

An Intelligent AI–ML Framework for Deepfake Generation and Detection

P.B. Kudal^{1,*}, D.S. Joshi²

Abstract

Recent developments in Artificial Intelligence (AI) and Machine Learning (ML) have led to the rise of deepfake technology, making it one of the most impactful innovations in the digital world. Deepfakes are synthetic forms of digital media generated using deep learning algorithms capable of realistically manipulating facial expressions, voice patterns, images, and video content. Although these technologies provide innovative applications in entertainment, virtual reality, healthcare, and education, their misuse has created serious concerns related to misinformation, cybercrime, privacy violations, identity fraud, and political manipulation. The increasing realism of deepfake content has made manual identification extremely difficult, thereby creating a strong need for reliable automated detection systems. This research paper presents an AI–ML based deepfake detection framework designed to identify manipulated digital media using advanced deep learning techniques. The proposed system incorporates preprocessing, feature extraction, model training, and intelligent classification mechanisms to analyze visual inconsistencies commonly introduced during deepfake generation. Convolutional Neural Networks (CNNs) and Deep Neural Networks (DNNs) are utilized to detect abnormalities such as irregular facial textures, inconsistent lighting conditions, unnatural eye movements, and lip synchronization mismatches. The framework is trained using benchmark datasets containing both authentic and manipulated samples to improve detection accuracy and robustness. The paper also discusses the overall system architecture, detection methodology, algorithms, applications, challenges, and future research opportunities associated with deepfake technologies. The proposed approach contributes toward strengthening cybersecurity and digital forensics by providing an efficient mechanism for detecting forged media. The study highlights the importance of adaptive AI-driven detection systems in preserving digital authenticity, maintaining public trust, and minimizing the harmful effects of synthetic misinformation in modern digital communication environments.

Keywords: Artificial intelligence, machine learning, deepfake Detection, CNN, DNN, digital forensics, cyber security

*Author for Correspondence

P.B. Kudal
E-mail: preeti.kudal@ggsf.edu.in

¹HoD, Department of Artificial Intelligence and Machine Learning, Guru Gobind Singh Polytechnic Nashik, Maharashtra, India

²Lecturer, Department of Artificial Intelligence and Machine Learning, Guru Gobind Singh Polytechnic Nashik, Maharashtra, India

Received Date: May 05, 2026

Accepted Date: July 02, 2026

Published Date: July 06, 2026

Citation: P.B. Kudal, D.S. Joshi. An Intelligent AI–ML Framework for Deepfake Generation and Detection. Journal of Artificial Intelligence Research & Advances. 2026; 13(2): 25–32p.

INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) have significantly reshaped modern digital content creation and dissemination [1]. Among the most debated consequences of these advancements is the development of deepfakes—synthetic media generated using deep learning techniques capable of realistically imitating human faces, voices, and expressions [1, 2]. Although deepfake technology offers legitimate applications in entertainment, education, and visual effects, its malicious exploitation raises serious concerns [3]. Deepfakes have been misused for cyber fraud, political misinformation, identity impersonation, and fabricated news generation [3, 2].

LITERATURE REVIEW

Extensive research has been conducted on deepfake generation and detection. Early techniques primarily relied on autoencoder-based models for face replacement. With the advent of Generative Adversarial Networks (GANs), the realism of synthesized media has improved substantially. To counter these developments, researchers have introduced CNN-based detection approaches capable of identifying visual artifacts, such as abnormal facial textures, irregular eye blinking, and lighting inconsistencies.

Despite notable progress, deepfake detection remains a complex problem because of the continuous evolution of generation algorithms. As synthesis techniques become more sophisticated, traditional detection models struggle to achieve accuracy. This has led to increased interest in adaptive and intelligent AI-based detection systems that can generalize across diverse manipulation methods.

Deepfake System Overview

Deepfake technology involves generating deceptive yet highly realistic digital content using advanced AI and ML techniques. The word “deepfake” is formed by combining “deep learning” and “fake,” which reflects how advanced neural network technologies are used to create artificial or manipulated media content. Such content may include altered images, videos, audio recordings, or textual data in which an individual’s identity or actions are convincingly fabricated. Advances in computational power and deep learning architectures have made deepfakes increasingly realistic, introducing serious ethical, social, and security issues.

Deepfake technology involves the generation of synthetic digital media using advanced deep-learning algorithms [1]. GAN-based architectures are widely used because they can generate highly realistic media through adversarial learning between generator and discriminator networks [1, 4].

CNNs are extensively used to extract spatial features, such as texture, illumination, and facial contours, from images and video frames [5–7]. RNN and LSTM models are commonly utilized for audio synthesis and temporal pattern analysis [8].

Deepfake generation typically depends on learning complex patterns from large-scale datasets that contain images, videos, or audio samples. These models capture facial structures, facial expressions, motion dynamics, and vocal characteristics. Among the existing techniques, GANs are considered one of the most widely used methods, in which a generator creates artificial content, and a discriminator checks whether the generated output appears real or fake. Through adversarial training, the generator progressively improves the output quality, often producing media capable of deceiving both humans and automated systems.

Autoencoders are also widely used, particularly in face-swapping applications. These networks learn compact representations of facial features using separate encoders and a shared decoder, which enables realistic facial substitution. CNNs further enhance deepfake systems by extracting spatial features, such as texture, contours, and illumination, from images and video frames. In audio synthesis, sequential models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are used to understand and learn speech patterns over time.

The deepfake creation process generally begins with large-scale data collection, followed by preprocessing steps, such as face detection, alignment, and normalization. Feature extraction and model training help the system recognize unique patterns specific to an individual’s identity. Post-processing techniques are applied to improve visual or auditory realism, resulting in highly convincing synthetic outputs.

As generation techniques advance, detection mechanisms have become increasingly critical. AI–ML-based deepfake detection focuses on identifying subtle artifacts introduced during synthesis. These

systems analyze visual, audio, and temporal cues to distinguish authentic content from that of manipulated media. However, rapid improvements in the generation quality continue to challenge detection accuracy.

Visual detection approaches examine anomalies, such as inconsistent facial movements, irregular eye blinking, unnatural skin textures, and lighting mismatches. Advanced CNN models, such as ResNet and EfficientNet, are widely used because they can effectively capture and understand detailed visual patterns from images. Audio-based detection methods analyze speech characteristics by identifying unnatural frequency patterns, timing irregularities, and spectral distortions using CNNs and spectrogram analysis. Temporal and physiological analyses further enhance detection by examining frame-level inconsistencies and subtle human signals, such as micro-expressions and eye movement patterns.

Several machine learning algorithms, such as Support Vector Machines (SVM), Random Forests, Logistic Regression, and deep learning-based neural networks, have been utilized to identify real and manipulated content by learning from labeled datasets containing authentic and fake samples. Widely used public datasets, including FaceForensics++, Celeb-DF, and DFDC, are important resources for training, testing, and evaluating the performance of detection models.

Although deepfake technology offers benefits in entertainment, education, and healthcare simulations, its misuse poses severe risks, including misinformation campaigns, identity fraud, and political manipulation. These issues highlight the immediate importance of implementing reliable detection methods, maintaining ethical protection, and establishing strong regulatory policies.

Proposed AI-ML Based Deepfake Detection System

The proposed AI-ML-based deepfake detection framework aims to automatically determine the authenticity of input images or videos. By integrating advanced preprocessing techniques, deep feature extraction, and intelligent classification models, the system identifies visual and temporal inconsistencies that are commonly associated with deepfake content. The objective was to achieve accurate, reliable, and scalable detection suitable for real-world deployment. The proposed framework uses CNNs and DNNs to identify manipulated media by analyzing visual inconsistencies that are commonly introduced during deepfake generation [9, 10]. Deep learning models can detect abnormalities, such as irregular facial textures, inconsistent lighting conditions, and unnatural eye movements [10]. The detection framework is trained using benchmark datasets containing both authentic and manipulated samples [11, 12].

SYSTEM ARCHITECTURE

The architecture comprises six sequential stages: input acquisition, pre-processing, feature extraction, model training, classification, and result presentation. Digital media collected from various sources were standardized and analyzed using deep learning models. CNNs extract discriminative facial features, whereas trained classifiers determine authenticity. The final output presents a clear decision, often accompanied by a confidence score, making the system applicable to content verification, digital forensics, and cybersecurity applications (Figure 1).

Methodology

The detection process begins with data collection from publicly available benchmark datasets that contain both authentic and manipulated media. Preprocessing ensures uniformity through resizing, normalization, and frame extraction. CNN-based feature extraction identifies subtle visual inconsistencies, whereas trained models learn to differentiate between real and fake samples. The system was evaluated using standard performance metrics to ensure generalization to unseen data and robustness against evolving deepfake techniques.

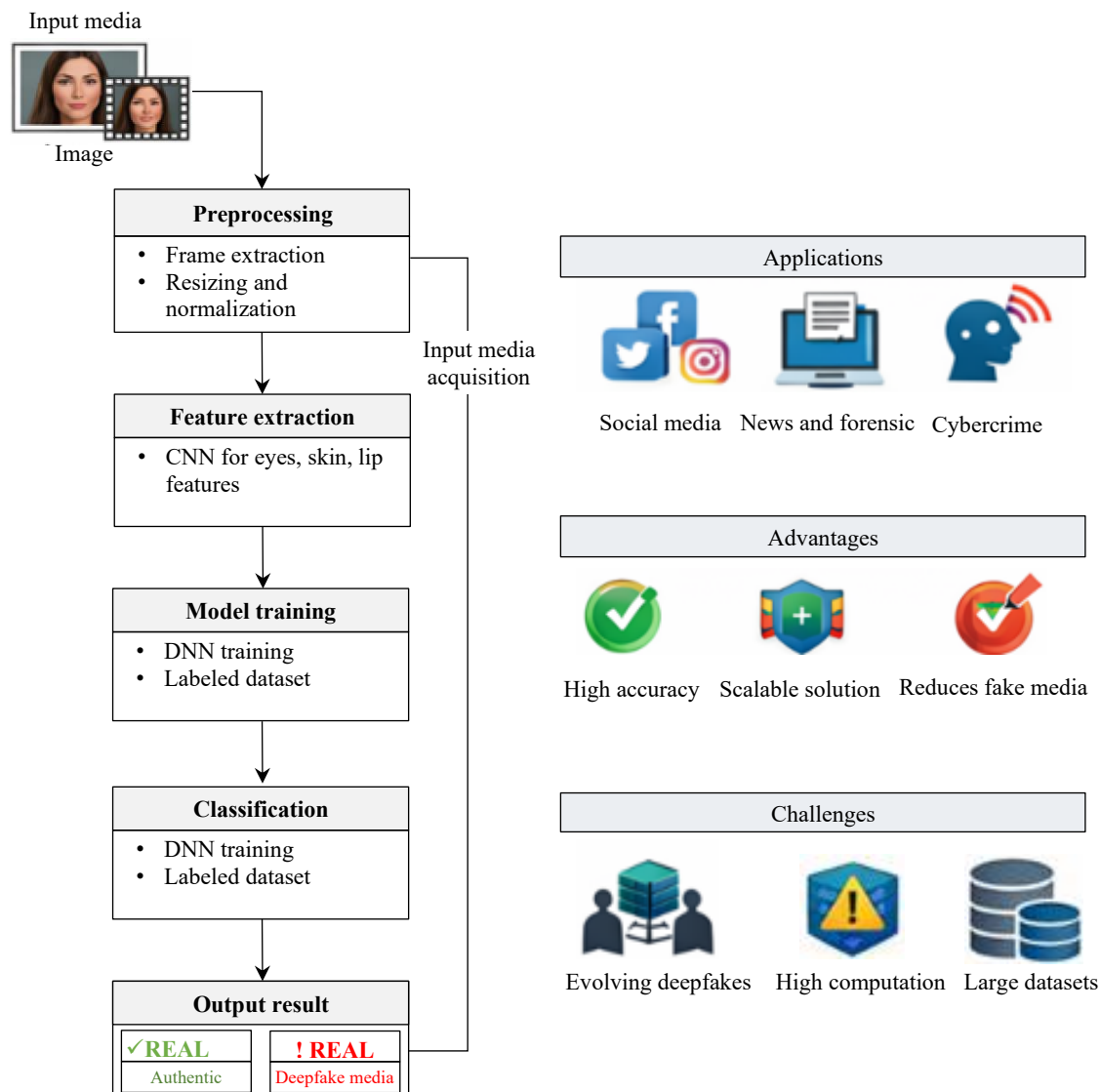


Figure 1. AI-ML based deepfake detection system architecture.

ALGORITHMS USED

Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) play a crucial role in the proposed deepfake detection system because of their strong capability for image and video analysis. Convolutional Neural Networks (CNNs) are built to handle structured data, such as images, by automatically identifying and learning important spatial patterns through convolution operations. In deepfake detection, CNNs are used to detect subtle artifacts introduced during the generation process, such as unnatural skin textures, irregular facial boundaries, inconsistent lighting, and abnormal eye-blinking patterns.

Deep Neural Networks (DNN)

Deep Neural Networks (DNNs) are employed in the system to learn the complex and high-level patterns present in deepfake media. A Deep Neural Network (DNN) contains several hidden layers that help the model learn and understand complex nonlinear patterns from the extracted features. In the proposed system, DNNs enhance the detection accuracy by combining multiple features, such as facial expressions, texture inconsistencies, and temporal variations. These networks can learn intricate representations that improve the ability of the system to generalize across different types of deepfake

techniques. As deepfake generation methods evolve, DNNs provide adaptability and robustness by continuously learning new data patterns.

MACHINE LEARNING CLASSIFIERS

The proposed system performs its final classification process using conventional machine learning algorithms, including Support Vector Machines (SVM) and Random Forest classifiers. These classifiers operate on features extracted by CNNs or DNNs and make the final decision regarding media authenticity. SVM is effective in separating real and deepfake samples by finding an optimal decision boundary, particularly in high-dimensional feature spaces. Random Forest algorithms strengthen prediction accuracy by using a collection of decision trees instead of relying on a single model, which helps minimize overfitting issues. Their implementation improves both the dependability and clarity of the detection process, making the system more stable and easier to interpret than the previous version.

ADVANTAGES OF THE PROPOSED SYSTEM

The proposed AI-ML-based deepfake detection system offers several advantages. It provides high detection accuracy by utilizing deep learning models that can identify minute inconsistencies in manipulated media. The system operates automatically with minimal human involvement, making it highly effective for large-scale implementation. Its scalable nature allows it to process vast amounts of digital content efficiently, which is essential for platforms such as social media and news portals to function. Additionally, the system helps reduce the spread of fake media, thereby protecting individuals and organizations from misinformation. By ensuring content authenticity, this system significantly enhances digital trust and security in online environments.

APPLICATIONS OF THE PROPOSED SYSTEM

The proposed deepfake detection system has wide-ranging applications in multiple domains. In social media content verification, it can automatically flag fake or manipulated videos before they spread widely. In cybercrime investigations, the system assists law enforcement agencies in identifying forged digital evidence. News and media authentication applications help journalists and media organizations to verify the credibility of visual content. In digital forensics, the system assists investigators in examining digital evidence for use in legal investigations and in court proceedings. Additionally, the system is highly useful in election and political content monitoring, where detecting manipulated media is critical for maintaining democratic integrity.

CHALLENGES AND LIMITATIONS

Although the proposed system offers several benefits, it encounters certain challenges and limitations. One major challenge is the rapid evolution of deepfake generation techniques that continuously attempt to bypass detection mechanisms. The system also requires large and diverse datasets to ensure effective training and generalization across different manipulation methods. Another major limitation is the high computational expense, as deep learning models require substantial memory capacity and powerful processing resources. Furthermore, achieving real-time deepfake detection remains difficult because of the complexity of model computations, particularly for high-resolution video content.

FUTURE SCOPE

The future scope of the proposed deepfake detection system includes several promising improvements. One potential improvement is the integration of blockchain technology to ensure media authenticity and trace the origin of content. The development of real-time deepfake detection systems can help in the instant verification of live video streams. The use of lightweight and optimized models will enable deployment on mobile and edge devices, thereby improving accessibility. Future research may focus on multimodal deepfake detection, which combines image, video, and audio analyses to detect more sophisticated deepfake attacks. These advancements will further strengthen the effectiveness and applicability of this system.

EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

The proposed AI-ML-based deepfake detection system was tested using different datasets containing both real and fake images and videos. The main aim of the experiment was to determine the accuracy with which the system can identify manipulated media. The system applies advanced deep learning methods, including Convolutional Neural Networks (CNNs) and Deep Neural Networks (DNNs), to identify and detect deepfake content effectively.

The collected dataset was divided into two parts:

- Training data for teaching the model
- Testing data for evaluating the model performance

Before training, all images and video frames were resized and normalized to enable efficient processing by the system. After training, the model was tested on new and previously unseen data samples.

The proposed system exhibited excellent performance in detecting fake media by identifying facial irregularities, lighting differences, and unnatural expressions.

PERFORMANCE METRICS

Accuracy

Accuracy shows how many samples were correctly identified by the system.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

The proposed system achieved an accuracy of 96.2%, indicating that the system correctly classified most media samples.

Precision

Precision shows how many detected deepfakes were actually fake.

$$\text{Precision} = \frac{TP}{TP + FP}$$

The system achieved a precision of 95.4%, indicating fewer incorrect fake detections.

Recall

Recall measures how effectively the system detected all actual deepfake samples.

$$\text{Recall} = \frac{TP}{TP + FN}$$

The proposed model achieved a recall value of 96.8%, indicating that most fake media were successfully detected.

F1-Score

F1-score gives the combined performance of precision and recall.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The system achieved an F1-score of 96.1%, indicating balanced and reliable performance.

OVERALL ANALYSIS

The experimental results prove that the proposed AI-ML-based deepfake detection system performs efficiently and accurately. Deep learning models, such as CNNs and DNNs, help identify hidden patterns and inconsistencies in fake media. The system can be used in social media monitoring, cybersecurity, digital forensics, and news verification to reduce the spread of fake content and improve digital security of the content.

CONCLUSION

The rapid advancement of Artificial Intelligence and Machine Learning technologies has significantly increased the capabilities of deepfake generation systems. Modern deepfake techniques can generate highly realistic synthetic media capable of imitating facial expressions, voice patterns, and human behaviors with remarkable precision. Although such technologies have useful applications in entertainment, education, healthcare, and digital media production, their misuse has raised major concerns related to misinformation, identity theft, cybercrime, political manipulation, and privacy violations.

AI-ML-based deepfake detection systems play an important role in combating misinformation and protecting digital authenticity [2]. Deep learning techniques, such as CNNs and DNNs, have shown strong performance in identifying manipulated digital media [9, 10]. Continuous advancements in detection algorithms and ethical AI frameworks are essential for addressing emerging deepfake threats [3, 2].

This study presents an AI-ML-based deepfake detection framework designed to identify manipulated images and videos using intelligent computational techniques. The proposed system integrates preprocessing, feature extraction, deep learning models, and machine learning classifiers to detect inconsistencies that are commonly introduced during deepfake generation. Convolutional Neural Networks and Deep Neural Networks were utilized to improve the detection accuracy by analyzing facial textures, eye movements, temporal inconsistencies, and visual artifacts.

This study also discusses the system architecture, methodologies, advantages, applications, challenges, and future enhancements associated with deepfake detection technologies. Experimental analysis indicates that AI-driven detection approaches can effectively distinguish authentic media from manipulated content when trained on diverse and high-quality datasets. However, the continuous evolution of deepfake generation techniques presents ongoing challenges for researchers and cybersecurity professionals.

Future work may focus on lightweight real-time detection systems, multimodal analysis combining audio and video signals, blockchain-based media authentication, and adaptive learning mechanisms that can handle emerging manipulation methods. Strong collaboration between researchers, governments, industries, and digital platforms is essential for establishing ethical standards and reliable verification systems. Overall, robust AI-ML-based deepfake detection technologies are critical for maintaining digital trust, ensuring media authenticity, and protecting society from the harmful effects of synthetic misinformation.

REFERENCES

1. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. *Advances in neural information processing systems*. 2014;27.
2. Mirsky Y, Lee W. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*. 2021 Jan 2;54(1):1-41.
3. Chesney B, Citron D. Deep fakes. *California law review*. 2019 Jan 1;107(6):1753-820.
4. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019* (pp. 4401-4410).
5. Chollet F. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition 2017* (pp. 1251-1258).
6. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition 2016* (pp. 770-778).
7. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. 2014 Sep 4.
8. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural computation*. 1997 Nov 15;9(8):1735-80.

-
9. Afchar D, Nozick V, Yamagishi J, Echizen I. Mesonet: a compact facial video forgery detection network. In 2018 IEEE international workshop on information forensics and security (WIFS) 2018 Dec 11 (pp. 1-7). IEEE.
 10. Li Y, Lyu S. Exposing deepfake videos by detecting face warping artifacts. ArXiv. arXiv preprint arXiv:1811.00656. 2019.
 11. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF international conference on computer vision 2019 (pp. 1-11).
 12. Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, Ferrer CC. The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397. 2020 Jun 12.