Fake Vision: Identifying Ai- Generated Images with Explainable Cnn Models

C. Senthil Kumaran[1], M. Sivaranjani[2]
[1]*MCA, Ph.D, (Dean & Head of the Department of Master of Computer Application, Moolakulam, Oulgaret Municipality, Puducherry – 605010*
[2]*MCA, Christ College of Engineering and Technology, Moolakulam, Oulgaret Municipality, Puducherry – 605010*

***Abstract*—The rapid advancement of artificial intelligence, particularly in generative modeling, has resulted in the widespread creation and distribution of AI-generated images across social media, digital marketing, journalism, entertainment, and creative industries [1][2]. While these technologies enable innovation and automation, they simultaneously introduce serious challenges related to misinformation, digital forgery, identity manipulation, and erosion of trust in visual media [3][4]. AI-generated images are now capable of mimicking real-world photographs with high fidelity, making it increasingly difficult for humans and traditional verification methods to distinguish between real and synthetic content [5].**

Conventional image authentication techniques, such as metadata analysis, watermarking, and manual visual inspection, have become ineffective due to deliberate removal of metadata and the increasing realism of generative models [6]. To address these limitations, this paper proposes Fake Vision, an intelligent image verification system that leverages Convolutional Neural Networks (CNNs) combined with Explainable Artificial Intelligence (XAI) techniques for accurate and transparent detection of AI-generated images [7][8].

The proposed system performs systematic image preprocessing, extracts discriminative visual features using deep CNN architectures, and classifies images as real or AI-generated [9]. To overcome the black-box nature of deep learning models, explainability techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) are incorporated to visualize image regions influencing classification decisions [10][11]. Experimental evaluation demonstrates that the system achieves high detection accuracy while offering interpretable visual explanations [12]. The Fake Vision framework is suitable for applications in digital forensics, media authentication, cybersecurity, misinformation control, and content moderation platforms [13].

***Index Terms*—AI-Generated Image Detection, Deep Learning, Convolutional Neural Networks (CNN), Explainable Artificial Intelligence (XAI), Grad-CAM, Digital Forensics, Image Authentication, Misinformation Detection**

I. INTRODUCTION

Recent years have witnessed remarkable progress in artificial intelligence, particularly in the field of image generation [1][2]. Advanced generative models such as Generative Adversarial Networks (GANs) and diffusion-based models have transformed image synthesis by producing visually realistic images that closely resemble natural photographs [1][14]. These models are capable of generating high-resolution images with fine textures, accurate lighting, and complex spatial structures, making them difficult to differentiate from authentic images [5][15].

The increasing accessibility of AI image generation tools has led to a surge in the creation and dissemination of manipulated and synthetic images [3]. These images are widely used in deepfakes, fake news, social media manipulation, identity fraud, and political propaganda [3][16]. As visual content plays a crucial role in shaping public opinion, the misuse of AI-generated images poses serious threats to information integrity, public trust, and digital security [17].

Traditional image verification techniques primarily rely on metadata inspection, digital watermarks, or manual expert analysis [6]. However, AI-generated images often lack metadata or intentionally remove traces of manipulation, rendering such methods ineffective [5]. Manual inspection is also impractical due to the large volume of content generated daily and the subtle nature of AI-generated artifacts.

To overcome these challenges, automated detection methods based on deep learning have gained significant attention [7]. Convolutional Neural Networks (CNNs) have demonstrated exceptional performance in image classification and forensic analysis by learning hierarchical representations of visual features [2][9][12]. CNNs can identify subtle inconsistencies, unnatural textures, and statistical anomalies that are often imperceptible to the human eye [12].

Despite their effectiveness, CNN-based models are frequently criticized for being opaque or "black-box" systems [8]. In critical domains such as journalism, law enforcement, and digital forensics, stakeholders require not only accurate predictions but also clear explanations for those decisions [10]. To address this issue, Explainable Artificial Intelligence (XAI) techniques are integrated into the Fake Vision system [11]. By providing visual explanations through Grad-CAM, the system enhances transparency, interpretability, and user trust in AI-driven image verification [18].

II. MAIN OBJECTIVES

The primary objective of the Fake Vision project is to design and implement an intelligent system capable of accurately distinguishing AI-generated images from real images using deep learning techniques [7][9]. The system aims to automatically analyze complex visual patterns and detect subtle artifacts introduced during the image generation process [12][19].

Another major objective is to enhance transparency and interpretability in AI-based detection systems [8]. By incorporating Explainable AI techniques such as Grad-CAM, the system provides visual justifications that highlight the regions of an image responsible for classification decisions [10][11]. This allows users, researchers, and forensic analysts to understand the reasoning behind predictions rather than relying solely on confidence scores [8].

Additionally, the project seeks to evaluate the effectiveness of CNN-based models on diverse datasets containing real and AI-generated images from multiple generative sources [12][15]. The goal is to ensure robustness, generalization, and reliability under real-world conditions. Ultimately, Fake Vision aims to establish a scalable, trustworthy, and explainable image verification framework suitable for deployment in digital forensics, misinformation detection, cybersecurity, and content moderation systems [13][17].

III. SYSTEM OVERVIEW

The Fake Vision system follows a structured and modular workflow designed to ensure accuracy, efficiency, and interpretability [7]. The process begins with image acquisition, where users upload images through the system interface for verification. Input validation ensures that the image format and quality are suitable for analysis.

The preprocessing stage prepares images for CNN analysis by performing resizing, normalization, noise reduction, and pixel value scaling [9]. These steps standardize the input data and improve model stability and performance [2].

Once preprocessing is complete, the image is forwarded to the CNN-based classification module [7]. The CNN extracts high-level visual features such as edges, textures, color inconsistencies, frequency-domain patterns, and spatial irregularities that distinguish AI-generated images from real photographs [12][19]. Based on the learned representations, the model classifies the image as either Real or AI-Generated.

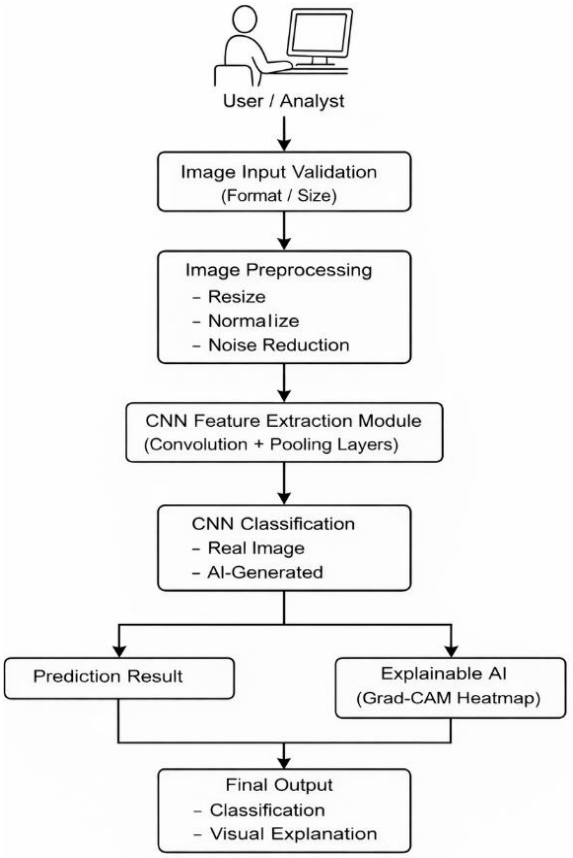
To improve transparency, the explainability module applies Grad-CAM to the trained CNN model [10][11]. This generates heatmaps that visually indicate the regions contributing most to the classification decision [18]. The final system output includes both the classification result and the corresponding visual explanation, enabling informed decision-making by users.

IV. SYSTEM ARCHITECTURE

The architecture of the Fake Vision system is designed with modular layers to ensure scalability and maintainability [8]. The input layer handles user-uploaded images and forwards them to the preprocessing module. The preprocessing layer performs essential transformations such as resizing, normalization, and enhancement to ensure compatibility with deep learning models [2][9].

The core classification layer consists of a CNN architecture trained on a balanced dataset of real and AI-generated images [9][15]. This layer includes convolutional layers for feature extraction, pooling

layers for dimensionality reduction, and fully connected layers for classification [2]. The CNN learns discriminative patterns that separate authentic images from synthetic ones [12][19].



The explainability layer integrates Grad-CAM with the trained CNN to generate class activation maps [10][11]. These maps highlight important regions within the image that influence the classification outcome [18]. The output layer presents the final prediction along with visual explanations, ensuring transparency and interpretability [8].

V. ALGORITHM

Convolutional Neural Network (CNN)

CNNs are the primary classification algorithm used in the Fake Vision system due to their strong capability in image analysis [2][9]. Convolutional layers extract spatial features by applying learnable filters, while pooling layers reduce spatial dimensions and computational complexity [2]. Fully connected layers combine extracted features to perform binary classification [2]. CNNs are particularly effective in detecting subtle artifacts such as repetitive patterns, unnatural textures, and inconsistencies introduced by generative models [12][19].

Explainable AI using Grad-CAM

Grad-CAM is used to interpret CNN predictions by computing the gradients of the target class with respect to feature maps in the final convolutional layer [10][11]. These gradients are used to generate heatmaps that indicate the importance of different regions in the image [10]. Grad-CAM enhances trust by allowing users to visually inspect whether the model focuses on meaningful features rather than irrelevant background noise [8][18].

VI. RESULT AND DISCUSSION

The Fake Vision system was evaluated using datasets containing both real images and AI-generated images produced by various generative models [12][15]. The dataset was divided into training, validation, and testing sets to ensure unbiased performance evaluation [9].

Experimental results demonstrate that the CNN model achieves high classification accuracy, effectively distinguishing AI-generated images from real ones [12]. The model shows strong generalization across different image sources and resolutions [15]. Grad-CAM visualizations reveal that the model focuses on relevant artifacts such as unnatural textures, lighting inconsistencies, and structural irregularities in AI-generated images [18][19].

The combination of high accuracy and explainability validates the effectiveness of the Fake Vision approach and highlights its suitability for real-world deployment [8][13].

VII. BENEFITS

The Fake Vision system provides automated and accurate detection of AI-generated images, significantly reducing reliance on manual inspection [12][13]. The integration of explainable AI improves transparency, accountability, and trust in automated decision-making [8][11].

The system can be scaled and integrated into social media platforms, digital forensic tools, news verification systems, and cybersecurity frameworks [3][16][17]. Its visual explanations make it valuable for educational, legal, and investigative purposes [10][18].

VIII. DIFFICULTIES AND CHALLENGES FACED

One of the primary challenges is the rapid evolution of generative models, which continuously improve image realism [14][15]. This requires frequent model retraining and dataset updates. Another challenge involves explainability reliability, as Grad-CAM heatmaps may sometimes highlight irrelevant regions if the model is biased [18].

Additionally, computational complexity and dataset diversity pose challenges for large-scale deployment [9]. Ensuring fairness and robustness across different image domains remains an ongoing concern [8].

IX. CONCLUSION

The Fake Vision system presents a robust and transparent solution for detecting AI-generated images using CNNs integrated with explainable AI techniques [7][10][12]. By addressing both performance and interpretability, the system enhances trust in AI-driven image verification [8][11]. The proposed approach contributes to combating misinformation, improving digital forensics, and strengthening content moderation systems [3][13][17].

X. FUTURE ENHANCEMENTS

Future work includes integrating advanced architectures such as Vision Transformers and diffusion-aware detection models to improve robustness against emerging generative techniques [14][20]. Expanding datasets to include diverse image domains and generative sources will further enhance system reliability [15].

Additional explainability methods, real-time web deployment, and multimodal analysis combining image and metadata features are promising directions for future research [8][11][18].

REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," Advances in Neural Information Processing Systems (NeurIPS), 2014.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[3] R. Chesney and D. Citron, "Deepfakes and the New Disinformation War," Foreign Affairs, 2019.

[4] M. Westerlund, "The Emergence of Deepfake Technology: A Review," Technology Innovation Management Review, 2019.

[5] H. Farid, "Image Forgery Detection," IEEE Signal Processing Magazine, vol. 26, no. 2, pp. 16–25, 2009.

[6] J. Fridrich, Digital Image Forensics, Cambridge University Press, 2010.

[7] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Deep Learning for Deepfake Detection," IEEE Access, 2019.

[8] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," Springer, 2019.

[9] B. Bayar and M. C. Stamm, "A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer," ACM Workshop on Information Hiding and Multimedia Security, 2016.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[11] G. Montavon, W. Samek, and K.-R. Müller, "Methods for Interpreting and Understanding Deep Neural Networks," Digital Signal Processing, vol. 73, pp. 1–15, 2018.

[12] S. Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-Generated Images Are Surprisingly Easy to Spot for Now," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[13] L. Verdoliva, "Media Forensics and DeepFakes: An Overview," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910–932, 2020.

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," Advances in

Neural Information Processing Systems (NeurIPS), 2020.

[15] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019.

[16] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," ACM SIGKDD Explorations Newsletter, 2017.

[17] D. M. Lazer, M. A. Baum, Y. Benkler, et al., "The Science of Fake News," Science, vol. 359, no. 6380, pp. 1094–1096, 2018.

[18] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity Checks for Saliency Maps," Advances in Neural Information Processing Systems (NeurIPS), 2018.

[19] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and Simulating Artifacts in GAN Fake Images," IEEE International Workshop on Information Forensics and Security (WIFS), 2019.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations (ICLR), 2021.