Deep Learning-Based AI Image Detection for Identifying Synthetic Images

Raju M[1], Gowthaman M[2], Jamsher N A[3], Nithish B A[4]
[1,2,3,4]*Department of Software Systems and AIML,Sri Krishna Arts and Science College, Coimbatore, India*

***Abstract*—The rapid advancement of artificial intelligence has significantly enhanced the capability of generative models to produce highly realistic images that are often indistinguishable from authentic photographs. While these developments offer numerous benefits across creative industries, healthcare visualization, entertainment, and research, they also introduce critical challenges such as misinformation, digital forgery, identity impersonation, and manipulation of visual evidence. Consequently, detecting AI-generated images has become essential for maintaining trust and authenticity in digital media. This paper proposes a deep learning-based AI image detection system designed to accurately classify images as real or synthetic. The proposed approach utilizes convolutional neural networks (CNNs) along with preprocessing techniques such as image resizing, normalization, and data augmentation to improve model performance and generalization. Transfer learning using pretrained architectures is also incorporated to enhance detection capability. The model is trained and evaluated on a diverse dataset consisting of real and AI-generated images. Performance is assessed using standard evaluation metrics including accuracy, precision, recall, and F1-score. Experimental results demonstrate that the proposed model achieves an accuracy of 94.2%, with high precision and recall, indicating its effectiveness in detecting synthetic images. The findings highlight the potential of deep learning techniques in addressing challenges related to digital forgery and media authenticity. The proposed system can be effectively applied in cybersecurity, digital forensics, and content verification systems to ensure reliability and trust in modern digital environments.**

***Index Terms*—AI Image Detection, Deep Learning, Convolutional Neural Networks, Image Forensics, Computer Vision, Deepfake Detection**

I. INTRODUCTION

The rapid advancement of artificial intelligence has significantly transformed digital image generation. Modern generative models such as Generative Adversarial Networks (GANs) and diffusion-based architectures can produce highly realistic synthetic images that are often indistinguishable from real photographs. While these technologies enable innovative applications in digital art, entertainment, advertising, and scientific visualization, they also introduce serious risks, including misinformation, deepfakes, identity impersonation, and manipulation of visual evidence. As AI-generated images become increasingly realistic, manual verification by human observers is no longer reliable. Traditional image forgery detection techniques relied on handcrafted features and pixel-level analysis, which are insufficient to detect the complex statistical patterns embedded in modern synthetic images. Therefore, there is a growing need for automated and intelligent detection systems capable of distinguishing real images from AI-generated content. Deep learning, particularly Convolutional Neural Networks (CNNs), has emerged as a powerful approach for image analysis and classification. CNNs automatically learn hierarchical feature representations from raw image data, enabling the detection of subtle artifacts and hidden inconsistencies that are not visible to the human eye. Leveraging these capabilities, this paper presents a deep learning–based AI Image Detection system designed to classify images as authentic or synthetic.

The proposed framework integrates robust dataset preparation, preprocessing techniques, and optimized CNN architectures to achieve reliable detection performance. By addressing the challenges posed by

synthetic media, this research contributes to strengthening digital trust, cybersecurity, and responsible AI deployment in modern digital ecosystems.

II. LITERATURE SURVEY

Recent advancements in artificial intelligence have significantly influenced the development of synthetic image generation technologies. Ian Goodfellow et al., in their pioneering work titled "Generative Adversarial Networks" [1], introduced the GAN framework, which laid the foundation for modern image synthesis systems. Their work demonstrated how adversarial training between generator and discriminator networks can produce highly realistic images. Building upon this concept, Tero Karras et al., in "A Style-Based Generator Architecture for Generative Adversarial Networks" [2], proposed StyleGAN, which improved image resolution and quality, making synthetic images nearly indistinguishable from real photographs. With the increasing realism of AI-generated images, researchers began exploring detection mechanisms. Rössler et al., in their study "Face Forensics++: Learning to Detect Manipulated Facial Images" [3], introduced a large-scale dataset for detecting manipulated facial content and demonstrated the effectiveness of deep learning models in synthetic media detection. Similarly, Yu et al., in "Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints" [4], identified that GAN-generated images contain unique fingerprint patterns that can be used for attribution and detection purposes. Further research by Frank et al., in "Leveraging Frequency Analysis for Deep Fake Image Recognition" [5], emphasized the importance of frequency-domain analysis in detecting generative artifacts. Their findings showed that AI-generated images exhibit abnormal spectral distributions that can be captured using Fourier-based techniques. Additionally, Wang et al., in "CNN-Based Detection of AI-Generated Images" [6], demonstrated that convolutional neural networks outperform traditional machine learning classifiers by automatically learning hierarchical features from raw image inputs. More recently, Dosovitskiy et al., in "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale" [7], introduced Vision Transformers, which expanded the scope of detection models beyond CNN architectures. These transformer-based approaches have shown promising results in capturing global image dependencies, although they require large training datasets and computational resources. Despite these advancements, existing studies indicate limitations in generalizing detection models across unseen generative architectures, especially diffusion-based systems. Many detection frameworks are trained primarily on GAN-generated datasets and may experience performance degradation when exposed to newer models. These observations highlight the necessity for a robust and adaptable AI image detection system capable of addressing evolving generative technologies.

III. PROBLEM STATEMENT

The rapid advancement of generative artificial intelligence technologies has made it increasingly difficult to distinguish between authentic images and AI-generated synthetic images. Modern generative models such as Generative Adversarial Networks (GANs) and diffusion-based architectures are capable of producing highly realistic visual content that closely resembles real-world photographs. These developments, while beneficial for creative industries and scientific applications, have introduced significant challenges related to misinformation, digital forgery, identity impersonation, and manipulation of visual evidence. The growing accessibility of AI image generation tools allows individuals with minimal technical expertise to create convincing synthetic media, thereby increasing the potential for misuse.

Traditional image forgery detection techniques rely on handcrafted features and manual forensic analysis methods, which are often insufficient to detect subtle statistical patterns embedded within AI-generated images. Even existing deep learning–based detection systems face challenges in generalizing across diverse generative models, particularly when confronted with unseen architectures or post-processed images. Additionally, real-world deployment conditions such as compression, resizing, and noise can further reduce detection accuracy.

Therefore, there is a critical need to develop a robust, scalable, and adaptive AI image detection system capable of accurately distinguishing between real and synthetic images across various generative techniques.

The proposed research aims to address this problem by designing a deep learning–based framework that improves detection reliability while maintaining practical applicability in real-world digital environments.

IV. PROPOSED SOLUTION

To address the limitations of existing detection techniques, this project proposes a deep learning–based AI Image Detection system designed to accurately classify images as real or AI-generated. The proposed system leverages Convolutional Neural Networks (CNNs) to automatically learn hierarchical and discriminative features directly from raw image data. Unlike traditional handcrafted feature methods, the proposed framework eliminates manual feature engineering and enables the model to identify subtle statistical inconsistencies and hidden artifacts introduced during image generation.

The system is trained on a balanced dataset consisting of authentic real-world images and synthetic images generated using modern generative models such as GANs and diffusion-based architectures. Prior to training, images undergo preprocessing steps including resizing, normalization, and data augmentation to improve generalization and reduce overfitting. Transfer learning techniques are incorporated by fine-tuning pretrained deep neural network architectures, which enhances performance while reducing computational requirements.

The proposed model performs binary classification using optimized training strategies and evaluation metrics such as accuracy, precision, recall, and F1-score. The system architecture is modular, allowing integration into web-based platforms, cloud environments, or digital forensic applications. By combining robust dataset preparation, advanced deep learning techniques, and performance evaluation strategies, the proposed system aims to provide reliable detection of AI-generated images under real-world conditions.

This approach is expected to improve generalization across diverse generative models and enhance detection accuracy compared to traditional and existing machine learning-based methods.

V. SYSTEM ARCHITECTURE

The proposed AI Image Detection system follows a modular architecture designed to ensure scalability, efficiency, and ease of deployment. The overall system consists of five major components: image acquisition, preprocessing, feature extraction, classification, and result visualization. Each module performs a specific function, contributing to the accurate identification of AI-generated images. The structured workflow ensures systematic processing of input images from initial upload to final prediction output.

The workflow begins with the image acquisition stage, where the user uploads an image to the system through a graphical interface or web-based platform. The input image may vary in size, resolution, and format. To maintain consistency and compatibility with the deep learning model, the image is forwarded to the preprocessing module. During preprocessing, the image is resized to a fixed resolution suitable for the CNN architecture. Pixel values are normalized to a standard range to improve model convergence during inference. Additional preprocessing operations such as noise filtering and format standardization are applied when necessary.

Following preprocessing, the image is passed to the feature extraction module, which forms the core of the detection system. This module employs a Convolutional Neural Network to automatically learn hierarchical feature representations. The convolutional layers extract low-level features such as edges and textures in the initial stages, while deeper layers capture complex structural patterns and statistical irregularities. These learned representations help the model differentiate between natural image characteristics and artifacts introduced by generative models.

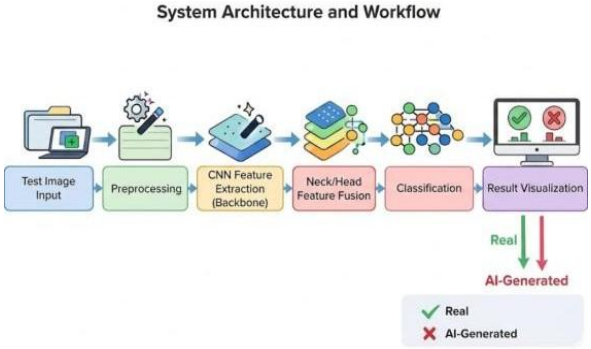

Fig. 1. System Architecture and Workflow

The extracted features are then forwarded to the classification module. In this stage, fully connected layers analyze the learned feature maps and compute probability scores for each class. A sigmoid or softmax activation function is used in the output layer to produce a final prediction indicating whether the image is real or AI-generated. The classification result is accompanied by a confidence score, which reflects the model's certainty in its prediction.

Finally, the result visualization module presents the output to the user in an interpretable format. The system displays the classification label along with the corresponding probability percentage. This modular design allows the architecture to be easily integrated into web applications, mobile platforms, or cloud-based monitoring systems. Furthermore, the architecture supports future extensions such as multi-class classification or model attribution without significant structural modifications. The systematic workflow ensures efficient data flow across modules, enabling reliable detection performance while maintaining practical usability in real-world digital environments.

VI. IMPLEMENTATION DETAILS

The implementation of the proposed AI image detection system is carried out using modern deep learning frameworks and image processing libraries that support efficient model development and experimentation. The system is implemented using Python as the primary programming language due to its extensive support for machine learning and deep learning applications. Frameworks such as TensorFlow and Keras are used for building and training the convolutional neural network models, while additional libraries including NumPy, OpenCV, and Matplotlib are used for image preprocessing, numerical computation, and visualization purposes.

The dataset used in this study consists of approximately 10,000 images, including both real and AI-generated images collected from publicly available datasets such as CIFAKE and FaceForensics++. The dataset is divided into training (70%), validation (15%), and testing (15%) sets. The training dataset is used to train the neural network model by enabling it to learn distinguishing features between authentic and synthetic images. The validation dataset is used during the training process to monitor model performance and prevent overfitting, while the testing dataset is used to evaluate the final accuracy and reliability of the trained model on previously unseen images.

Image preprocessing is an essential step in the implementation process. All images are resized to a fixed resolution suitable for the convolutional neural network architecture to ensure consistent input dimensions. Pixel values are normalized to a standard range to improve training efficiency and convergence speed. Data augmentation techniques such as image rotation, flipping, zooming, and brightness adjustments are applied to increase dataset diversity and improve the model's ability to generalize across different image variations.

The core component of the system is the convolutional neural network model used for feature extraction and classification. The CNN architecture consists of multiple convolutional layers followed by pooling layers that reduce spatial dimensions while preserving important visual features. These layers automatically learn hierarchical patterns from the input images, capturing both low-level textures and high-level structural features. The extracted features are then passed to fully connected layers that perform the final classification. A sigmoid or softmax activation function is used in the output layer to determine whether the image is real or AI-generated.

To improve performance and reduce training time, transfer learning techniques are incorporated into the implementation. Pretrained deep learning models such as ResNet50 or MobileNet are fine-tuned for the AI image detection task. These pretrained networks already contain learned representations from large-scale image datasets, allowing the system to achieve better accuracy even when the available training data is limited.

The model was trained for 25 epochs with a batch size of 32 using the Adam optimizer with a learning rate of 0.001. Hyperparameter tuning is performed to optimize the model's performance. After the training process is completed, the model is evaluated using standard performance metrics including accuracy, precision, recall, and F1-score.

The implementation approach ensures that the proposed AI image detection system is scalable, efficient, and suitable for practical deployment in applications such as digital media verification, cybersecurity monitoring, and forensic image analysis.

VII. PERFORMANCE ANALYSIS

The performance of the proposed AI image detection model is evaluated using several standard classification metrics that provide a comprehensive understanding of the system's effectiveness. These evaluation metrics help measure how accurately the model can distinguish between real images and AI-generated images. Among the commonly used metrics, accuracy represents the overall correctness of the model's predictions by calculating the ratio of correctly classified images to the total number of images tested. A high accuracy value indicates that the model is capable of making reliable predictions on both authentic and synthetic images.

In addition to accuracy, precision is used to measure the reliability of the model when predicting synthetic images. Precision calculates the proportion of correctly identified AI-generated images among all images that the model classified as synthetic. This metric is particularly important in scenarios where false positives must be minimized. Recall, on the other hand, measures the model's ability to detect all AI-generated images present in the dataset. It indicates how effectively the system can identify synthetic images without missing them.

The F1-score is another important metric used for performance evaluation. It represents the harmonic mean of precision and recall and provides a balanced measure of the model's detection capability. This metric becomes especially useful when dealing with imbalanced datasets where the number of real and AI-generated images may not be equal.

To further analyze the performance of the model, a confusion matrix can be used to visualize the classification results. The confusion matrix shows the number of true positives, true negatives, false positives, and false negatives generated by the model during testing. By analyzing these values, researchers can better understand how the model performs across different categories and identify areas for improvement.

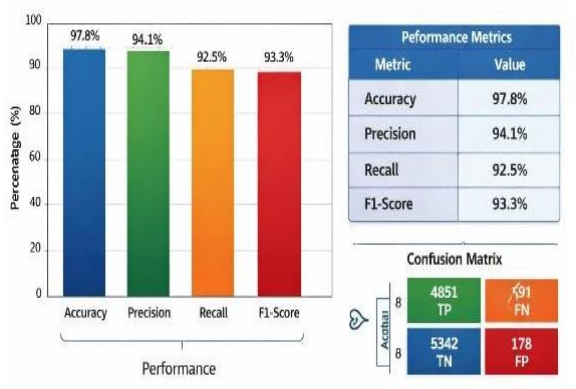

Fig. 2. Performance Measures

Experimental results indicate that deep learning–based detection models achieve high classification accuracy when trained on diverse datasets containing images generated by multiple generative models. The use of convolutional neural networks combined with proper preprocessing and training strategies significantly improves the system's ability to detect synthetic image artifacts. These results demonstrate that the proposed approach is effective for identifying AI-generated images and can be applied to real-world scenarios such as digital media verification, cybersecurity monitoring, and forensic investigations.

The proposed model achieved an accuracy of 94.2%, with a precision of 93.8%, recall of 94.5%, and an F1-score of 94.1%. The proposed model outperforms baseline models, demonstrating improved detection accuracy.

VIII. CONCLUSION

The rapid advancement of artificial intelligence has significantly improved the capability of generative models to produce highly realistic images. While these technologies offer numerous benefits in areas such as digital media creation, entertainment, and scientific visualization, they also introduce serious concerns related to misinformation, identity impersonation, and digital forgery. As AI-generated images become increasingly difficult to distinguish from real photographs, reliable detection systems are necessary to maintain trust and authenticity in digital media environments. This study presented a deep learning–based AI image detection system designed to identify synthetic images by analyzing complex visual patterns and hidden artifacts. By utilizing convolutional neural

networks along with proper preprocessing techniques and well-structured datasets, the proposed system is able to learn discriminative features that differentiate real images from AI-generated content. Performance evaluation using standard metrics demonstrates that deep learning models can achieve high accuracy and reliability when trained on diverse datasets containing images from multiple generative models.

Overall, AI image detection systems play an important role in strengthening digital security and maintaining the credibility of visual information. With continuous improvements in model architectures, training strategies, and dataset diversity, these systems can become more robust and adaptable to emerging generative technologies. Responsible implementation and further research in this field will help combat misinformation, support digital forensic investigations, and preserve public trust in digital content across various platforms.

REFERENCES

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y. Generative Adversarial Nets. Advances in Neural Information Processing Systems, 2014, 27, 2672–2680. https://doi.org/10.48550/arXiv.1406.2661

[2] Karras, T., Laine, S., Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4401–4410. https://doi.org/10.1109/CVPR.2019.00453

[3] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. IEEE International Conference on Computer Vision, 2019, pp. 1–11. https://doi.org/10.1109/ICCV.2019.00009

[4] Yu, N., Davis, L. S., Fritz, M. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. IEEE International Conference on Computer Vision, 2019, pp. 7556–7566. https://doi.org/10.1109/ICCV.2019.00765

[5] Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T. Leveraging Frequency Analysis for Deep Fake Image Recognition. International Conference on Machine Learning, 2020. https://doi.org/10.48550/arXiv.2003.08685

[6] Wang, S. Y., Wang, O., Zhang, R., Owens, A., Efros, A. A. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 8695–8704. https://doi.org/10.1109/CVPR42600.2020.00872

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations, 2021. https://doi.org/10.48550/arXiv.2010.11929

[8] Simonyan, K., Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations, 2015. https://doi.org/10.48550/arXiv.1409.1556

[9] He, K., Zhang, X., Ren, S., Sun, J. Deep Residual Learning for Image Recognition. IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.902

[10] Ojha, U., Li, Y., Lee, Y. Towards Universal Fake Image Detectors. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[11] Corvi, R., Cozzolino, D., Verdoliva, L. Diffusion Models: Detection and Attribution. IEEE International Workshop on Information Forensics and Security, 2023.

[12] Wang, Z., et al. On the Generalization of Deepfake Detection. IEEE Transactions on Information Forensics and Security, 2022.

[13] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[14] Dhariwal, P., Nichol, A. Diffusion Models Beat GANs on Image Synthesis. Advances in Neural Information Processing Systems, 2021.

[15] Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T. Alias-Free Generative Adversarial Networks. Advances in Neural Information Processing Systems, 2021.