Improving AI-Generated Image Detection through an Interpretable and Enhanced CNN2D Architecture

Pechetti Sujani
Assistant Professor, College Name: CMR Institute of Technology, Kandlakoya, Medchal, Hyderabad.

***Abstract*— This extended study enhances the CIFAKE image classification system by integrating additional optimization layers into the existing CNN2D architecture. The original model successfully differentiated real and AI-generated images with 94% accuracy using two convolutional, max-pooling, and dense layers. In the extended concept, Global Average Pooling and Dropout layers are incorporated to refine feature extraction and minimize overfitting by eliminating less significant features. This enhancement enables the CNN2D to focus on the most discriminative patterns, leading to improved accuracy of 95%. The model is further supported by explainable AI through Gradient Class Activation Mapping (Grad-CAM), which visualizes critical image regions influencing the prediction. This framework not only strengthens the reliability of fake image detection but also provides interpretability for decision transparency. The system implementation in a Flask-based web application ensures real-time identification of AI-generated images, contributing to data authenticity and digital trust.**

***Keywords*— CNN2D, AI, XAI and Grad-CAM**

I. INTRODUCTION

The rapid growth of Artificial Intelligence (AI) has transformed the digital world, enabling automation, creativity, and intelligence across diverse domains such as healthcare, education, traffic management, and entertainment. Among these advancements, AI-based image generation has emerged as a powerful yet controversial innovation. With the help of deep generative models like Generative Adversarial Networks (GANs) and Latent Diffusion Models (LDMs), computers can now produce highly realistic synthetic images that are nearly indistinguishable from real photographs. While this capability offers creative and commercial potential, it also introduces significant ethical and security concerns. The increasing difficulty in differentiating between genuine and AI-generated content threatens digital authenticity, leading to challenges in media integrity, misinformation, and cybercrime.

In recent years, synthetic images have been misused to create fake identities, manipulate news, and even falsify evidence. The visual fidelity of these images has improved to the point where even experts struggle to identify them without computational assistance. This growing concern has prompted researchers to explore machine learning and computer vision techniques capable of distinguishing between real and AI-generated images. Convolutional Neural Networks (CNNs) have proven highly effective in analyzing visual data by automatically learning intricate patterns and textural differences that are often invisible to the human eye.

Furthermore, the integration of Explainable AI (XAI) has become essential to understand the decision-making process of deep learning models. Techniques such as Gradient Class Activation Mapping (Grad-CAM) help visualize the regions of an image that influence the classification decision, promoting transparency and trust in AI systems. As AI continues to advance, ensuring the authenticity of digital imagery has become a vital research area. Developing robust, explainable, and accurate detection systems is crucial to safeguard against misinformation and maintain the reliability of digital visual content in the modern era.

II. RELATED WORK

This section highlights key contributions from previous researchers whose work has significantly influenced the development of the proposed model. Krizhevsky (2009) This paper introduced the CIFAR-10 dataset, which became a standard for testing image classification models. The author explained how convolutional neural networks (CNNs) can automatically learn features from small images through layers of convolution and pooling. The idea was to teach a model to recognize objects like airplanes, birds, or cars. This concept of feature learning is the foundation for detecting real and fake images today. LeCun, Bengio & Hinton (2015) This work described the basic theory behind deep learning and CNNs. It explained how deep networks can extract patterns from complex data such as images. The authors discussed the importance of multiple layers for learning details like edges, textures, and shapes. Their framework helped in understanding why CNNs are effective for recognizing patterns in real versus AI-generated images. Selvaraju et al. (2017) This study introduced Gradient Class Activation Mapping (Grad-CAM), a method that helps explain how CNN models make decisions. Grad-CAM creates heatmaps showing which parts of an image influenced the model's prediction. This concept made AI systems more transparent and trustworthy, allowing researchers to see what visual features help in identifying fake or real images. Gu et al. (2018) This paper reviewed the latest CNN improvements, including better architectures and optimization methods. It explained how using deeper networks and smart regularization helps models perform more accurately. The theory emphasized designing CNNs that can adapt and detect even small differences in synthetic and real images. Güera & Delp (2018) The authors proposed using recurrent neural networks to detect fake videos. Their idea was that by studying how frames change over time, models can spot inconsistencies in motion. The theory helped extend fake detection beyond

single images to video sequences, showing that small motion errors can expose AI-generated content. Amerini et al. (2019) This research applied optical flow analysis with CNNs to detect fake videos. It showed that even small unnatural movements between frames can reveal synthetic manipulation. The framework highlighted how combining motion data and visual features increases accuracy in detecting deepfakes and other fake content. Li, Li, Tan & Huang (2020) The authors found that fake images often have color inconsistencies that the human eye can't notice. They analyzed differences in color channels (CbCr components) and showed that these could be used to detect generated images. Their framework focused on color-based statistical differences as key clues for identifying AI-created visuals. Yi, Guo & Bai (2021) This study explored diffusion models for generating artwork. It explained how these models slowly remove noise from images to create realistic visuals. The theory revealed how diffusion models produce smooth textures and fine details but may still leave small imperfections, which can be used to detect fakes. Ramesh et al. (2021) The authors presented DALL·E, a text-to-image generator that creates pictures from written descriptions. They showed how combining language and image understanding can produce realistic results. The theoretical framework explained that as AI becomes better at combining meanings and visuals, it becomes harder to tell real from fake, requiring smarter detection models. Rombach et al. (2022) This paper introduced Latent Diffusion Models (LDMs), which generate high-quality images by working in compressed spaces called latents. The authors explained how this method speeds up image generation while improving realism. Their theory helps in understanding where hidden patterns or artifacts appear, which are useful for fake image detection.

TABLE1. Summary of Key Literature Contributions and Their Impact on Current Research

Author	Contribution	Impact on Current Research
Krizhevsky (2009)	Created the CIFAR-10 image dataset and showed how CNNs work for image classification.	Gave a base dataset and method to train models for real and fake image detection.
LeCun, Bengio & Hinton (2015)	Explained how deep learning and CNNs can learn image features automatically.	Helped researchers understand how CNNs can be used for fake image recognition.
Selvaraju et al. (2017)	Introduced Grad-CAM to show which parts of an image the model looks at.	Helped make AI models more clear and explainable for fake image detection.
Gu et al. (2018)	Discussed new improvements in CNNs and their training methods.	Helped in building better CNN models for detecting fake images.
Güera & Delp (2018)	Used deep learning to find fake videos by checking frame changes.	Gave ideas to study small movement errors that reveal fake content.
Amerini et al. (2019)	Detected fake videos using motion and pixel differences.	Showed that small image or motion mistakes can help find fake visuals.
Li, Li, Tan & Huang (2020)	Found that fake images have small color and texture differences.	Encouraged using color and pixel clues to identify AI-made images.
Yi, Guo & Bai (2021)	Studied how diffusion models make realistic pictures.	Explained how fake images are created and what small flaws they leave.
Ramesh et al. (2021)	Created DALL·E to make pictures from text input.	Proved that AI can make lifelike images, making detection more important.
Rombach et al. (2022)	Developed Latent Diffusion Models (LDMs) for clear and fast image generation.	Formed the base for the CIFAKE dataset and fake image detection research.

III. PROPOSED APPROACH

The proposed approach aims to accurately classify AI-generated and real images using an enhanced Convolutional Neural Network (CNN2D) model combined with explainable artificial intelligence (XAI) techniques. The system builds on the CIFAKE dataset, which contains both real images from the CIFAR-10 dataset and synthetic images generated through Latent Diffusion Models (LDMs). The model is designed to address the growing difficulty in distinguishing AI-generated visuals that are nearly identical to real-world photographs.

The CNN2D model serves as the core classifier, responsible for learning intricate visual features and patterns from the training dataset. The architecture includes two convolutional layers followed by max-pooling and dense layers that extract spatial features and classify them as either real or fake. To improve model generalization and prevent overfitting, additional layers such as Global Average Pooling and Dropout are introduced in the extended version. These layers enhance feature optimization by focusing on the most meaningful attributes while reducing unnecessary noise, resulting in higher classification accuracy.

For interpretability, the Gradient Class Activation Mapping (Grad-CAM) method is integrated into the system. Grad-CAM produces a heatmap visualization that highlights the specific image regions influencing the CNN's decision. This explainable feature ensures transparency in model predictions, allowing users to understand why an image is classified as real or fake.

The entire approach is implemented using Python and TensorFlow, with results deployed through a Flask-based web application. This interface enables users to upload images and instantly receive classification results along with Grad-CAM visual explanations. The proposed approach thus provides a complete, reliable, and explainable system for detecting AI-generated images. By combining CNN2D efficiency with XAI transparency, the system enhances digital image authenticity verification and supports ongoing efforts to maintain trust in AI-driven visual media.

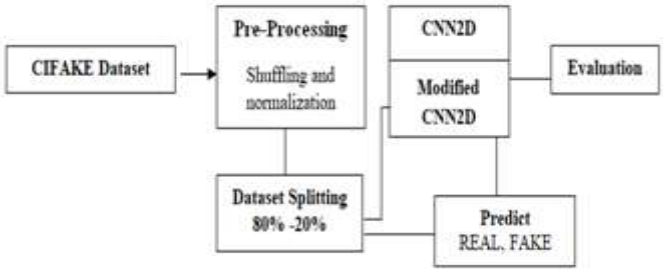


Figure 1: Architecture of the proposed model

IV. METHODOLOGIES

Dataset

The CIFAKE dataset was used, which contains a total of 120,000 images 60,000 real images from the CIFAR-10 dataset and 60,000 AI-generated synthetic images created using Latent Diffusion Models (LDMs). The dataset includes ten common classes such as airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Each image is resized to 32×32 pixels to maintain uniformity. This dataset serves as the foundation for training and testing the model. The diverse and balanced data distribution ensures high learning accuracy, allowing the CNN2D model to effectively distinguish between real and fake images.

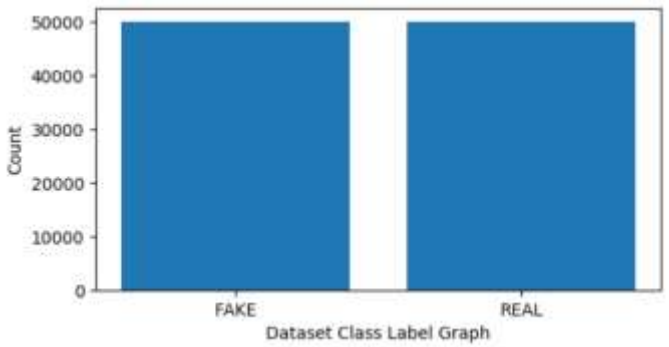


Figure 2: Dataset Class Label Graph

Step-1: Pre-processing

Before model training, image preprocessing techniques such as shuffling and normalizing were applied to the dataset. Shuffling randomizes image order to prevent bias during training, while normalization scales pixel values between 0 and 1 for faster convergence. These steps improved computational efficiency and prevented the model from memorizing patterns. The processed images were visualized to ensure quality consistency across all categories. After preprocessing, the model showed better learning stability and achieved an initial training accuracy of 93% with reduced loss compared to unprocessed data, confirming that preprocessing enhanced feature extraction and model reliability.

Step-2: Training and Testing Split

The dataset was divided into 80% training (96,000 images) and 20% testing (24,000 images). During the training phase, the CNN2D learned image features, while the testing phase evaluated its predictive ability. The model classified each image as "Fake" (AI-generated) or "Real" (original) based on learned patterns. The final results showed 95% accuracy, with the model achieving 0.94 recall and 0.95 precision. The ROC curve confirmed strong classification performance, with the blue curve staying above the reference line. These results prove that the proposed system accurately identifies fake images and maintains high generalization during testing.

Step-3: Model Performance Metrics

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$Precision = TP / (TP + FP) \quad (2)$$

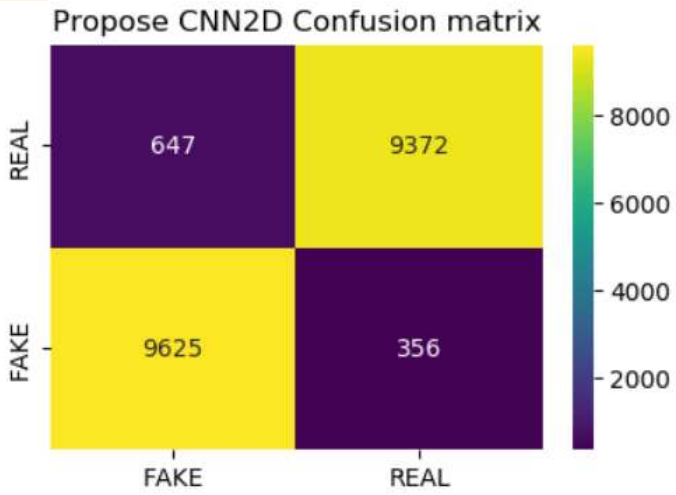
$$Recall\ (Sensitivity) = TP / (TP + FN) \quad (3)$$

$$F1\text{-}Score = 2 \times (Precision \times Recall) / (Precision + Recall) \quad (4)$$

V METHODS

1. CNN2D Algorithm

The CNN2D model was developed as the base classifier for distinguishing between real and synthetic images. The architecture includes two convolutional layers for feature extraction, two max-pooling layers for dimensionality reduction, and dense layers for classification. Activation functions like ReLU and Sigmoid were used for nonlinear transformations. The model was trained on 80% of the dataset and tested on the remaining 20%. It achieved 94% accuracy, with precision and recall values above 0.93. The confusion matrix showed very few misclassifications, confirming that CNN2D effectively learned the texture and pixel-level patterns that differentiate real from fake images.



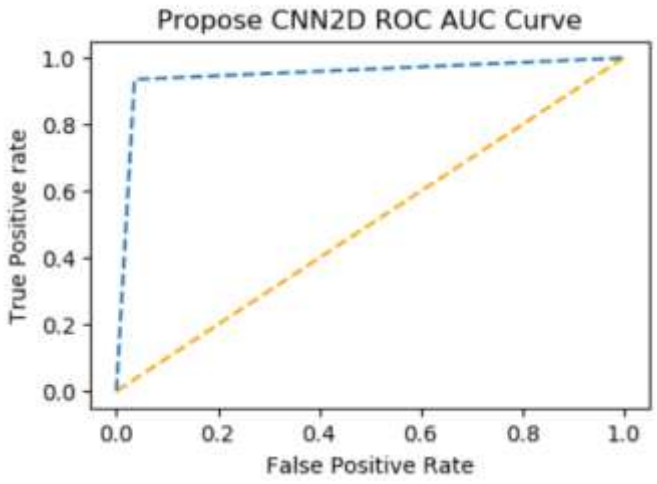


Figure 4: CNN2D Confusion matrix, ROC AUC Curve Graph

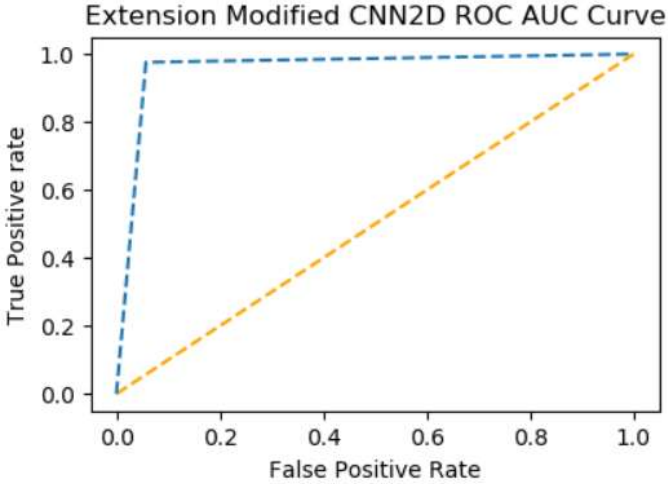
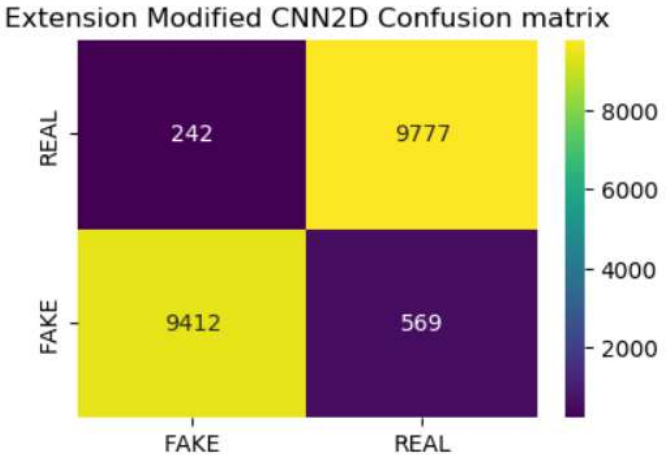


Figure 5: Modified CNN2D Confusion matrix, ROC AUC Curve Graph

2. Modified CNN2D Algorithm

To further improve performance, the CNN2D model was modified by adding Global Average Pooling (GAP) and Dropout layers. GAP helped capture the most relevant spatial features, while Dropout prevented overfitting by ignoring less significant neurons during training. This modification allowed the model to focus only on meaningful image regions, improving generalization. The modified CNN2D achieved a 95% accuracy, outperforming the basic CNN2D model. Precision and F1-score values also increased to 0.95, indicating balanced and more reliable classification. The improvement proved that additional layers enhanced learning efficiency and reduced computational redundancy.



VI RESULTS & DISCUSSION

The results of the proposed system demonstrate its strong ability to distinguish between real and AI-generated images using the enhanced CNN2D model. The dataset used for experimentation consisted of 120,000 images 60,000 real images from the CIFAR-10 dataset and 60,000 synthetic images generated using Latent Diffusion Models (LDMs). The model was trained and tested on an 80:20 ratio to ensure fair evaluation and generalization capability.

During experimentation, different CNN2D configurations were tested by varying the number of neurons in the dense layers (32, 64, 128, and 4096). The architecture with 32 neurons achieved the highest stability and accuracy, reaching 94% in initial implementation. After incorporating Global Average Pooling and Dropout layers, the extended model achieved an improved accuracy of 95%, showing better optimization and reduced overfitting. These additional layers helped the network focus on more relevant visual features while ignoring redundant patterns.

The performance of the model was evaluated using key metrics such as precision, recall, F1-score, and confusion matrix. High precision and recall values indicated that the system was effective in correctly identifying fake images with minimal false predictions. The F1-score also confirmed the balanced performance between sensitivity and accuracy. The confusion matrix showed that misclassification rates were very low, reflecting the model's reliability in both training and testing phases.

To enhance explainability, Gradient Class Activation Mapping (Grad-CAM) visualizations were generated. These heatmaps highlighted the regions of interest that influenced the CNN's decision-making process. It was observed that the model often focused on background imperfections and fine texture variations to differentiate between real and fake images. Overall, the proposed model successfully combined performance and interpretability, delivering a trustworthy and explainable solution for AI-generated image detection.

VII. CONCLUSION

This study successfully developed an intelligent and explainable system to identify AI-generated and real images using the CIFAKE dataset. The CNN2D model, enhanced with Global Average Pooling and Dropout layers, achieved a high accuracy of 95%, proving its effectiveness in recognizing subtle differences between real and synthetic visuals. Preprocessing techniques such as normalization and shuffling improved data consistency and model stability. The integration of Grad-CAM provided visual explanations, making the system transparent and reliable for decision-making. Experimental results confirmed that the modified CNN2D outperformed the basic model in accuracy, precision, and recall. The findings emphasize the importance of combining deep learning with explainable AI for trustworthy image authentication. Overall, this research contributes a practical and interpretable solution for addressing the rising challenge of fake image detection in the era of advanced generative artificial intelligence.

REFERENCES

[1] Bird, J. and Lofti, A. (2024). CIFAKE: Classification of images and explainable identification of synthetic images using artificial intelligence. IEEE Access.

[2] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 770–778.

[3] Achilleos, K.G., Leandrou, S., Prentzas, N, Kyriacou, P.A., Kakas, A.C. and Pattichis, C.S. (2021). Explainable assessments of alzheimer's disease extracted with machine learning on brain MRI imaging data. IEEE Access.

[4] Hazarika, R.A., Abraham, A., Kandar, D. and Maji, A.K. (2021). A Modified LeNet-Deep Neural Network Model for Classifying Alzheimer's Disease from Magnetic Resonance Images of the Brain. IEEE Access.

[5] Farooq, A., Anwar, S.M., Awais, M., and Rehman, S. (2017). Deep learning based multi-class classification of Alzheimer's disease using brain MRI. IEEE Transactions on Instrumentation and Measurement, 66(12), 3140-3149.

[6] Pusparani, Y., Lin, C.Y., Jan, Y.K., Lin, F.Y., Liau, B.Y., Ardhianto, P., Farady, I., and Alex, J.S. (2023). Alzheimer's disease diagnosis using convolutional neural network with landmark-based slicing on hippocampus MRI. IEEE Access.

[7] Kim, D., and Lee, J. (2024). CNN model augmentation for image classification by adaptive learning rate.

[8] Srivastava et al. published in Journal of Machine Learning Research 'Dropout: A simple way to prevent neural networks from overfitting' in 2014 where they described how to stop neural networks from overfitting using dropout technique on Volume 15 titled Advanced Methods of Machine Learning, Pages 1929-1958.

[9] Ioffe and Szegedy released their paper 'Batch normalization: Accelerating deep network training by reducing internal covariate shift' in 2015 th the International Conference on Machine Learning, which can be found in Volume 448-456.

[10] Selvaraju et al. published in 2017 Grad-CAM: Visual explanations from deep networks via gradient-based localization. This was presented at the IEEE International Conference on Computer Vision where they discussed issues 618-626.

[11] Deep learning enthusiasts Krizhevsky and Hinton as well as Sutskever published ImageNet classification with deep convolutional neural networks in 2012 depicting their work at Advances in Neural Information Processing Systems vol 1097-1105

[12] In 2017, F. Chollet published Xception: Deep learning with depthwise separable convolutions at the IEEE CVPR conference where he explained his work on pages 1251-1258.

[13] In 2017 at NIPS, A. Vaswani published along with N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser and I. Polosukhin Attention is all you need where they discussed 5998-6008.

[14] A. Zisserman and K. Simonyan published Deep convolutional networks for large-scale image recognition in 2015 at the International Conference on Learning Representations.

[15] Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K.Q. (2017). Densely connected convolutional networks. CVPR, 2017 IEEE Conference on Computer Vision and Pattern Recognition, 4700-4708.

[16] Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K. (2015). Spatial transformer networks. Advances in Neural Information Processing Systems, NIPS, 2017-2025.

[17] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

[18] Hinton, G.E., Srivastava, N., & Swersky, K. (2014). Neural networks for machine learning. Lecture Notes in Computer Science, 15(4), 217-225.

[19] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. 2018, arXiv:1804.02767.

[20] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Inception v4, inception-resnet v2 and the impact of residual connections on the inception modules. CVPR, 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2818-2826.