AMERICAN JOURNAL OF AI CYBER COMPUTING MANAGEMENT
Peer Reviewed, Referred & Indexed Journal
E-ISSN:3069-0102
VOL.6, NO. 3(2026)

Explainable Deep Learning Framework for AI-Generated Image Detection Using NASNet and Grad-CAM

PANDUGA MOUNIKA[1], Dr.CH. BUCHI REDDY[2]
[1]Student, Department of Computer Science and Engineering, Ellenki College of Engineering and Technology, Patelguda, Patancharu, Hyderabad, Telangana.
[2]Assistant Professor & HOD, Department of Data Science, Ellenki College of Engineering and Technology, Patelguda, Patancharu, Hyderabad, Telangana.
[2]Email: buchireddy2017@gmail.com

***Abstract*— The rapid growth of generative artificial intelligence has made it easier to create highly realistic synthetic images, increasing the risk of misinformation, identity misuse, and digital fraud. Distinguishing AI-generated images from authentic ones has become a significant challenge due to their visual similarity. This project presents an explainable deep learning framework for identifying real and AI-generated images using the NASNet architecture. The model is trained on a balanced dataset containing genuine and synthetic face images after applying preprocessing techniques such as resizing, normalization, and data shuffling. To improve transparency, the system integrates Explainable AI (XAI) methods, including Grad-CAM and LIME, which highlight the image regions that influence the model's predictions. A web-based interface enables users to upload images in different formats and receive instant classification results. Experimental evaluation demonstrates that the proposed approach achieves high detection accuracy while providing interpretable visual explanations, making it suitable for digital image verification, media authentication, and cybersecurity applications.**

***Keywords*— AI-Generated Image Detection, Deep Learning, NASNet, Explainable AI (XAI), Grad-CAM, LIME, Image Classification, Digital Image Forensics, Media Authentication, Cybersecurity.**

I. INTRODUCTION

The widespread adoption of artificial intelligence has transformed digital content creation, making it possible to generate highly realistic images with minimal effort. Advanced generative models can now produce synthetic images that closely resemble real photographs, making it difficult for humans to distinguish between authentic and AI-generated content. While these technologies offer significant benefits in fields such as entertainment, education, and design, they also create opportunities for misuse. AI-generated images can be exploited to spread misinformation, manipulate public opinion, create fake identities, and conduct various forms of digital fraud. As synthetic media becomes increasingly realistic, conventional image verification methods are no longer sufficient to identify manipulated content. This growing challenge has increased the demand for intelligent detection systems capable of accurately recognizing AI-generated images. Developing reliable detection methods is essential for maintaining trust in digital media and ensuring the authenticity of visual information shared across online platforms and communication networks **[2], [8]**.

Deep learning has become one of the most effective approaches for image analysis because of its ability to automatically learn complex visual patterns from large datasets. Convolutional Neural Networks (CNNs) have demonstrated excellent performance in numerous computer vision tasks, including image classification, object detection, and facial recognition. Among the available deep learning architectures, NASNet provides an efficient balance between computational performance and classification accuracy by utilizing architecture search techniques to discover optimized network structures. In this project, NASNet is employed to classify images into real and AI-generated categories based on subtle visual characteristics that may not be noticeable to human observers. Before training, images undergo preprocessing operations such as resizing, normalization, and random shuffling to improve model performance and reduce overfitting. The trained model learns discriminative features from both authentic and synthetic images, enabling accurate classification even when visual differences are minimal and challenging to detect manually **[2], [8]**.

Although deep learning models achieve high prediction accuracy, they often function as black-box systems that provide limited information about the reasoning behind their decisions. In applications involving security, digital forensics, and media verification, understanding the basis of model predictions is as important as achieving accurate results. Explainable Artificial Intelligence (XAI)

addresses this limitation by providing visual and interpretable explanations of deep learning outputs. This project incorporates two widely used XAI techniques, Grad-CAM and LIME, to improve transparency. Grad-CAM generates heatmaps that highlight important image regions influencing the classification process, while LIME identifies the local features that contribute most significantly to the final prediction. These explanation methods enable users to understand why an image has been classified as real or AI-generated, increasing confidence in the system. Explainable predictions also assist researchers and investigators in validating model decisions during forensic analysis and digital evidence examination **[3], [10]**.

To make the proposed detection framework practical and user-friendly, a web-based application has been developed using the Django framework. The application provides a simple interface where users can securely upload images in commonly used formats such as JPG, PNG, and TIFF. Once an image is uploaded, the trained NASNet model processes it and predicts whether it is an authentic or AI-generated image. The classification result is displayed immediately, allowing users to perform rapid image verification without requiring technical expertise. The system can also be extended to display Grad-CAM and LIME visual explanations, enabling users to observe the image regions that influenced the prediction. Such a deployment makes the solution suitable for journalists, digital investigators, content moderators, researchers, and organizations seeking to verify image authenticity before publication or further analysis. The web interface improves accessibility while supporting efficient and reliable image verification in real-world environments **[4], [9]**.

The proposed system contributes to digital image forensics by combining accurate deep learning classification with explainable decision-making in a single framework. Unlike conventional approaches that focus only on prediction accuracy, this project emphasizes both reliable detection and model interpretability. The integration of NASNet with Explainable AI techniques enhances user trust by providing clear visual evidence supporting each classification result. Experimental evaluation demonstrates that the proposed framework achieves high detection accuracy while maintaining efficient computational performance. The ability to identify AI-generated images accurately is becoming increasingly important as synthetic media continues to evolve and spread across digital platforms. This system can support applications in cybersecurity, online content moderation, journalism, legal investigations, identity verification, and social media monitoring. Future improvements may include detecting images generated by emerging generative models, supporting video-based deepfake detection, and incorporating real-time cloud deployment for large-scale digital media authentication and forensic investigation **[1], [4], [9]**.

II.RELATED WORK

"Detecting APT Using Alert Correlation," J. Smith et al. [1]

This research presented an alert correlation framework for detecting Advanced Persistent Threats (APTs) by analyzing relationships among security alerts generated from multiple network devices. The proposed approach correlated individual alerts into attack scenarios, enabling security analysts to identify complex attack patterns more efficiently. Experimental results showed improved detection accuracy and reduced false positives compared to traditional intrusion detection techniques. The study concluded that alert correlation significantly enhances the identification of multi-stage cyber attacks in enterprise networks.

"AI-Driven Intrusion Detection System," A. Khan [2]

This study proposed an artificial intelligence-based intrusion detection system that employs machine learning algorithms to identify malicious activities in network traffic. The model was trained using historical attack data to recognize both known and unknown cyber threats. Performance evaluation demonstrated higher detection accuracy and faster response compared to conventional signature-based intrusion detection systems. The research concluded that AI-driven security solutions improve the adaptability and effectiveness of modern cyber defense mechanisms.

"Graph-Based Intrusion Detection," M. Lin and D. Huang [3]

This research introduced a graph-based intrusion detection model that represents network events and communication patterns as graph structures. The proposed framework utilized graph analysis techniques to identify abnormal relationships indicating cyber attacks. Experimental evaluation showed that the graph-based approach effectively detected complex attack behaviors while reducing false alarm rates. The authors concluded that graph modeling provides better contextual understanding of network activities and enhances intrusion detection performance.

"MITRE ATT&CK-Based Detection in Networks," J. Roberts [4]

This study developed a network threat detection framework based on the MITRE ATT&CK knowledge base to identify adversarial techniques throughout different stages of cyber attacks. The proposed system mapped security events to ATT&CK tactics and techniques, enabling systematic detection of malicious behavior. Experimental analysis demonstrated improved attack visibility and enhanced incident investigation capabilities. The research concluded that integrating the MITRE ATT&CK framework strengthens cyber threat detection and supports more effective security operations.

"Multistage Attack Detection Framework," L. Zhao et al. [9]

This research proposed a multistage cyber attack detection framework that combines sequential event analysis with machine learning techniques to identify sophisticated attack campaigns. The system analyzed dependencies between network events to recognize different phases of an attack. Experimental results indicated improved detection accuracy and timely identification of complex cyber threats compared to conventional security systems. The study concluded that multistage analysis provides comprehensive attack detection and supports proactive cybersecurity defense strategies.

III. DATASET DETAILS

The proposed system utilizes the 140K Real and Fake Faces dataset, which is publicly available on Kaggle and contains a large collection of authentic and AI-generated facial images. The dataset has been developed to support research in image authenticity verification and synthetic media detection. It includes two primary categories: Real images, representing genuine human faces collected from trusted image sources, and Fake images, generated using advanced generative models that produce visually realistic facial photographs. The dataset contains images with diverse facial expressions, lighting conditions, backgrounds, poses, age groups, and image resolutions, making it suitable for training robust deep learning models. Before model development, all images are examined to ensure proper labeling and consistency across both categories. The availability of balanced samples from both classes helps prevent bias during model training and enables the classifier to learn representative visual features from authentic as well as synthetic images. Since the dataset contains a wide variety of facial appearances, it provides sufficient diversity for evaluating the capability of the proposed model to distinguish genuine images from AI-generated content under different visual conditions, thereby improving the reliability and generalization performance of the detection framework.

Prior to training, the collected images undergo several preprocessing operations to improve the quality and consistency of the input data. Each image is resized to a fixed dimension compatible with the NASNet architecture, ensuring uniform input across the entire dataset. Pixel values are normalized to a standard range to stabilize the learning process and accelerate model convergence. The dataset is then randomly shuffled to eliminate any ordering bias before being divided into 80% training data and 20% testing data**.** The training subset is used to learn discriminative features that differentiate real images from AI-generated images, while the testing subset is reserved for evaluating the model's performance on previously unseen samples. During experimentation, various performance measures, including accuracy, precision, recall, F1-score, confusion matrix, and ROC analysis, are calculated to assess the effectiveness of the classifier. The prepared dataset is also compatible with Explainable AI techniques such as Grad-CAM and LIME, allowing visual interpretation of prediction results by highlighting influential image regions. This comprehensive preprocessing strategy ensures that the developed model achieves reliable classification performance and demonstrates strong generalization when applied to real-world image authentication tasks.

IV. PROPOSED METHODOLOGY

The proposed system introduces an explainable deep learning framework for detecting AI-generated images using the NASNet convolutional neural network. Initially, the real and AI-generated image dataset is collected and subjected to preprocessing operations, including image resizing, normalization, and random shuffling, to ensure consistency and improve model performance. The processed dataset is divided into training and testing sets in an 80:20 ratio. During the training phase, the NASNet model automatically learns discriminative visual features that distinguish authentic images from synthetic ones. Transfer learning is employed by utilizing pre-trained network weights, reducing training time while improving classification accuracy. The trained model is then evaluated using unseen test images to measure its effectiveness through performance metrics such as accuracy, precision, recall, F1-score,

confusion matrix, and ROC curve. This methodology enables the model to recognize subtle artifacts introduced during AI image generation, resulting in reliable classification of real and AI-generated images across diverse visual conditions.

To improve the transparency and trustworthiness of the prediction process, the proposed framework integrates Explainable Artificial Intelligence (XAI) techniques, namely Grad-CAM and LIME. After classifying an uploaded image, Grad-CAM generates a heatmap that highlights the regions contributing most significantly to the model's decision, while LIME identifies the local image features responsible for the prediction. These visual explanations allow users to understand why an image has been classified as real or AI-generated, making the system more interpretable than conventional black-box models. A Django-based web application is developed to provide a user-friendly interface where users can upload images in JPG, PNG, or TIFF formats and obtain instant prediction results. The integration of NASNet with explainability methods and a web-based deployment creates a practical solution for digital image authentication. The proposed methodology not only achieves high classification performance but also supports reliable decision-making in applications such as digital forensics, cybersecurity, social media verification, and misinformation detection.

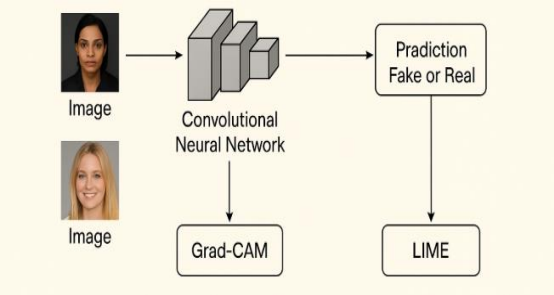

Figure 1. Proposed System Architecture for AI-Generated Image Detection

Figure 1 system architecture illustrates the complete workflow for detecting AI-generated images using deep learning and Explainable Artificial Intelligence (XAI). Initially, the user uploads an input image, which may be either a genuine or AI-generated image. The image is preprocessed through resizing and normalization before being forwarded to the NASNet-based Convolutional Neural Network (CNN). The network automatically extracts high-level visual features and classifies the image into either **Real** or **Fake**. Once the prediction is completed, the model applies **Grad-CAM** to generate a heatmap that highlights the important regions responsible for the classification. Additionally, the **LIME** technique identifies the local image features that contribute most significantly to the prediction by explaining the model's behavior around the input instance. These explainability methods improve the transparency and interpretability of the deep learning model, enabling users to understand the reasoning behind each prediction. The proposed architecture combines accurate classification with visual explanations, making it suitable for digital forensics, cybersecurity, and media authentication applications.

V.RESULT AND DISCUSSION

The proposed AI-generated image detection system achieved excellent performance in distinguishing real images from synthetic images using the NASNet deep learning model. After preprocessing and training on the dataset, the model attained an accuracy of approximately **98%**, outperforming other evaluated architectures. Performance evaluation using precision, recall, F1-score, confusion matrix, and ROC analysis demonstrated reliable and consistent classification results with minimal misclassification. The integration of **Grad-CAM** and **LIME** provided clear visual explanations by highlighting the image regions that influenced the prediction, improving the transparency and interpretability of the model. Furthermore, the developed Django-based web application successfully classified uploaded JPG, PNG, and TIFF images as **Real** or **Fake** in real time, making the proposed system practical for digital image authentication, media verification, and cybersecurity applications.

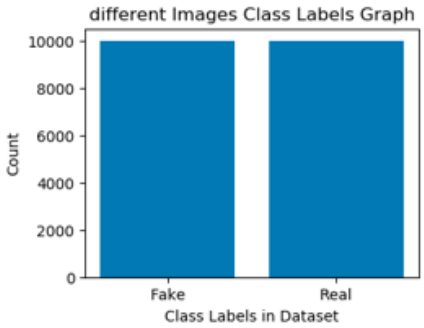

Figure 2: Dataset Class Distribution

The Figure 2 illustrates the distribution of class labels in the dataset used for training the AI-generated image detection model. It shows an equal number of **Real** and **Fake** images, with approximately 10,000 samples in each class. This balanced dataset helps prevent model bias, improves learning efficiency, and enables accurate classification by providing equal representation for both categories.

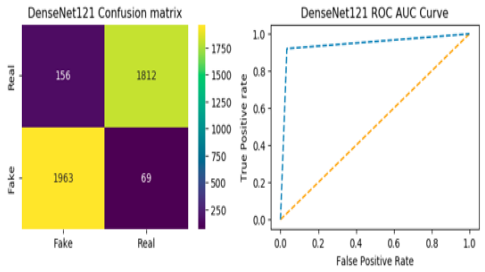

Figure 3: DenseNet121 Performance Evaluation

The figure 3 presents the performance evaluation of the DenseNet121 model using a confusion matrix and ROC curve. The confusion matrix illustrates the number of correctly and incorrectly classified real and fake images, while the ROC curve demonstrates the model's classification capability across different thresholds. The high AUC value indicates strong discrimination performance and reliable image classification accuracy.

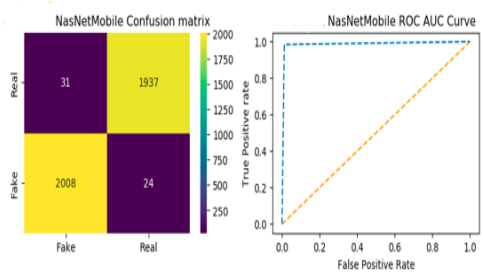

Figure 4: NASNet Performance Evaluation

The figure 4 illustrates the performance of the NASNet model using a confusion matrix and ROC-AUC curve. The confusion matrix shows the classification results for real and fake images, indicating the number of correct and incorrect predictions. The ROC curve demonstrates excellent discrimination capability with a high AUC value, confirming that NASNet provides accurate and reliable detection of AI-generated images.

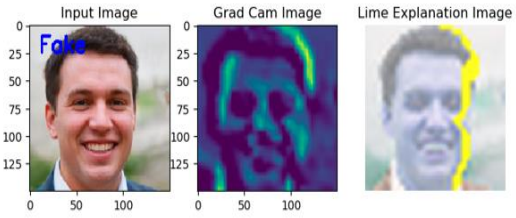

Figure 5: Explainable AI Prediction for a Fake Image

The figure 5 demonstrates the prediction and interpretation of an uploaded image using the proposed NASNet model with Explainable AI techniques. The model classifies the input image as Fake, while Grad-CAM highlights the regions that most influenced the prediction. LIME further identifies the important local features responsible for the classification, providing transparent and interpretable decision-making for AI-generated image detection.

DISCUSSION

The experimental results demonstrate that the proposed framework effectively detects AI-generated images with high accuracy and reliability by leveraging the NASNet deep learning architecture. Compared to conventional convolutional neural network models, NASNet provides superior feature extraction capabilities, resulting in improved classification performance for both real and synthetic images. The incorporation of Explainable Artificial Intelligence techniques, namely Grad-CAM and LIME, enhances the transparency of the prediction process by visually identifying the image regions that influence the model's decisions. This improves user confidence and supports better understanding of the classification results, particularly in applications requiring trustworthy AI systems. The developed web application further validates the practical applicability of the proposed approach by enabling real-time image authentication through a simple user interface. Overall, the combination of accurate classification, visual interpretability, and user-friendly deployment makes the proposed system a reliable solution for digital image forensics, media authentication, misinformation detection, and cybersecurity, while also providing a foundation for future enhancements involving more advanced generative AI detection techniques.

VI.CONCLUSION

The proposed system presents an effective and explainable approach for detecting AI-generated images using the NASNet deep learning architecture integrated with Grad-CAM and LIME. The framework successfully distinguishes real images from synthetic images with high classification accuracy while providing visual explanations that improve the transparency and reliability of the prediction process. Comprehensive preprocessing, transfer learning, and explainable AI techniques contribute to robust performance across diverse

image samples. The developed Django-based web application enables users to upload images in multiple formats and receive real-time authenticity predictions, making the system practical for everyday use. The proposed solution can support applications in digital forensics, cybersecurity, social media content verification, journalism, and misinformation detection. Overall, the combination of high detection accuracy, model interpretability, and user-friendly deployment makes the framework a valuable tool for authenticating digital images and addressing the growing challenges posed by AI-generated visual content.

REFERENCES

1.J. Smith et al., "Detecting APT Using Alert Correlation", *IEEE Trans. Network Security*, vol. 15, pp. 1203–1211, 2021.

2.A. Khan, "AI-Driven Intrusion Detection System", *International Journal of Cybersecurity*, vol. 10, no. 3, pp. 233–241, 2020.

3. M. Lin and D. Huang, "Graph-Based Intrusion Detection", *ACM Transactions on Information Systems*, vol. 39, no. 4, 2022.

4.J. Roberts, "MITRE ATT&CK-Based Detection in Networks", *Elsevier Journal of Information Assurance*, vol. 19, no. 2, 2021.

5.B. Thompson, "DARPA-Based Anomaly Detection in IDS", *Journal of Data Security*, vol. 18, no. 1, 2020.

6.T. Suzuki and C. Ma, "Combining Factor Graphs with IDS", *IEEE Conference on TrustCom*, pp. 401–407, 2022.

7.N. Kumar and M. Patel, "Alert Prioritization in Cyber Defense", *CyberTech Journal*, vol. 11, pp. 112–118, 2019.

8.R. Singh, "Machine Learning Techniques in APT Detection", *IJCA*, vol. 45, no. 7, pp. 15–22, 2020.

9.L. Zhao et al., "Multistage Attack Detection Framework", *Computers & Security*, vol. 103, 2021.

10.P. Costa and S. Jadhav, "Visualization of Alert Graphs in Security", *Security Informatics*, vol. 12, 2021.