Explainable Detection of AI-Generated Synthetic Images Through the CIFAKE Hybrid Architecture

Dr G. SESHA PHANEENDRA BABU[1], S. SIVA SAI VIVEK[2]

[1]Associate Professor, Dept. of C.S.E, Anantha Lakshmi Institute of Technology and sciences Anantapur – 515721

[2]PG Scholar, Dept. of C.S.E, Anantha Lakshmi Institute of Technology and sciences Anantapur- 515721

Abstract: The rapid advancement of artificial intelligence has led to a significant increase in the generation of synthetic images across social media, journalism, advertising, and digital communication platforms. While AI-generated images offer numerous benefits, they also pose serious challenges related to misinformation, identity misuse, digital trust, and content authenticity. Traditional image verification methods based on human visual inspection, metadata analysis, and forensic examination are often subjective, time-consuming, and ineffective against highly realistic synthetic content. To address these limitations, this paper presents an automated AI-generated image detection framework based on Convolutional Neural Networks (CNNs) and Explainable Artificial Intelligence (XAI). The proposed system is trained using real images from the CIFAR-10 dataset and synthetic images from the CIFAKE dataset to learn discriminative visual features that distinguish authentic images from AI-generated counterparts. The CNN model performs binary classification, identifying images as either real or fake with high accuracy and efficiency. Furthermore, explainable AI techniques provide visual interpretations of model decisions, enhancing transparency and user trust. Experimental results demonstrate that the proposed approach effectively detects synthetic images while reducing analysis time and improving scalability compared to manual methods. The framework contributes to strengthening digital content verification, supporting reliable media authentication, and promoting trustworthiness in modern digital ecosystems.

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

1.Introduction

Image classification is a fundamental computer vision task that automatically categorizes images into predefined classes based on their visual characteristics. With the rapid growth of digital data from social media, smartphones, medical imaging, and surveillance systems, automated image classification has become essential for efficiently analysing large volumes of image data. Recent advances in deep learning, particularly Convolutional Neural Networks (CNNs), have significantly improved image classification performance. CNNs can automatically learn complex visual patterns from images and have been successfully applied in areas such as object recognition, facial

recognition, medical diagnosis, and autonomous systems, achieving high accuracy and reliability.

Despite their effectiveness, deep learning models often lack transparency and operate as "black boxes." To address this challenge, Explainable Artificial Intelligence (XAI) techniques provide interpretable explanations for model predictions, improving trust and accountability. This project combines image classification with explainable identification to deliver accurate predictions while highlighting the key features influencing decisions, making the system more reliable and suitable 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.

3. Proposed System

The proposed system focuses on detecting AI-generated synthetic images using a 2D Convolutional Neural Network (CNN2D) enhanced with dropout regularization for improved generalization. The model automatically learns hierarchical features from images, capturing both low-level textures and high-level semantic patterns to distinguish real and synthetic content. Dropout layers are used to reduce overfitting and improve robustness, ensuring reliable performance on unseen data for accurate image

4 System Architecture

The CIFAKE system architecture utilizes a hybrid framework to detect and explain AI-generated synthetic images. Input images are first pre-processed through resizing and normalization, followed by feature extraction using deep learning and forensic analysis techniques.

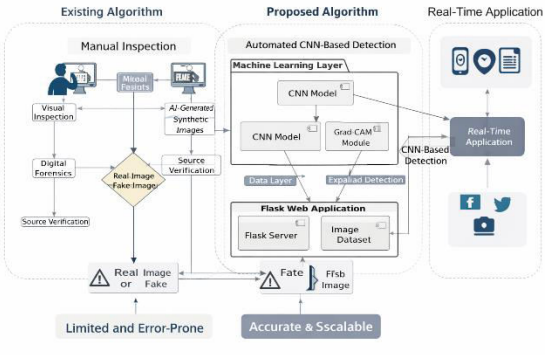

Fig 1: System Architecture

The extracted features are fused and passed to a classification module that identifies images as real or synthetic. An explain ability module then highlights the key image regions and features influencing the prediction, providing transparent and

interpretable results for reliable synthetic image detection.

5.Methodology

The CIFAKE system follows a deep learning pipeline where real (CIFAR-10) and synthetic (CIFAKE) images are pre-processed, normalized, and split into training and testing sets. A CNN model is trained using convolutional, pooling, and fully connected layers with Adam optimizer and binary cross-entropy loss to classify images as real or fake. Grad-CAM is applied to provide interpretability by highlighting the regions influencing the model's predictions.

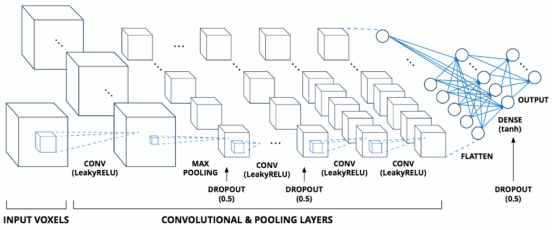

Fig 2: CNN2D Architecture

The CNN2D model learns hierarchical image features, ranging from simple patterns to complex visual structures, through multiple convolutional layers. During training, dropout layers improve generalization by reducing overfitting, while optimization algorithms update model parameters to accurately classify real and AI-generated synthetic images. The resulting model remains robust across variations in image quality, style, and resolution.

6. Design and Construction

The CIFAKE system is designed as a modular and scalable framework for detecting AI-generated synthetic images. It employs a 2D Convolutional Neural Network (CNN2D) with dropout layers to extract discriminative image features and accurately classify images as real or synthetic. Preprocessing techniques such as resizing, normalization, and augmentation enhance data quality and improve model performance and generalization.

i) 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.

ii) 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.

iii) 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.

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

v) 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.

The system is implemented using a Flask-based web application that enables real-time image upload and prediction, enhanced with Grad-CAM for visual explanation of model decisions. Performance is evaluated using accuracy, precision, recall, F1-score, confusion matrix, and ROC-AUC to ensure reliable and effective classification of real and AI-generated images.

7. Results and Discussion

The CIFAKE system achieved high accuracy in distinguishing AI-generated synthetic images from real images by

effectively learning discriminative visual features. The integration of explainable AI techniques provided interpretable insights into the classification process, enhancing the transparency, reliability, and practical applicability of the proposed framework.

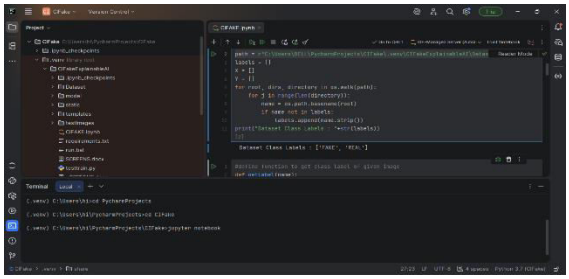

Fig 3: Python Environment & Dataset Loading

The fig 3 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.

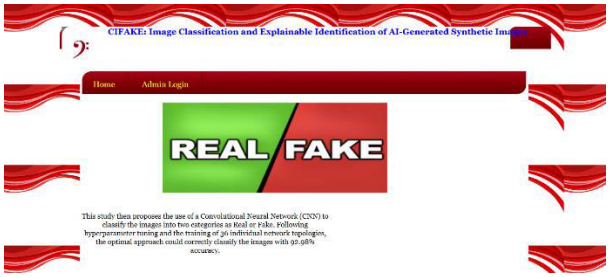

Fig 4: Home Page for Classification

The fig 4 illustrates the CIFAKE home interface, where a CNN-based model classifies images as real or AI-generated with a simple and user-friendly visualization of the prediction process.

Fig 5: Admin Login Screen & Authentication Process

The fig 5 depicts the Admin Login interface, where secure authentication verifies user credentials to ensure authorized access and protect system data.

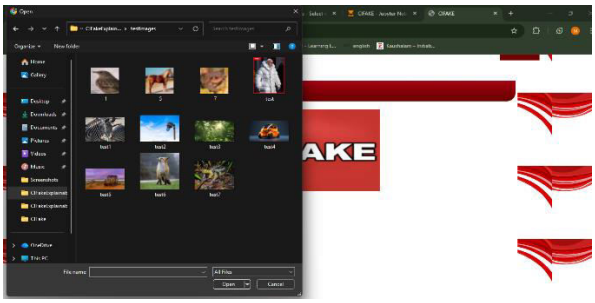

Fig 6: Image Upload Selection & Input Process

The fig 6 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 fig 7 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.

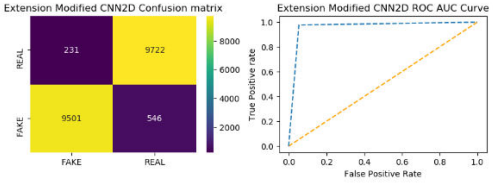

Fig 8: Modified CNN2D Confusion Matrix & ROC Curve

The fig 8 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.

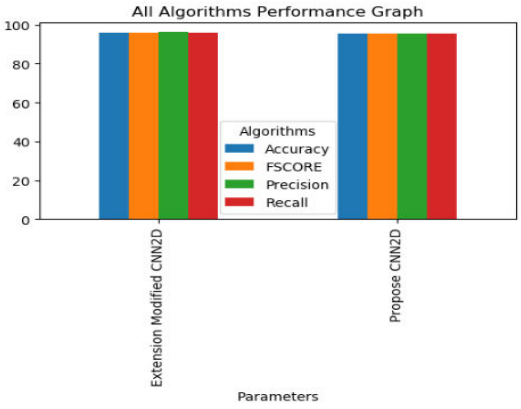

Fig 9: Overall Performance Comparison of Algorithms

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

8. Conclusion and Future Scope

The research demonstrates 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.

References

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

[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.