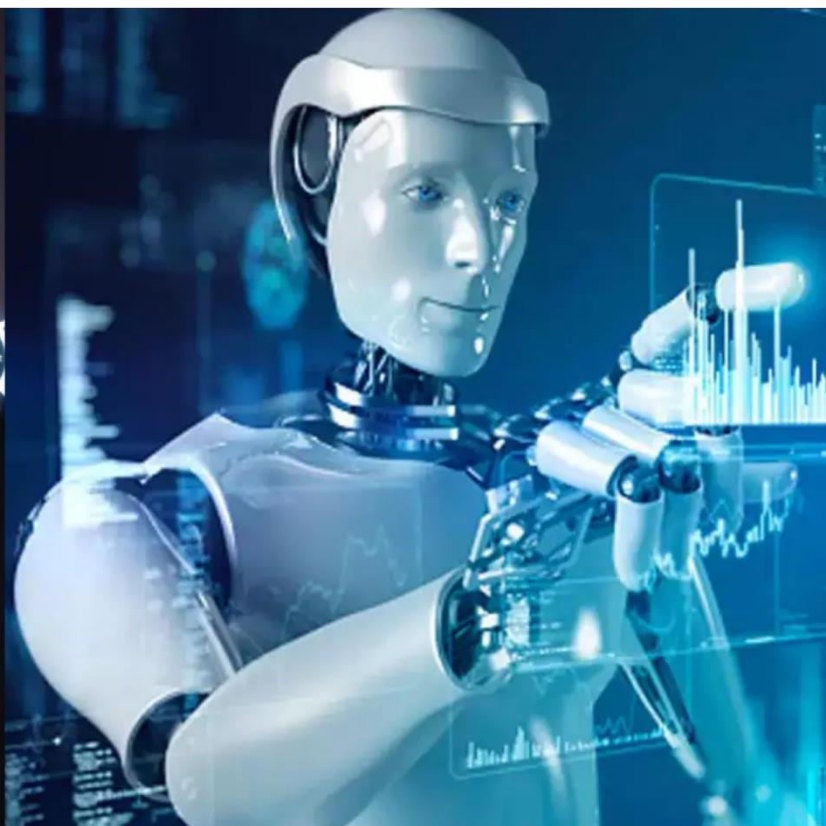




International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 8, August 2026

Deepfake Detection using AI

Rethik Rosan M.N¹, Saswinth J.C², Sanjeevee K.C³, Surya Kumar S⁴, Shiva S⁵, Rithish T⁶

Department of Computer Science, KPR College of Arts Science and Research, Coimbatore Dt., Tamil Nadu, India

ABSTRACT: The rapid advancement of artificial intelligence and deep learning has made it possible to generate highly realistic synthetic media, commonly referred to as deepfakes. Deepfake technology can manipulate facial identities, expressions and speech in images and videos, making fabricated content increasingly difficult to distinguish from authentic media. Although deepfake generation has legitimate uses in entertainment, education and digital content production, malicious use can contribute to misinformation, identity impersonation, fraud, cyberbullying and reputational damage. This paper presents an AI-based framework for detecting manipulated facial videos using deep learning. The proposed pipeline performs video preprocessing, frame sampling, face detection and extraction, image normalization, deep spatial feature extraction and binary classification. Xception and EfficientNet are considered for spatial representation learning, while an optional Long Short-Term Memory (LSTM) stage is used to model temporal relationships between consecutive frames. The study considers FaceForensics++, Celeb-DF and the DeepFake Detection Challenge (DFDC) datasets and emphasizes both within-dataset performance and cross-dataset generalization. Evaluation is defined using accuracy, precision, recall, F1-score and ROC-AUC. The paper deliberately distinguishes reported benchmark results from proposed experiments: no new experimental accuracy is claimed unless produced by an actual run. Prior benchmark evidence shows that deep learning detectors can perform extremely well under controlled conditions, while compression and distribution shifts can reduce performance. Therefore, robustness, generalization and interpretable evidence are treated as central requirements for practical deployment.

KEYWORDS: Deepfake Detection; Artificial Intelligence; Deep Learning; Convolutional Neural Network; XceptionNet; EfficientNet; LSTM; Computer Vision; Digital Forensics; FaceForensics++; Celeb-DF; DFDC; Video Classification.

I. INTRODUCTION

The increasing availability of artificial intelligence has transformed the way digital images, audio and videos are created and distributed. Modern generative systems can synthesize realistic human faces and manipulate existing videos with visual quality that can make ordinary human inspection unreliable. A deep fake is a form of synthetic media in which identity, facial appearance, expression or other characteristics are altered using machine-learning techniques. Deepfakes are commonly associated with face swapping and facial re-enactment. These techniques can be used for legitimate purposes such as film production, visual effects, education and entertainment, but they can also be misused to impersonate people, spread misinformation, manipulate public opinion, damage reputations or facilitate fraud. The central forensic question is therefore whether a digital video represents a genuine recording or an artificially manipulated one. Deepfake detection is a digital media-forensics problem. Useful evidence may occur in spatial appearance, temporal consistency, frequency characteristics, compression behavior or other artifacts. Earlier approaches investigated manually designed clues such as facial landmarks, eye blinking, head-pose consistency, image noise and facial boundaries. As synthesis methods improve, however, fixed handcrafted clues can become less reliable. Deep learning has changed this field by allowing models to learn discriminative representations directly from data. CNNs can learn spatial features from face crops, while sequential models can analyze relationships between neighboring frames. FaceForensics++ established an important benchmark containing multiple manipulation techniques and more than 1.8 million manipulated images, enabling standardized comparisons of learned detectors. A major challenge is that a detector can learn dataset-specific artifacts instead of general properties of manipulation. Consequently, a model that performs strongly on a single benchmark may degrade when it encounters another dataset, a different synthesis method or stronger compression. This paper therefore emphasizes cross-dataset evaluation and robustness rather than treating a single high accuracy number as sufficient evidence of practical reliability.

II. RELATED WORK

A. Traditional Deepfake Detection

Initial research examined manually identifiable inconsistencies in manipulated videos, including facial landmarks, head poses, eye blinking, facial expressions, image artifacts and boundary inconsistencies. Optical-flow approaches also explored inter-frame motion as evidence of manipulation. Such methods can be computationally efficient and interpretable, but their dependence on specific artifacts makes them vulnerable when generation pipelines change.

B. CNN-Based Detection

CNNs became a major approach because they automatically learn visual patterns from training images. FaceForensics++ reported strong performance from learned detectors and demonstrated that domain-specific face preprocessing can be important.

XceptionNet, which uses depthwise separable convolutions, became a widely used baseline for facial manipulation detection.

C. Face Forensics++

Face Forensics++ contains 1,000 original video sequences and manipulated versions generated using DeepFakes, Face2Face, Face Swap and Neural Textures. The benchmark includes more than 1.8 million manipulated images and different compression settings. These properties make it useful for controlled comparisons of detectors and for studying the effect of compression.

D. Celeb-DF

Celeb-DF was introduced to provide a more challenging high-quality deepfake benchmark. The dataset contains 5,639 high-quality DeepFake videos involving celebrities. Its generation process was designed to reduce obvious visual defects, making subtle forensic cues more important.

E. DeepFake Detection Challenge (DFDC)

The full DFDC dataset contains 124,000 videos and eight facial modification algorithms. The dataset was created for large-scale deepfake detection research and contains diverse real-world recording conditions. Its scale makes it valuable for testing whether a detector remains useful when content varies across subjects, lighting, pose and compression.

F. Benchmarking and Generalization

Deepfake Bench was proposed to address inconsistent data processing and evaluation practices. It integrates multiple datasets and detection methods under standardized protocols. This direction is important because reproducible evaluation and cross-dataset testing are necessary to determine whether a model has learned transferable forensic evidence.

III. METHODOLOGY

A. Video Preprocessing

An input video is decoded into a sequence of frames. Rather than processing every frame, representative frames can be sampled at a consistent interval. Sampling reduces computational cost while preserving a temporal sequence for optional LSTM analysis.

B. Face Detection and Extraction

Faces are detected in sampled frames and cropped so that the classifier concentrates on the facial region. Alignment and consistent cropping reduce irrelevant background variation and help the CNN focus on facial appearance. The face crops are then resized and normalized before model inference.

C. Spatial Feature Extraction

Xception and EfficientNet are considered as CNN backbones for learning spatial representations. The goal is to identify subtle texture, blending, boundary and other visual evidence that may distinguish authentic from manipulated facial content. Transfer learning can be used where appropriate, followed by fine-tuning on the selected benchmark data.

D. Temporal Feature Modeling

Frame-level CNN features can be arranged as a sequence and passed to an LSTM. The recurrent model is intended to capture temporal inconsistencies that may not be visible in an isolated frame. The temporal branch is optional because it introduces additional computation and should be retained only if it produces meaningful gains in recall, F1-score or robustness.

E. Binary Classification

The final classifier produces a binary prediction: authentic (REAL) or manipulated (FAKE). A probability score can accompany the class decision so that a threshold can be selected according to the intended application. The study does not claim that a particular threshold is universally optimal.

F. Evaluation Protocol

Experiments should include both within-dataset and cross-dataset evaluation. A within-dataset experiment measures performance when training and testing distributions are similar. A cross-dataset experiment is more demanding: for example, a model trained on one benchmark can be tested on another to reveal sensitivity to dataset-specific artifacts. Compression conditions should also be considered where benchmark versions permit such analysis.

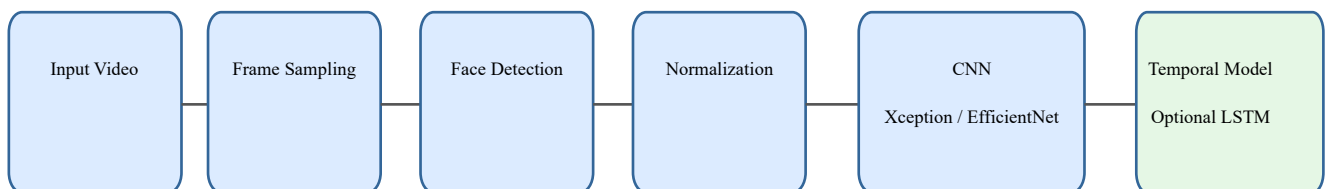


Fig.1. Proposed AI-based deepfake detection pipeline The diagram is an original schematic derived from the methodology described in this paper; it is not a claimed experimental result.

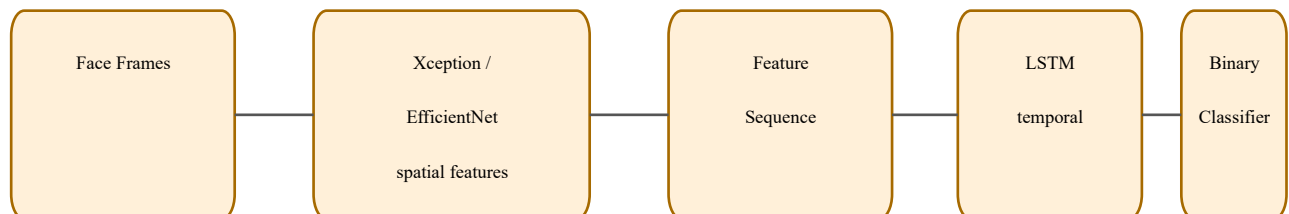


Fig.2. Spatial-temporal model concept The CNN extracts frame-level spatial features; an optional LSTM models the sequence before video-level binary classification. This is a proposed architecture diagram, not a report of a completed training run.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

- Because the supplied study material describes a research framework rather than a completed experimental run, this paper does not fabricate new accuracy, precision, recall, F1-score or ROC-AUC values.
- The experimental section instead defines the evaluation procedure and reports benchmark findings from cited literature as reference points. FaceForensics++ provides an important controlled reference.
- The original study reported an XceptionNet binary detection accuracy of approximately 99.26% on raw data, 95.73% on high-quality compressed data and 81.00% on low-quality compressed data.
- These figures are benchmark results from the cited FaceForensics++ study, not results obtained by the present student authors.
- The compression trend demonstrates why benchmark accuracy should not be interpreted as universal real-world performance.
- A detector may rely on visual or compression artifacts that are easier to learn in a particular dataset.
- Cross-dataset evaluation is therefore essential for estimating practical robustness. Recent benchmarking work has further emphasized standardized evaluation.

- DeepfakeBench integrates multiple datasets, detection methods and evaluation protocols to make comparisons more reproducible.
- Research on explainable XceptionNet–LSTM frameworks also indicates increasing interest in combining spatial features, temporal information and visual explanations such as Grad-CAM.
- For the proposed study, the principal comparison should be between spatial-only CNN classification and the optional spatial–temporal CNN-LSTM configuration.
- If the temporal model improves recall or F1-score consistently across datasets, the additional computation may be justified. If the gain is small, a simpler CNN detector may be preferable for efficient deployment.

Benchmark Datasets Considered

Benchmark	Size / characteristic
FaceForensics++	1,000 original videos; 4 manipulation methods; 1.8M+ manipulated images
Celeb-DF	639 deepfake videos; high-quality synthesis
DFDC	124k full-dataset videos; 8 manipulation algorithms

Fig.3. Benchmark datasets considered in this study Dataset sizes and characteristics are drawn from the cited FaceForensics++, Celeb-DF and official DFDC sources.

A. Evaluation Metrics

Accuracy measures the proportion of correctly classified samples. Precision measures how many samples predicted as manipulated are actually manipulated. Recall measures how many manipulated samples are correctly identified. F1-score provides a harmonic balance between precision and recall. ROC-AUC summarizes discrimination across classification thresholds and is useful when comparing detectors independently of one selected operating threshold. Cross-dataset generalization evaluates whether a detector transfers to unseen data distributions and manipulation methods.

B. Robustness Analysis

A strong detector should be evaluated under distribution changes, including different manipulation methods, compression levels, subject identities and recording conditions. The key research question is not simply whether a model can memorize a benchmark, but whether it learns forensic evidence that remains useful when the source distribution changes. This perspective is particularly important for practical digital-media verification.

Evaluation dimension	Recommended strategy
Within-dataset	Train/test from same benchmark
Cross-dataset	Train on one, test on another
Compression	Raw / HQ / LQ where available
Metrics	Accuracy • Precision • Recall • F1 • ROC-AUC

Fig.4.Recommended evaluation strategy The emphasis is on reproducibility, robustness and cross-dataset generalization rather than a single benchmark score.

V. DISCUSSION

The literature indicates a recurring trade-off between benchmark performance and generalization. Models trained and tested under closely matched conditions can achieve very high scores, while performance may decrease when compression or dataset distribution changes. This suggests that the model can exploit shortcuts or artifacts that are

correlated with a particular benchmark rather than learning universal indicators of manipulation. Another important consideration is temporal information. A single face frame can contain subtle blending or texture evidence, but a sequence may reveal inconsistent motion, expression transitions or other frame-to-frame irregularities. However, temporal modeling increases computational cost and does not automatically guarantee better generalization. It should therefore be evaluated experimentally rather than assumed to be superior. Interpretability is also important for digital forensics. A binary label alone gives limited evidence to a human investigator. Future implementations can consider explanation methods such as Grad-CAM to highlight influential facial regions, while clearly distinguishing model attention from definitive proof of manipulation. Finally, responsible deployment is essential. A detector should be treated as a decision-support system rather than an infallible authenticity oracle. False positives can wrongly label genuine content as fake, while false negatives can allow manipulated content to pass undetected. Threshold selection, confidence reporting, human review and documented evaluation conditions are therefore important.

VI. CONCLUSION AND FUTURE WORK

Deepfake technology represents an important challenge for digital media forensics because artificial intelligence can now generate highly realistic manipulated videos. This paper presented an AI-based framework that combines video preprocessing, face extraction, CNN-based spatial feature learning and optional LSTM-based temporal modeling. FaceForensics++, Celeb-DF and DFDC provide complementary benchmarks for evaluating detection performance, visual realism and large-scale diversity. The central finding from prior research is that high within-dataset accuracy does not by itself establish practical reliability; compression and cross-dataset distribution shifts can substantially affect performance. Accordingly, robustness and generalization should be treated as primary evaluation objectives. Future work should implement the proposed pipeline and perform controlled experiments across the three datasets, compare Xception and EfficientNet, quantify the contribution of the LSTM branch, evaluate compression robustness, and report accuracy, precision, recall, F1-score and ROC-AUC with clearly defined train/test splits. Further extensions can investigate frequency-domain features, multimodal audio-visual evidence, explainable predictions, calibration and efficient deployment. Any future numerical results should be reported only after actual experiments are completed and reproducibly documented.

REFERENCES

1. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1–11.
2. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. Canton Ferrer, "The DeepFake Detection Challenge (DFDC) Preview Dataset," arXiv:1910.08854, 2019.
3. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. Canton Ferrer, "The DeepFake Detection Challenge Dataset," arXiv:2006.07397, 2020.
4. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 3207–3216.
5. Z. Yan, Y. Zhang, X. Yuan, S. Lyu, and B. Wu, "DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection," arXiv:2307.01426, 2023.
6. A. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake Video Detection through Optical Flow Based CNN," Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.
7. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Page 4 Recognition Workshops, 2019, pp. 80–87.
8. Y. Yu, X. J. J. W. et al., "A Survey on Deepfake Video Detection," IET Biometrics, 2021.
9. "Deepfake Detection Using Deep Feature Stacking and Meta-Learning," Heliyon, vol. 10, 2024, e25933.
10. "An Ensemble of CNNs with Self-Attention Mechanism for DeepFake Video Detection," Neural Computing and Applications, 2023.
11. "Deepfake Detection via Spatial–Temporal Deep Networks: Leveraging CNNs and LSTMs for Enhanced Accuracy," Discover Applied Sciences, 2026.
12. "Decoding Deception: State-of-the-Art Approaches to Deep Fake Detection," Frontiers in Big Data, 2025.

13. “Deepfake Detection Based on the Adaptive Fusion of Spatial-Frequency Features,” International Journal of Intelligent Systems, 2024.
14. M. Kumar, A. Kumar, V. Yadav et al., “A Hybrid Spatial-Frequency Attention-Based Algorithm Using EfficientNet for Robust and Interpretable Deepfake Detection,” Scientific Reports, 2026.
15. “Deepfake Face Video Detection and Interpretation Using an Explainable XceptionNet and Long Short-Term Memory Framework,” Discover Applied Sciences, 2026.
16. “Multi-Level Feature Fusion for Generalized Face Forgery Detection,” Neurocomputing, vol. 653, 2025, article 131235. Source and integrity note: The four figures in this paper are original schematic diagrams created for this manuscript from the cited factual material. No fabricated experimental photographs, model outputs, accuracy tables or student-generated performance claims are presented. Benchmark numbers are explicitly identified as literature-reported values.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details