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. 14, Issue 11, November 2025

DOI: 10.17148/IJARCCE.2025.141112

Advances in AI and ML for Face-Swap Deepfake Detection: A Comprehensive Review

Shriya Arunkumar[1], Aaradhana. R[2], Sadiya Noor[3], Sanskriti Raghav[4],

Dr. Kushal Kumar B N[5]

Dept of CSE-ICB, KSIT, Karnataka, India[1-4]

HOD and Assoc. Prof., Dept of CSE-ICB, KSIT, Karnataka, India[5]

Abstract: Deepfake images, particularly those generated through face-swapping techniques, have become increasingly realistic and widespread, raising serious concerns about digital trust, personal privacy, and public safety. As these manipulated visuals grow more sophisticated, detecting them reliably has become a pressing challenge. This paper proposes a deep learning-based approach for the automatic detection of face-swapped deepfakes using Convolutional Neural Networks (CNNs). Our method focuses on identifying subtle visual cues and inconsistencies introduced during the face manipulation process—artifacts that are often imperceptible to the human eye. To enhance detection accuracy and robustness, we integrate advanced image preprocessing, feature extraction, and data augmentation techniques. The model is trained and evaluated on widely used benchmark datasets containing a mix of authentic and manipulated images. Experimental results demonstrate high accuracy and generalization capability, reinforcing the practical value of the proposed solution for real-world applications in digital content verification. By automating the detection process, this work contributes meaningfully to the field of media forensics and supports ongoing efforts to preserve the authenticity and integrity of visual media in the age of synthetic content.

Keywords: Deepfake Detection, Face-Swapping, Convolutional Neural Networks (CNN), Image Forensics, Synthetic Media, Digital Content Verification, Image Preprocessing, Media Integrity, AI-generated Images, Feature Extraction.

I. INTRODUCTION

Recent advancements in AI-generated synthetic media have prompted the development of several techniques for detecting face-swapped (DeepFake) content. Existing research in this domain has largely focused on identifying spatial artifacts in static images or leveraging background modeling techniques to distinguish manipulated facial regions. In our work, we explored a combination of AIML-based methods—including object tracking, identity verification, and classification models—to enhance the detection of face-swapped media, particularly in video sequences.

Traditional approaches often rely on detecting inconsistencies in facial alignment or blending artifacts. For instance, many studies have attempted to separate forged facial features by comparing them to a single or multiple background models, enabling the detection of unnatural face movements. However, these models typically lack robustness when faced with complex video dynamics.

To address temporal inconsistencies, we incorporated frame-wise tracking of facial regions to monitor motion coherence and expression continuity key indicators of synthetic manipulation. Fan et al. proposed a method that significantly reduced false positive rates by employing a finite state machine along with a single background model, allowing for the detection of short-lived facial manipulations. While effective in certain cases, their approach struggled under conditions involving dynamic lighting or nuanced facial expressions.

Inspired by advances in other domains, Omrani introduced a stereovision-based model tailored for object differentiation in underwater scenes. Their technique utilized stereo imaging and depth estimation to distinguish real objects from synthetic ones, which holds potential for adaptation in face-swap detection under dynamic backgrounds. Drawing from this, our work integrated biometric signals such as facial landmarks, blink patterns, and gaze direction with deep learning models to improve detection reliability across diverse scenarios. This holistic approach aims to reduce false positives and improve generalization against evolving face manipulation technologies.

II. OVERVIEW

The rapid advancement of artificial intelligence and machine learning has revolutionized the way digital media is created and consumed. Among the most striking—and potentially dangerous—applications of this technology is the rise of

DOI: 10.17148/IJARCCE.2025.141112

deepfakes. These AI-generated manipulations often rely on face-swapping techniques to superimpose one person's face onto another's body with high realism. While originally developed for entertainment and creative use, deepfakes have increasingly been exploited for malicious purposes, including political misinformation, identity theft, cyberbullying, and reputational harm.

This ongoing project seeks to address the growing threat of deepfake media by developing a robust AI/ML-based detection system specifically designed to identify face-swap manipulations in both images and videos. Our proposed framework combines computer vision and deep learning techniques to automatically analyze facial features, temporal coherence, lighting anomalies, and mismatched audio-visual cues that typically indicate tampering.

Key components of the system include:
1.Face Detection and Tracking for segmenting facial regions and monitoring consistency across frames.
2.Visual Realism Analysis to detect unnatural lighting, rendering artifacts, and irregular blinking patterns.
3.Behavioral and Expression Consistency Checks to verify lip-sync accuracy and facial movement against typical human behavior.
4.Supervised Learning Models trained on curated datasets of real and manipulated content to improve classification accuracy.
5.Real-Time Detection Capabilities aimed at enabling rapid screening of media content for authenticity.
While significant progress has been made in designing and integrating these modules, the system is currently under active development. Current efforts are focused on improving detection precision, ensuring generalizability across diverse datasets, and enhancing resistance to adversarial attacks. Additionally, audio-visual synchronization verification is being explored to strengthen robustness against highly realistic deepfakes.
As the project moves forward, future phases will involve extensive testing, performance evaluation, and optimization, with the ultimate goal of delivering a practical and reliable tool for deepfake detection in real-world scenarios.

III. METHODOLOGY

A. Overview

The goal of this systematic literature review was to find, assess, and summarize the body of knowledge regarding AI and machine-learning-based face-swap deepfake detection solutions. The three primary steps of the methodology were data extraction and synthesis, screening and selection, and literature search.

B. Literature Search

Using keywords like "deepfake detection," "face-swap identification," "AI-based forgery detection," and "machine learning for media forensics," the search was conducted across several major databases, including IEEE Xplore, ScienceDirect, SpringerLink, and arXiv. In order to incorporate recent developments in CNN, transformer, and GAN-based detection models, the time frame was limited to studies released between 2020 and 2025.

C. Screening and Selection

The selection of publications was based on their applicability to the detection of face-swap deepfakes, their use of AI or ML algorithms, and their inclusion of performance metrics. Excluded were studies with no empirical support or that only addressed deepfakes that weren't face-swapping (such as text or audio). Following abstract review and duplicate removal, a total of n papers were shortlisted.

D. Data Extraction and Analysis

Important information was extracted, including detection methods, model architectures (CNN, transformer, hybrid), datasets used (e.g., FaceForensics++, Celeb-DF, DFDC), evaluation metrics, and attained performance metrics. The reviewed works were then subjected to comparative analysis in order to find methodological advancements, limitations, and trends.

IV. RESEARCH STUDIES

Chen et al. [1] introduced DefakeHop, a lightweight detector that uses successive subspace learning and channel-wise Saab transforms. It has only 42,845 parameters and gets 100% AUC on UADFV and 90.56% on Celeb-DF v2. The technique employs feature distillation alongside spatial dimension reduction and soft classification to produce distinctive facial representations. Mishkhal et al. [2] undertook a thorough assessment of deep learning methodologies for facial swap detection, scrutinizing their advantages and drawbacks within contemporary deepfake datasets. The research tackles essential difficulties in identifying facial swaps amidst occlusion and minor modifications, highlighting

DOI: 10.17148/IJARCCE.2025.141112

considerable deficiencies in existing detection methodologies and datasets. Both works help improve deepfake detection by suggesting effective frameworks and thorough analyses of detection methods.

Amate et al. [3] investigated voice and video clone detection through deep learning, employing Domain Adversarial Training (DAT), biometric verification, and facial landmark detection to distinguish authentic from fabricated content by analyzing speech patterns, pitch, cadence, and temporal facial consistency. Sun et al. [4] introduced FakeTracer, a proactive defense method that embeds sustainable traces (STrace) and erasable traces (ETrace) into training faces, modifying the encoder-decoder face-swap model's learning process such that generated deepfakes retain only sustainable traces while erasable traces are removed, enabling authentication through marker detection. FakeTracer represents a paradigm shift from passive detection to proactive defense, requiring no architectural modifications or training process alterations while demonstrating robust efficacy across diverse DeepFake model architectures and post-processing perturbations, as validated on the Celeb-DF dataset. Both approaches address complementary deepfake challenges—Amate et al. focusing on reactive detection through temporal and acoustic feature analysis, while FakeTracer implements preventive protection through embedded forensic signatures. These methodologies collectively advance multimodal deepfake detection by combining detection-based and prevention-based defense mechanisms for comprehensive media authentication

Gupta et al. [5] created a system for detecting AI-based deepfake face manipulation using a Convolutional Neural Network and the FaceForensics++ dataset. It had a validation accuracy of 85% and a test accuracy of 77% across a range of manipulation techniques. Arshed et al. [6] put forward a Vision Transformer-based method for detecting multiple types of deepfakes using patch-wise analysis. This method solves problems that arise with Stable Diffusion and StyleGAN2 technologies. The Vision Transformer method got an F1 score of 99.90% on a dataset that was ready for multiple classes. This shows that it works better than traditional CNN architectures like ResNet-50 and VGG-16. Both approaches stress how important it is to have strong detection systems to stop the spread of fake media and facial manipulation in digital content. These contributions signify substantial progress in the creation of effective defense systems against the advancing techniques of deepfake generation.

Zhang and Zhao [b7] proposed a face-swap deepfake detection method combining Error Level Analysis (ELA) with a shallow two-layer convolutional neural network to identify JPEG compression artifacts indicative of manipulated facial regions, achieving 97% accuracy on the MUCT dataset while enhancing computational efficiency. Doan et al. [8] introduced BTS-E, an audio deepfake detection framework utilizing breathing-talking-silence encoders that segment audio into distinct acoustic components using Gaussian Mixture Models to extract natural bioacoustic markers like breathing patterns, which are challenging for synthetic speech to replicate. The BTS-E model demonstrated perfect classification accuracy with 1.0 AUPRC and significantly improved spoofing countermeasures. Together, these methods provide robust multimodal deepfake detection by leveraging compression artifacts in images and physiological acoustic features in audio for comprehensive authentication

Through comparative model analysis, Vajpayee et al. [9] achieved notable classification accuracy on deepfake human face images by proposing an efficient deepfake detection method that uses transfer learning with the EfficientNetV2 architecture. Compared to training from scratch, the method uses pre-trained weights and fine-tuning mechanisms to improve model generalization while using less computing power.

Mawa and Kabir [10] used pre-trained architectures—ResNet50, DenseNet201, and InceptionV3—tuned with global average pooling, ReLU activation, and dropout regularization (0.4 rate) to study deepfake face detection using transfer learning and ensemble neural network fusion. By utilizing complementary feature extraction capabilities, the weighted average ensemble methodology greatly increases detection accuracy by combining predictions from separate models. Joshi and Sinha [11] extracted statistical texture descriptors like contrast, correlation, homogeneity, energy, and dissimilarity from spatial pixel relationships at various angles (0°, 45°, 90°, and 135°) by combining deep learning with Grey Level Co-occurrence Matrix (GLCM) texture analysis for CelebDF(v2) detection. Their hybrid approach takes advantage of the fact that deepfake generation introduces distinctive texture anomalies by combining GLCM texture features with handcrafted and deep learning features.By combining texture-driven forensic analysis and ensemble-based feature fusion, both approaches enhance multimodal deepfake detection and offer supplementary strategies for strong authentication.

By employing dual recognition networks to analyze identity differences between facial regions and the surrounding context, Nirkin et al. [12] proposed detecting face swaps and achieved state-of-the-art results on the FaceForensics++, Celeb-DF-v2, and DFDC benchmarks. The technique takes advantage of the finding that manipulation only modifies certain facial features while maintaining contextual consistency, resulting in broadly applicable differences between

DOI: 10.17148/IJARCCE.2025.141112

invisible manipulation techniques. Using the Adam optimizer with categorical cross entropy loss, Badale et al. [13] created a deepfake detection system that combined layers of convolutional and dense neural networks, achieving 91% accuracy on 900 deepfake videos. The method uses pooling layers and convolutional operations to extract frame-level features for the binary classification of authentic and fraudulent video content. Both approaches make substantial contributions to pixel-level manipulation detection using neural networks and deepfake detection using contextual analysis.

The Visual Realism Assessment (VRA) model by Sun et al. [14] evaluates the perceived realism of face-swap videos using eleven machine learning and deep learning models trained on the DFGC-2022 dataset, which includes Mean Opinion Scores (MOS) annotations, leveraging handcrafted, deep features and Support Vector Regression to predict visual quality and deception risk.

Wang et al. [15] presented M2TR, a Multi-modal Multi-scale Transformer architecture that uses frequency domain information and multi-scale patch analysis through cross-modality fusion to capture manipulation artifacts across spatial levels. With the help of the SR-DF dataset, which contains 4,000 high-quality deepfake videos, the method obtained 99.76% accuracy on Celeb-DF and 99.50% accuracy on the FaceForensics++ raw dataset. Following a thorough comparison of deepfake detection techniques, such as feature-based, temporal-based, and deep feature-based methods, John et al. [16] suggested a semi-supervised GAN architecture that blends supervised and unsupervised discriminator training. Using multiclass classification with unlabeled data integration, the semi-supervised model obtained an accuracy of 92.30% on a dataset of 40,000 images. By using transformer-based multi-modal analysis and semi-supervised learning paradigms for reliable synthetic content identification, both approaches improve deepfake detection.

Nath et al. [17] conducted a comparative study on fake news detection using machine learning models such as Logistic Regression, Naïve Bayes, Decision Trees, SVM, and Random Forest against deep learning models, particularly LSTM networks, showing LSTM's superior ability to capture contextual dependencies but with higher computational cost. Haleev [18] proposed an AI-based swapped face detection method utilizing facial landmark discrepancies and geometric feature extraction to evaluate various face swap algorithms, identifying facial inconsistencies introduced during manipulation. The study highlights the complementary importance of linguistic context analysis for fake news detection and geometric landmark analysis for swapped face detection, advancing multimodal authenticity verification methods in media forensics.

Using transfer learning and custom convolutional neural networks, Kumar et al. [19] suggested deepfake image detection that achieves high accuracy on benchmark datasets by utilizing pre-trained models. The technique maintains detection performance on both real and manipulated facial images while lowering computational requirements by combining CNN feature extraction with transfer learning fine-tuning.

Ding et al. [20] proposed a swapped face detection approach utilizing deep convolutional neural networks augmented with subjective human assessments, focusing on perceptual and contextual anomalies that highlight inconsistencies introduced by face-swap algorithms. By combining automated feature extraction with expert judgment, their system flags subtle manipulations, yielding enhanced accuracy and robustness over traditional forensic methods. Korshunov and Marcel [21] conducted foundational research assessing deepfakes' impact on facial recognition benchmarks, revealing alarmingly high false acceptance rates (up to 95%) for GAN-generated synthetic faces, demonstrating significant vulnerabilities in conventional biometric authentication systems. Their work underscored the urgent need for advanced forensic detectors and standardized evaluation protocols for combating deepfake threats. Collectively, both Ding et al. and Korshunov and Marcel advocate integrating deep learning with perceptual and contextual analysis to counter emerging deepfake threats in biometric security and multimedia forensics.

V. LIMITATIONS

1. Methodology Gap: Current methods are fragmented. Many papers rely solely on standard CNNs, which have limited ability to capture diverse forgery artifacts. There is a lack of advanced, multi-stream fusion models that intelligently combine different types of features (e.g., spatial and frequency).

2. Generalization Gap: Models often show high accuracy on specific, known datasets but struggle to generalize when faced with new or different types of deepfakes. This is a major challenge for real-world applications.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025
DOI: 10.17148/IJARCCE.2025.141112

3. Efficiency Gap: Many high-performance deep learning models are computationally expensive, making the too slow for real-time detection on standard hardware.

4. Novelty Gap: The vast majority of research focuses on reactive detection (finding a forgery after it's made), with very little exploration into proactive solutions.

VI. CONCLUSION

The proposed methodology provides a structured and well-defined framework for developing an AI/ML-based system to accurately identify face-swap deepfakes. By systematically organizing the process into data acquisition, pre-processing, feature extraction, model training, and evaluation, it ensures consistency, scalability, and reproducibility. The integration of hybrid deep learning approaches—such as CNNs for spatial feature extraction and LSTMs for temporal sequence analysis—enables the detection of intricate inconsistencies that are imperceptible to humans. Furthermore, the use of benchmark datasets like FaceForensics++, Celeb-DF, and DFDC ensures standardization and facilitates objective performance evaluation.

Overall, this methodology provides a solid technical foundation for effectively detecting manipulated media across diverse formats and qualities. The evaluation metrics and validation strategies further emphasize model robustness and real-world applicability. Additionally, the inclusion of a deployment phase highlights the system's potential for practical integration into digital content verification tools. This structured approach not only enhances the system's accuracy and generalization but also contributes significantly to combating the increasing sophistication of face-swap deepfakes in digital media environments.

REFERENCES

[1]. H.-S. Chen, M. Rouhsedaghat, H. Ghani, S. Hu, S. You, and C. C. J. Kuo, "DefakeHop: A Light-Weight High-Performance Deepfake Detector," in Proc. IEEE Int. Conf. Multimedia and Expo (ICME), pp.1–6, 2021.
[2]. I. Mishkhal, N. Abdullah, H. H. Saleh, N. I. R. Ruhaiyem, and F. H. Hassan, "Facial Swap Detection Based on Deep Learning: Comprehensive Analysis and Evaluation," Iraqi J. Comput. Sci. Math., vol. 6, no.1, pp. 109–124, 2025.
[3]. S. Amate and A. Sarnaik, "Detecting Voice and Video Clones through Deep Learning and Artificial Intelligence: A Study on the Effectiveness of Techniques against Deep Fakes," Int. J. Adv. Eng. Manage., vol. 6, no. 6, pp. 568–574, June 2024.
[4]. P. Sun, H. Qi, Y. Li, and S. Lyu, "FakeTracer: Catching Face-swap DeepFakes via Implanting Traces in Training," IEEE Trans. Emerg. Topics Comput., 2024.
[5]. R. Gupta, S. Singh, and H. Kaur, "AI-Based Deep Fake Face Manipulation Detection As News," Int. J. Eng. Res. Technol., vol. 13, no. 6, June 2024.
[6]. M. A. Arshed, S. Mumtaz, M. Ibrahim, C. Dewi, M. Tanveer, and S. Ahmed, "Multiclass AI-Generated Deepfake Face Detection Using Patch-Wise Deep Learning Model," Comput., vol. 13, no. 1, p. 31, Jan. 2024.
[7]. Z. Zhang and J. Zhou, "Exposing Face-Swap Images Based on Deep Learning and ELA Detection," in Proc. IEEE Int. Conf. Image Process. (ICIP), pp. 1–5, 2019.
[8]. T. P. Doan, L. N. Vu, S. Jung, and K. Hong, "BTS-E: Audio Deepfake Detection Using Breathing-Talking-Silence Encoder," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), pp. 1–5, May 2023.
[9]. V. Vajpayee, R. Pal, and M. Dutta, "Detecting Deepfake Human Face Images Using Transfer Learning: A Comparative Study," in Proc. IEEE Int. Conf. Imaging Syst. Techn. (IST), pp. 1–6, 2023.
[10]. M. Mawa and S. Kabir, "Study on deepfake face detection using transfer learning approach," International Journal of Computer Applications, vol. 186, no. 37, pp. 1–10, 2024.
[11]. K. Joshi and A. Sinha, "Integrating GLCM Texture Analysis for Improved Deepfake Detection on CelebDF (v2) Dataset," in Proc. Asian Conf. Intell. Technol. (ACOIT), pp. 1–6, 2024.
[12]. Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "DeepFake Detection Based on Discrepancies Between Faces and Their Context," arXiv preprint arXiv:2008.12262, 2020.
[13]. A. Badale, L. Castelino, C. Darekar, and J. Gomes, "Deep Fake Detection using Neural Networks," International Journal of Engineering Research Technology (IJERT), vol. 10, no. 2, pp. 569–573, 2021.
[14]. X. Sun, B. Dong, C. Wang, B. Peng, and J. Dong, "Visual Realism Assessment for Faceswap Videos," arXiv preprint arXiv:2302.00918, 2023.
[15]. J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, S.-N. Lim, and Y.G. Jiang, "M2TR: Multimodal Multi-scale Transformers for Deepfake Detection," in Proc. ACM Int. Conf. Multimedia Retrieval (ICMR), pp. 450–458, 2022.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 11, November 2025
DOI: 10.17148/IJARCCE.2025.141112

[16]. J. John and B. V. Sherif, ”Comparative Analysis on Different DeepFake Detection Methods and Semi-Supervised GAN Architecture for DeepFake Detection,” in Proc. International Conference on Computer Science and Information Technology (ICCSIT), pp. 516–521, 2021.
[17]. A. Nath, S. Sharma, A. Ahuja, and R. Katarya, ”Study of Fake News Detection using Machine Learning and Deep Learning Classification Methods,” International Journal of Information Technology, vol. 13, no. 4, pp. 1665–1673, 2021.
[18]. M. Haleev,”Swapped Face Detection: AI-Based Method and Evaluation for Different Face Swap Algorithms,” in Proc. Conf. Open Innovations Assoc. FRUCT, pp. 564–570, 2023.
[19]. N. Kumar, P. Pranav, V. Nirney, and V. Geetha,”Deepfake Image Detection using CNNs and Transfer Learning,” Image and Video Processing, no. 9, pp. 1–6, 2024.
[20]. X. Ding, Z. Raziei, E. C. Larson, E. V. Olinick, P. S. Krueger, and M. Hahsler, ”Swapped Face Detection using Deep Learning and Subjective Assessment,” EURASIP Journal on Information Security, vol. 2020, no. 9, pp. 1–14, 2020.
[21]. P. Korshunov and S. Marcel, ”DeepFakes: a New Threat to Face Recognition? Assessment and Detection,” arXiv preprint arXiv:1812.08685, 2018.

