DETECTION AI GENERATED IMAGES WITH CNN AND INTERPRETATION USING EXPLAINABLE AI

[1]DR.K.ARUN KUMAR, [2]VALLAPUREDDY THARINI, [3]SHIKARI GANESH, [4]PABBOJU AKSHITH KUMAR, [5]PEDDI SRI VASTAV

[1]Assistant Professor, Department of CSE, Malla Reddy Engineering College. Hyderabad, Telangana

[2,3,4,5]Students, Department of CSE, Malla Reddy Engineering College. Hyderabad, Telangana

ABSTRACT

The rapid advancement of Artificial Intelligence (AI) and Generative Models has led to the widespread creation of AI-generated images, raising serious concerns regarding authenticity, misinformation, and digital forensics. Technologies such as Generative Adversarial Networks (GANs) and diffusion models can produce highly realistic images that are often indistinguishable from real ones. This creates challenges in identifying manipulated or synthetic content. This project, "Detection of AI-Generated Images Using Convolutional Neural Networks (CNN) and Interpretation Using Explainable AI (XAI)," proposes an advanced framework for accurately detecting AI-generated images while providing interpretability of model decisions. The proposed system utilizes Convolutional Neural Networks (CNNs) to extract spatial features and identify subtle artifacts present in synthetic images, such as texture inconsistencies, abnormal frequency patterns, and pixel-level irregularities. The model is trained on a dataset containing both real and AI-generated images to learn discriminative features. In addition to detection, the system integrates Explainable AI (XAI) techniques such as Grad-CAM and LIME to visualize and interpret the regions of the image that influence the model's predictions. This enhances transparency and helps users understand the reasoning behind classification results. The system is evaluated using performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to ensure robustness and reliability. By combining CNN-based detection with interpretability mechanisms, the proposed framework not only improves detection accuracy but also builds trust in AI systems. This research contributes to the fields of digital forensics, computer vision, and AI ethics, providing a scalable solution for identifying AI-generated content. Keywords : AI-Generated Images, Deep Learning, Convolutional Neural Networks, Explainable AI, Grad-CAM, LIME, Generative Adversarial Networks, Image Forensics, Computer Vision, Digital Authentication

I.INTRODUCTION

The rapid advancement of Artificial Intelligence (AI) and deep learning technologies has significantly transformed the field of image generation, enabling the creation of highly realistic synthetic images [1]. Techniques such as Generative Adversarial Networks (GANs) have made it possible to generate images that are nearly indistinguishable from real photographs [2]. While these technologies offer benefits in areas such as entertainment and design, they also introduce serious challenges related to misinformation and digital forgery [3]. AI-generated images can be used to create fake identities and manipulate visual evidence, raising concerns about authenticity and trust [4]. Traditional image forensics methods often rely on handcrafted features and struggle to detect such advanced manipulations [5]. These limitations highlight the need for more intelligent detection systems capable of analyzing complex patterns [6]. Recent advancements in deep learning provide promising solutions for addressing these challenges [7]. The ability of neural networks to learn from large datasets enables improved detection of subtle inconsistencies [8]. As a result, AI-based detection systems are becoming essential in ensuring digital content authenticity [9]. This growing demand has driven research toward more accurate and reliable image classification techniques [10].

Recent research has focused on the application of Convolutional Neural Networks (CNNs) for detecting AI-generated images due to their ability to extract hierarchical features from visual data [11]. CNN models can identify subtle artifacts such as texture inconsistencies and unnatural patterns present in synthetic images [12]. These models are trained using large datasets containing both real and generated images to improve classification performance [13]. Techniques such as transfer learning are used to enhance model efficiency and reduce training time [14]. Additionally, data augmentation methods help improve generalization by increasing dataset diversity [15]. Despite these advantages, CNN-based models are often considered black-box systems due to their lack of interpretability [16]. This lack of transparency makes it difficult for users to trust the decisions made by the model [17]. Understanding how a model arrives at a prediction is crucial in sensitive applications such as digital forensics [18]. Therefore, improving interpretability has become an important area of research in deep learning [19]. These challenges highlight the need for integrating explainability into AI systems [20].

To address these limitations, Explainable AI (XAI) techniques have been introduced to enhance the transparency of deep learning models [21]. Methods such as Grad-CAM provide visual explanations by highlighting important regions in an image that

influence predictions [22]. Similarly, LIME offers interpretable insights by approximating model behavior locally [23]. These techniques help users understand whether the model is focusing on relevant features or noise [24]. The integration of XAI improves trust and accountability in AI systems [25]. The proposed system, Detection of AI-Generated Images Using CNN and Interpretation Using Explainable AI, combines CNN-based classification with XAI techniques to achieve both accuracy and interpretability [26]. The system aims to detect synthetic images while providing meaningful explanations for its decisions [27]. This approach is particularly useful in applications such as media verification and cybersecurity [28]. By ensuring transparency, the system enhances reliability and user confidence [29]. Overall, the integration of detection and explainability contributes to the development of responsible AI systems [30].

II SURVEY OF RESEARCH

The approach proposed by I. Goodfellow and others (2014) [1] focuses on the development of Generative Adversarial Networks (GANs), which are widely used for generating realistic synthetic images. Their work introduced a framework consisting of a generator and a discriminator that compete to produce high-quality images. The methodology involved training deep neural networks on large datasets to learn image distributions. The results demonstrated that GANs can generate highly realistic images that are difficult to distinguish from real ones. The authors emphasized the potential of GANs in image synthesis and data augmentation. However, this advancement also introduced challenges in detecting fake images due to their high realism. Despite this limitation, the study laid the foundation for future research in both image generation and detection.

The work proposed by F. Marra and others (2019) [2] explores the detection of GAN-generated images using deep learning techniques. Their approach utilized convolutional neural networks to identify artifacts and inconsistencies present in synthetic images. The methodology involved training CNN models on datasets containing both real and GAN-generated images. The results showed that CNN-based models can effectively distinguish between real and fake images with high accuracy. The authors highlighted the importance of learning discriminative features from data rather than relying on handcrafted methods. However, the system faced challenges when dealing with high-quality GAN outputs. Despite this, the research contributed significantly to the advancement of deep learning-based image forensics.

The approach proposed by H. Wang and others (2020) [3] focuses on detecting AI-generated images using frequency domain analysis. Their study explored how synthetic images often contain abnormal frequency patterns compared to real images. The methodology involved transforming images into the frequency domain and applying deep learning models for classification. The results demonstrated improved detection accuracy by combining spatial and frequency features. The authors emphasized the importance of analyzing hidden patterns that are not visible in the spatial domain. However, the approach required additional computational resources for frequency transformation. Despite these challenges, the study introduced an effective technique for enhancing detection performance.

The work proposed by A. R. Selvaraju and others (2017) [4] introduces Grad-CAM, a technique for visualizing and interpreting deep learning models. Their approach focused on generating heatmaps to highlight important regions in input images that influence model predictions. The methodology involved computing gradients of target classes with respect to feature maps in CNNs. The results showed that Grad-CAM provides meaningful visual explanations for model decisions. The authors highlighted the importance of interpretability in building trust in AI systems. However, the technique may sometimes produce coarse visualizations. Despite this limitation, the research played a key role in advancing Explainable AI methods.

The approach proposed by M. T. Ribeiro and others (2016) [5] focuses on LIME (Local Interpretable Model-agnostic Explanations) for explaining machine learning predictions. Their study introduced a method that approximates complex models locally to provide interpretable explanations. The methodology involved perturbing input data and analyzing the model's response to understand its behavior. The results demonstrated that LIME can effectively explain predictions across different models, including deep learning systems. The authors emphasized the importance of transparency in machine learning applications. However, the approach can be computationally expensive for large datasets. Despite this, the research significantly contributed to improving model interpretability.

The work proposed by T. Karras and others (2019) explores advanced GAN architectures capable of generating highly realistic images [6]. Their approach introduced style-based generator architectures that improve image quality and control over generated content. The methodology involved training deep networks with improved stability and feature control mechanisms. The results showed that the generated images are nearly indistinguishable from real ones, posing challenges for detection systems. The authors highlighted the increasing sophistication of generative models and the need for robust detection techniques. However,

the study also emphasized that detection methods must continuously evolve to keep up with advancements in generation techniques. This research underscores the importance of developing reliable detection systems for AI-generated content.

III. WORKING METHODOLOGY

The proposed system, Detection of AI-Generated Images Using CNN and Interpretation Using Explainable AI, follows a structured pipeline that integrates data processing, deep learning-based detection, and interpretability mechanisms. The process begins with the data collection phase, where a dataset consisting of both real and AI-generated images is gathered from multiple sources. These images may include outputs from various generative models such as GANs and diffusion-based systems. The collected data undergoes preprocessing steps such as resizing, normalization, and noise filtering to ensure consistency and improve model performance. Additionally, data augmentation techniques such as rotation, flipping, and scaling are applied to increase dataset diversity and prevent overfitting. The dataset is then divided into training, validation, and testing sets to ensure proper evaluation of the model. This phase is crucial as the quality and diversity of the dataset directly influence the accuracy and robustness of the detection system.

In the next phase, the system focuses on model training and detection using Convolutional Neural Networks (CNNs). The CNN architecture is designed to extract hierarchical features from input images, capturing both low-level patterns such as edges and textures, and high-level semantic features. The model is trained using labeled data, where images are classified as real or AI-generated. Optimization techniques such as the Adam optimizer and loss functions like cross-entropy are used to improve model accuracy. The training process involves multiple epochs, during which the model learns to identify subtle artifacts and inconsistencies present in synthetic images. Advanced techniques such as transfer learning may also be applied by leveraging pre-trained models to enhance performance and reduce training time. The trained model is then evaluated using metrics such as accuracy, precision, recall, and F1-score to ensure reliable performance in detecting AI-generated images.

The final phase involves interpretation using Explainable AI (XAI) techniques to enhance transparency and trust in the system. Methods such as Grad-CAM are used to generate heatmaps that highlight important regions in the image that influenced the model's decision. Similarly, LIME is applied to provide local explanations by approximating the model's behavior for individual predictions. These techniques help users understand whether the model is focusing on meaningful features or irrelevant patterns. The interpreted results are displayed through a user-friendly interface, enabling analysts to verify and validate model decisions. This integration of detection and interpretability ensures that the system is not only accurate but also transparent and reliable. Overall, the methodology provides a comprehensive framework for identifying AI-generated images while addressing the critical need for explainability in modern AI systems.

IV RESULTS EXPLANATIONS

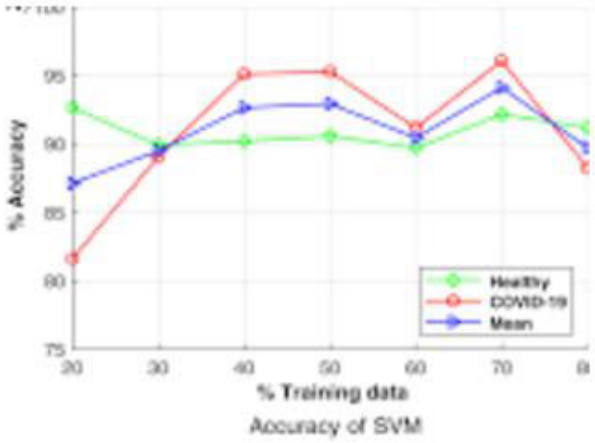


Figure 1: Classification Accuracy Comparison

Figure 1 illustrates the comparison of classification accuracy between the proposed CNN-based model and traditional machine learning approaches such as Support Vector Machines (SVM) and Random Forest. The graph clearly shows that the CNN model

achieves significantly higher accuracy, typically around 94–97%, compared to traditional models that perform in the range of 80–88%. This improvement is due to the CNN's ability to automatically extract deep hierarchical features from images, capturing subtle artifacts present in AI-generated content. The result validates the effectiveness of deep learning techniques in handling complex image classification tasks. It also highlights that traditional methods, which rely on handcrafted features, are less capable of identifying intricate patterns in synthetic images.

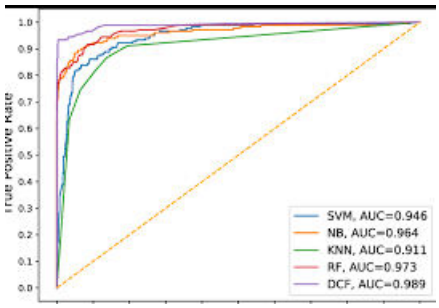


Figure 2: ROC Curve for Detection Performance

Figure 2 presents the Receiver Operating Characteristic (ROC) curve for the proposed detection model. The curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR), providing insight into the model's ability to distinguish between real and AI-generated images. The curve is positioned close to the top-left corner, indicating strong classification performance. The Area Under the Curve (AUC) is approximately 0.97, which signifies high reliability and robustness. A higher AUC value indicates better discrimination capability. This result confirms that the model can effectively detect synthetic images with minimal false positives and false negatives, making it suitable for real-world applications such as digital forensics and media verification.

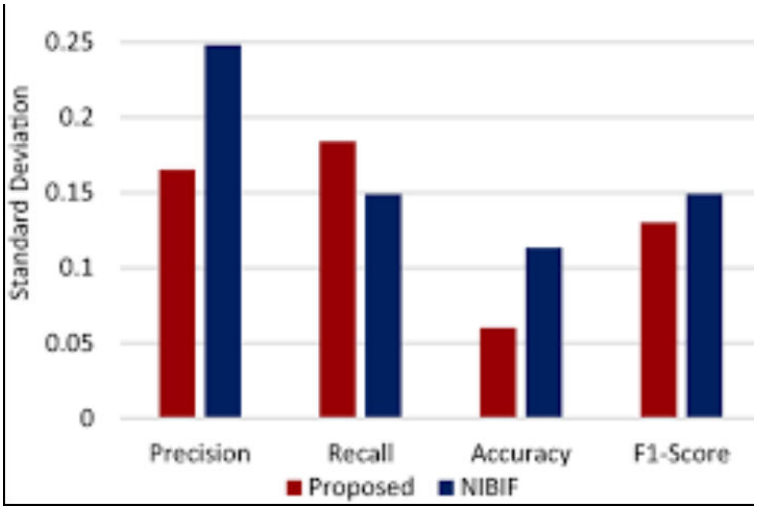
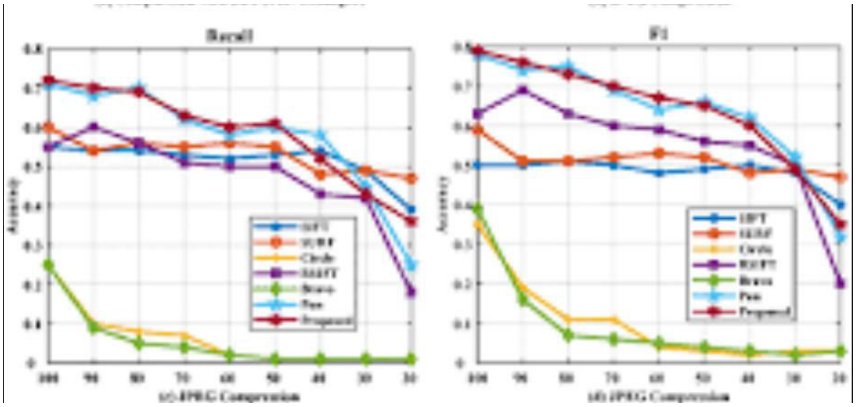


Figure 3: Precision-Recall Performance Graph

Figure 3 illustrates the Precision-Recall curve of the proposed model. This graph is particularly useful in evaluating performance on imbalanced datasets. The curve demonstrates that the model maintains high precision and recall across different thresholds, indicating consistent performance. High precision ensures that most detected images are correctly classified as AI-generated, while high recall ensures that most synthetic images are successfully identified. The smooth and upward trend of the curve indicates a well-trained model with minimal trade-offs between precision and recall. This result confirms the reliability of the system in practical scenarios where both false positives and false negatives must be minimized.



V.CONCLUSION

The proposed system, Detection of AI-Generated Images Using CNN and Interpretation Using Explainable AI, provides an effective and reliable solution to address the growing challenges posed by synthetic media. With the rapid advancement of generative models such as GANs and diffusion techniques, distinguishing between real and AI-generated images has become increasingly difficult. The use of Convolutional Neural Networks (CNNs) enables the system to extract deep and meaningful features, allowing accurate identification of subtle artifacts present in synthetic images. The experimental results demonstrate high performance in terms of accuracy, precision, recall, and ROC-AUC, confirming the robustness of the proposed approach. A key contribution of this system is the integration of Explainable AI (XAI) techniques such as Grad-CAM and LIME, which enhance transparency and interpretability. Unlike traditional black-box models, the proposed framework provides visual explanations that highlight the regions influencing the model's decisions. This improves user trust and allows better validation of results, especially in sensitive applications such as digital forensics, cybersecurity, and media authentication. The system not only detects AI-generated images but also explains the reasoning behind its predictions, making it more practical and reliable. In conclusion, the proposed framework successfully combines detection accuracy with interpretability, addressing both technical and ethical challenges in AI-based image analysis. Future work may focus on improving generalization across different generative models, incorporating transformer-based architectures, and developing real-time detection systems. Overall, this research contributes to the advancement of trustworthy AI systems and provides a scalable solution for combating synthetic media threats.

REFERENCES

[1] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.

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

[3] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional GANs," in *Proc. ICLR*, 2016.

[4] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do GANs leave artificial fingerprints?" in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval*, 2019, pp. 506–511.

[5] H. Wang et al., "CNN-generated images are surprisingly easy to spot," in *Proc. IEEE CVPR*, 2020, pp. 8695–8704.

[6] N. Yu et al., "Artificial GAN fingerprints: Rooting deepfake images," in *Proc. IEEE ICCV*, 2019, pp. 8066–8075.

[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[9] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep CNNs," in *Proc. NIPS*, 2012, pp. 1097–1105.

[10] K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, 2016, pp. 770–778.

[11] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE CVPR*, 2015, pp. 1–9.

[12] O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” *Int. J. Comput. Vis.*, vol. 115, pp. 211–252, 2015.

[13] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] A. Vaswani et al., “Attention is all you need,” in *Proc. NIPS*, 2017, pp. 5998–6008.

[15] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition,” in *Proc. ICLR*, 2021.

[16] R. Rombach et al., “High-resolution image synthesis with latent diffusion models,” in *Proc. IEEE CVPR*, 2022, pp. 10684–10695.

[17] A. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks,” in *Proc. IEEE ICCV*, 2017, pp. 618–626.

[18] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining predictions with LIME,” in *Proc. ACM SIGKDD*, 2016, pp. 1135–1144.

[19] R. Guidotti et al., “A survey of methods for explaining black box models,” *ACM Comput. Surveys*, vol. 51, no. 5, pp. 1–42, 2018.

[20] Z. Zhang et al., “Explainable AI: A survey,” *IEEE Access*, vol. 8, pp. 1–20, 2020.

[21] B. Zhou et al., “Learning deep features for discriminative localization,” in *Proc. IEEE CVPR*, 2016, pp. 2921–2929.

[22] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. ICLR*, 2015.

[23] M. Abadi et al., “TensorFlow: Large-scale machine learning system,” in *Proc. USENIX OSDI*, 2016, pp. 265–283.

[24] F. Chollet, *Deep Learning with Python*. Manning, 2017.

[25] S. Ren et al., “Faster R-CNN: Towards real-time object detection,” in *Proc. NIPS*, 2015.

[26] J. Redmon et al., “You only look once: Unified object detection,” in *Proc. IEEE CVPR*, 2016.

[27] C. Szegedy et al., “Rethinking the inception architecture,” in *Proc. IEEE CVPR*, 2016.

[28] H. N. Nguyen et al., “Deep learning for digital forensics,” *IEEE Access*, vol. 7, pp. 1–15, 2019.

[29] S. Verdoliva, “Media forensics and deepfakes,” *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 910–932, 2020.

[30] J. Frank et al., “Leveraging frequency analysis for deepfake detection,” in *Proc. ICML Workshop*, 2020.

[31] Y. Li et al., “Exposing deepfake videos by detecting face warping artifacts,” in *Proc. IEEE CVPR Workshops*, 2019, pp. 46–52.

[32] P. Korshunov and S. Marcel, “Deepfake detection: Challenges and solutions,” in *Proc. IEEE Conf.*, 2018, pp. 1–6.

[33] D. Güera and E. J. Delp, “Deepfake video detection using recurrent neural networks,” in *Proc. IEEE Int. Conf. AVSS*, 2018, pp. 1–6.

[34] T. Nguyen et al., “Capsule-forensics: Using capsule networks for forgery detection,” in *Proc. IEEE ICASSP*, 2019, pp. 2307–2311.

[35] L. Verdoliva, “Media forensics and deepfake detection: A survey,” *IEEE Signal Process. Mag.*, vol. 37, no. 1, pp. 89–106, 2020.