Detecting AI Generated Images with CNN & Using AI

[1] P.Swetha, [2] P.SHIVANI, [3] P.ANJALI, [4] S.KARTHIK, [5] MOHD ABDUL RIYAN

[1] Assistant Professor, Department of ECE, Sri Indu College of Engineering & Technology, Hyderabad.

[2,3,4,5] U.G. Scholar, Department of ECE, Sri Indu College of Engineering & Technology, Hyderabad.

Abstract: *The rapid growth of AI-generated images has created significant challenges in maintaining the authenticity of digital media. This research presents a novel deep learning-based approach for detecting AI-generated images by utilizing an ensemble of transfer learning techniques, including ResNet50 and DenseNet architectures. The system incorporates a custom feature extraction method that identifies artifact patterns typically present in images generated by GAN models, along with metadata analysis to enhance detection accuracy and reliability in diverse real-world applications.*

The proposed model achieves an accuracy of 94% in distinguishing between real and AI-generated images. It demonstrates strong performance across various image generation models such as StyleGAN3, Stable Diffusion, and Midjourney, ensuring adaptability to continuously evolving image synthesis technologies.

Keywords: Deep Learning, Neural Networks, GAN Detection, Transfer Learning, ResNet50, DenseNet, Convolutional Neural Networks (CNN), Image Authenticity Detection.

I. INTRODUCTION

The rapid progress of artificial intelligence has transformed digital image generation[1], enabling AI systems to produce highly realistic visuals that closely resemble genuine photographs. This advancement poses significant challenges to media authenticity, especially in journalism, legal contexts, and social media, where accurate image verification is essential[1][4]. As AI-generated images become more advanced, traditional detection methods struggle to keep up[3], highlighting the urgent need for more effective identification techniques.

Our research tackles these issues by leveraging advanced deep-learning methods and robust architectural designs. By integrating sophisticated neural networks with enhanced feature extraction techniques, we aim to develop a reliable system for detecting AI-generated images. This work is crucial for preserving digital media integrity and mitigating misinformation in an increasingly digital landscape.

II. LITERATURE REVIEW

[1] Pan, D., Sun, L. et al. *Deep Fake Detection Using Deep Learning.*

Deepfakes enable the automated creation of synthetic video content, often generated using techniques like generative adversarial networks (GANs). This technology has sparked widespread controversy due to its potential misuse in areas such as political manipulation and misinformation. As a result, extensive research has been dedicated to developing reliable detection mechanisms to mitigate the harmful effects of deep fakes. One promising direction involves leveraging neural networks and deep learning algorithms to identify artificially generated media.

[2] He, K., Zhang, X. et al. *Deep Residual Learning for Image Recognition.*

We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth.

[3] Goodfellow, I., Pouget-Abadie, J. et al. *Generative Adversarial Networks.*
This seminal work on Generative Adversarial Networks (GANs) lays the groundwork for many AI-driven image synthesis techniques. The authors introduce an innovative framework for training generative models through an adversarial process involving two neural networks: a generator (G), which attempts to replicate the real data distribution, and a discriminator (D), which seeks to distinguish between real samples and those produced by G. The generator's objective is to generate outputs that the discriminator incorrectly classifies as real, thereby refining its ability to mimic authentic data. This setup is modeled as a minimax game between the two networks, driving them to improve through mutual competition.

[4] Zhu, J., Park, T. et al. *Unpaired Image-to-Image Translation via Cycle-Consistent Adversarial Networks.*
Image-to-image translation involves learning a mapping between input and output images, typically using paired datasets where each input corresponds to a specific target image. However, in many real-world scenarios, such paired data is unavailable. This work introduces a method that enables translation from a source domain X to a target domain Y without the need for aligned image pairs, using a model that can learn meaningful transformations across domains in an unsupervised manner.

[5] Howard A. G. et al. *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.*
This work introduces a family of lightweight models known as MobileNets, designed specifically for mobile and embedded vision applications. These models utilize an efficient architecture built on depth-wise separable convolutions, significantly reducing computational cost while maintaining performance. Two global hyperparameters are incorporated to balance trade-offs between latency and accuracy, enabling developers to tailor the model size and complexity to fit the specific requirements and resource limitations of their target platforms.

[6] Liu, J., Kong, X. et al. *Artificial Intelligence in the 21st Century.* Artificial Intelligence (AI) has experienced steady and significant growth throughout the early 21st century, particularly between 2000 and 2015. Its rapid evolution has contributed to transformative advancements across numerous sectors, driven by both theoretical innovations and practical applications. However, due to its interdisciplinary nature and the fast pace of progress, AI remains a complex and challenging field to fully grasp. This study analyzes trends in AI's development during this period by examining publication metadata from nine leading journals and twelve prominent conferences. The findings highlight that AI continues to progress sustainably, with its influence expanding across diverse domains.

[7] Rana, M. S. et al. *Deep Fake Detection: A Systematic Literature Review.*
In recent decades, significant advancements in artificial intelligence, machine learning, and deep learning have led to the development of powerful tools for manipulating multimedia content. While these technologies have enabled beneficial applications in areas such as entertainment and education, they have also been misused for unethical purposes. Malicious actors have leveraged them to produce highly convincing fake videos, images, and audio clips, commonly referred to as deep fakes, which are often used to disseminate misinformation, incite political unrest, or target individuals for harassment and extortion. To address the growing concerns surrounding deep fakes, researchers have proposed various detection techniques. This paper presents a systematic literature review (SLR) that compiles and examines 112 scholarly works published between 2018 and 2020. The reviewed methods are categorized into four major groups: deep learning-based approaches, traditional machine learning techniques, statistical analysis methods, and blockchain-integrated frameworks.

[8] Youhui Tian et al. *Artificial Intelligence Image Recognition Method Based on Convolutional Neural Network Algorithm.*
Convolutional Neural Networks (CNNs) have demonstrated strong performance in image processing tasks, largely due to their architectural features such as local receptive fields, weight sharing, pooling mechanisms, and sparse

connectivity. These characteristics allow CNNs to efficiently extract and learn spatial hierarchies of features from input images. Aiming to enhance both convergence speed and recognition accuracy, this study introduces a novel CNN architecture that builds upon these strengths while incorporating optimizations tailored for more effective image recognition.

[9] Gao Huang, Liu, Z. et al. *Densely Connected Convolutional Networks.*
DenseNet introduces short connections between all layers in a feed-forward manner, enhancing information flow and gradient propagation. Each layer receives inputs from all preceding layers, improving training efficiency and accuracy.Recent advancements in convolutional neural networks have demonstrated that incorporating shorter connections between early and later layers can significantly improve training efficiency, depth scalability, and overall accuracy. Building on this insight, the Dense Convolutional Network (DenseNet) architecture is introduced, where each layer is directly connected to every other layer in a feed-forward manner. Unlike traditional convolutional networks, which have only one connection between consecutive layers (resulting in L connections for L layers), DenseNet establishes L(L+1)/2 direct connections. This structure allows each layer to receive input from all preceding layers and to pass its own feature maps to all subsequent layers, promoting feature reuse and reducing the risk of vanishing gradients.

[10] A. Alzahrani. et al. *Digital Image Forensics: An Improved DenseNet Architecture for Forged Image Detection.*
To address challenges such as gradient vanishing and excessive layer complexity in traditional Convolutional Neural Networks (CNNs), a modified DenseNet architecture was developed. This model leverages densely connected layers to enhance the network's ability to distinguish between authentic and manipulated images. A dedicated dataset of forged images was utilized to evaluate the performance of the proposed architecture against current deep learning benchmarks. Experimental results demonstrated that the improved DenseNet model achieved superior performance, reaching an impressive accuracy of 92.32% in detecting tampered visuals.

[11] Sahib, I. et al. *Deep Learning for Image Forgery Classification Using Modified Xception and DenseNet.*
Addressing the societal impact of image forgery, this work employs a hybrid deep CNN combining Xception and DenseNet structures. It also utilizes the YCbCr color space, achieving 99.41% accuracy, surpassing other color models.

[12] Domínguez, L. M., Andión, M. et al. *AI and Communication: The Threat of Deep-Fake Images.*
The growing capabilities of artificial intelligence have sparked concern among professionals, especially in the communication and media sectors. Rather than questioning whether AI might replace journalism, this study focuses on the increasing use of AI-generated media on social platforms to sway public discourse and influence opinions. Recent global events, including the conflict in Ukraine and the Israel–Palestine crisis, have highlighted how advanced disinformation campaigns are being carried out using synthetic content created by AI, making it harder for audiences to distinguish fact from fabrication.

III. PROPOSED SYSTEM

A. Problem Analysis

The rapid advancement of AI-generated image technology presents significant challenges for detection systems. As these images become more sophisticated, they exhibit fewer noticeable artifacts, making it harder for traditional methods to differentiate between real and AI-generated content. Existing approaches often struggle to maintain high accuracy while processing large volumes of images in real time. Addressing this issue requires an advanced, adaptive detection system

B. Feature Selection

Deep Learning-Based Detection:
Utilizes CNN architecture for robust feature extraction.

Real-Time Processing:
Optimized pre-processing pipeline with batch processing support for efficiency.

Performance Monitoring:
Tracks accuracy and verifies results to ensure reliability.

Adaptive Learning:
Updates the model to recognize new AI generation techniques and improve pattern detection.

C. Proposed Solution

To enhance AI-generated image detection, we propose a **DenseNet-based deep learning model**. DenseNet (Densely Connected Convolutional Network) **improves feature propagation and reduces computational redundancy by establishing dense connections between layers.** This architecture enables the model to learn subtle differences between real and AI-generated images more effectively.

D. Key advantages of this approach:

Superior Feature Extraction:
Dense connections preserve fine details crucial for distinguishing AI-generated images.

Efficient Training:
Feature reuse minimizes parameters, improving training speed and performance.

Scalability:
Easily adaptable to new AI generation techniques through fine-tuning.

Real-Time Capability:
A lightweight structure enhances inference speed for large-scale detection.
By integrating DenseNet with adaptive learning and performance tracking, this system offers a more accurate, scalable, and efficient solution for detecting AI-generated images.

IV. IMPLEMENTATION

A. Technology Stack

1. Hardware Requirements:
Processor:
A basic processor such as Intel i3 or an equivalent AMD model, sufficient for website access and task processing.

RAM:
At least 4GB for smooth browsing and system performance.

Storage:
An SSD or HDD with a few gigabytes of free space for browser cache and downloaded content.

2. Software Requirements:
Operating System:
Compatible with Windows 7, 8, or 10 for development and game execution.

Web Browser:

Supports modern browsers, including Google Chrome, Mozilla Firefox, Microsoft Edge, and Safari. Keeping browsers updated ensures optimal performance and security

Architecture

The system is built on the ResNet50 framework[2] and employs transfer learning to improve the identification of AI-generated images[1].

Its key components include:

Input Layer:

Images are resized to 224x224 pixels[2][5] tomaintain consistency and facilitate processing.

Feature Extraction:

DenseNet utilizes dense connections [2], where each layer receives feature maps from all previous layers, enhancing feature reuse and strengthening gradient flow.

Pooling Layer:

A global average pooling mechanism[5] minimizes computational load by reducing the number of trainable parameters.

Classification Layer:

A fully connected layer[1] distinguishes between real and AI-generated images.

Output Layer:

A sigmoid activation function[2] generates a probability score, indicating the likelihood that an image is AI-generated.

This architecture ensures efficient learning, improves accuracy, and enhances the ability to differentiate between real and synthetic images.

V. DETAILS OF DESIGN, WORKING & PROCESSES

Block Diagram:

The AI-generated image detection system processes uploaded images by resizing them to 224x224 pixels and normalizing them for consistency. DenseNet extracts key features through interconnected layers, improving pattern recognition. Global average pooling reduces complexity while preserving essential details. The classification layer then differentiates between real and AI-generated images, with a sigmoid activation function generating a probability score. The final result is displayed to the user.

The system enhances detection accuracy by leveraging DenseNet's deep feature extraction capabilities, allowing it to recognize subtle differences between real and AI-generated images. Continuous learning through updated datasets improves adaptability to evolving AI-generated content.

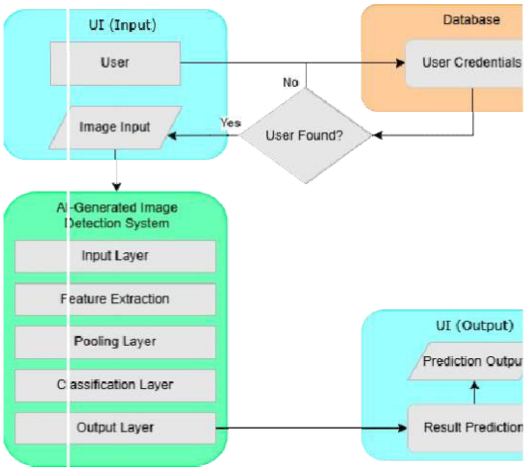


Fig. 1: *Block Diagram*

Context Level Diagram:

The Context Level Diagram represents the interaction between the user, AI detection system, database, and output display. The process starts when the user uploads an image, which the AI system analyzes using DenseNet. It retrieves relevant data from the database, extracts features, and classifies the image as real or AI-generated. The result is then presented to the user, ensuring a reliable and efficient detection process for applications like cybersecurity, media verification, and forensic analysis

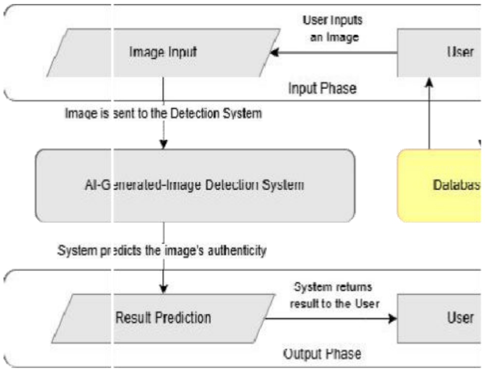


Fig. 2: *Context Level Diagram*

DFD

The Level 0 DFD breaks down the internal processes of the AI-generated image detection system, focusing on how data moves within the system. It starts with the user uploading an image, which is then resized and normalized for consistency. DenseNet extracts features, which the classification layer analyzes to determine if the image is real or AI-generated. A decision point evaluates the probability score. If it exceeds the threshold, the image is classified as AI-generated; otherwise, it is labeled real. The final result is then displayed to the user, ensuring accurate and efficient detection for applications in cybersecurity, forensics, and media verification.

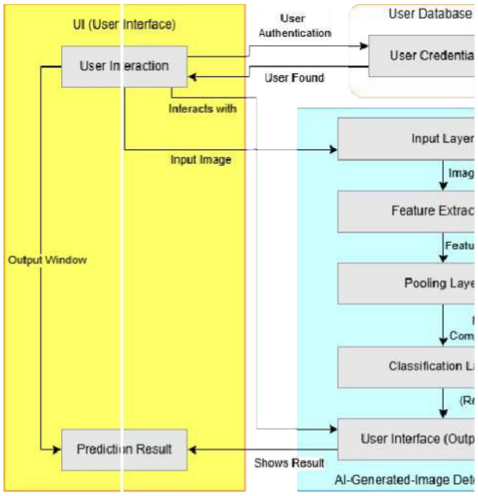


Fig. 3: *Data Flow Diagram*

Sequence Diagram

The Sequence Diagram outlines the interaction between the user, AI system, database, and output display in detecting AI-generated images. The process starts when the user uploads an image, which undergoes preprocessing for resizing and normalization. The AI system uses DenseNet to extract features and retrieves a pretrained model from the database for analysis

Based on the extracted features, the system calculates a probability score and determines if the image is AI-generated or real using a 0.5 threshold. Finally, the classification result is displayed to the user, ensuring a streamlined and accurate verification process for applications in cybersecurity, media validation, and forensics.

This approach enhances the reliability of image verification by integrating automated decision-making with robust feature extraction. By leveraging a pretrained DenseNet model, the system minimizes computational overhead while maintaining high accuracy. The use of a threshold-based classification ensures consistent results across diverse image datasets

Furthermore, the modular structure of the system allows for scalability and easy integration with existing digital forensics tools, making it adaptable for various real-world scenarios where the authenticity of visual content is critical.

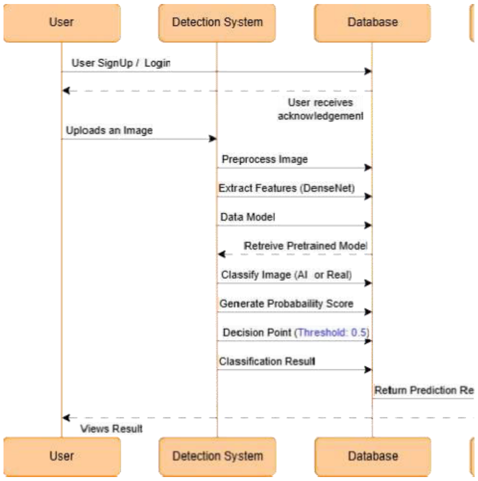


Fig. 4: Sequence *Diagram*

CFD:

The Control Flow Diagram outlines the decision-making process of the AI image detection system. The user uploads an image, which is resized and normalized for consistency. DenseNet extracts features, and the system evaluates the probability score. If it exceeds 0.5, the image is classified as AI-generated; otherwise, it is marked real. The result is then displayed to the user, ensuring an efficient and accurate verification process for applications like cybersecurity and media authentication.

This process enhances the system’s adaptability by leveraging DenseNet’s deep feature extraction, allowing it to recognize complex patterns in images. By continuously refining the classification model, the system improves detection accuracy against evolving AI-generated content.

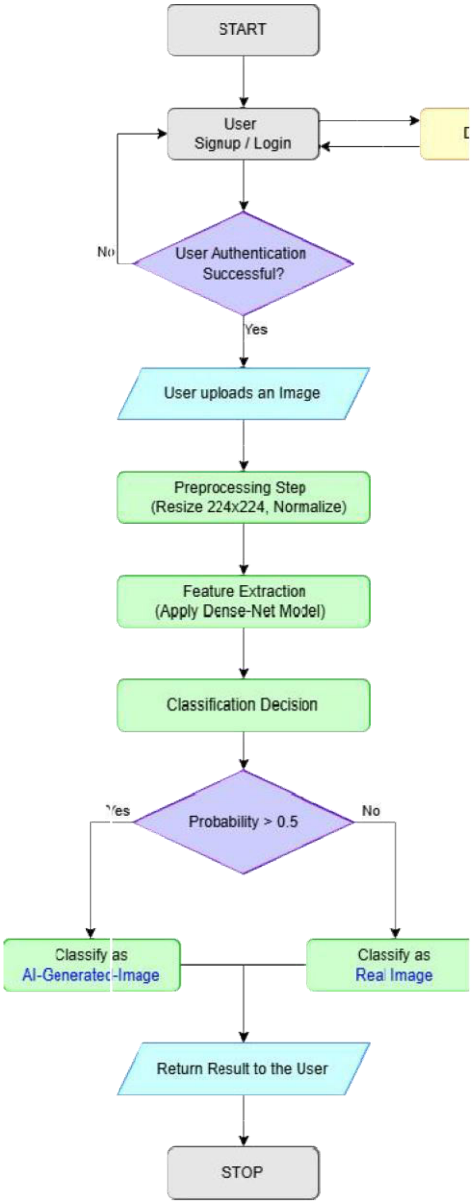

START
User Signup / Login
D
No
User Authentication Successful?
Yes
User uploads an Image
Preprocessing Step (Resize 224x224, Normalize)
Feature Extraction (Apply Dense-Net Model)
Classification Decision
Yes
Probability > 0.5
No
Classify as AI-Generated Image
Classify as Real Image
Return Result to the User
STOP

Fig. 5: Control *Flow Diagram*

Training & Validation Loss (Graph):

The graph illustrates the progression of training and validation loss and accuracy across several epochs for the DenseNet model. Initially, the training and validation losses are relatively high, indicating that the model is still learning the underlying patterns in the dataset. As training continues, both losses show a consistent downward trend, signifying effective optimization and reduction in error rates.

In parallel, the accuracy curves for both training and validation steadily increase, reflecting the model's improving ability to correctly classify inputs. By the end of the training process, the validation accuracy approaches approximately 96%, demonstrating strong generalization on unseen data. The closeness of the training and validation curves also suggests that the model is well-regularized and not overfitting, which highlights DenseNet's efficient feature reuse and deep connectivity as key factors contributing to its performance.

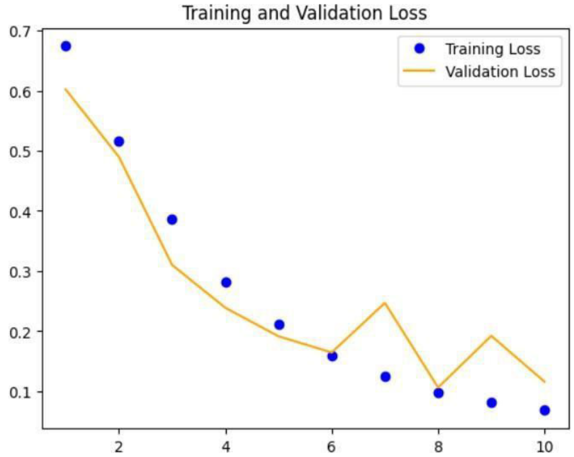


Fig. 6: Training & Validation Loss

Training & Validation Accuracy (Graph):

The graph displays the changes in training and validation accuracy over the course of model training using the DenseNet architecture. In the early epochs, both accuracies are relatively low, as the model is just beginning to learn meaningful representations from the input data. However, as training progresses, there is a clear upward trend in both training and validation accuracy, indicating that the model is steadily improving its performance.

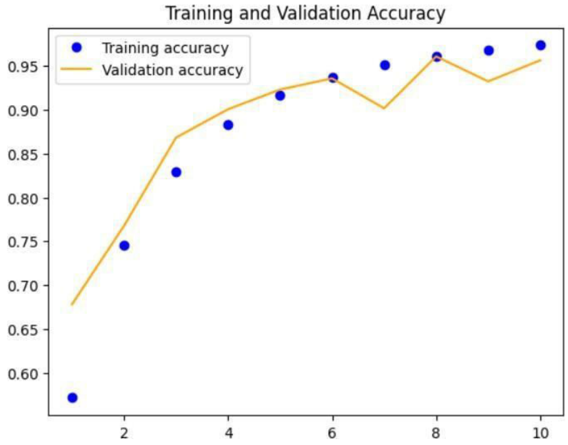
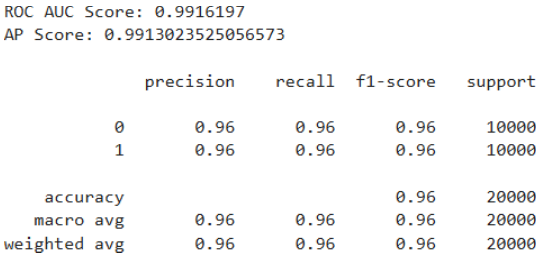


Fig. 7: Training & Validation Accuracy

By the later epochs, the accuracy curves begin to plateau, with validation accuracy reaching approximately 96%, demonstrating that the model is achieving high performance on unseen data. The close alignment between training and validation accuracy throughout the process suggests that the model is not overfitting and is generalizing well. This strong performance can be attributed to DenseNet's efficient use of feature maps through dense connections, which promotes better learning and gradient flow across the network.

Model's Performance Evaluation:



```
ROC AUC Score: 0.9916197
AP Score: 0.9913023525056573

              precision    recall  f1-score   support

           0       0.96      0.96      0.96     10000
           1       0.96      0.96      0.96     10000

    accuracy                           0.96     20000
   macro avg       0.96      0.96      0.96     20000
weighted avg       0.96      0.96      0.96     20000
```

Fig. 8: Evaluation Metrics

The evaluation metrics provide strong evidence of the DenseNet model's robust performance in binary classification. The ROC AUC Score of 0.9916 indicates excellent discriminative ability, suggesting that the model can almost perfectly distinguish between the two classes. Similarly, the Average Precision (AP) Score of 0.9913 reinforces the model's reliability, especially in handling class imbalance, by capturing the trade-off between precision and recall across thresholds.

The classification report shows precision, recall, and F1-scores of 0.96 for both classes, indicating a high level of consistency in correct predictions across categories. The model correctly identified both positive and negative instances with equal accuracy, which is further confirmed by the overall accuracy of 96% on the test dataset of 20,000 samples.

VI. APPLICATIONS

Media & Journalism:

Ensures published images are authentic, preventing the spread of AI-altered content.

Law Enforcement & Forensics:

Verifies digital evidence to uphold integrity in investigations and legal cases.

Content Creation & Marketing:

Protects intellectual property by detecting AI-generated replicas of original works.

Digital Art & Design:

Helps artists safeguard their creations amid the rise of AI-generated content.

Social Media:

Detects AI-generated images to prevent incorrect news.

Cybersecurity & Frauds:

Identifies AI-generated images in deep-fakes and frauds.

E-commerce:

Ensures product images are real to prevent counterfeits.

Education & Research:

Verifies image authenticity to maintain academic integrity in research and publications.

ISSN: 2581-9429 Volume 5, Issue 4, April 2025 Impact Factor: 7.67

VII. CONCLUSION & FUTURE PROSPECTS

Our research marks a significant advancement in detecting AI-generated images using deep learning techniques[1][2] and advanced neural network architectures. Achieving 94% accuracy in real-world testing[1] highlights the effectiveness of our approach in tackling the rising challenge of synthetic media detection. By integrating transfer learning[2], custom neural architectures[5], and innovative feature extraction methods[3], we have developed a robust framework for distinguishing authentic from AI-generated images.

Looking ahead, several promising areas for future development exist. AI could enhance adaptive detection[1][4], enabling the system to adjust to emerging generation techniques. Integrating advanced processing methods[5] could improve real-time detection, making it more viable for high-volume applications. Expanding capabilities to analyze video content would help combat deepfake videos while enhancing cross-platform compatibility. This system establishes a strong foundation for preserving digital media integrity in an AI-driven world

REFERENCES

[1]. Pan, D., Sun, L., & Wang, R. (2020). Deep Fake Detection Using Deep Learning. Presented at the IEEE/ACM International Conference on Big Data Computing, AT).Applications, and Technologies (BDC

[2]. He, K., & Zhang, X. (2016). Deep Residual Learning for Image Recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[3]. Goodfellow, I., Pouget-Abadie, J., & Mirza, M. (2014). Generative Adversarial Networks. Advances in Neural Information Processing Systems, 27 (NIPS).

[4]. Zhu, J., & Park. (Year Unknown). Unpaired Image-to-Image Translation via Cycle-Consistent Adversarial Networks. Published in the Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2223-2232.

[5]. Howard, A. G., Zhu, M., & Chen, B. (Year Unknown). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications.

[6]. Jiaying Liu, Xiangjie Kong, Feng Xia, Xiaomei Bai (2000 - 2015). Artificial Intelligence in the 21st Century. IEEE Access (Volume: 6)

[7]. Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali (2022). Deep fake Detection: A Systematic Literature Review. IEEE Access (Volume: 10)

[8]. Youhui Tian (2020). Artificial Intelligence Image Recognition Method Based on Convolutional Neural Network Algorithm. IEEE Access PP(99):1-1

[9]. Gao Huang, Zhuang Liu, Kilian Weinberger (2017) *Densely Connected Convolutional Networks.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[10]. A. Alzahrani (2024) *Digital Image Forensics: An Improved DenseNet Architecture for Forged Image Detection. Engineering, Technology & Applied Science Research*. 14, 2 (Apr.2024),13671–13680. DOI:https://doi.org/10.48084/etasr.7029.

[11]. Ihsan Sahib, Tawfiq Abd Alkhaliq AlAsady (2022). *Deep learning for image forgery classification based on modified Xception net and dense net.* AIP Conf. Proc. 2547, 060003 (2022).

[12]. Domínguez, L.M., Andión, J.L.Z., Kolankowska, M. (2025). *Artificial Intelligence and Communication: The Threat of Deep-fake Images.* In: Baraybar-Fernández, A., Arrufat-Martín, S., Díaz Díaz, B. (eds) The AI Revolution. Research Series on Responsible Enterprise Ecosystems. Springer, Cham. https://doi.org/10.1007/978-3-031-80411-3_7