iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VI **Month of publication:** June 2025

DOI: https://doi.org/10.22214/ijraset.2025.72291

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VI June 2025- Available at www.ijraset.com

FakeVision AI: Detecting and Explaining AI-Generated Images with Deep Learning

Janakiraman. S[1], Poovarasan K[2]
[1]*Assistant Professor, II MCA,* [2]*Department of Master of Computer Applications, Er.Perumal Manimekalai College of Engineering, Hosur,*

Abstract: ***The advancement of artificial intelligence (AI) has led to the development of powerful generative models such as StyleGAN, DALL·E, and Stable Diffusion, which are capable of creating highly realistic synthetic images. The CIFAKE dataset serves as a benchmark for training deep learning models to distinguish between real and AI-generated images. In this study, we propose an AI-based framework for detecting synthetic imagery using deep learning and explainable AI (XAI) methods. Our approach incorporates Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) to classify images as either authentic or artificially generated. Additionally, explainability tools such as Grad-CAM and SHAP are employed to highlight the most influential features contributing to the model's predictions. This system promotes greater transparency in AI decision processes and strengthens the trustworthiness of digital content authentication.***
Keywords: Generative Models, Synthetic Images, StyleGAN.

I. INTRODUCTION

In recent times, the field of synthetic image creation has seen remarkable progress, largely due to breakthroughs in deep learning—especially with the emergence of Generative Adversarial Networks (GANs). These models are capable of producing highly convincing visual content that can deceive both human observers and automated systems. Although such technology offers valuable uses in areas like entertainment, education, and enhancing training datasets, it also raises significant concerns related to ethics and security.

II. PROPOSED WORK

CIFAKE is an AI-driven architecture developed to identify computer-generated imagery by merging image classification techniques with explainable AI (XAI) methodologies.
The system utilizes a Convolutional Neural Network (CNN) to sort images into two groups: genuine or artificially generated.
The training process utilizes the CIFAKE dataset, which includes 60,000 computer-generated visuals and 60,000 real-world images, all originating from the CIFAR-10 image collection.
The model achieves a high accuracy rate of 92.98% in recognizing the difference between real and AI-generated images.

1) Interpretability Module: This component enhances transparency by explaining the rationale behind each prediction made by the classifier.
2) Image Storage System: A well-organized database is used to archive details and attributes associated with each image.

III. MODULES

A. Dataset Collection

Gathers an extensive collection of synthetic (FAKE) and genuine (REAL) images from the CIFAKE dataset, maintaining an equal distribution between the two categories.

B. Image Preprocessing

Every image is scaled to 224 by 224 pixels and standardized to ready it for input into the CNN. To enhance performance, images are processed in batches.

C. Model Training

A convolutional neural network model is trained using the prepared dataset. The model conducts binary classification to distinguish between authentic and synthetic images, utilizing binary cross-entropy as the loss function for optimization.

D. Model Evaluation

Assessment is carried out using performance measures like accuracy, precision, recall, and the F1-score. Visualization tools such as graphs and confusion matrices support the evaluation process.

E. Prediction Interface

Enables prediction on unseen test images and provides visual explanations of the results to enhance user comprehension.

IV. RESULT

FakeVision AI effectively identifies images created by artificial intelligence through the use of sophisticated deep learning methods. The model demonstrates strong accuracy in differentiating between real and computer-generated images and integrates explainable AI techniques to offer transparent visual explanations of its predictions. This strategy not only boosts detection accuracy but also promotes clarity, helping users gain confidence in and comprehend the system's classification decisions.

V. CONCLUSION

The swift progress in generative AI technologies has made it increasingly challenging to differentiate between authentic and synthetic images. Image classification has become essential for verifying the credibility and trustworthiness of digital media.

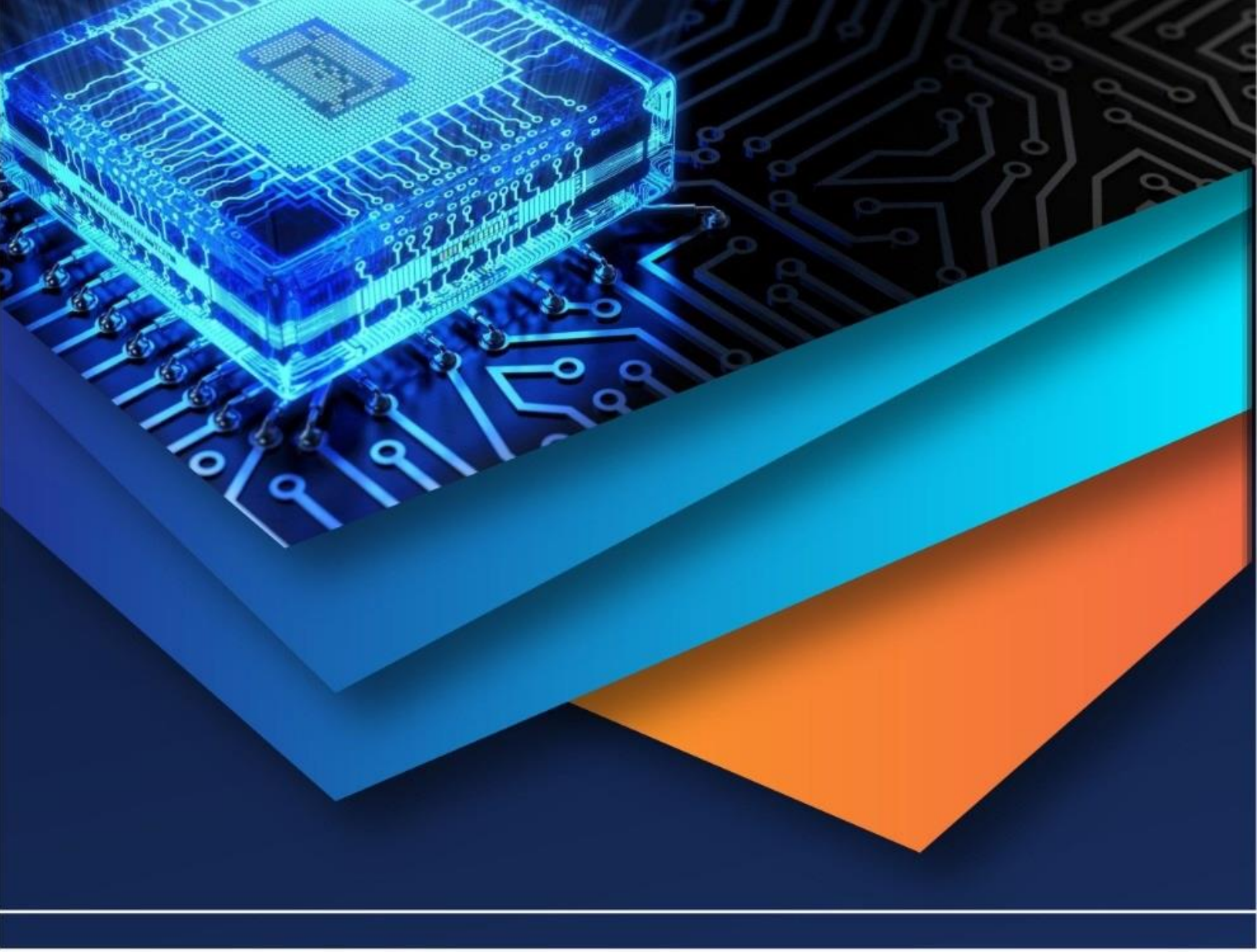
By utilizing advanced deep learning architectures such as Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), along with explainable AI methods, detection accuracy can be significantly enhanced while maintaining transparency in AI-driven decisions. This strategy not only tackles present difficulties but also establishes a strong foundation for defending against emerging AI-generated visual content. In summary, CIFAKE serves as an effective tool in combating misinformation by offering reliable detection and interpretability, holding great promise for fostering truth and integrity in the digital space.

VI. ACKNOWLEDGMENT

The authors confirm that this research was carried out independently, with no external funding or assistance warranting acknowledgment.

REFERENCES

[1] Wang, Z., and colleagues (2021). CIFAKE: A Dataset Designed for Synthetic Image Classification. arXiv preprint arXiv:2103.07948.
[2] Selvaraju, R. R., and team (2017). Grad-CAM: Gradient-Based Localization for Visual Explanations from Deep Neural Networks. Presented at ICCV.
[3] Dosovitskiy, A., et al. (2020). Transformers for Scaled Image Recognition: An Image is Worth 16x16 Words. arXiv preprint.
[4] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining Classifier Predictions. Presented at KDD.
[5] Lundberg, S. M., & Lee, S.-I. (2017). A Unified Framework for Model Interpretation. Presented at NeurIPS.
[6] Chollet, F. (2017). Xception: Deep Neural Networks Using Depthwise Separable Convolutions. Presented at CVPR.
[7] Tan, M., & Le, Q. (2019). EfficientNet: Revisiting Model Scaling for CNNs. Presented at ICML.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY