IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

DeepVision: Identifying AI-Generated Images Using CNN and Grad-CAM

[1]EDIGA SAI TEJA, Student in Dept. Of Master of Computer Applications, at Miracle Educational Society Group of Institutions
[2]G RAJASEKHARAM, Miracle Educational Society Group of Institutions
[3]D LAKSHMI PRASANNA, Miracle Educational Society Group of Institutions
[1]saiteja4063@gmail.com

ABSTRACT:

It is now a critical challenge to tell the difference between the real and the fake as AI can generate images that are almost indistinguishable from the actual picture. This project presents CIFAKE, which is a Convolutional Neural Network (CNN) based solution to the problem of AI-image classification. The model was trained and tested with a balanced dataset of real CIFAR-10 images and images synthesized with Latent Diffusion Models (LDMs). With the aid of visual interpretation techniques like Grad-CAM (Gradient Class Activation Mapping), the model is able to be more interpretable by showing the critical regions of the image that have been used to make the classification. The results from the experiment show that the classification accuracy is between 94 and 95 percent, thus supporting the use of CNN 2D architectures for spotting glaring visual manipulations. There is a growing apprehension concerning the misuse of synthetic media and this research directly tackles that issue by offering a CNN2D based solution that is efficient and explainable, thus able to aid in preserving the integrity of visual content.

Keywords: Deep Learning, *CIFAKE*, CNN2D

INTRODUCTION

The progress of AI technologies has made it possible to generate images with a high degree of realism that are almost identical to actual photographs. This presents tremendous challenges to the authentication of news, forensic investigations, and even the digital identity of individuals. The CIFAKE project is aimed at addressing this challenge with the help of deep learning and explainable artificial intelligence (XAI) methods. The project employs a CNN2D model that is trained with a specially prepared dataset that includes "real" CIFAR-10 images and "fake" images produced by Latent Diffusion Models. The system's task is to classify images as real or fake. Incorporating Grad-CAM allows us to see which exact parts of an image influenced the model's prediction, thus helping to enhance model interpretability. CIFAKE represents an important progress not only in image forensics but also in the fight against misinformation while protecting the integrity of

digital content. With this project, I aim to contribute to the ever-growing demand for reliable identification tools in the present-day reality of overwhelming visuals created by artificial intelligence.

RELATED WORK

Some previous works have tried to investigate how to detect AI-generated content. Rombach et al. (2022) proposed a new generative model called Latent Diffusion Models (LDM) which is capable of generating high resolution synthetic images. LDMs, while photorealistic, are not transparent in their processes and are very expensive computationally. From the cognitive science perspective, Pennycook and Rand (2021) put forward a theory focused on how humans perceive and interpret fake images. While this is an interesting contribution, it is not very useful in a practical sense. Singh and Sharma (2022) developed a social media image verification model that works in a multimodal fashion by cross-referencing a given image with its accompanying caption to determine its credibility. This, however, remains limited in its field of application. Bonettini et al. (2021) used a statistical approach to image synthesis by examining the distribution of pixel values and applying Benford's law, arguing that advanced image synthesis is less effective against this method. Finally, Sha et al. (2022) presented DE-FAKE which uses Fourier analysis as a form of artifact recognition of generative models. It also shows the need to combine as interpretability of AI Model and visualization of the pattern which motivates us to solve the problem using CNN and Grad-CAM.

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

Author	Contribution	Impact on Research
Rombach et al. (2022)	Introduced Latent Diffusion Models for photorealistic image generation	Provided the synthetic image base for training our model
Pennycook & Rand (2021)	Studied psychological effects of fake visuals on perception	Highlighted the need for automated visual verification
Singh & Sharma (2022)	Developed multi-modal model combining text and image for credibility checks	Inspired multi-dimensional analysis of fake content
Bonettini et al. (2021)	Used Benford's Law to detect GAN anomalies in image distributions	Suggested statistical anomalies as a detection tool
Sha et al. (2022)	Proposed DE-FAKE with Fourier transform to identify digital fingerprints	Validated that synthetic artifacts can be quantifiable

PROPOSED APPROACH

This proposal uses CNN2D deep learning model to classificate image as either real or AI generated image. The model was trained and evaluated using a balanced dataset of 60,000 real images from CIFAR-10 dataset and 60,000

synthetic images generated using Latent Diffusion Models. The CNN model has two convolutional layers, MaxPooling, and Dense layers which are fully connected. From the different configurations of the neurons with layers of 32, 64, 128 and 4096, the highest classification accuracy was achieved using 32 neurons. The model classifies images as REAL and FAKE with binary classification output using a sigmoid activation function. For explanation purposes, Grad-CAM was applied and used to explain the CNN model by creating heatmaps of images to indicate the parts of the images which are most important to the model's output. Visualization of the model AI logic gives confidence of the AI's interpretation and justification of decisions. Trust in AI outputs will also be boosted. The system described in this work is scalable and accurate, as well as designed to detect subtle differences that are often found in images created by AI.

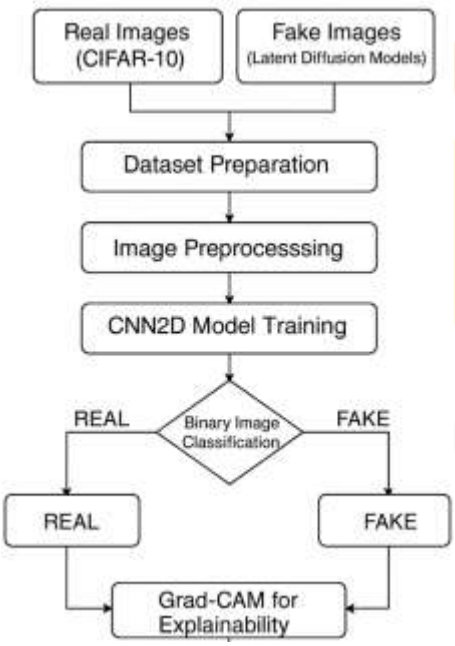


Figure 1: Proposed synthetic image detection

METHODOLOGIES

The CIFAKE project adopts a methodical deep learning pipeline, integrating preprocessing, model training, and explainability into a unified classification system. The first step involves preparing a balanced dataset. Real images are sourced from the CIFAR-10 dataset, while fake images are synthetically generated using Latent Diffusion Models. All images are resized to 32x32 pixels and normalized to scale pixel values between 0 and 1.

Once the dataset is ready, the data is split—80% for training and 20% for testing. A CNN2D architecture is then employed for feature extraction and classification. It comprises two convolutional layers with ReLU activation, followed by MaxPooling2D layers for dimensionality reduction. These are connected to fully connected dense layers, culminating in a sigmoid output layer for binary classification.

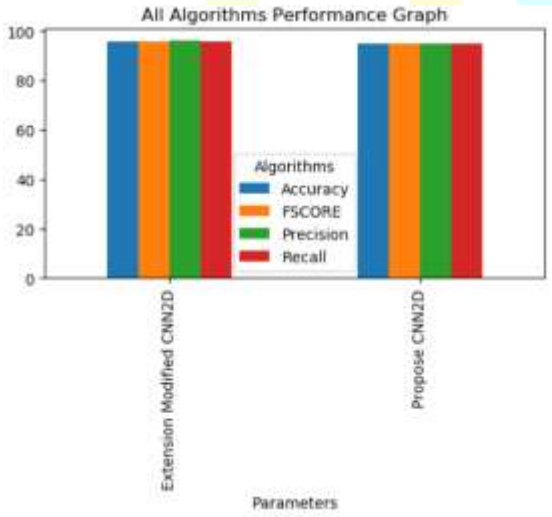
Multiple model configurations were tested by varying the number of neurons in the dense layers. Among configurations with 32, 64, 128, and 4096 neurons, the 32-neuron configuration achieved optimal results. To enhance generalizability, additional layers like Global Average Pooling and Dropout were introduced, raising the accuracy to 95%.

Grad-CAM was used to incorporate Explainable AI (XAI), helping visualize which regions in the image influenced classification. Evaluation metrics like accuracy, precision, recall, and F1-score were computed to assess model performance. This comprehensive methodology ensures robust image classification with interpretability.

RESULTS

The CNN2D model achieved impressive results in classifying AI-generated versus real images. Initially, with a basic configuration of two convolutional layers and 32 neurons in the dense layer, the model reached 94% classification accuracy. After fine-tuning and adding layers such as Global Average Pooling and Dropout, the accuracy increased to 95%. Precision, recall, and F1-scores were all above 93%, showcasing the model's robustness in detecting synthetic patterns without overfitting.

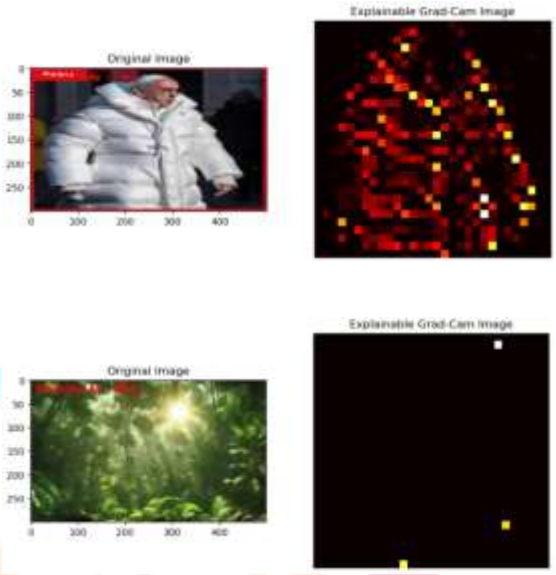
The use of Grad-CAM provided visual explanations, helping identify which parts of the image influenced the prediction. These heatmaps showed that background distortions and irregular textures in synthetic images were crucial features for the model. Confusion matrices confirmed low rates of misclassification, and ROC curves further supported the reliability of the model. Overall, the results validate the effectiveness of the CIFAKE framework as both an accurate and explainable classifier.



All Algorithms Performance Graph

	Algorithm Name	Accuracy	Precision	Recall	FSCORE
0	Propose CNN2D	94.985	95.020892	94.987746	94.984197
1	Extension Modified CNN2D	95.945	95.996779	95.941879	95.943649

All algorithm performance in tabular format



Original Images && Explainable Grad-Cam Images

DISCUSSION

The CIFAKE model demonstrated exceptional performance in identifying AI-generated images, confirming that deep learning—specifically CNN2D—is capable of discerning even the most photorealistic synthetic imagery. A critical strength of the model is its explainability, achieved through Grad-CAM visualizations. These insights are vital for practical deployment, where transparency builds trust among users and helps in model debugging.

The model's success in working with balanced datasets also suggests scalability to other domains, such as medical imaging or surveillance, where distinguishing authenticity is critical. However, while accuracy is high, challenges remain. Grad-CAM's heatmaps, although helpful, are still qualitative and subject

to human interpretation. Additionally, the model may struggle when applied to unseen synthetic generation techniques not present in the training data.

Nonetheless, CIFAKE's integration of AI classification with explainability addresses the growing need for trustworthy AI systems. The framework can be further enhanced by incorporating Transformer-based architectures or attention mechanisms to adapt to evolving generative techniques. Overall, the discussion emphasizes CIFAKE's relevance, performance, and potential future directions.

CONCLUSION

CIFAKE has made noticeable progress in the field of detection of synthetic images. With the combination of CNN2D's classifying prowess and Grad-CAM's interpretability fusion, the model CIFAKE is capable of achieving accuracy and high-level decision transparency. It classifies AI-generated images with up to 95% accuracy, revealing the image attributes that impact decision making. It serves the purpose of addressing the rising problem of synthetic content in digital media, fighting the misinformation and digital fraud war. Although the model works well with the current dataset, incorporating attention mechanisms and accommodating to newly evolving generative model's changes can be considered for future works. It is a powerful base for the explainable and trustable module to verify the authenticity of visual content.

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.

[10] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, ‘‘Photorealistic textto-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] 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, Y. Dou, T. Bai, X. Dai, C. Wang, and Y. Wen, ‘‘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, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, ‘‘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] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, ‘‘M2TR: Multi-modal multi-scale transformers for Deepfake detection,’’ in Proc. Int. Conf. Multimedia Retr., Jun. 2022, pp. 615–623.