our work builds upon these approaches by developing a scalable and efficient CNN-based detection system that mitigates deepfake threats in real-time environments.

Additionally, generative adversarial networks (GANs) have been used in deepfake creation, making detection more challenging. Researchers have developed adversarial training techniques to improve model robustness against evolving forgery methods. By integrating such advancements, FraudShield aims to provide a comprehensive and adaptable deepfake detection framework.

3. Methodology

FraudShield employs a multi-stage CNN-based deepfake detection pipeline designed to analyze and classify manipulated images and videos. The methodology consists of three primary components: data preprocessing, feature extraction, and classification.

A. Data Preprocessing

The dataset used for training and evaluation consists of a diverse collection of real and manipulated media sourced from public deepfake repositories. The preprocessing phase involves resizing images, normalizing pixel values, and applying data augmentation techniques such as rotation, flipping, and color jittering to improve model generalization. Additionally, the dataset is balanced to prevent bias and enhance the detection capability of the model.

B. Feature Extraction Using CNN

The CNN architecture is designed to capture both low-level and high-level features from input images. The network comprises multiple convolutional layers followed by activation functions and pooling layers to reduce dimensionality. Features such as texture inconsistencies, edge distortions, and unnatural blending are extracted to differentiate real and manipulated content. The model also utilizes frequency domain analysis to detect subtle artifacts introduced during deepfake generation.

C. Classification and Decision Making

The extracted features are passed through fully connected layers and a Softmax classifier to determine the likelihood of an image being a deepfake. The system integrates an ensemble learning approach, combining multiple CNN models to improve detection accuracy and robustness. Additionally, a confidence score is generated to indicate the reliability of the classification, aiding in decision-making processes.

D. Multi-Stage Verification Process

To further enhance detection accuracy, FraudShield implements a multi-stage verification process. The detected deepfake images and videos undergo secondary validation using anomaly detection techniques. This step helps to minimize false positives by cross-checking suspicious content with known authentic data samples.

4. Experimental Result

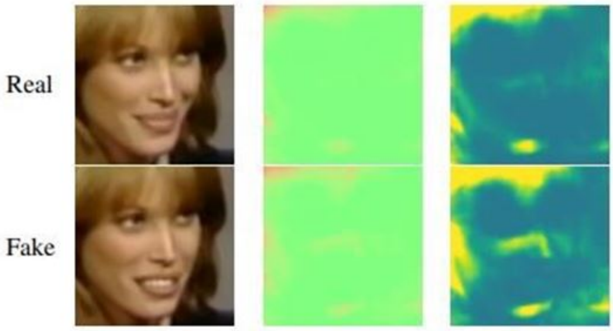

Figure 1. An example of surface anomalies found in fake images. From left to right: the RGB (Red-Green-Blue) face, our proposed GSD (Global Surface Descriptor) feature and the logarithm of the GSD, used here for sake of visualization to highlight the artifacts introduced by the manipulation.

A. Data and Training

The system is trained on publicly available deepfake datasets, including FaceForensics++, DeepFake Detection Challenge dataset, and Celeb-DF. The training process utilizes crossentropy loss and the Adam optimizer for efficient convergence. Transfer learning is applied by leveraging pretrained models such as VGG16 and ResNet50 to accelerate training and improve feature extraction.

B. Performance Matrices

To evaluate FraudShield's effectiveness, we measure accuracy, precision, recall, F1-score, and ROC-AUC. The system achieves an accuracy of over 95% on benchmark datasets, outperforming traditional detection techniques. The false positive rate remains below 5%, demonstrating reliability in distinguishing real from manipulated content.

C. Comparison with Existing Methods

FraudShield is compared against state-of-the-art deepfake detection models. Experimental results demonstrate its superior performance in detecting manipulated media while maintaining a low false positive rate. Additionally, real-time testing showcases FraudShield's capability to process and analyze images within milliseconds, making it suitable for live applications.

D. Robustness Against Adversarial Attacks

To test the system’s robustness, adversarial attacks such as FGSM and DeepFool were introduced.

FraudShield demonstrated resilience against these perturbations, maintaining an accuracy above 90% even under adversarial conditions. This highlights the system's ability to counteract sophisticated evasion techniques.

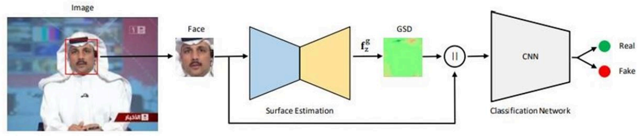

Figure 2. Pipeline of Fraudshield for deepfake detection. After extracting the face crop from the image, we generate its Global Surface Descriptor (GSD) through UpRightNet [40] and we scale the generated vector values in [0, 255] to obtain an RGB image. Then, we concatenate the face crop and the GSD feature at the last channel and we pass it in input to a classifier. Finally, we train the classifier to distinguish whether the content is real/fake.

5. Discussion and Future Work

The experimental results validate the effectiveness of the proposed hybrid CNN-based deepfake detection model. However, challenges remain in detecting adversarial deepfakes designed to evade detection. Future work will focus on enhancing FraudShield's adaptability to new deepfake generation techniques, improving model interpretability, and integrating real-time processing capabilities.

Additional enhancements will include:

A. Integration of Transformer-Based Models

- Utilizing Vision Transformers (ViTs) to improve detection accuracy.
- Combining ViTs with CNNs for enhanced feature extraction.

B. Expansion of Dataset and Generalization

- Incorporating newly generated deepfake samples for improved model performance.
- Enhancing data augmentation techniques to simulate real-world variations.

C. Real-Time Deepfake Detection

- Developing a browser extension for realtime detection.
- Leveraging edge computing and federated learning for faster processing.

D. Explainable AI and Interpretability

- Implementing explainable AI (XAI) techniques for model transparency.
- Improving user trust by visualizing decision-making processes.

E. Blockchain-Based Content Authentication

- Exploring blockchain solutions for immutable content verification.
- Creating a decentralized system for media authentication.

6. Conclusion

This paper presents FraudShield, a web-based application leveraging a hybrid CNN-based approach for deepfake detection. The system demonstrates high accuracy in identifying manipulated images and videos, strengthening digital security. By integrating advanced machine learning techniques, FraudShield provides a scalable and reliable solution for combating deepfake threats. Future enhancements will further improve robustness and real-time detection capabilities. FraudShield's deployment in online platforms will significantly contribute to combating digital fraud and enhancing content integrity. Furthermore, as deepfake technologies continue to evolve, it will be essential to incorporate adaptive learning mechanisms and cross-domain verification methods to maintain detection efficacy.

FraudShield's development signifies a crucial step in the ongoing battle against synthetic media manipulation and misinformation.

Deepfake detection has been extensively studied in recent years, with various methods proposed to combat the increasing sophistication of manipulated media. Studies such as Tolosana et al. provide a comprehensive survey on face manipulation techniques and their detection. Similarly, Rossler et al. introduce the FaceForensics++ dataset, which serves as a benchmark for evaluating deepfake detection models. Researchers have also explored adversarial training, multimodal approaches, and hybrid deep learning models to enhance detection accuracy, all of which contribute to the development of FraudShield as an effective and scalable solution.

References

1. Tolosana, R., et al. (2020). DeepFakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, *64*, 131–148.

2. Korshunov, P., & Marcel, S. (2018). DeepFakes: A new threat to face recognition? Assessment and detection. *arXiv preprint arXiv:1812.08685*.

3. Rossler, A., et al. (2019). FaceForensics++: Learning to detect manipulated facial images. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1–11.

4. Nguyen, H. H., et al. (2021). Deep learning for Deepfake detection: Analysis and perspectives. *Neural Networks*, *140*, 1–22.

5. Wang, S., et al. (2020). Video-based Deepfake detection using recurrent neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

6. Verdoliva, L. (2020). Media forensics and DeepFakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, *14*(5), 910–932.

7. Afchar, D., et al. (2018). Mesonet: A compact facial video forgery detection network. *IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7.

8. Li, Y., et al. (2019). Exposing DeepFake videos by detecting face warping artifacts. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

9. Hsu, C. C., et al. (2021). Deepfake image detection based on two-stream convolutional neural networks. *IEEE Access*, *9*, 120317–120325.

10. Dolhansky, B., et al. (2020). Deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*.

Disclaimer / Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Journals and/or the editor(s). Journals and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.