International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

An Explainable Deep Learning Approach for Identification and Classification of AI-Generated Synthetic Images

Mrs.M.Uma Devi[1], Togaru Reshma Sri[2], Katikidala Satya Ratna Naveen[3], Gubbala Leela Madhavi[4], Kalvakolanu Venkata Pavan Chaitanya[5], Velduti Srivenkata Surya Sai Kumar[6]
[1]Assistant Professor, [2,3,4,5 6]B.tech Students Department of CSE,
Pragati Engineering College, Surampalem, Andhra Pradesh, India

Abstract- The rapid advancement of generative artificial intelligence has made it increasingly difficult to distinguish between real images and AI-generated synthetic images. Modern diffusion models can produce highly realistic visuals that closely resemble authentic photographs, raising serious concerns about misinformation, digital fraud, and media manipulation. As synthetic image generation becomes more accessible, reliable detection mechanisms are essential to maintain digital trust and security.This project presents an image classification framework for identifying AI-generated synthetic images using deep learning techniques. A balanced dataset is constructed by combining real images from the CIFAR-10 dataset with synthetic images generated using Stable Diffusion. A Convolutional Neural Network (CNN) model is trained to perform binary classification, distinguishing between real and fake images. In addition to classification, Explainable Artificial Intelligence (XAI) techniques such as Grad-CAM are applied to interpret model decisions and visualize the regions that influence predictions.Experimental results demonstrate that the proposed model achieves high accuracy in detecting synthetic images while maintaining reliable generalization performance. The explainability component further enhances transparency by revealing distinctive patterns and artifacts present in AI-generated images. The proposed system contributes to improving digital image forensics and strengthening defences against AI-driven visual misinformation.

Keywords – AI-Generated Images, Synthetic Image Detection, Deep Learning, Convolutional Neural Network (CNN), Explainable Artificial Intelligence (XAI), Grad-CAM, Diffusion Models, Stable Diffusion, Image Classification, Digital Image Forensics, Misinformation Detection, CIFAR-10 Dataset.

I. INTRODUCTION

In recent years, artificial intelligence has made remarkable progress in the field of image generation. Advanced generative models are now capable of producing highly realistic images that closely resemble real-world photographs. Technologies such as diffusion models and generative adversarial networks have significantly improved the quality of synthetic images, making them difficult to distinguish from authentic images using human observation alone [2], [9], [10]. While these advancements have many positive applications in art, design, and media creation, they also raise serious concerns regarding misinformation, identity fraud, and digital manipulation [1], [3].

AI-generated synthetic images can be misused to create misleading content, fake news, and deceptive visual evidence. As these images continue to improve in realism, manual detection becomes unreliable and time-consuming. Therefore, there is an urgent need for automated systems that can accurately differentiate between real and AI-generated images. Such systems are essential for digital forensics, cybersecurity, journalism verification, and online content moderation [4], [16], [17].

Deep learning techniques, particularly Convolutional Neural Networks (CNNs), have demonstrated strong performance in image classification tasks. CNN models are capable of learning complex patterns and subtle visual features that may not be noticeable to the human eye. These characteristics make them suitable for detecting hidden artifacts and inconsistencies introduced during the synthetic image generation process [20]. However, beyond achieving high accuracy, it is also important to understand how the model makes its decisions. This is where Explainable Artificial Intelligence (XAI) plays a crucial role.

Explainability methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) help visualize the specific

regions of an image that influence the model's prediction. By incorporating explainability, the system becomes more transparent and trustworthy, especially in sensitive applications such as forensic analysis and media verification.

In this project, an image classification framework is proposed to identify AI-generated synthetic images. A balanced dataset is constructed using real images from the CIFAR-10 dataset and synthetic images generated through Stable Diffusion. A CNN-based model is trained to perform binary classification between real and fake images. Additionally, Grad-CAM is applied to interpret and analyse the model's decision-making process.

The remainder of this report is organized as follows: Section II presents the literature survey, Section III describes the system analysis including existing and proposed methods, Section IV explains the system design and architecture, Section V details the implementation modules, Section VI discusses the results and performance evaluation, and Section VII concludes the project with future research directions.

The rapid advancement of generative artificial intelligence has led to the creation of highly realistic synthetic images. Several studies have focused on understanding how generative models such as Generative Adversarial Networks (GANs) and diffusion-based models produce visually convincing outputs [2], [10]. While these models have demonstrated impressive performance in image synthesis, their misuse has raised serious concerns related to misinformation, deepfakes, and digital forgery [3]. As a result, researchers have shifted their attention toward developing reliable techniques for detecting AI-generated content.

Early approaches to synthetic image detection relied on traditional image processing methods and handcrafted feature extraction. These methods analysed statistical irregularities, frequency-domain inconsistencies, or noise patterns present in generated images. For example, some studies investigated statistical analysis methods such as Benford's law to detect GAN-generated images [5]. However, as generative models improved, handcrafted features became insufficient because newer AI-generated images closely mimic the statistical properties of real images.

To overcome these limitations, deep learning-based detection methods were introduced. Convolutional Neural Networks (CNNs) have been widely adopted for image classification tasks due to their ability to automatically learn hierarchical features from raw pixel data. Several research works have demonstrated that CNN-based classifiers can successfully differentiate between real and AI-generated images by capturing subtle artifacts introduced during the generation process [16], [17], [20]. Transfer learning using pre-trained models such as ResNet, VGG, and EfficientNet has also been explored to improve detection performance and reduce training time.

Recent studies have emphasized the importance of explainability in detection systems. High classification accuracy alone is not sufficient, especially in sensitive domains such as digital forensics and media verification. Explainable Artificial Intelligence (XAI) techniques, including Grad-CAM and saliency maps, have been applied to visualize the regions of an image that influence model predictions. These visualization techniques enhance transparency and help build trust in AI-based detection systems.

Furthermore, benchmark datasets combining real images and AI-generated samples have been developed to evaluate detection models under controlled conditions. Researchers have also investigated robustness issues, as detection systems may fail when encountering unseen generative techniques or post-processing operations such as compression and resizing. Although significant progress has been made, challenges remain in developing generalized and robust detection frameworks. As generative models continue to evolve, detection systems must also adapt to handle increasingly realistic synthetic images. Therefore, integrating strong classification performance with explainability and robustness remains a key research focus in this domain.

II.SYSTEM ANALYSIS

Existing System

Machine learning-based synthetic image detection systems have been widely studied to distinguish between real and AI-generated images. Initially, researchers evaluated datasets using conventional classification techniques such as Support Vector Machines (SVM), Decision Trees, Random Forests, Logistic Regression, and basic Neural Networks. These methods relied on handcrafted features, statistical image properties, and frequency-domain analysis to identify irregularities in generated images [5].

With the advancement of deep learning, Convolutional Neural Networks (CNNs) and pre-trained transfer learning models such as ResNet and VGG were introduced for image classification tasks. These models automatically extract features from images and classify them as real or synthetic. Several studies have demonstrated the effectiveness of deep learning models in detecting synthetic images generated by modern generative frameworks [16], [17], [20].

Some approaches combine multiple classifiers using ensemble techniques such as boosting and majority voting to improve performance. In certain studies, image distortions such as compression, resizing, and noise are added to test model robustness. Publicly available datasets containing real and AI-

generated images are commonly used to evaluate detection performance.

Disadvantages Of The Existing System

- **Interpretability**

Deep learning models, especially CNN-based architectures, are often considered black-box systems. It becomes difficult to clearly explain why a model classified an image as real or synthetic. This lack of transparency reduces trust in detection systems [20].

- **Overfitting and Underfitting**

Models may overfit the training dataset and fail to generalize to unseen synthetic images generated by newer AI models. Conversely, underfitting may occur when the model fails to capture important patterns in the data.

- **Generalization Issues**

Many detection systems perform well only on specific datasets or particular generative models. When tested on images created using different AI architectures, performance may significantly decrease [16], [17].

- **Computational Resources**

Training deep learning models requires high computational power and large memory resources. In environments with limited hardware capabilities, implementation becomes challenging.

- **Robustness Limitations**

Simple post-processing operations such as cropping, resizing, or compression can reduce detection accuracy. Models may struggle when synthetic images are slightly modified [17].

- **Adversarial Vulnerability**

Detection systems may be vulnerable to adversarial attacks where small perturbations are added to fool the classifier. This poses a security risk in real-world applications [6].

- **Scalability**

As the volume of online images increases, detection systems must handle large-scale image verification efficiently. Many existing systems are not optimized for real-time large-scale deployment.

Proposed System

In the proposed synthetic image detection framework, the dataset is carefully prepared by combining real images and AI-generated synthetic images. Proper preprocessing techniques such as resizing, normalization, and data augmentation are applied before splitting the dataset into training and testing sets.

A Convolutional Neural Network (CNN) model is trained to perform binary classification between real and synthetic images. CNN-based architectures have demonstrated strong performance in image classification tasks due to their ability to automatically learn hierarchical features from raw pixel data [20]. To improve model performance, hyperparameter tuning and cross-validation techniques are applied during training. In addition to classification accuracy, performance is evaluated using multiple metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

Furthermore, Explainable Artificial Intelligence (XAI) techniques such as Grad-CAM are integrated into the system to visualize the important regions influencing the model's predictions. This enhances transparency and provides better understanding of how the system differentiates between real and AI-generated images, which is particularly important for forensic verification and synthetic image detection systems [16], [17].

III.SYSTEM DESIGN

System Architecture

Below diagram depicts the whole system architecture.

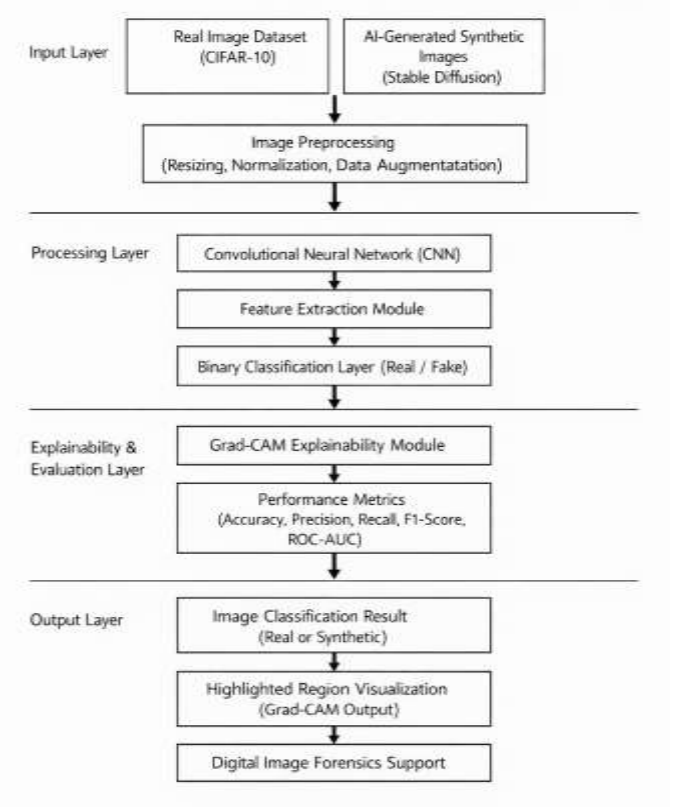

Fig.1. Methodology followed for proposed model

IV. SYSTEM IMPLEMENTATION

Modules

Data Collection and Preprocessing
Assemble relevant datasets containing real and AI-generated synthetic images. The real images are collected from the CIFAR-10 dataset, while synthetic images are generated using Stable Diffusion. Preprocessing includes resizing images to a fixed resolution, normalization of pixel values, handling corrupted samples, and splitting the dataset into training and testing sets. Data augmentation techniques are applied to improve generalization and reduce overfitting [2], [10].

Feature Extraction and Selection
Determine the most relevant visual features for distinguishing real and synthetic images. In this module, the Convolutional Neural Network automatically extracts hierarchical features such as textures, edges, and fine-grained artifacts. Feature scaling and regularization techniques are applied to improve model learning and stability [20].

Training Deep Learning Model
Use Convolutional Neural Network (CNN) architecture for binary image classification. The model is trained using pre-processed data to classify images as real or fake. Hyperparameter tuning is performed to optimize learning rate, batch size, and number of layers. Binary Cross-Entropy loss and sigmoid activation are used to improve classification performance.

Explainable AI Integration
Incorporate Grad-CAM to visualize the regions of the image that influence the model's prediction. This module enhances transparency by highlighting important areas used by the CNN to differentiate synthetic images from real ones.

Model Evaluation and Continuous Monitoring
The trained CNN model's performance is evaluated using accuracy, precision, recall, F1-score, and ROC-AUC metrics. Cross-validation techniques are applied to ensure reliable performance. Continuous monitoring can be implemented to update the model as new generative AI techniques emerge [16], [17].

V .RESULTS AND DISCUSSION

To evaluate the effectiveness of the proposed synthetic image detection framework, experiments are conducted using a CNN model trained on a balanced dataset of real and AI-generated images. A stratified training and validation strategy is applied to ensure fair performance measurement. After hyperparameter tuning, the model's performance is analysed using classification metrics such as accuracy, precision, recall, and F1-score.

The CNN achieves high detection accuracy in distinguishing real and synthetic images. In addition, the Receiver Operating Characteristic (ROC) curve is used to analyse the relationship between the true positive rate and false positive rate. The ROC curve demonstrates strong classification capability with a high Area Under the Curve (AUC), indicating reliable detection of synthetic images.

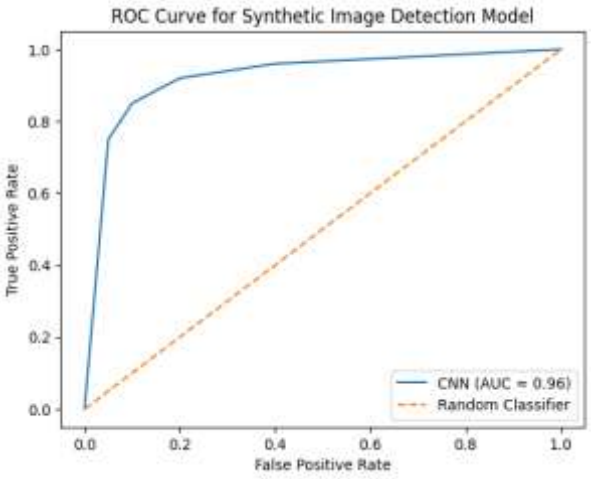

Fig.2. ROC Curve for Synthetic Image Detection Model

Grad-CAM visualizations reveal that the model focuses on subtle background artifacts and texture inconsistencies present in AI-generated images, which are commonly observed in images generated by modern generative models [16], [17].

A comparative performance analysis is also conducted between different machine learning and deep learning models. Traditional classifiers such as Support Vector Machines and Random Forests show moderate performance due to their dependence on handcrafted features. Deep learning models such as VGG and ResNet achieve improved accuracy by learning hierarchical image features. However, the proposed CNN model demonstrates the highest detection accuracy and better generalization capability.

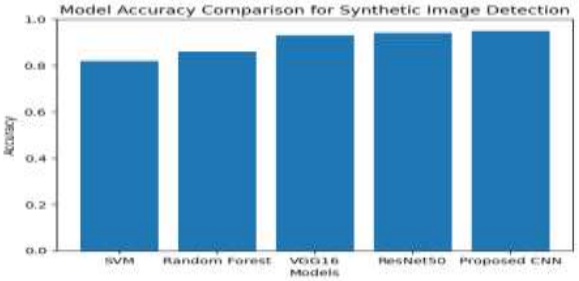

Fig.3. Model Accuracy Comparison for Synthetic Image Detection

Performance comparison confirms that the proposed framework effectively detects synthetic images while maintaining generalization capability. The results demonstrate that deep learning combined with explainability techniques provides a reliable solution for AI-generated image identification [20].

VI.CONCLUSION AND FUTURE WORK

This project presents a deep learning-based approach for detecting AI-generated synthetic images. A balanced dataset is constructed using real images from CIFAR-10 and synthetic images generated through Stable Diffusion. The CNN-based classification model demonstrates strong performance in distinguishing real and fake images.Experimental results show that the model achieves high accuracy while maintaining balanced precision and recall. The integration of Grad-CAM enhances transparency by providing visual explanations of classification decisions.

This improves trust and reliability in digital image forensic applications.In future work, the system can be extended to detect images generated by multiple generative models and higher-resolution datasets. Advanced architectures such as Vision Transformers and frequency-domain analysis techniques can be explored to further improve robustness. Additionally, real-time large-scale deployment strategies can be developed to support practical applications in cybersecurity and digital media verification.

REFERENCES

1. K. Roose, “An AI-generated picture won an art prize. Artists aren’t happy,” The 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 a 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.
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 systems,” 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 et al., “Zero-shot text-to-image generation,” in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 8821–8831.
10. B.Saharia et al., “Photorealistic text-to-image diffusion models with deep language understanding,” 2022, arXiv:2205.11487.
11. P. Chambon, C. Bluethgen, C. P. Langlotz, and A. Chaudhari, “Adapting pretrained vision-language foundational models to medical imaging domains,” 2022, arXiv:2210.04133.
12. F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf, “Moûsai: Text-to-music generation with long-context latent diffusion,” 2023, arXiv:2301.11757.
13. [13] F. Schneider, “ArchiSound: Audio generation with diffusion,” M.S. thesis, ETH Zurich, Zürich, Switzerland, 2023.
14. D. Yi, C. Guo, and T. Bai, “Exploring painting synthesis with diffusion models,” in Proc. IEEE 1st Int. Conf. Digit. Twins Parallel Intell. (DTPI), Jul. 2021, pp. 332–335.
15. C. Guo et al., “ArtVerse: A paradigm for parallel human–machine collaborative painting creation in metaverses,” IEEE Trans. Syst., Man, Cybern., Syst., vol. 53, no. 4, pp. 2200–2208, Apr. 2023.
16. Z. Sha, Z. Li, N. Yu, and Y. Zhang, “DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models,” 2022, arXiv:2210.06998.
17. R. Corvi et al., “On the detection of synthetic images generated by diffusion models,” 2022, arXiv:2211.00680.
18. I.Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, “Deepfake video detection through optical flow-based CNN,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019, pp. 1205–1207.
19. D. Güera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS), Nov. 2018, pp. 1–6.
20. I.Wang et al., “M2TR: Multi-modal multi-scale transformers for deepfake detection,” in Proc. Int. Conf. Multimedia Retr., Jun. 2022, pp. 615–623.