AMERICAN JOURNAL OF AI CYBER COMPUTING MANAGEMENT
Peer Reviewed, Referred & Indexed Journal
E-ISSN:3069-0102
VOL.6, NO. 2(2026)

CIFAKE: A Hybrid Approach for Detecting and Explaining AI-Generated Synthetic Images

Mrs. K.Anusha[1]|Mohammad Bushra Fathima[2]|Sukamanchi Chandu Sai[3]|Panduri Bhumesh[4]|Nelluri Vamsi Gopala Krishna[5].

[1,2,3,4,5]Department of CSE-AI, Chalapathi Institute of Engineering and Technology,

LAM, Guntur, Andhra Pradesh, India

Abstract: With the rapid growth of artificial intelligence in India, synthetic media generation has increased across social media, journalism, advertising, and cybercrime. India recorded thousands of deepfake-related complaints in recent years, especially involving misinformation and identity misuse. As AI-generated images become photorealistic, traditional visual inspection fails. This raises serious concerns for digital trust, national security, and data authenticity, motivating research into automated AI-generated image detection using computer vision techniques.The objective is to automatically distinguish real images from AI-generated images using CNN-based classification and explainable AI, ensuring image authenticity, reliability, and trustworthiness in digital ecosystems. In the manual system, image authenticity is verified through human visual inspection, expert judgment, metadata analysis, and basic digital forensic techniques. Analysts examine image quality, lighting inconsistencies, compression artifacts, and contextual information. Sometimes watermark verification or source validation is used to determine whether an image is real or artificially generated. Manual analysis is time-consuming, subjective, and error-prone. Human observers cannot reliably detect high-quality synthetic images, especially diffusion-generated content. The process does not scale for large datasets, lacks consistency, and fails when metadata is removed or manipulated, making it ineffective against modern AI-generated images. The motivation is to overcome the limitations of manual inspection by introducing an automated, scalable, and objective system. Machine learning can detect subtle visual artifacts invisible to humans, improve accuracy, reduce analysis time, and provide explainable insights, addressing reliability and scalability challenges present in traditional methods. The proposed system uses a Convolutional Neural Network (CNN) to classify images as Real or Fake by learning discriminative visual patterns from both CIFAR-10 real images and synthetically generated images (CIFAKE dataset).

Keywords: AI-generated image detection, Deepfake detection, Convolutional Neural Network (CNN), Computer vision, Explainable AI (XAI), Image classification, Synthetic media, Digital forensics,

1. INTRODUCTION

Image classification is a core area of computer vision that focuses on automatically identifying and categorizing images into predefined classes based on their visual features. With the rapid advancement of digital technologies, vast amounts of image data are generated daily from sources such as smartphones, social media platforms, medical imaging devices, satellite systems, and surveillance cameras. Manually analysing and organizing this data is inefficient and often impractical, which has led to the widespread adoption of automated image classification techniques powered by artificial intelligence (AI) and machine learning.

In recent years, deep learning methods particularly convolutional neural networks (CNNs) have revolutionized image classification by enabling systems to learn complex patterns directly from raw image data. These models have achieved remarkable success in tasks such as object recognition, facial recognition, medical diagnosis, traffic monitoring, and autonomous driving. Their ability to process large-scale image datasets and produce highly accurate predictions has made them an essential component of modern intelligent systems.However, despite their high performance, many deep learning–based image classification models function as “black boxes.” While they can produce accurate predictions, they often fail to provide understandable explanations for how or why a particular decision was made. This lack of interpretability poses significant challenges, especially in critical application areas such as healthcare, law enforcement, and finance, where incorrect or biased decisions can have serious consequences. Users and stakeholders increasingly demand transparency, accountability, and trust in AI-driven systems.Explainable identification, commonly associated with Explainable Artificial Intelligence (XAI), aims to address this issue by enhancing the interpretability and transparency of image classification models. By providing clear and human-interpretable explanations, these methods make AI systems more reliable and easier to validate.The integration of image classification with explainable identification enhances trust, improves decision-making, and helps detect biases or errors in models. This project aims to build an accurate image classification system using ML and deep learning while providing transparent explanations of predictions. By highlighting key image features influencing decisions, it ensures reliability, interpretability, and suitability for real-world applications.

2. LITERATURE SURVEY

The ability to distinguish between real images and those generated by machine learning models is critically important for several reasons. Identifying real data helps confirm the authenticity and originality of an image. For example, a fine-tuned Stable Diffusion Model (SDM) could be used to generate a synthetic photograph of an individual committing a crime or, conversely, provide false evidence as an alibi for a person who was actually elsewhere. Misinformation and fake news represent a major modern challenge, and machine-generated images can be exploited to manipulate public opinion [3], [4]. When synthetic imagery is used within fake news, it can enhance false credibility and lead to serious consequences [5].

This approach is rapidly evolving but remains relatively young, and as a result, the existing literature on the topic is limited. These models introduce a new paradigm in generative modelling, and only a small number of applications have been explored so far. Notable examples include DALL-E by OpenAI [9], Imagen by Google, and the open-source equivalent Stable Diffusion Model developed by Stability AI [2]. These models have significantly advanced image quality, both in terms of realism and artistic capability. Consequently, they have sparked extensive debate regarding the professional, social, ethical, and legal implications of such technologies [1]. Much of the research in this field is cutting-edge and is often published as preprints or recent academic theses.

3 SYSTEM ARCHITECTURE

System Architecture defines the overall structure of the proposed system by illustrating the interaction between hardware, software components, and data flow. In this system, the architecture consists of a user interface, a Flask-based web application, a deep learning CNN model, and a data storage layer. The user uploads an image through the web interface, which is forwarded to the Flask server for processing. The server performs image pre-processing and sends the data to the trained CNN model for classification. The CNN model predicts the class label and generates feature maps using the Grad-CAM module for explain ability, The system delivers prediction results along with visual explanations through a web interface, enabling clear and interpretable decision-making. Its architecture supports real-time, accurate classification with low latency while remaining scalable and modular for future enhancements. This approach enhances reliability, transparency, and overall user experience by combining deep learning with visual interpretability.

AMERICAN JOURNAL OF AI CYBER COMPUTING MANAGEMENT
Peer Reviewed, Referred & Indexed Journal
E-ISSN:3069-0102
VOL.6, NO. 2(2026)

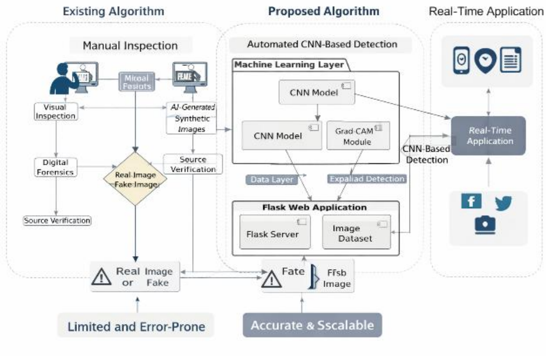

Fig1: System Architecture

3.1 Methodology

The methodology of the CIFAKE system follows a structured deep learning pipeline for detecting AI-generated synthetic images. Initially, A balanced dataset of real (CIFAR-10) and AI-generated (CIFAKE) images is pre-processed by resizing to 32×32 pixels, normalizing values, and applying augmentation techniques. It is then split into training and testing sets using an 80:20 ratio for effective model evaluation. A Convolutional Neural Network (CNN2D) with multiple convolutional, pooling, and fully connected layers is designed to automatically learn hierarchical features such as edges, textures, and complex patterns. Dropout layers are incorporated to prevent overfitting and improve model robustness. The model is trained using a binary cross-entropy loss function and optimized using the Adam optimizer through forward and backward propagation. After training, the model classifies input images as real or synthetic based on learned features. For interpretability, Grad-CAM is applied to generate highlight important regions influencing predictions.

3.2 Dataset Acquisition:A balanced dataset of real images (CIFAR-10) and AI-generated images is collected and labelled into two classes, ensuring diversity and quality for effective training and generalization.

3.3 Data Pre-processing: Images are resized to 32×32, normalized, and augmented, then shuffled and split into training and testing sets (80:20) to improve model performance and reduce bias.

3.4 Model Development:A CNN2D model with dropout layers is designed to extract features and classify images, using ReLU, max-pooling, and softmax for accurate and robust prediction.

3.5 Model Performance: The CNN with dropout outperforms the standard model by reducing overfitting and achieving higher accuracy, precision, recall, and overall reliability.

3.6 Real-Time Detection: The system enables instant image classification via a web interface, providing predictions and confidence scores with optional visual explanations for better transparency.

4. DESIGN AND CONSTRUCTION

The design and construction of the CIFAKE system focus on building an

Received: 04-03-2026 | Accepted: 18-03-2026 | Published: 03-04-2026 | www.ajaccm.com

efficient, scalable, and user-friendly framework for detecting AI-generated synthetic images. The system is designed with a modular architecture consisting of data collection, preprocessing, model development, and deployment components. 2D Convolutional Neural Network (CNN2D) with dropout layers is constructed to learn hierarchical image features and perform accurate classification between real and synthetic images. The preprocessing module ensures uniform input through resizing, normalization, and augmentation, while the training module optimizes the model using appropriate loss functions and optimizers. For implementation, the system is integrated into a Flask-based web application that allows users to upload images and receive real-time predictions, explainable AI techniques such as Grad-CAM are incorporated to visualize important regions influencing the model's decisions, enhancing transparency and trust., the system is carefully designed to ensure high performance, reliability, and practical usability in real-world applications.

5. RESULTS AND DISCUSSION

The proposed image classification system demonstrated effective performance in distinguishing AI-generated synthetic images from real images, achieving high classification accuracy on the test dataset. The use of deep learning models enabled the system to learn discriminative features such as texture inconsistencies and unnatural patterns commonly present in synthetic images. Explainable identification techniques, including visualization methods like heatmaps, provided clear insights into the model's decision-making process by highlighting image regions that influenced classification outcomes. These explanations helped validate that the model focused on meaningful visual cues rather than irrelevant background information. the results confirm that integrating explainable AI with image classification enhances transparency, reliability, and trust, making the system suitable for real-world applications involving the detection of AI-generated synthetic images**.**

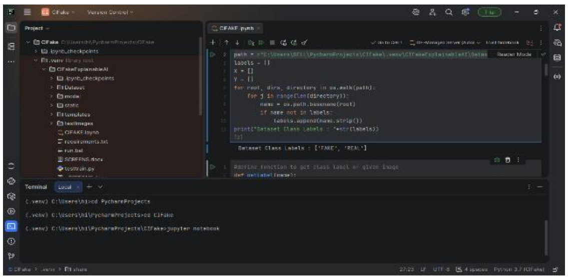

Fig 2: Python Environment & Dataset Loading

The figure shows the project setup in a Python environment Where the dataset is loaded and processed to extract class labels (FAKE, REAL), which are used for training and prediction.

Received: 04-03-2026 | Accepted: 18-03-2026 | Published: 03-04-2026 | www.ajaccm.com

Fig 3: Home Page for Classification

The figure shows the home interface of the CIFAKE system designed for identifying AI-generated images. It highlights the classification process where a CNN model distinguishes between real and fake images with high accuracy. The interface provides easy navigation and a clear visualization of the prediction concept.

Fig 4: Admin Login Screen & Authentication Process

The figure shows the Admin Login interface of the system designed for secure access control. It highlights the authentication process where user credentials are verified to allow authorized entry. The interface ensures data security by preventing unauthorized access and provides a simple and user-friendly login experience.

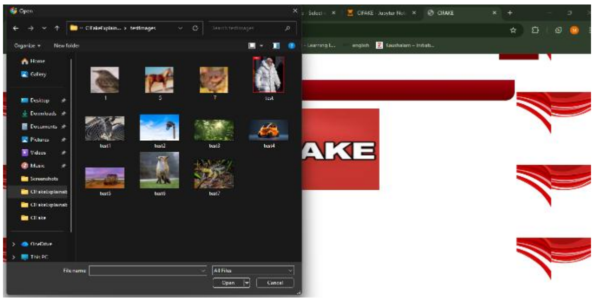

Fig 6: Image Upload Selection & Input Process

The figure shows the file selection interface where users choose an image from the system for classification. It highlights the input process of selecting test images that will be analysed.

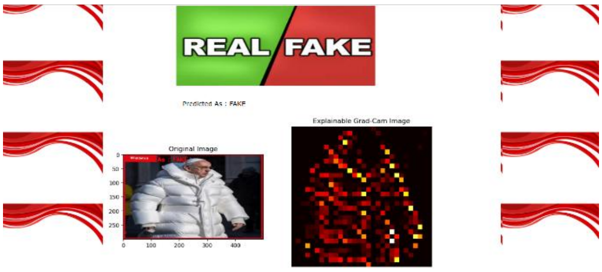

Fig 7: Prediction Result & Explainable Grad-CAM Visualization

The figure shows the output of the classification system where the uploaded image is predicted as fake. It highlights the use of Grad-CAM visualization to identify important regions influencing the model's decision. This improves transparency by explaining how the model arrives at its prediction.

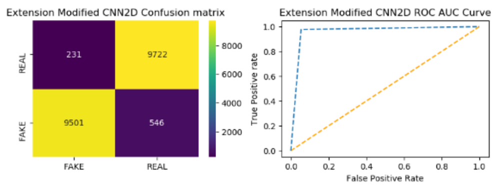

Fig 8: Modified CNN2D Confusion Matrix & ROC Curve

Received: 04-03-2026 | Accepted: 18-03-2026 | Published: 03-04-2026 | www.ajaccm.com

The figure presents the performance of the modified CNN2D model, showing improved classification results between real and fake images. The confusion matrix highlights prediction accuracy, while the ROC curve indicates strong model performance with a high true positive rate.

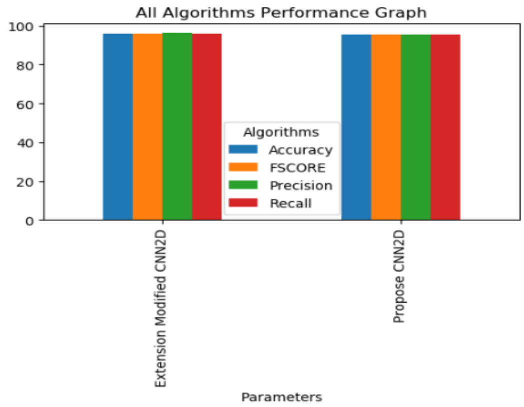

Fig 9: Overall Performance Comparison of Algorithms

The graph compares the performance of the proposed CNN2D and the modified CNN2D models using accuracy, F1-score, precision, and recall. Both models show consistently high performance across all metrics, indicating their effectiveness in image classification

6. CONCLUSION

The researchdemonstrates a robust and effective approach to distinguishing between real and AI-generated images. By leveraging the hierarchical feature learning capability of convolutional neural networks, the proposed model successfully extracts both low-level and high-level features from image data. The integration of dropout layers plays a critical role in preventing overfitting, ensuring that the network generalizes well to unseen images. Performance evaluation using metrics such as accuracy, precision, recall, F1-score,and confusion matrices confirms that the proposed model outperforms the baseline CNN2D architecture., Grad-CAM visualizations provide interpretability by highlighting important regions influencing the predictions, making the model more transparent and trustworthy. Overall, the research validates that combining deep feature extraction with regularization techniques can significantly enhance classification performance for AI-generated synthetic images, offering a reliable solution for real-world applications.

FUTURE SCOPE: Future work can explore the use of Vision Transformer (ViT) models to capture global image dependencies and improve detection of subtle patterns in AI-generated images. Integrating ViT with hybrid architectures and explainable AI techniques can further enhance accuracy, interpretability, and robustness for real-world applications.

REFERANCES

[1] K. Roose, “An AI-generated picture won an art prize. Artists aren’t happy,” New York Times, vol. 2, p. 2022, Sep. 2022.

Received: 04-03-2026 | Accepted: 18-03-2026 | Published: 03-04-2026 | www.ajaccm.com

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 10684–10695.

[3] G. Pennycook and D. G. Rand, “The psychology of fake news,” Trends Cogn. Sci., vol. 25, no. 5, pp. 388–402, May 2021.

[4] B. Singh and D. K. Sharma, “Predicting image credibility in fake news over social media using multi-modal approach,” Neural Comput. Appl., vol. 34, no. 24, pp. 21503–21517, Dec. 2022.

[5] N. Bonettini, P. Bestagini, S. Milani, and S. Tubaro, “On the use of Benford’s law to detect GAN-generated images,” in Proc. 25th Int. Conf. Pattern Recognit. (ICPR), Jan. 2021, pp. 5495–5502.

[6] D. Deb, J. Zhang, and A. K. Jain, “AdvFaces: Adversarial face synthesis,” in Proc. IEEE Int. Joint Conf. Biometrics (IJCB), Sep. 2020, pp. 1–10.

[7] M. Khosravy, K. Nakamura, Y. Hirose, N. Nitta, and N. Babaguchi, “Model inversion attack: Analysis under gray-box scenario on deep learning based face recognition system,” KSII Trans. Internet Inf. Syst., vol. 15, no. 3, pp. 1100–1118, Mar. 2021.

[8] J. J. Bird, A. Naser, and A. Lotfi, “Writer-independent signature verification: Evaluation of robotic and generative adversarial attacks,” Inf. Sci., vol. 633, pp. 170–181, Jul. 2023.

[9] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, “Zero-shot text-to-image generation,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 8821–8831.

Received: 04-03-2026 | Accepted: 18-03-2026 | Published: 03-04-2026 | www.ajaccm.com