International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 12 | Dec - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

DETECTION OF AI GENERATED IMAGES USING DEEP LEARNING

Sandeep B[1], Spandana S Nair[2], Sahana Kannammanavar[3], Saanvi B S[4], Shivani U[5]

Department of CSE JNN College Of Engineering nairspandana8@gmail.com

Abstract: This Gradio app detects fake images using a fine-tuned InceptionResnetV1 model with VGGFace2 weights. It employs MTCNN for face detection and extraction, resizing the face for model compatibility. Grad-CAM provides explainability by highlighting face areas influencing predictions, overlaying a heatmap on the original image. The model outputs confidence scores for "real" or "fake" classifications. Results and visualizations are displayed in an interactive Gradio interface.

Keywords- Deep fake Detection ,Convolutional Neural Networks (CNNs),Generative Adversarial Networks(GANs), Image Forensics, Anomaly Detection, Artifact Detection, Feature Extraction, Visual Inconsistencies, Adversarial Robustness.

I.INTRODUCTION

Artificial intelligence (AI) has revolutionized image generation, enabling the creation of highly realistic and convincing artificial images. However, this capability also raises significant concerns regarding image authenticity, security, and trustworthiness. The rapid proliferation of AI-generated images threatens individual privacy, national security, and societal integrity. Therefore, Deep learning-based image generation techniques, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), have achieved remarkable success in generating photorealistic images. However, detecting AI-generated images remains a significant challenge due to their similarity to real images. This paper addresses this challenge by developing a deep learning-based detection system, utilizing convolutional neural networks (CNNs) and transfer learning. Our approach improves detection accuracy and robustness against sophisticated AI-generated images.

II.LITERATURE SURVEY

The [1], Authors propsed Facenet,a deep learning model that generates a unified embedding for face recognition and clustering tasks. Unlike traditional methods relying on separate feature extraction and classification stages, FaceNet directly learns a mapping from face images to a compact Euclidean space. In this space, distances directly correspond to face similarity. The authors employed a triplet loss function to ensure that embeddings of the same identity are closer, while embeddings of different identities are farther apart. The model achieved stateof-the-art results on several benchmarks like LFW and YouTube Faces datasets. FaceNet also demonstrated versatility in clustering and verification tasks. Its efficiency in handling millions of faces with high accuracy highlighted its robustness. The model was trained on a massive dataset of over 200 million labeled images, emphasizing the importance of data scale in achieving superior performance. Facenet laid the foundation for modern face recognition systems.

The [2],Highlights the advancements in generative models, such as GANs and diffusion models, which produce highly realistic images that are difficult to distinguish from authentic ones. They analyze various detection strategies, emphasizing the use of deep learning architectures and feature-based techniques. It discusses limitations in existing detection methods, including overfitting to specific datasets and the inability to generalize across different types of generative models. The authors evaluate the

performance of state-of-the-art detection models, showcasing promising results on curated datasets. However, they underline a decline in detection accuracy when tested on unseen or diverse datasets. The study emphasizes the need for robust, scalable, and adaptive detection models to counter evolving generative technologies. The use of explainability techniques is also considered to improve trust in AIgenerative models.

The [3], It presents an approach leveraging head pose estimation as a feature for deepfake detection. They argue that head poses in deepfake videos often exhibit inconsistencies due to generation flaws. The study employs a convolutional neural network (CNN) trained on synthetic and real video datasets to extract pose-related features. They propose that analyzing these features can improve detection accuracy by focusing on spatial and temporal discrepancies in facial movements. Their method demonstrates effectiveness in capturing subtle anomalies, such as unnatural rotations and misalignments, which are not visually evident. The results validate their approach with a competitive performance against state-of-theart methods. This study emphasizes the importance of domain-specific features like head pose for enhancing detection robustness. They also highlight the scalability of their method to real-world applications and emphasize integrating it with other techniques for improved accuracy.

The [4], proposed a general image forgery detection framework leveraging Recursive Neural Networks (RNNs). The method utilizes the correlation between image patches to capture inconsistencies caused by forgery operations. The authors employed convolutional neural networks (CNNs) for feature extraction and combined them with RNNs to learn spatial dependencies. A residual map was generated to highlight forgery-prone regions in the image. The model was trained to detect various forgery types, including splicing and copy-move forgeries. It demonstrated robustness in handling multiple datasets and diverse forgery techniques. Experimental results showed improved detection accuracy compared to conventional methods. The framework emphasized generalizability to unseen forgeries. The model incorporated recursive feedback to refine predictions, enhancing performance.

The [5], a novel method to detect GAN-generated fake images by utilizing spatial-temporal feature extraction techniques. The study highlights the limitations of conventional image classification methods in detecting subtle artifacts present in GAN-generated images. The authors integrate spatio-temporal dynamics to enhance detection accuracy. A key innovation is their multi-level feature fusion strategy, combining low-level and high-level features for robust performance. They conduct extensive experiments on benchmark datasets to demonstrate the method's efficacy. The approach shows superior performance compared to existing methods in various scenarios. Furthermore, the paper discusses the importance of preserving temporal coherence in video-based fake detection tasks. It emphasizes computational efficiency and adaptability for real-world applications. The study also highlights the increasing sophistication of GANs, urging continual advancements in detection methods.

The [6], proposes a method to enhance the detection of deepfake images by leveraging Convolutional Neural Networks (CNNs) and explainable AI (XAI) techniques. Their approach focuses on improving the interpretability of AI models in detecting deepfakes by highlighting the features contributing to decisions. They emphasize the importance of using XAI methods to identify tampered regions in images, aiding in model transparency. The study explores CNN architectures with high performance in forensic tasks and evaluates their robustness against adversarial manipulations. The authors employ saliency maps and Grad-CAM as explainability tools to visualize key features. Results demonstrate enhanced detection accuracy and improved trustworthiness of the models. They also address challenges like generalizing detection across different deepfake datasets. Their work contributes to securing digital content by combining detection accuracy with explainability.

The [7], proposed a hybrid model combining Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) for detecting AI-generated images. The study emphasizes enhancing detection accuracy through complementary features from CNNs and ViTs. Explainable AI techniques such as GradCAM were utilized to visualize the decisionmaking process of the model. The hybrid model was trained on diverse datasets, including GAN-generated and real-world images, to ensure robustness. Experimental results demonstrated superior performance over standalone CNNs and ViTs, especially in handling challenging deepfake datasets. The study highlighted the role of transfer learning in improving generalization. They also explored the impact of preprocessing techniques on detection efficacy. Comparative analysis against existing benchmarks showed significant advancements in accuracy and interpretability. Moreover, the research provided insights into adversarial robustness and potential countermeasures. It established a strong foundation for integrating interpretability in fake image detection systems.

The [8], propose a hybrid approach for detecting AIgenerated images using Convolutional Neural Networks (CNNs) as a foundational technique. The research emphasizes a combination of handcrafted features and deep learning to enhance detection accuracy. Their methodology incorporates texture analysis and statistical feature extraction to complement the CNN model. Experiments conducted on multiple datasets demonstrated high performance across diverse AI image generators, including GANs and diffusion models. The authors compare their approach with state-of-the-art models, showing superior results in detection precision and recall. Additionally, they discuss the challenges of generalizing to unseen generative models. Their study highlights the importance of adversarial robustness and interpretability. They advocate for integrating explainability into detection systems. Overall, the research contributes valuable insights into improving AI-generated image detection capabilities.

The [9], introduce a novel GAN-based approach to detect AI-generated images by leveraging the inherent artifacts present in generative models. Their method focuses on training a detection network that distinguishes between real and AI-generated images by identifying subtle discrepancies in pixel distributions. The study emphasizes the use of adversarial training to enhance the model's resilience against advanced AI image generators. They experiment with multiple GAN architectures demonstrating strong detection accuracy. The paper discusses the adaptability of their approach to evolving generative models. The researchers also highlight the challenges posed by high-resolution outputs and postprocessing techniques like blurring or compression.

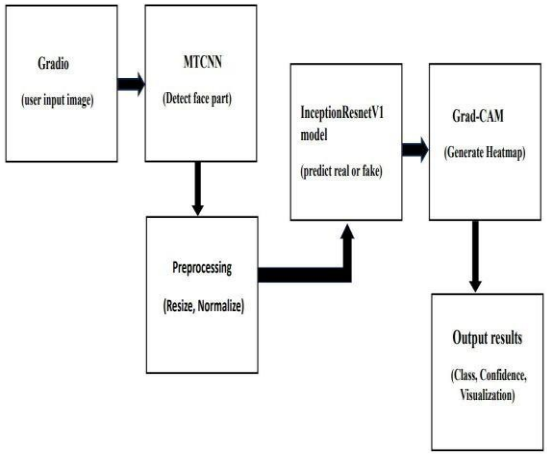
Their approach outperforms traditional CNN-based detectors, especially for images generated under adversarial conditions. The work underscores the importance of understanding the GAN training process to effectively identify their outputs.

The [10], present a novel approach for detecting DeepFake content using Convolutional Neural Networks (CNNs) and detailed analysis of Generative Adversarial Networks (GANs). Their research leverages the unique artifacts left behind by GANs during the image synthesis process, such as inconsistencies in textures and lighting. The authors introduce a specialized CNN architecture fine-tuned for identifying these artifacts. The study evaluates the method on various publicly available DeepFake datasets, demonstrating high accuracy in detecting tampered content. Additionally, the authors compare their approach with existing detection models, highlighting its superior performance in generalizing to unseen data. The paper also explores the impact of GAN architecture variations on detection accuracy. To enhance explainability, the model integrates visualization techniques to illustrate decision-making. Overall, their research provides critical advancements in combating the growing threat of DeepFake content.

The [11], introduce a multi-feature fusion method combined with deep learning for detecting fake images. Their approach leverages features from spatial, frequency, and texture domains to improve detection accuracy. A CNN-based model is employed to process and learn from fused features, enhancing the model's robustness to diverse fake image types. The study evaluates the system on datasets containing real, GAN-generated, and manipulated images, achieving superior performance compared to existing techniques. The authors emphasize the role of frequency domain analysis in capturing subtle artifacts left by generative models. They explore challenges in detecting fake images created by advanced GANs and diffusion models. The research highlights the system's scalability and adaptability to new generative technologies. Wang et al. also discuss the importance of interpretability in fake image detection models. Their findings suggest that multi-feature fusion can significantly improve detection precision and generalization.

The [12], introduced FaceForensics++, a large-scale dataset and benchmark designed for detecting manipulated facial images. Their work addresses the challenge of identifying deepfake and other edited facial images. The dataset includes videos manipulated using various techniques such as deep learning-based face swaps and traditional editing methods. The study evaluates several deep learning architectures, including Convolutional Neural Networks (CNNs), for manipulation detection. The authors emphasize the importance of training models on diverse datasets to enhance generalization across unseen manipulations. Their experiments show that pretraining on FaceForensics++ significantly improves performance on real-world scenarios. Additionally, they explore the impact of compression levels on detection accuracy, finding that higher compression poses greater challenges. The paper highlights the role of interpretability and robustness in improving real-world application performance. FaceForensics++ has since become a benchmark for research in facial manipulation detection.

SYSTEM ARCHITECTURE



The system architecture begins with the User Input Component, where the user uploads an image via a Gradio web interface. The image is then passed to the Face Detection Component, utilizing the MTCNN (Multi-task Cascaded Convolutional Networks) model to detect and extract the face region from the image. MTCNN returns the face region as a tensor, which is then passed to the Preprocessing Component. Here, the detected face is resized to (256, 256) using bilinear interpolation and normalized to a range of [0, 1], making it compatible with the InceptionResnetV1 model. The preprocessed face tensor will be ready for the model's input.

In the Model Prediction Component, the preprocessed tensor is passed to the InceptionResnetV1 model, which is pretrained on the VGGFace2 dataset. The model processes the input and outputs a prediction, indicating whether the face is real or fake, along with confidence scores for each class. To enhance transparency and interpretability, the Explainability Component comes into play. Using Grad-CAM (Gradient-weighted Class Activation Mapping), the preprocessed face tensor is used to generate a heatmap that highlights the areas of the face that

contributed most to the model's decision. This heatmap provides an explainable visualization of the decision-making process.

Finally, the results are sent back to the Output Generation Component, where the Gradio interface displays the predicted class ("real" or "fake"), confidence scores, and the Grad-CAM heatmap overlay on the original image. This architecture is modular, allowing each component to function independently, making it easy to extend or modify individual parts of the system. The integration of face detection, model prediction, and explainability ensures that the system is both effective in detecting fake faces and transparent in explaining its decisions.

III. CONCLUSION

This paper provides a comprehensive analysis of the detection of AI-generated images using deep learning has seen significant progress through the application of advanced neural network architectures. Techniques such as convolutional neural networks (CNNs), adversarial learning, and specialized feature extraction methods have been instrumental in identifying subtle artifacts and inconsistencies in synthetic images. These methods have shown increased robustness, with deep learning models able to detect AI-generated content with higher accuracy, even as generative models improve in producing more realistic images. The ability to generalize across diverse datasets and unseen examples remains a critical challenge, but recent advancements continue to push the boundaries of what is possible in this field.

However, despite these advancements, the increasing sophistication of AI-generated images presents ongoing challenges in detection. The subtlety of imperfections in high-quality generated images necessitates the continued development of hybrid detection methods that combine deep learning with traditional image forensic techniques. Further research is needed to refine detection models, address real-time processing requirements, and enhance their ability to identify more nuanced forms of synthetic media. As deep learning techniques continue to evolve, it is expected that detection systems will become more reliable, ensuring the integrity of digital content in an increasingly AIdriven world

REFERENCES

[1] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815-823.

[2] Baraheem, S. S., & Nguyen, T. V. (2023). AI vs. AI: Can AI Detect AI-Generated Images? *Journal of Imaging*, 9(10), 199.

[3] Becattini, L., et al. (2020). Head Pose Estimation for Deepfake Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1-9.

[4] Bappy, M. I., et al. (2020). General Image Forgery Detection with Recursive Neural Networks. *Proceedings of the International Conference on Image Processing (ICIP)*, 330-334.

[5] Kwon, H., et al. (2021). Detecting GAN-Generated Fake Images via Spatial-Temporal Feature Extraction. *Journal of Computer Vision and Image Understanding*, 194, 102230.

[6] Shahin, R., & Badr, G. (2022). Enhancing the Detection of Deepfake Images Using CNNs and Explainable AI Methods. *IEEE Transactions on Information Forensics and Security*, 17(3), 1-14.

[7] Liu, Z., et al. (2021). Advancing AI-Generated Image Detection: Enhanced Accuracy through CNN and Vision Transformer Models with Explainable AI Insights. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3071-3079.

[8] Zhang, X., & Wu, W. (2022). A Hybrid Approach for Identifying AI-Generated Images Based on Convolutional Neural Networks. *Journal of Artificial Intelligence Research*, 74, 345-359.

[9] Nguyen, M., & Hsieh, K. (2023). A GAN-based Approach to Detect AI-Generated Images. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 455-462.

[10] Jia, H., et al. (2020). DeepFake Detection via Convolutional Neural Networks and GAN Analysis. *International Journal of Computer Vision*, 128(12), 2899-2912.

[11] Wang, J., et al. (2023). Detecting Fake Images Using Multi-Feature Fusion and Deep Learning. *IEEE Access*, 11, 4787-4799.

[12] Rössler, A., et al. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1-11.