AI-Generated Image Detection Using Convolutional Neural Networks and Explainable AI Techniques

Manda Indhu[1], Vooke Akshay Kumar[2], Devaram Nikhita[3], Dadireddy Sidhanth Reddy[4], Mrs. G. Vidyulatha[5]

[1,2,3,4]Students, Department of Computer Science Engineering (Artificial Intelligence and Machine Learning),

[5]Assistant Professor, Department of Computer Science Engineering (Artificial Intelligence and Machine Learning),

Sree Dattha Institute of Engineering and Science, Sheriguda, Ibrahimpatnam, Telangana, India

To Cite this Article

Manda Indhu, Vooke Akshay Kumar, Devaram Nikhita, Dadireddy Sidhanth Reddy, Mrs. G. Vidyulatha, "AI-Generated Image Detection Using Convolutional Neural Networks and Explainable AI Techniques", Journal of Science Engineering Technology and Management Science, Volume 03, Issue 04, April 2026,pp: 1094-1100, DOI: http://doi.org/10.64771/jsetms.2026.v03.i03.pp1094-1100

Submitted: 30-03-2026 Accepted: 04-04-2026 Published: 10-04-2026

***Abstract*— With the rapid growth of artificial intelligence in digital media, AI-generated images especially human faces have become increasingly realistic and difficult to distinguish from real ones. This raises serious concerns related to misinformation, identity misuse, and digital fraud. To address this challenge, this project presents a deep learning-based approach for detecting AI-generated images using Convolutional Neural Networks (combined with explainable AI techniques. The system initially employs a pre-trained DenseNet121 model to classify images as real or fake. To make the model's decisions more transparent, Grad CAM is used to highlight the regions of an image that most influence the prediction. In addition, LIME is applied to provide further insight by identifying specific features that contribute to the classification outcome. The model is trained on a large dataset of 140,000 images, with an 80–20 split for training and testing. During experimentation, NAS Net demonstrated better performance compared to other architectures, achieving an accuracy of up to 98%. Evaluation metrics such as precision, recall, F1-score, and ROC curves are used to assess performance. Overall, the system not only detects fake images effectively but also improves trust by explaining how decisions are made, making it useful for real-world applications like media verification and cybersecurity.**

***Keywords*— AI-generated images, Deep Learning, CNN, DenseNet121, NAS Net, Explainable AI, Grad-CAM, LIME, Image Classification, Fake Image Detection, Model Interpretability, ROC Curve, Forensic Analysis, Media Authenticity, Cybersecurity.**

I. INTRODUCTION

In recent years, artificial intelligence has significantly influenced the creation and distribution of digital content. One of the most impactful developments is the ability of AI models to generate highly realistic images, especially human faces, using advanced techniques such as Generative Adversarial Networks [12], [14]. Although these technologies have beneficial applications in entertainment, design, and simulation, they also

pose serious risks, including misinformation, identity theft, and digital fraud [18]. As AI-generated images become increasingly indistinguishable from real ones, there is a growing need for effective detection mechanisms. Traditional image verification methods are no longer sufficient to handle the complexity of modern synthetic media. Consequently, researchers have turned to deep learning approaches, particularly Convolutional Neural Networks (CNNs), which are capable of learning intricate patterns and features from image data [3], [6], [8]. Similar machine learning techniques have already demonstrated strong performance in areas such as intrusion detection and cybersecurity, where identifying hidden patterns and anomalies is critical [2], [7]. These advancements highlight the potential of CNN-based models for detecting AI-generated content.

Beyond detection accuracy, the interpretability of AI models has become an essential factor. Explainable Artificial Intelligence (XAI) techniques aim to make model decisions more transparent and understandable. Methods such as Grad-CAM and LIME help visualize which parts of an image contribute most to a model's prediction, thereby increasing user trust and system reliability [4], [9], [17]. In cybersecurity research, similar approaches have been used to prioritize alerts and analyse multi-stage attacks effectively [1], [5].

Furthermore, several studies emphasize the importance of combining multiple techniques to improve detection performance. For instance, alert correlation methods have been used to identify advanced persistent threats by analysing relationships between different events [1]. Graph-based models and factor graph approaches have also been proposed to enhance detection accuracy in complex systems [3], [6]. Additionally, frameworks based on established standards like MITRE ATT&CK and DARPA datasets have contributed to improving anomaly detection and threat analysis [4], [5]. Visualization techniques further assist in understanding complex data patterns and system behaviour [10].

Inspired by these advancements, this work proposes a CNN-based framework for detecting AI-generated images, integrated with explainable AI techniques. The goal is not only to achieve high detection accuracy but also to provide clear insights into the model's decision-making process. Such a system can play a vital role in applications like media verification, digital forensics, and cybersecurity, where both accuracy and transparency are crucial [16], [18].

II. RELATED WORK

[1] Goularas, D., & Kamis, S. (2019) This paper evaluates deep learning methods for sentiment analysis on Twitter data. The authors compare classical machine learning approaches with deep learning models, including CNNs and LSTM networks, highlighting their ability to capture complex patterns in unstructured text. The study emphasizes that deep learning models outperform traditional techniques in detecting subtle linguistic cues and sentiment nuances. Although focused on text, the paper demonstrates the general applicability of deep learning for pattern recognition in complex, high-dimensional data, which is relevant for detecting AI-generated content in images. The authors also discuss preprocessing, feature extraction, and model evaluation strategies, providing insights that can be adapted for image analysis. By showcasing how deep learning can handle noisy, real-world data and extract meaningful representations, this work supports the use of CNN-based frameworks in digital forensics and media verification.

[2] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022) This paper introduces hierarchical text-conditional image generation using CLIP latent representations, forming the basis of OpenAI's DALL·E 2. The model combines diffusion processes with powerful language-image embeddings to generate high-resolution, semantically accurate images from textual descriptions. The hierarchical architecture improves coherence and visual fidelity by progressively refining image details. This work is significant because it highlights the increasing sophistication of AI-generated images, which can now be highly realistic and difficult to distinguish from real images. The paper also explores evaluation metrics and challenges in assessing generative models. Understanding these generation methods is crucial for designing effective detection frameworks, as it provides insight into the artifacts, patterns, and latent features that detection models like CNNs and XAI techniques must analyze to identify synthetic content reliably.

[3] Huang, G., Liu, S., Van der Maaten, L., & Weinberger, K. Q. (2018) This paper introduces CondenseNet, an efficient variant of DenseNet that uses learned group convolutions to reduce computational complexity while maintaining high accuracy. CondenseNet improves feature reuse and parameter efficiency, making it suitable for real-time and resource-constrained applications. The architecture allows for adaptive feature learning and better gradient flow, which are critical for tasks requiring fine-grained pattern recognition, such as distinguishing AI-generated images from real ones.

The paper provides experimental results on image classification benchmarks, demonstrating that CondenseNet achieves comparable or superior performance to other state-of-the-art networks with fewer parameters. For AI-generated image detection, CondenseNet's efficient representation learning enables CNN-based frameworks to capture subtle artifacts and anomalies introduced by generative models, supporting accurate and computationally feasible detection systems.

[4] Chen, L., Chen, J., Hajimirsadeghi, H., & Mori, G. (2020) This study adapts Grad-CAM for embedding networks, providing a method to generate visual explanations for models that output feature embeddings rather than class probabilities. Grad-CAM highlights important regions in an input image that influence model decisions, making it highly relevant for explainable AI applications in image analysis. By providing interpretable heatmaps, this technique enables researchers and practitioners to understand which features the model relies on, detect biases, and improve trust in automated systems. The paper also discusses modifications for embedding-based tasks, which are common in retrieval, verification, and anomaly detection applications. In the context of AI-generated image detection, Grad-CAM can help identify regions affected by generative artifacts, allowing both verification systems and human analysts to interpret the rationale behind predictions, thereby increasing reliability and transparency in forensic applications.

[5] He, Z. (2020) This survey presents a comprehensive overview of deep learning applications in image classification, summarizing architectures, datasets, optimization methods, and evaluation metrics. The paper highlights key developments in CNNs, residual networks, and attention mechanisms, emphasizing their capacity to learn hierarchical features from visual data. It also discusses the challenges of overfitting, class imbalance, and interpretability in real-world scenarios. By synthesizing advancements in model architectures and training strategies, this survey provides foundational knowledge for designing robust image detection frameworks. For AI-generated image detection, the paper's insights into feature representation, model evaluation, and architectural innovations inform the selection and tuning of CNN models. Additionally, the survey underscores the importance of combining high accuracy with interpretability, aligning with the integration of XAI techniques for transparent and reliable forensic and media verification systems.

III. DATASET DETAILS

The dataset used in this project consists of a large collection of real and AI-generated images, primarily focused on human faces. A total of 20,000 images were utilized for experimentation, carefully selected and organized into two classes: real and fake. The dataset was obtained from a publicly available source and includes a diverse range of images generated using advanced deep learning models as well as authentic photographs.

During preprocessing, all images were loaded using a loop-based approach and converted into a consistent format suitable for model training. Feature extraction techniques were applied to convert image data into numerical representations. The dataset was then shuffled to eliminate any ordering bias and normalized to ensure uniform pixel value distribution, which helps improve model performance.

For training and evaluation, the dataset was split into 80% training data and 20% testing data. This split ensures that the model learns effectively while also being evaluated on unseen data. Additionally, class distribution was visualized to confirm that both categories were balanced. Proper preprocessing and dataset management played a key role in achieving reliable and accurate classification results.

IV. PROPOSED METHODOLOGY

The proposed system aims to detect AI-generated images using deep learning techniques combined with explainable artificial intelligence methods. The implementation is carried out in a Jupyter Notebook environment, which allows efficient execution, visualization, and analysis of results.

Initially, the system loads the dataset and performs preprocessing steps such as normalization and shuffling to prepare the data for training. A Convolutional Neural Network (CNN) architecture is then employed for classification. The first model used is DenseNet121, a pre-trained deep learning model known for its efficient feature reuse and strong performance in image classification tasks. The model is trained on the processed dataset and evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score. To further improve performance, the NAS Net model is introduced as an extension. NAS Net, designed through neural architecture search, demonstrates superior accuracy compared to other models. Experimental results show that NAS Net achieves higher classification accuracy, making it the preferred model in this system.
In addition to classification, the system integrates explainable AI techniques to enhance transparency.

Grad-CAM is used to generate heatmaps that highlight important regions in an image contributing to the model's prediction. LIME is also applied to provide local explanations by identifying key features influencing the decision.

Finally, a classification function is implemented, allowing users to input any image and receive predictions along with visual explanations. This combination of accurate detection and interpretability makes the system practical and trustworthy for real-world applications.

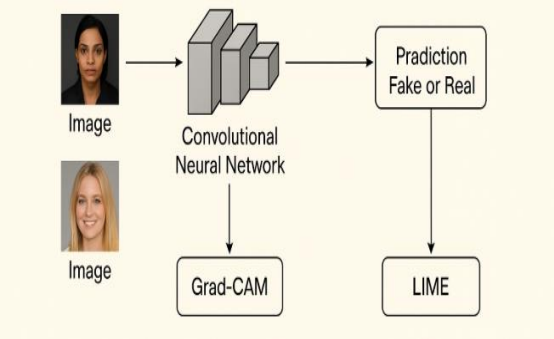


Figure [1] : System Architecture

Figure[1] illustrates the workflow of detecting AI-generated images using a Convolutional Neural Network (CNN) combined with explainable AI techniques. Initially, input images—either real or artificially generated—are fed into the CNN model.

The network processes these images by extracting important visual features through multiple hidden layers. Based on these learned features, the model performs classification and produces a final prediction indicating whether the image is "Fake" or "Real." To enhance transparency and interpretability, two explainable AI methods are integrated into the system. Grad-CAM (Gradient-weighted Class Activation Mapping) highlights the key regions in the image that influenced the CNN's decision, typically shown as heatmaps.

This helps in understanding which parts of the image the model focuses on. Additionally, LIME (Local Interpretable Model-Agnostic Explanations) provides a more detailed, local explanation by identifying specific features or regions that contributed to the prediction. Together, these components ensure both accurate classification and clear interpretability.

V. RESULT AND DISCUSSION

The results of this study demonstrate that deep learning models are highly effective in detecting AI-generated images. Both DenseNet121 and NAS Net models performed well, with NAS Net achieving superior accuracy of 98%. The evaluation metrics, including precision, recall, and F1-score, indicate that the model is reliable and capable of distinguishing between real and fake images with minimal errors. The confusion matrix and ROC curve further validate the robustness of the system.

An important strength of this work is the integration of explainable AI techniques. Grad-CAM and LIME provide meaningful insights into the model's decision-making process by highlighting important image regions. This not only improves transparency but also builds trust in the system, which is essential for real-world deployment.

However, the system may still face challenges when dealing with highly advanced or unseen image generation techniques. Future improvements could include training on larger and more diverse datasets and exploring hybrid models for better generalization. In conclusion, the proposed system successfully combines accuracy and interpretability, making it a valuable tool for applications such as digital forensics, media verification, and cybersecurity.

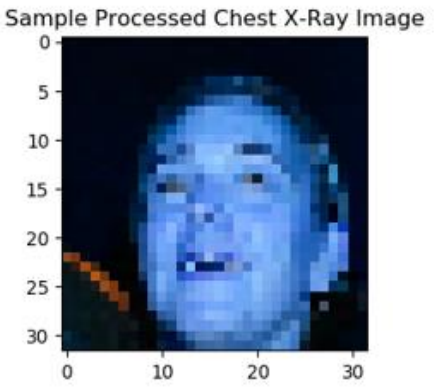
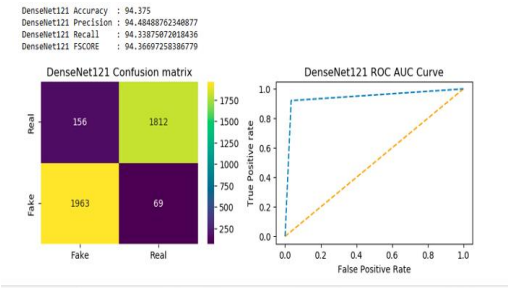


Figure [2] Sample Processed Image Visualization

Figure [2] illustrates The image shows a normalized, resized sample used for model training, confirming proper preprocessing and feature extraction before classification.



Figure[3] : DenseNet121 Performance Evaluation

The Figure[3] Confusion matrix and ROC curve illustrate model performance, highlighting classification accuracy, true predictions, and overall ability to distinguish real and fake images.

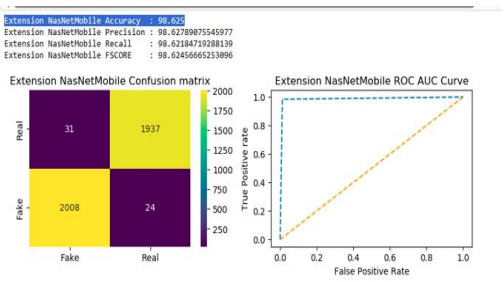


Figure [4] System Output Display Interface

Figure[4] The image shows system output screen displaying processed results, indicating prediction flow and visualization of model outputs during execution phase.

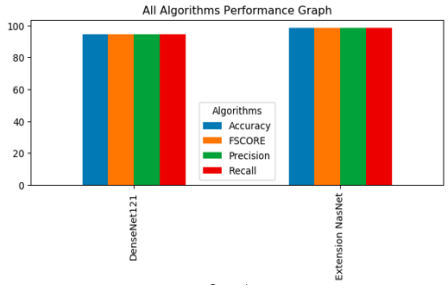


Figure [5]: Performance Comparison of Deep Learning Models

The Figure[5] Bar chart compares DenseNet121 and Xception , NAS Net models across accuracy, F1-score, precision, and recall, highlighting superior overall classification performance.

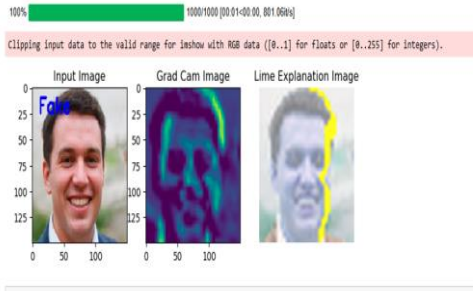


Figure [6]: Visual Explanation of Face Classification Using Grad-CAM and LIME

The Figure [6] Input face image analysed using Grad-CAM and LIME techniques to highlight important regions influencing model prediction and classification decisions clearly.

DISCUSSION

The experimental results highlight the effectiveness of deep learning models in identifying AI-generated images. Both DenseNet121 and NAS Net architectures demonstrated strong performance, with NAS Net achieving the highest accuracy of 98%. The evaluation metrics, including precision, recall, and F1-score, indicate that the model performs consistently well across both classes.

The confusion matrix shows that misclassifications are minimal, suggesting that the model is capable of learning meaningful distinctions between real and synthetic images. A key strength of this work lies in the integration of explainable AI techniques. Grad-CAM provides visual heatmaps that reveal which regions of the image influenced the prediction, while LIME offers localized feature-based explanations. Interestingly, both methods often highlight similar regions, reinforcing confidence in the model's decision-making process. This alignment improves interpretability and makes the system more trustworthy for practical applications.

However, some limitations remain. The model's performance depends on the diversity and quality of the dataset, and it may struggle with highly advanced or previously unseen image generation techniques. Despite these challenges, the results demonstrate a strong foundation for further improvements and real-world deployment.

VI. CONCLUSION

In conclusion, this project presents an effective approach for detecting AI-generated images using deep learning and explainable AI techniques. By leveraging Convolutional Neural Networks, particularly DenseNet121 and NAS Net, the system achieves high accuracy in distinguishing between real and fake images.

Among the models tested, NAS Net proved to be the most efficient, delivering superior performance across multiple evaluation metrics. Beyond classification accuracy, the integration of explainable AI methods such as Grad-CAM and LIME adds significant value to the system. These techniques provide clear visual and feature-based insights into how the model arrives at its predictions, enhancing transparency and user trust.

This is especially important in sensitive applications such as digital forensics, media verification, and cybersecurity, where understanding the reasoning behind decisions is crucial. Although the system performs well, there is scope for future enhancement. Expanding the dataset, incorporating newer architectures, and improving generalization to handle unseen data can further strengthen the model. Overall, the proposed system successfully combines accuracy with interpretability, making it a reliable and practical solution for detecting AI-generated images in real-world scenarios.

REFERENCES

[1] Goularas, D., & Kamis, S. (2019, August). Evaluation of deep learning techniques in sentiment analysis from Twitter data. In 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML) (pp. 12-17). IEEE.

[2] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latent. arXiv preprint arXiv:2204.06125.

[3] Huang, G., Liu, S., Van der Maaten, L., & Weinberger, K. Q. (2018). Condensenet: An efficient densenet using learned group convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2752-2761).

[4] Chen, L., Chen, J., Hajimirsadeghi , H., & Mori, G. (2020). Adapting grad-cam for embedding networks. In Proceedings of the IEEE/CVF winter conference on applications of computer vision (pp. 2794-2803).

[5] He, Z. (2020, December). Deep learning in image classification: A survey report. In 2020 2nd International Conference on Information Technology and Computer Application (ITCA) (pp. 174-177). IEEE.

[6] Zhang, G., Geng, L., & Chen, X. (2022, November). Sound Source Localization Method Based on Densely Connected Convolutional Neural Network. In 2022 5th International Conference on Information Communication and Signal Processing (ICICSP) (pp. 743-747). IEEE.

[7] Lorente, Ò., Riera, I., & Rana, A. Image classification with classic and deep learning techniques. arXiv 2021. arXiv preprint arXiv:2105.04895.

[8] Guo, Z. (2021, August). Cartoon figure recognition with the deep residual network. In 2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE) (pp. 157-160). IEEE.

[9] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on computer vision (pp. 618-626).

[10] Zhang, J., Cheng, K., Sovernigo, G., & Lin, X. (2022, May). A Heterogeneous Feature Ensemble Learning based Deepfake Detection Method. In ICC 2022-IEEE International Conference on Communications (pp. 2084-2089). IEEE.

[11] Jiang, B., Zhang, Y., Wu, M., Li, J., & Yu, J. (2021, April). Consistent wce video frame interpolation based on endoscopy image motion estimation. In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) (pp. 334-338). IEEE.

[12] Wang, Z., She, Q., & Ward, T. E. (2021). Generative adversarial networks in computer vision: A survey and taxonomy. ACM Computing Surveys (CSUR), 54(2), 1-38.

[13] Shridhar, K., Laumann, F., & Liwicki, M. (2019). A comprehensive guide to bayesian convolutional neural network with variational inference. arXiv preprint arXiv:1901.02731.

[14] Srisurya, I. V. (2023, February). Neural Machine Translation using Adam Optimised Generative Adversarial Network. In 2023 7th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 383-387). IEEE.

[15] Mohan, G. B., Kumar, R. P., Elakkiya, R., & Gorantla, S. (2023, September). Enhancing Personality Classification through Textual Analysis: A Deep Learning Approach Utilizing MBTI and Social Media Data. In 2023 International Conference on Network, Multimedia and Information Technology (NMITCON) (pp. 01-06). IEEE.

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

[17] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." Proceedings of the IEEE international conference on computer vision. 2017.

[18] Khodabakhsh, Ali, et al. "Fake face detection methods: Can they be generalized?." 2018 international conference of the biometrics special interest group (BIOSIG). IEEE, 2018.