IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 1, January 2026

DOI: 10.17148/IJARCCE.2026.15107

Deepfake Detection Using Convolutional Neural Networks (CNN)

Vijay Chakole[1], Akshita Lanjewar[2], Astha Jadhao[3], Pallavi Chikate[4], Mayuri Sawalakhe[5]

Assistant Professor, Department of Electronics & Telecommunication Engineering,

KDK College of Engineering, Nandanvan, Nagpur-440009, Maharashtra, India[1]

Student, Department of Electronics & Telecommunication Engineering,

KDK College of Engineering Nandanvan, Nagpur-440009, Maharashtra, India[2-5]

Abstract: Recent improvements in AI-driven content synthesis have made it possible to fabricate highly realistic human imagery and video sequences. These artificially produced visuals now resemble genuine recordings so closely that older inspection practices struggle to recognize what has been digitally altered. This situation has amplified concerns related to online authenticity, public trust, and misuse of sensitive digital content. Because manual or traditional forensic checks no longer keep pace with advanced manipulation tools, dependable automated detection systems have become essential. Models built using convolution-based learning strategies are frequently employed for this purpose because of their ability to uncover fine-grained visual abnormalities introduced during fabrication. This article reviews studies exploring such models for identifying altered still images, detecting AI-produced visuals, and analyzing videos using both spatial patterns and motion information. While these approaches show promising performance on curated datasets, they often face difficulties under real-world conditions. The review highlights current limitations—such as softness under compression, dataset dependency, and weak generalization—and identifies the need for stronger multimodal and adversarial-resistant solutions.

Keywords: Deepfake Detection, Convolutional Neural Networks, AI-Generated Images, Video Forgery Analysis, Multimedia Forensics, Digital Authenticity

I. INTRODUCTION

Synthetic media that resemble real human appearances have become increasingly common due to the growth of advanced generative algorithms such as adversarial networks and encoder–decoder systems. These methods can modify facial behavior, recreate someone's likeness, or even invent entirely new identities, resulting in visuals that appear authentic to untrained observers. Although such technology originally served entertainment and creative purposes, it is now linked to activities such as impersonation, falsified narratives, privacy violations, and digital manipulation for harmful intent.As these generation tools improve, distinguishing untouched media from fabricated content becomes progressively more challenging. Earlier verification techniques—such as manually inspecting frames, checking metadata, or examining basic artifacts—are no longer reliable when faced with high-quality synthetic media. The rapid spread of manipulated content on social platforms further intensifies the need for automated detection mechanisms. Convolution-based recognition models have gained attention because they can capture layered spatial cues and identify features that do not align with natural visual patterns. Work in this field largely falls into three categories:

a. Image-level identification of altered visuals using convolutional classifiers

b. Recognition of general AI-produced images beyond facial manipulation

c. Video-level analysis that merges spatial information with motion-based learning through temporal architectures

Despite steady improvement, these techniques continue to face barriers involving robustness, dataset diversity, adaptability, and vulnerability to adversarial changes. This motivates further investigation into more resilient and broadly applicable detection frameworks.

II. RELATED WORK

A. Image-Based Deepfake Detection Using CNN Tobing et al. [1] developed a CNN-based approach for detecting manipulated facial images by learning spatial inconsistencies introduced during synthesis. Their results showed that CNNs can reliably distinguish forged images from authentic ones without using handcrafted features.

B. Survey of CNN-Based Deepfake Detection Techniques Kumar et al. [2] conducted a detailed survey of CNN-driven deepfake image detection methods. The study compared various architectures and training strategies, highlighting strong detection performance but also noting limitations such as dataset dependency and poor generalization to unseen manipulations.

C. Detection of Fully AI-Generated Images Nayak et al. [3] explored the identification of images generated entirely by AI systems using deep learning models. Their findings indicated that CNNs outperform traditional classifiers by capturing texture-level artifacts, while transfer learning significantly improves accuracy and training efficiency.

D. Ensemble CNN Models for Robust Detection Sharma et al. [4] proposed a GANCNN ensemble framework to improve robustness against evolving deepfake generation techniques. By incorporating generative replay, the model reduced catastrophic forgetting and achieved better performance under compression, though with increased computational cost.

E. Temporal Feature-Based Video Deepfake Detection Alagi and Patil [5] reviewed CNN-based deepfake detection techniques that integrate temporal information. Their study emphasized that combining spatial CNN features with temporal models such as LSTMs and GRUs is essential for accurately detecting manipulated videos.

F. Physiological Signal-Based Detection Agarwal et al. [6] introduced a deepfake video detection method based on heart rate estimation from facial color variations. While the approach effectively exploited physiological inconsistencies, its performance was sensitive to lighting conditions and video quality.

G. Head Pose Inconsistency Analysis Yang et al. [7] proposed a CNN-based technique that detects deepfake videos by identifying abnormal head pose relationships. Their method demonstrated high accuracy, particularly for early-generation deepfakes that failed to preserve natural geometric constraints.

H. Impact of Deepfakes on Biometric Systems Korshunov and Marcel [8] examined the vulnerability of face recognition systems to deepfake attacks. Their work highlighted deepfakes as a significant threat to biometric security and stressed the importance of reliable CNN-based detection mechanisms.

I. Feature Fusion-Based Deepfake Detection Fernandez et al. [9] developed a CNN-based deepfake detection system that combines multiple spatial features using feature fusion techniques. The integrated approach improved robustness and accuracy compared to single-feature CNN models.

J. Benchmark Datasets for Deepfake Detection Rossler et al. [10] introduced the FaceForensics++ dataset, which serves as a standard benchmark for evaluating deepfake detection algorithms. Their evaluation showed that CNNs perform well on manipulated images but suffer performance degradation under heavy compression.

K. Frequency-Domain Analysis for Deepfake Detection Qi and Zheng [11] proposed a hybrid detection approach combining frequency-domain features with CNN classifiers. Their method demonstrated improved robustness against compression and noise compared to purely spatial CNN-based techniques.

L. Behavioral Cues for Video Deepfake Detection Li et al. [12] presented an early deepfake detection method based on abnormal eye blinking behavior. Using CNN-based feature extraction, the study showed that early synthetic videos failed to reproduce natural blinking patterns, though later deepfakes reduced this limitation.

III. DISCUSSION AND RESEARCH GAP

Although CNN-based deepfake detection methods have achieved high accuracy on benchmark datasets, their effectiveness often diminishes in real-world environments. A key limitation is their strong dependence on training data characteristics, which causes models to learn dataset-specific patterns rather than general manipulation cues. Consequently, performance degrades when detectors encounter unseen generation techniques, varying compression levels, or post-processing operations commonly applied on social media platforms.Another significant challenge is the limited ability of many CNN-based approaches to capture temporal inconsistencies in video content. While spatial feature extraction is effective for identifying frame-level artifacts, it fails to model dynamic facial behaviors such as eye movement, head motion, and expression continuity. These temporal patterns are crucial for detecting advanced video-based deepfakes, where individual frames may appear visually authentic.

Furthermore, most existing detection systems rely exclusively on visual information, neglecting other potentially informative modalities such as audio coherence, lip–speech synchronization, physiological signals, and metadata analysis. The absence of multimodal integration restricts the robustness and reliability of current solutions, particularly in complex real-world scenarios.

Model transparency also remains an unresolved issue. Many CNN-based detectors operate as black-box systems, providing only binary classification outputs without explaining the underlying decision process. For forensic, legal, and security applications, the lack of interpretability reduces trust and limits practical adoption.Overall, the limitations related to generalization, temporal modeling, multimodal fusion, and explainability highlight critical research gaps. Addressing these challenges is essential for developing scalable, robust, and trustworthy deepfake detection systems suitable for real-world deployment.

DOI: 10.17148/IJARCCE.2026.15107

IV. PROPOSED METHODOLOGY

The primary purpose of this work is to design and analyze a robust deepfake detection framework capable of identifying manipulated images and videos under diverse real-world conditions. With the rapid evolution of AI-driven content synthesis, existing detection models often fail when exposed to unseen manipulation techniques, compression artifacts, or low-quality media. This study aims to address these challenges by combining spatial feature learning with temporal modeling to improve generalization and reliability.

Specifically, the objectives of this work are:

- To analyze visual artifacts introduced by deepfake generation using convolution-based feature extraction
- To capture temporal inconsistencies present in manipulated video sequences
- To evaluate detection performance across multiple datasets and compression levels
- To highlight the limitations of purely spatial models and motivate hybrid spatial–temporal solutions

By achieving these goals, the proposed approach seeks to contribute toward more adaptable and scalable deepfake detection systems suitable for practical deployment.

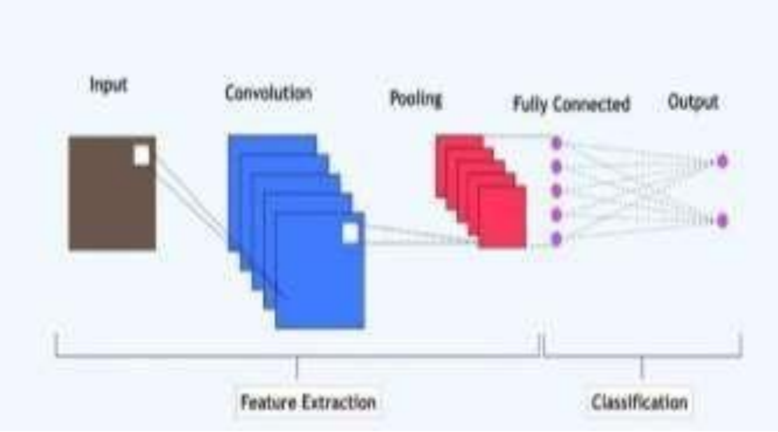

Figure 1.1. CNN Architecture Used for Feature Extraction and Classification

The proposed methodology follows a structured pipeline designed to detect both image-based and video-based deepfakes effectively. The overall workflow of the system is illustrated in Figure 1.1, which depicts the CNN-based architecture used for feature extraction and classification.

Dataset Collection

To ensure diversity and reduce dataset bias, multiple publicly available datasets are utilized, including CIFAKE, Face Forensics++, DFDC, Celeb-DF, and Deeper Forensics. These datasets contain a balanced mix of real and manipulated media generated using different synthesis techniques, resolutions, and compression levels.

Preprocessing and Data Augmentation

Input media undergoes preprocessing steps such as face detection, alignment, resizing to a fixed resolution, and pixel normalization. To enhance robustness and reduce overfitting, extensive data augmentation is applied, including rotation, horizontal flipping, brightness variation, and noise injection.

Spatial Feature Extraction

A convolutional neural network or transfer-learning backbone (such as ResNet or Xception) is employed to extract discriminative spatial features. These features capture subtle texture inconsistencies, illumination mismatches, and structural anomalies commonly introduced during deepfake generation. This stage corresponds to the feature extraction block shown in Figure 1.1.

Temporal Modeling for Video Analysis

For video-based detection, frame-level features are passed to a temporal learning module such as an LSTM, GRU, or 3D convolutional layer. This component captures motion-based irregularities including abnormal blinking patterns, unnatural head movements, and inconsistent facial dynamics across frames.

DOI: 10.17148/IJARCCE.2026.15107

Classification and Training

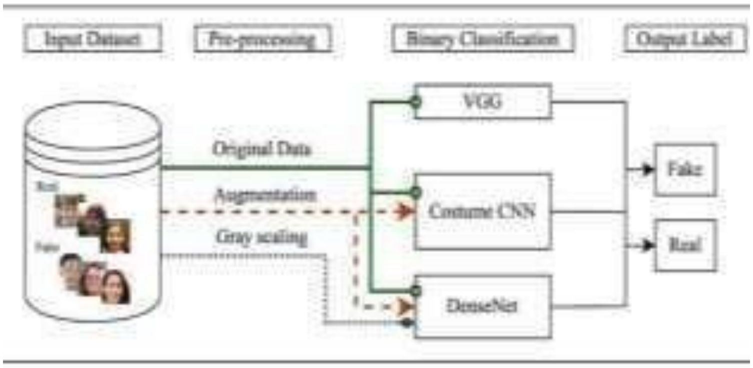

Figure 1.2. Overview of the Model Training

The combined spatial and temporal features are fused and passed to a classification layer, as illustrated in Figure 1.2, which shows the overall training process. A Softmax classifier is used for video classification, while a Sigmoid-based output is employed for binary image classification.

Evaluation Metrics

Model performance is evaluated using accuracy, precision, recall, F1-score, and AUC. Additional testing is conducted under varying compression levels to assess robustness and real-world applicability.

V. CONCLUSION

Convolution-based learning techniques continue to play a vital role in the detection of manipulated digital media due to their effectiveness in extracting fine-grained spatial features. However, the rapid advancement of synthetic content generation—particularly in video-based manipulations—has exposed the limitations of approaches that rely solely on spatial analysis. Robust deepfake detection requires not only detailed frame-level inspection but also an understanding of temporal dynamics and resilience to variations in data quality.The reviewed literature indicates that although CNN-based models demonstrate strong performance on controlled datasets, their effectiveness is often reduced when confronted with dataset shifts, compression artifacts, adversarial manipulation, and limited model transparency. To address these challenges, this study emphasizes a hybrid spatial–temporal framework that integrates convolutional feature extraction with sequence-level learning to better capture motion inconsistencies present in manipulated videos.

While the proposed approach shows promising results across diverse scenarios, detecting highly realistic or severely compressed deepfakes remains challenging. Future research should focus on incorporating multimodal information such as audio–visual consistency, leveraging advanced transfer learning techniques, utilizing frequency-domain representations, and improving interpretability through explainable AI methods.Overall, the development of reliable, scalable, and interpretable deepfake detection systems is essential for maintaining trust in digital media and safeguarding individuals, organizations, and society against the malicious use of synthetic content.

REFERENCES

[1]. L. M. S. Tobing, M. Wahyudi and D. R. I. M. Wibowo, “Deepfake Detection Using Convolutional Neural Network,” International Journal of Electrical Engineering and Computer Science, vol. 31, no. 2, pp. 1– 10, 2023.

[2]. R. Kumar, S. Kumar and P. R. S. Kumar, “A Survey on Deep Fake Image Detection Using CNN Model,” International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), vol. 12, no. 6, pp. 45–51, 2025.Varun, S. et al., “True Single-Phase Clock (TSPC) Flip-Flop and Shift Register Designs — Comparative Study,” (2025).

[3]. P. Nayak, S. R. Patro and S. Das, “Detection of AI Generated Images Using Machine Learning and Deep Learning Models,” International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), vol. 13, no. 4, pp. 112–118, 2024.S. Sanadhya and S. Sharma, “DC-DB PFAL Adiabatic Logic for LowPower Shift Registers,” (2025).

[4]. Sharma, P., Kumar, M., & Sharma, H. K. (2024). GANCNN Ensemble: A Robust Deepfake Detection Model of Social Media Images Using Minimized Catastrophic Forgetting and Generative Replay Technique. Procedia Computer Science, 235, 948-960.F. D’Aniello, M. Tettamanti, S. A. A. Shah, and A. Baschirotto, “Single-Event

DOI: 10.17148/IJARCCE.2026.15107

Upset Characterization of a Shift Register in 16 nm Bulk FinFET Technology," *Electronics*, vol. 14, no. 7, 2025, DOI: 10.3390/electronics14071421.

[5]. B. Alagi and R. Patil, "Literature Survey: Deepfake Detection Using CNN and Temporal Feature," International Journal of Scientific Research in Engineering and Technology (IJSRET), vol. 11, no. 3, pp. 250–260, 2025.

[6]. S. Agarwal, H. Farid, Y. Gu, and M. He, "Detecting DeepFake Videos from Heart Rate Estimates," IEEE International Conference on Image Processing (ICIP), pp. 1–5, 2020.

[7]. X. Yang, Y. Li, and S. Lyu, "Exposing Deep Fakes Using Inconsistent Head Poses," IEEE ICASSP, pp. 8261–8265, 2019.

[8]. P. Korshunov and S. Marcel, "Deepfakes: A New Threat to Face Recognition? Assessment and Detection," arXiv preprint arXiv:1812.08685, 2018.

[9]. A. Fernandex, R. Bharati, and P. Saikia, "Deepfake Video Detection Based on CNN and Feature Fusion Approach," IEEE Symposium Series on Computational Intelligence (SSCI), pp. 2109–2114, 2022.

[10]. J. Rossler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1–12, 2019.

[11]. Z. Qi and Z. Zheng, "Deepfake Image Detection Using Frequency Domain Analysis and CNN Classifier," IEEE Transactions on Information Forensics and Security, vol. 17, pp. 195–207, 2022.

[12]. Y. Li, M. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Eye Blinking," IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7, 2018.