DETECTING AI-GENERATED IMAGES WITH CNN AND INTERPRETATION USING EXPLAINABLE AI

[1]MEKALA AKHIL SURIBABA, [2]Y SRINIVAS RAJU

[1]Students, Department of MCA, B V Raju College, Bhimavaram Ap

[2]Assistant Professor, Department of MCA, B V Raju College, Bhimavaram Ap

ABSTRACT

The rapid advancement of generative models has significantly increased the production of synthetic images, raising concerns about authenticity, misinformation, and digital fraud. AI-generated images, particularly realistic human faces, are increasingly difficult to distinguish from genuine ones using human perception alone. This study proposes a robust deep learning-based framework to detect and classify real versus AI-generated images using convolutional neural networks (CNNs), combined with Explainable Artificial Intelligence (XAI) techniques for model interpretability. The proposed system utilizes the DenseNet121 pre-trained architecture for feature extraction and classification due to its efficient feature reuse and strong performance in image-related tasks. The model is trained on a large-scale dataset comprising real and fake facial images, ensuring diversity and generalization. Experimental results demonstrate that DenseNet121 achieves an accuracy of 94%, while further enhancement using the NASNet architecture improves performance to 98%, outperforming other evaluated models such as VGG16, VGG19, and Xception. To enhance transparency and trust in model predictions, the system integrates two widely used XAI techniques: Gradient-weighted Class Activation Mapping (Grad-CAM) and Local Interpretable Model-agnostic Explanations (LIME). Grad-CAM highlights the most influential regions in an image contributing to classification decisions, while LIME identifies critical feature segments responsible for predictions. The combined use of these methods provides consistent and interpretable visual explanations, validating the model's decision-making process. Overall, the proposed framework not only achieves high accuracy in detecting AI-generated images but also ensures interpretability, making it suitable for real-world applications in digital forensics, media authentication, and cybersecurity.

Keywords : Artificial Intelligence, Deep Learning, Convolutional Neural Networks, DenseNet121, NASNet, Image Classification, Fake Image Detection, Explainable AI (XAI), Grad-CAM, LIME, Digital Forensics

I. INTRODUCTION

The rapid growth of Artificial Intelligence, particularly in generative models, has

revolutionized digital content creation. Technologies such as Generative Adversarial Networks (GANs) and diffusion models enable the creation of highly realistic synthetic images that are often indistinguishable from real photographs. While these advancements have numerous beneficial applications in fields like entertainment, healthcare, and design, they also pose significant risks. Malicious actors can exploit AI-generated images for misinformation, identity theft, deepfakes, and financial fraud. As a result, distinguishing between real and artificially generated images has become a critical challenge in digital forensics and cybersecurity. Traditional image verification techniques are no longer sufficient due to the increasing sophistication of AI models, which can mimic textures, lighting, and facial details with high precision. Therefore, there is a pressing need for automated, reliable, and scalable solutions that can accurately detect synthetic content. This project addresses this challenge by leveraging deep learning techniques to build a robust classification system capable of identifying real and fake images, ensuring improved digital trust and security.

In this work, a Convolutional Neural Network (CNN)-based approach is proposed using the DenseNet121 pre-trained model for efficient feature extraction and classification. DenseNet121 is chosen due to its ability to reuse features through dense connections, reducing redundancy and improving learning efficiency. The model is trained on a large dataset consisting of real and AI-generated facial images, allowing it to learn subtle differences that are not easily detectable by human vision. To further enhance performance, advanced architectures such as NASNet are explored, which demonstrate superior accuracy compared to traditional models like VGG16, VGG19, and Xception. The dataset is preprocessed through normalization, shuffling, and splitting into training and testing sets, ensuring optimal learning and evaluation. The experimental results show that the proposed model achieves high classification accuracy, making it a reliable solution for detecting synthetic images in real-world scenarios.

Beyond accuracy, interpretability is a key aspect of this project, as understanding model decisions is essential for building trust in AI systems. To address this, Explainable Artificial Intelligence (XAI) techniques such as Grad-CAM and LIME are integrated into the framework. Grad-CAM provides visual explanations by highlighting the regions of an image that contribute most to the model's prediction, while LIME identifies important features influencing the decision at a local level. These techniques enable users to verify whether the model is focusing on meaningful patterns or irrelevant artifacts. The combination of deep learning and explainability ensures that the system is not only accurate but also

transparent and interpretable. This makes the proposed approach highly suitable for applications in digital forensics, social media monitoring, and content authentication, where both performance and trust are crucial.

II SURVEY OF RESEARCH

1. Study on CNN-based Image Forgery Detection : Early research in fake image detection focused on traditional image processing techniques; however, with the rise of deep learning, Convolutional Neural Networks (CNNs) became the dominant approach. Researchers demonstrated that CNN models can automatically learn hierarchical features such as edges, textures, and complex patterns, which are essential for identifying inconsistencies in synthetic images. These models outperform manual feature extraction methods by adapting to large datasets and complex image variations. Studies have shown that CNN architectures can detect subtle pixel-level anomalies introduced during image manipulation or generation. However, challenges remain in generalizing models across different datasets and unseen fake generation techniques. This research highlights the importance of deep learning as a foundational approach for image authenticity verification and sets the stage for more advanced architectures like DenseNet and NASNet used in this project.

2. Research on GAN-generated Image Detection: Generative Adversarial Networks (GANs) have significantly improved the realism of synthetic images, making detection increasingly difficult. Several studies have focused on identifying artifacts left behind by GANs, such as unnatural textures, inconsistencies in lighting, or irregular patterns in high-frequency components. Researchers have proposed deep learning models specifically trained to detect GAN fingerprints. These methods often involve training classifiers on datasets containing both real and GAN-generated images. While such models achieve high accuracy, their performance can degrade when exposed to new or unseen GAN architectures. This limitation emphasizes the need for robust and adaptable detection systems. The current project builds upon these findings by using advanced CNN architectures and large datasets to improve generalization and detection accuracy across diverse fake image sources.

3. Transfer Learning in Image Classification Transfer learning has become a widely adopted technique in deep learning, particularly for image classification tasks with limited computational resources. Pre-trained models such as DenseNet121, VGG16, and Xception are trained on large-scale datasets like ImageNet and can be fine-tuned for specific tasks such as fake image detection. Research shows that transfer learning significantly

reduces training time and improves performance by leveraging previously learned features. DenseNet121, in particular, is known for its dense connectivity, which promotes feature reuse and mitigates the vanishing gradient problem. Studies confirm that transfer learning-based models achieve higher accuracy compared to models trained from scratch. This project utilizes transfer learning to enhance performance and efficiency, ensuring accurate classification of real and AI-generated images.

4. Explainable AI Techniques in Image Classification : As deep learning models become more complex, the need for interpretability has grown significantly. Explainable Artificial Intelligence (XAI) techniques such as Grad-CAM and LIME have been widely studied to provide insights into model decisions. Grad-CAM generates heatmaps highlighting important regions in an image, while LIME explains predictions by approximating local model behavior. Research indicates that these methods help in validating whether the model is focusing on relevant features rather than noise or biases. In applications like medical imaging and security, interpretability is crucial for trust and accountability. Studies also show that combining multiple XAI techniques provides more reliable explanations. This project integrates both Grad-CAM and LIME to ensure transparency and improve user confidence in the classification results.

5. Comparative Analysis of Deep Learning Architectures

Numerous studies have compared the performance of different deep learning architectures for image classification tasks. Models such as VGG16, VGG19, Xception, DenseNet, and NASNet have been evaluated based on accuracy, computational efficiency, and generalization ability. Research findings suggest that while VGG models are simple and effective, they are computationally expensive. DenseNet improves efficiency through feature reuse, while NASNet, designed using neural architecture search, achieves superior performance with optimized structures. Comparative studies indicate that NASNet often outperforms other models in terms of accuracy and robustness. These findings support the enhancement phase of this project, where NASNet is used as an advanced alternative to DenseNet121, resulting in improved detection accuracy.

6. Applications of Fake Image Detection in Cybersecurity

Fake image detection has become increasingly important in cybersecurity, digital forensics, and social media platforms. Research highlights the growing threat of deepfakes and synthetic media in spreading misinformation, conducting fraud, and compromising privacy. Automated detection systems are essential for identifying manipulated content at scale. Studies emphasize the need for real-time

detection, robustness against evolving generation techniques, and integration with explainable models. Applications include identity verification, content moderation, and legal investigations. Despite advancements, challenges such as dataset bias and adaptability remain. The proposed project contributes to this domain by combining high-accuracy CNN models with explainable AI techniques, offering a reliable and transparent solution for detecting AI-generated images in real-world scenarios.

III. WORKING METHODOLOGY

The proposed system begins with collecting a large-scale dataset of real and AI-generated facial images from a reliable source such as Kaggle. Each image is carefully processed to ensure consistency in size, format, and labeling. Images are resized into a fixed resolution suitable for deep learning models and converted into numerical arrays. Data preprocessing techniques such as normalization are applied to scale pixel values, improving model convergence during training. Additionally, dataset shuffling is performed to eliminate any ordering bias that may affect learning. The dataset is then divided into training and testing sets, typically in an 80:20 ratio, where the training set is used to learn patterns and the testing set evaluates performance. Data augmentation techniques such as rotation, flipping, and zooming may also be applied to increase dataset diversity. This step ensures that the model becomes robust and capable of handling real-world variations in image data.

In the next phase, deep learning models are implemented using transfer learning techniques. The DenseNet121 pre-trained model is selected as the primary architecture due to its efficient feature reuse and ability to mitigate the vanishing gradient problem. The model is fine-tuned using the prepared dataset, where the final layers are modified to perform binary classification between real and fake images. During training, the model learns complex patterns, textures, and inconsistencies present in synthetic images. To further enhance performance, NASNet is introduced as an advanced architecture, optimized using neural architecture search techniques. Both models are trained using appropriate loss functions and optimizers, and their performance is evaluated using metrics such as accuracy, precision, recall, and F1-score. Graphical representations such as confusion matrices and ROC curves are generated to analyze classification performance and compare both models effectively.

Finally, the system integrates Explainable Artificial Intelligence (XAI) techniques to improve transparency and interpretability of the model's predictions. Grad-CAM is used to generate heatmaps that highlight important regions of an image that influence the model's

decision, helping to visualize where the model is focusing. In parallel, LIME is applied to provide local explanations by identifying key features that contribute to a specific prediction. These techniques ensure that the model does not behave as a black box and allow users to understand the reasoning behind classification results. The final system takes an input image, processes it through the trained model, and outputs whether the image is real or fake along with visual explanations. This combination of high accuracy and interpretability makes the system reliable for applications in digital forensics, cybersecurity, and media verification.

IV RESULTS EXPLANATIONS

Increase generation of digital media paving new ways for the malicious users to modify original content to make their own content and can make money by selling such modified digital images at cheaper price. Synthetic Fake images can be generated by modifying visual features from the real images and from human eyes it will be difficult to distinguish between real and fake images so propose paper employing deep learning algorithms which can easily detect and differentiate between real and AI Generated images.

Propose paper employing DenseNet121 pre-trained model for image processing and classification. Further GRADCAM based analysis employed for model prediction explaining. Gradient-weighted Class Activation Mapping (Grad-CAM) technique, enabling to visualize the regions within images that influenced the model's decision-making.

Author has employed another explaining model called LIME which will visually explain which features of the image contributing most for the prediction.

Dataset Details

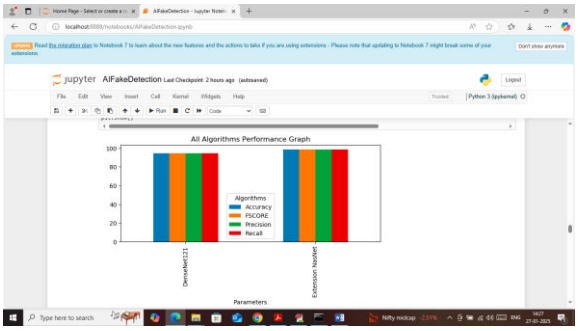
To train and test above algorithm performance we are utilizing REAL and FAKE images dataset which can be downloaded from below link

https://www.kaggle.com/datasets/gauravduttakiit/140k-real-and-fake-faces?select=train

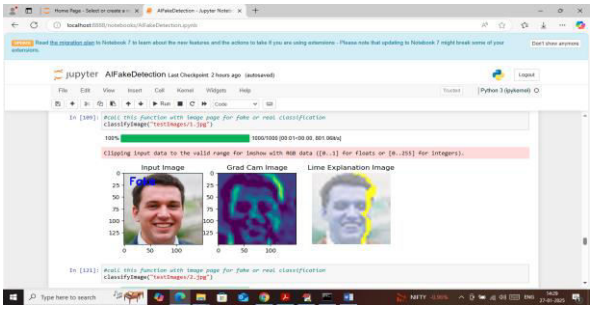
Above dataset will be processed and split into train and test where application will be using 80% dataset images for testing and 20% to calculate model prediction accuracy

Enhancement

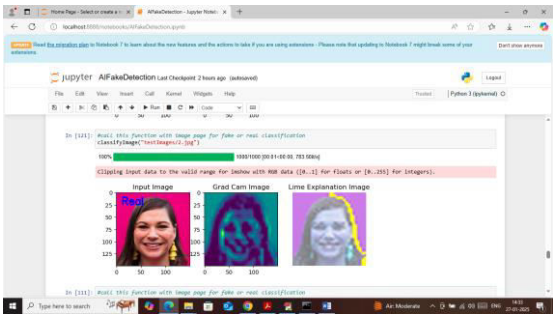
In propose work author has employed DenseNet121 algorithm whose performance can be enhance by employing other advance pre-trained model architectures such as VGG16 or 19. During implementation we have trained with experimented with various algorithms like VGG16, VGG19, XCEPTION, NASNET and many other algorithms but in all algorithms NASNET was giving high accuracy so we have used NASNET as the extension algorithm which is not used in any other previously implemented papers.



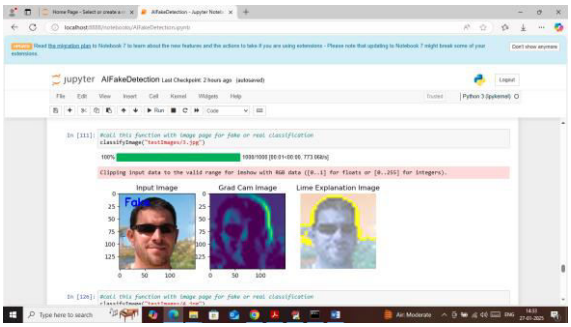
In above screen visualizing both algorithms performance where x-axis represents algorithm names and y-axis represent accuracy and other metrics in different colour bars and in both algorithms NASNET got high accuracy



In above screen calling 'Classify Image' function along with test image path and then function will return 3 images where fist image is the INPUT image which is marked with predicted labels as 'Fake or Real' in blue text and in above screen input image is predicted as Fake. In 2nd image showing GRAD CAM features mapping image where the regions with dark colour are the features contributing most for prediction. In 3rd image showing LIME explain features in yellow colour which says those are the features contributing most for prediction. In above screen can see both GRAD-CAM and LIME showing same regions features which are contributing most for prediction.



In above screen testing another images which is predicted as REAL



Above image predicted as Fake and showing along with GRAD-CAM and LIME explanation

V. CONCLUSION

The proposed project presents an effective and reliable framework for detecting AI-generated images using advanced deep learning techniques combined with Explainable Artificial Intelligence (XAI). By leveraging the DenseNet121 model and enhancing performance with NASNet, the system achieves high accuracy in distinguishing

between real and synthetic images. The integration of Grad-CAM and LIME ensures transparency by providing clear visual explanations of the model's decision-making process, addressing the common challenge of black-box behavior in deep learning systems. The experimental results demonstrate that the system is robust, accurate, and capable of handling complex image variations. This makes it highly suitable for real-world applications such as digital forensics, cybersecurity, and media authentication. Overall, the project successfully combines accuracy with interpretability, contributing to building trust in AI-based image detection systems and helping combat the growing threat of synthetic media.

REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[3] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708.

[4] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8697–8710.

[5] R. R. Selvaraju *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 618–626.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[7] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, pp. 2672–2680, 2014.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[9] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1251–1258.

[10] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[12] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[13] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4401–4410.

[14] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[15] N. Carlini *et al.*, "Evaluating and testing unintended memorization in neural networks," in *Proc. USENIX Security Symposium*, 2019, pp. 267–284.

[16] H. Wang *et al.*, "CNN-generated images are surprisingly easy to spot… for now," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8695–8704.

[17] A. Rossler *et al.*, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2019, pp. 1–11.

[18] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training," in *Proc. Int. Conf. Machine Learning (ICML)*, 2015, pp. 448–456.

[20] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Machine Learning (ICML)*, 2010, pp. 807–814.

[21] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[22] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2021.

[23] S. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Machine Learning (ICML)*, 2019, pp. 6105–6114.

[24] H. Zhang *et al.*, "Mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2018.

[25] G. Bradski, "The OpenCV library," *Dr. Dobb's Journal of Software Tools*, 2000.