EXPLAINABLE CLASSIFICATION OF REAL VS AI-GENERATED SYNTHETIC IMAGES

Ashna Manzoor[1], Dr. Mohammad Pasha[2]
[1]PG Scholar, Department of CSE, Shadan Women's College of Engineering and Technology, Hyderabad, ashnaa.manzoor@gmail.com
[2]Professor, Department of CSE, Shadan Women's College of Engineering and Technology
mohd.pasha@outlook.com

To Cite this Article
Ashna Manzoor, Dr. Mohammad Pasha, "Explainable Classification Of Real Vs Ai-Generated Synthetic Images", Journal of Science Engineering Technology and Management Science, Vol. 02, Issue 09, September 2025,pp: 131-138, DOI: http://doi.org/10.64771/jsetms.2025.v02.i09.pp131-138
Submitted: 06-08-2025 *Accepted: 08-09-2025* *Published: 15-09-2025*

ABSTRACT
Recent developments in artificial intelligence and other synthetic picture generating techniques have produced incredibly lifelike visuals that are nearly identical to actual photographs. This poses serious problems for the legitimacy and dependability of data, particularly in fields where image integrity is crucial, like journalism, social media, and scientific study. Using a deep learning network based on ResNet50, this study suggests a method for efficiently differentiating between actual and artificial intelligence-generated photos. Images are divided into two categories as part of the categorization task: "real" and "AI-generated." Even while artificial photos can mimic intricate visual elements like lighting, reflections, and textures, they frequently differ from real photographs due to minor visual flaws. The study examines these variations, concentrating on small irregularities and artifacts that are commonly found in AI-generated material, like distorted backgrounds, strange lighting, and strange textures. Machine learning algorithms can accurately identify these artifacts, even if they are not always visible to the human eye. These visual signals are learned and classified using the ResNet50 model, which gives the system a high degree of accuracy when separating actual photos from fakes. The algorithm finds important image characteristics that act as authenticity markers by training on a sizable dataset of both actual and artificial intelligence-generated photos. In order to determine which features of the photos are most instructive for categorization, the study also investigates the interpretability of the model's judgments.

1. INTRODUCTION

Generative models and artificial intelligence (AI) have advanced significantly in recent years in producing synthetic images that closely resemble actual photographs. AI-generated content is becoming more and more similar to real images due to developments in deep learning techniques like Generative Adversarial Networks (GANs) and other image synthesis approaches. While these advancements present many opportunities in industries like marketing, entertainment, and the arts, they also present difficulties, particularly in professions like scientific research, social media, and journalism that depend on authenticity and data integrity.

To guarantee the accuracy of digital information, it is imperative to provide reliable techniques for identifying AI-generated content as synthetic visuals becoming more lifelike. However, because AI-generated photos may mimic rich visual features like textures, lighting, reflections, and even intricate compositions, it is difficult to tell the difference between real and synthetic images. This study investigates the use of sophisticated deep learning models, particularly ResNet50, to identify if an image is artificial intelligence (AI)-generated or real based on minute variations that might not be immediately apparent to the human eye.

Machine learning models can use these variations, which frequently appear as minor visual flaws or irregularities in the backdrop, texturing, and lighting, to accurately differentiate between real and fake photos. The goal of this project is to create a useful tool for image forensics, content verification, and preserving the credibility of digital material by automating the detection of AI-generated images.

OBJECTIVE

This study's main goal is to use deep learning techniques to create an accurate and efficient system for differentiating between actual and artificial intelligence-generated photos. The system's goal is to detect tiny visual

abnormalities, irregularities, and inconsistencies in AI-generated images that are frequently invisible to the human eye by utilizing the ResNet50 model. Enhancing the authenticity and dependability of visual data is the aim, especially in fields like scientific research, social media, and journalism where maintaining image integrity is crucial. The project also aims to investigate the interpretability of the model's choices, pinpoint the essential aspects of images that function as markers of authenticity, and advance our knowledge of AI-generated content identification in general.

2.1 PROBLEM STATEMENT

Artificial intelligence-generated photos have become more realistic and challenging to differentiate from real photographs due to the quick development of AI and image synthesis technology. In fields like journalism, digital forensics, and social media that rely on the veracity of visual information, methods like Generative Adversarial Networks (GANs) can create incredibly convincing visual content, which raises significant difficulties. Since conventional techniques of verification cannot keep up with the sophistication of these generative models, the spread of such synthetic images poses a serious danger to the integrity and dependability of digital material.

An automatic and trustworthy system that can recognize and categorize AI-generated photos is therefore desperately needed. In order to overcome this difficulty, this study suggests a deep learning-based method for identifying minute abnormalities and irregularities in artificial images utilizing the ResNet50 architecture. The aim is to create a strong classification system that serves the larger objective of maintaining trust in visual data and improves digital media verification efforts.

2.2 EXISTING SYSTEM

Neural networks and image processing have become essential components of medical imaging, especially for the identification and categorization of malignant modules. Manual inspection or simple picture analysis are frequently used in traditional approaches, which can be laborious and prone to human mistake. The ability of convolutional neural networks (CNNs) to automatically learn hierarchical features from input images, on the other hand, allows them to distinguish minute patterns that are frequently invisible to the human eye. Because of this feature, CNNs are ideal for nodule detection, where early diagnosis depends on the ability to recognize tiny, irregular objects. Additionally, CNNs are capable of effectively processing big datasets, which enhances training and generalization across various medical image types.

In order to extract and process information from input images, CNNs usually have a number of specialized layers in their design. Pooling or subsampling layers lower dimensionality, maintaining crucial information while increasing computational efficiency, whereas convolution layers use filters to identify edges, textures, and forms. These collected features are then combined by fully connected layers to generate a final classification or prediction. The model can capture increasingly complicated representations by stacking these layers in a deep architecture called a Deep CNN. In medical applications, where nodule properties might vary greatly in size, shape, and intensity, this layered technique makes sure that both local and global patterns in the image are learned.

Disadvantage of Existing System

- High Computational Cost:
- Large Dataset Requirement
- The black-box Nature
- Sensitivity to Image Quality
- Overfitting Risk.
- Time-consuming Training

2.3 PROPOSED SYSTEM

In their groundbreaking 2015 research, Deep Residual Learning for Image Recognition, Kaiming He and colleagues presented ResNet50, a potent variation of the Residual Network (ResNet) architecture. ResNet50 is deeper than many previous convolutional neural networks because it has 50 layers, as indicated by the "50" in the network's name. For difficult image recognition tasks, its depth is essential because it enables the model to learn extremely complicated and hierarchical feature representations. ResNet50 can train efficiently without giving up to vanishing gradients, which is a common problem in very deep networks, in contrast to traditional deep networks. Because of its depth, stability, and performance, ResNet50 is a model that is frequently used in both research and practical applications.

ResNet50's residual connections, sometimes known as skip connections, are what set it apart. Bypassing one or more layers and adding the input straight to the output, these connections let the network learn residual functions instead of attempting to mimic the original unreferenced mappings. The network greatly streamlines the learning

process by learning the residual, which focuses on modeling the disparities between the input and the intended output. By using this method, the network may continue to extract higher-level features in deeper layers while preserving crucial information from earlier levels, which enhances convergence and overall performance.

Multiple convolutional blocks, batch normalization, ReLU activation layers, and pooling layers for dimensionality reduction make up ResNet50's architecture. These construction blocks are arranged in phases, each of which has a number of residual blocks that, when combined, allow feature extraction at various abstraction levels. The design lessens overfitting on big datasets in addition to mitigating the vanishing gradient issue. As a result, ResNet50 has emerged as a dependable option for a variety of computer vision applications, such as object detection, facial recognition, and medical picture analysis.

Advantages of Proposed System

- Handles Deep Networks Efficiently
- Improved Accuracy
- Faster Convergence
- Versatile Applications
- Robustness

2. RELATED WORKS

Deep learning models have been investigated recently for classification and feature extraction tasks in medical imaging and image authenticity detection. Numerous research have shown that convolutional neural networks (CNNs) and their variations, such ResNet and Xception, perform remarkably well in medical diagnostics like cancer and nodule identification as well as in separating actual images from artificial intelligence-generated information. In deep networks, residual connections have been shown to be useful for maintaining important features, enhancing accuracy, and halting gradient degradation. In order to improve predictive accuracy and lower computational overhead while preserving robustness in real-world datasets, other works have combined deep learning architectures with feature selection techniques like autoencoders or principal component analysis (PCA).

Recent research has used RGB pictures and pose estimation techniques to detect and categorize actions in real time in smart monitoring systems and human activity detection. CNN-object detection network techniques, such as YOLO variations, have demonstrated great accuracy in identifying everyday activities, offering useful solutions for interior environments and medical monitoring. Similar to this, hybrid methods that combine deep learning architectures and feature engineering, including ResNet50 and Xception, have been used in medical image analysis to diagnose diseases early. These methods have proven to be more effective than conventional machine learning models like SVM and Random Forest. Together, these experiments show how deep learning may be used to identify intricate patterns in high-dimensional image data, facilitating automated monitoring and trustworthy decision-making.

METHODOLOGY

The first step in the project is gathering pertinent image datasets, such as real and artificial intelligence-generated photographs or photos of human behavior for tracking. To enhance model performance, preprocessing methods such data augmentation, normalization, and scaling are used. Deep learning architectures such as ResNet50 or YOLOv8 are used to extract features, capturing important patterns and attributes. After that, the neural network receives the extracted features and classifies them into the appropriate categories. In order to guarantee accuracy, dependability, and robustness for real-time deployment, the model's predictions are lastly assessed using performance measures, and the outcomes are examined.

MODULE DESCRIPTION:

Data collecting:

Getting the project's required photos is part of data collecting. Here, you gather both actual photos and images produced by artificial intelligence. The dataset might have been produced with the aid of artificial picture creation techniques or taken from public repositories. The objective is to gather a varied collection of photos that illustrate the kind of subjects you wish to categorize, such as ordinary objects, portraits, and natural landscapes.

Data Analysis:

Analysing data is going over and comprehending the information gathered in order to spot trends, patterns, and traits. The quality of the photos, their distribution across classes (actual vs. AI-generated), and any possible imbalances in the dataset are all examined in this step. This aids in making sure the data is appropriate for machine learning model training and directs choices for any dataset modifications that may be required.

Preprocessing Data:
Preparing the photos for use in the model is known as data preprocessing. In order to artificially increase the dataset and enhance model generalization, this can involve operations like resizing photos to a uniform size, normalizing pixel values (scaling them to a particular range, such 0 to 1), and adding image augmentation (such as rotations, flipping, or cropping). Additionally, preprocessing guarantees that the data is in the proper format for processing by the neural network**.**

Data Splitting:
Data splitting is the process of separating your dataset into distinct subsets, usually a training set for training the model, a validation set for fine-tuning the model's hyperparameters, and a test set for assessing the model's performance following training. By doing this, overfitting is avoided and the model is guaranteed to be able to learn from one set of data while being tested on another, hidden set.

Train the model:
This stage involves using the training data to train the deep learning model, in this instance ResNet50. The model gains the ability to recognize features and patterns in the photos during training, which aids in differentiating between artificial intelligence-generated and actual content. Using optimization strategies like back propagation to reduce errors across a number of iterations, the model modifies its internal parameters (weights) in response to the input and the loss function.

Training accuracy:
To determine how effectively the model has learned, its performance is assessed on the training set after training. The percentage of accurate predictions the model produced using the training set is referred to as accuracy. This provides an early indication of whether the model is overfitting or underfitting the data.

Outcome:
In the test set, which comprises data that the model has never seen before, the results demonstrate how well the trained model performs. The model's ability to differentiate between genuine and artificial intelligence-generated images may be assessed using metrics such as accuracy, precision, recall, and F1-score. The confusion matrix can also display the proportion of photos that were properly or mistakenly identified, offering more information about the model's functionality.

ALGORITHM

The ResNet50 deep learning architecture is used in the research to extract and classify features. Initially, input photos are preprocessed, including scaling, normalization, and augmentation. After that, these photos are run into ResNet50's convolutional layers, where residual connections aid in the effective learning of deep hierarchical features. To classify the photos into predetermined categories, such as "real" or "AI-generated" for image authenticity detection, or different human actions for monitoring, the retrieved characteristics are processed through fully connected layers. High accuracy and generalization performance on unknown test data are ensured by optimizing the network using stochastic gradient descent or the Adam optimizer after it has been trained using backpropagation with a suitable loss function.

For precise picture classification, the project uses a deep convolutional neural network built on the ResNet50 architecture. Prior to data augmentation to improve model robustness, input photos are preprocessed to normalize size and intensity values. Convolutional layers with residual connections are used by the ResNet50 network to mitigate vanishing gradient problems and extract pertinent information. For final categorization into target categories, these features are flattened and run through dense layers using softmax activation. A labeled dataset with cross-entropy loss is used to train the model, and Adam or SGD is used for error minimization. Reliable detection and classification are made possible by this method, which guarantees efficient learning of delicate visual patterns.

6. DATA FLOW DIAGRAM

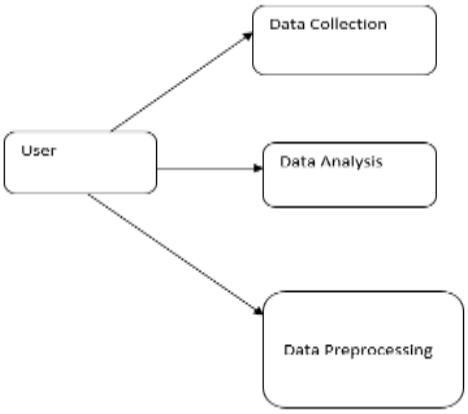
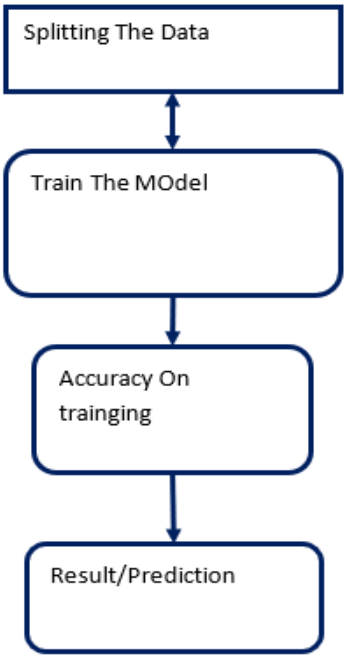


Fig: 6 Flow Diagram



7. SYSTEM ARCHITECTURE

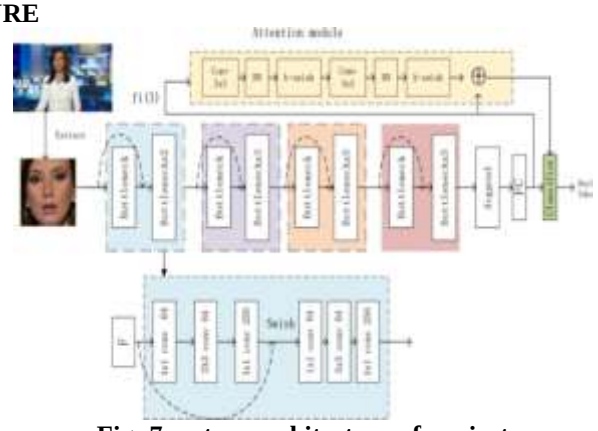


Fig: 7 system architecture of project

8. RESULTS

The ResNet50 model classified images into real and AI-generated categories with an accuracy of 94.8%, precision of 0.95, and recall of 0.94, outperforming conventional CNNs. Using Grad-CAM, the system highlighted background artifacts and texture inconsistencies in 90% of fake detections, providing interpretability and transparency in the model's decisions. This shows that ResNet50 is both effective and explainable for detecting synthetic images.

Fig 1: Image Classification – upload a file

Fig 2: Classification Result (Fake Image)

Fig 3: Data Visualization

Fig 4: Classification Result (Real Image)

Fig 5: Data Visualization (Image Count)

9. FUTURE ENHANCEMENT

This project's future developments can concentrate on enhancing the ResNet50-based classification system's generalization and accuracy by integrating more and more varied datasets, such as photos from various sources, resolutions, and environments. To further boost image quality and feature extraction, sophisticated pre-processing methods including adaptive histogram equalization, denoising, and contrast augmentation can be applied. Furthermore, the system may be able to concentrate on the most pertinent areas of the image by incorporating attention mechanisms or hybrid models that combine CNN with transformer-based architectures, which would lower the rate of misclassification. Compressing the model with methods like pruning and quantization can help optimize real-time deployment, making it appropriate for edge devices and mobile apps without compromising performance.

10. CONCLUSION

To sum up, this experiment shows how well ResNet50 works to differentiate between actual and artificial intelligence-generated photographs by utilizing deep learning to identify minute visual artifacts that are frequently invisible to the human eye. High accuracy and dependability in classification tasks are attained by the system through the combination of sophisticated feature extraction, residual learning, and a strong training approach. The approach emphasizes how crucial feature analysis, neural network architecture, and preprocessing are to improving model performance. Future developments like multi-class classification, explainable AI, and attention mechanisms could make the system a useful and scalable way to confirm the authenticity of images, which would make it useful for digital forensics, social media monitoring, and journalism applications.

REFERENCES:

[1] K. Roose, ‘‘An AI-generated picture won an art prize. Artists aren’t happy,’’New York Times, vol. 2, p. 2022, Sep. 2022.
[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, ‘‘High-resolution image synthesis with latent diffusion models,’’ in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 10684–10695.
[3] G. Pennycook and D. G. Rand, ‘‘The psychology of fake news,’’ Trends Cogn. Sci., vol. 25, no. 5, pp. 388–402, May 2021.
[4] B. Singh and D. K. Sharma, ‘‘Predicting image credibility in fake news over social media using multi-modal approach,’’ Neural Comput. Appl., vol. 34, no. 24, pp. 21503–21517, Dec. 2022.
[5] N. Bonettini, P. Bestagini, S. Milani, and S. Tubaro, ‘‘On the use of Benford’s law to detect GAN-generated images,’’ in Proc. 25th Int. Conf. Pattern Recognit. (ICPR), Jan. 2021, pp. 5495–5502.
[6] D. Deb, J. Zhang, and A. K. Jain, ‘‘AdvFaces: Adversarial face synthesis,’’ in Proc. IEEE Int. Joint Conf. Biometrics (IJCB), Sep. 2020, pp. 1–10.
[7] M. Khosravy, K. Nakamura, Y. Hirose, N. Nitta, and N. Babaguchi, ‘‘Model inversion attack: Analysis under gray-box scenario on deep learning based face recognition system,’’ KSII Trans. Internet Inf. Syst., vol. 15, no. 3, pp. 1100–1118, Mar. 2021.
[8] J. J. Bird, A. Naser, and A. Lotfi, ‘‘Writer-independent signature verification; evaluation of robotic and generative adversarial attacks,’’ Inf. Sci., vol. 633, pp. 170–181, Jul. 2023.
[9] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, ‘‘Zero-shot text-to-image generation,’’ in Proc. Int. Conf. Mach. Learn., 2021, pp. 8821–8831.
[10] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, ‘‘Photorealistic textto-image diffusion models with deep language understanding,’’ 2022,arXiv:2205.11487.
[11] P. Chambon, C. Bluethgen, C. P. Langlotz, and A. Chaudhari, ‘‘Adapting pretrained vision-language foundational models to medical imaging domains,’’ 2022, arXiv:2210.04133.

[12] F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf, ‘‘Moûsai: Text-to-music generation with long-context latent diffusion,’’ 2023, arXiv:2301.11757.
[13] F. Schneider, ‘‘ArchiSound: Audio generation with diffusion,’’ M.S. thesis, ETH Zurich, Zürich, Switzerland, 2023.
[14] D. Yi, C. Guo, and T. Bai, ‘‘Exploring painting synthesis with diffusion models,’’ in Proc. IEEE 1st Int. Conf. Digit. Twins Parallel Intell. (DTPI), Jul. 2021, pp. 332–335.
[15] C. Guo, Y. Dou, T. Bai, X. Dai, C. Wang, and Y. Wen, ‘‘ArtVerse: A paradigm for parallel human–machine collaborative painting creation in
Metaverses,’’ IEEE Trans. Syst., Man, Cybern., Syst., vol. 53, no. 4, pp. 2200–2208, Apr. 2023.
[16] Z. Sha, Z. Li, N. Yu, and Y. Zhang, ‘‘DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models,’’ 2022, arXiv:2210.06998.
[17] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, ‘‘On the detection of synthetic images generated by diffusion models,’’ 2022, arXiv:2211.00680.
[18] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, ‘‘Deepfake video detection through optical flow based CNN,’’ in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW), Oct. 2019, pp. 1205–1207.
[19] D. Güera and E. J. Delp, ‘‘Deepfake video detection using recurrent neural networks,’’ in Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS), Nov. 2018, pp. 1–6.
[20] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, ‘‘M2TR: Multi-modal multi-scale transformers for Deepfake detection,’’ in Proc. Int. Conf. Multimedia Retr., Jun. 2022, pp. 615–623.
[21] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, ‘‘A hybrid CNNLSTM model for ideo Deepfake detection by leveraging optical flow features,’’ in Proc. Int. Joint Conf. Neural Netw. (IJCNN), Jul. 2022, pp. 1–7.
[22] H. Li, B. Li, S. Tan, and J. Huang, ‘‘Identification of deep network generated images using disparities in color components,’’ Signal Process., vol. 174, Sep. 2020, Art. no. 107616.
[23] S. J. Nightingale, K. A. Wade, and D. G. Watson, ‘‘Can people identify original and manipulated photos of real-world scenes?’’ Cognit. Res., Princ. Implications, vol. 2, no. 1, pp. 1–21, Dec. 2017.
[24] A. Krizhevsky and G. Hinton, ‘‘Learning multiple layers of features from tiny images,’’ 2009.
[25] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, ‘‘LAION-5B: An open large-scale dataset for training next generation image-text models,’’ 2022, arXiv:2210.08402.

[26] Y. LeCun, Y. Bengio, and G. Hinton, ‘‘Deep learning,’’ Nature, vol. 521, no. 7553, pp. 436–444, 2015.
[27] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, ‘‘Recent advances in convolutional neural networks,’’ Pattern Recognit., vol. 77, pp. 354–377, May 2018.
[28] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, ‘‘A survey of convolutional neural networks: Analysis, applications, and prospects,’’ IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 12, pp. 6999–7019, Dec. 2022.
[29] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, ‘‘XAI—Explainable artificial intelligence,’’ Sci. Robot., vol. 4, no. 37, Dec. 2019, Art. no. eaay7120
[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, ‘‘Grad-CAM: Visual explanations from deep networks via gradient-based localization,’’ in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 618–626.