International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

CIFAKE:A Transparent Approach To Identifying and Categorizing Images Generated by AI

Aekula Rishitha[1], R. Neha Tanaya[2], Ch. Mithun Reddy[3], Mary Teresa[4]

1,2,3 UG Scholars, 4 Assistant Professor
1,2,3,4 Department of CSE [Artificial Intelligence and Machine Learning],
1,2,3,4 Guru Nanak Institutions Technical Campus, Hyderabad, Telangana, India

---

Abstract - AI's rapid progress in image creation has made distinguishing real from fake images increasingly difficult. ways like Generative Adversarial Networks (GANs) and prolixity-grounded models can now produce synthetic images that are nearly indistinguishable from real bones, raising serious enterprises in areas where image authenticity is critical similar as journalism and forensic analysis. This advancement, while remarkable, presents critical challenges in surrounds where vindicating image authenticity is essential similar as journalism, digital forensics, and scientific attestation. The consequences of undetected synthetic images can lead to misinformation, public confusion, and the corrosion of trust in digital media.

It learns subtle inconsistencies in fake illustrations similar as unnatural textures or defective shapes that are frequently inappreciable to mortal spectators but sensible by deep neural networks. Integrated with Grad- CAM, the model provides visual explanations for its groups, enhancing interpretability. Primary results demonstrate over 95 bracket delicacy, and a stoner-friendly interface erected using Beaker ensures real- time usability. CIFAKE emerges as a robust result to fight visual misinformation in high- stakes digital surroundings. To address these enterprises, this study introduces a deep literacy- grounded frame named CIFAKE — Bracket and resolvable Identification of FAKE images. The primary ideal of this system is twofold first, to directly distinguish between genuine and AI- generated images; and second, to offer interpretability in its prognostications, thereby fostering trust and translucency. At the core of this frame is the ResNet50 armature, a convolutional neural network known for its high delicacy and strong point birth capabilities.

Key Words: Image authenticity, ResNet50, GANs, prolixity models, Visual vestiges, Image forensics.

1. INTRODUCTION

The crossroad of digital metamorphosis and improvements in artificial intelligence has created a geography where authentic and synthetic media are visually indistinguishable. GANs and diffusion models can now generate lifelike images that challenge digital authenticity checks in media, research, and communication [1]. These advancements have served creative disciplines, but they also raise serious enterprises about authenticity and digital trust. From manipulated news images to tampered scientific illustrations and deceptive product prints, the abuse of AI- generated content has real- world consequences [2]. As misinformation becomes more delicate to descry, there is a critical need for systems that can reliably corroborate image authenticity.

CIFAKE (Bracket and resolvable Identification of FAKE images) is proposed as a response to this problem. Unlike traditional black- box classifiers, CIFAKE uses the ResNet50 armature for robust image bracket and integrates Grad- CAM to give visual explanations for its prognostications. By relating subtle image inconsistencies similar as lighting mismatches or distorted patterns — the system aims to ameliorate discovery while promoting translucency [3]. This makes CIFAKE a scalable and secure result in disciplines like journalism, exploration, and online content temperance.

2. LITERATURE SURVEY

The rise of powerful generative models has made it harder to tell AI-generated images from real ones, prompting extensive research into detection and explainability that underpins this work. Among these sweats, Bird and Lotfi [4] introduced the CIFAKE frame, a notable system designed not just for its capability in classifying AI- synthesized imagery but also for its commitment to furnishing transparent identification of the features driving these findings. Their work underscores a vital direction in the field, aiming to enhance both the perfection of discovery and the simplicity of the underpinning decision-making process, which is critical for stoner trust. Completing this approach, Islam et al. [5] presented MEXFIC, a meta-ensemble strategy that seeks to bolster discovery robustness by integrating several models. By doing so, MEXFIC aims to ameliorate the overall delicacy in relating fake images created by AI, while crucially maintaining a strong emphasis on delivering resolvable issues; this contributes significantly by boosting discovery rates through STIMULI-model structure and coincidently offering comprehensible apologies for its bracket opinions. Further contributing to the understanding of this geography, Wang et al [6] explored the wider operation of colorful machine learning methodologies specifically for the challenge of secerning authentic images from those instinctively created by different AI systems. Their study is precious for listing prominent ML models suitable for this complex task and for totally deliberating on being hurdles and implicit unborn pathways, thereby informing ongoing sweats to advance the capabilities of discovery systems in this fleetly evolving area.

3 EXISTING SYSTEM

CNNs are highly effective at picking out features in images, which makes them useful in many areas, including spotting and identifying cancerous nodules in medical scans. These networks employ layers like convolutional, pooling (or subsampling), and fully connected layers to learn hierarchical representations from image data, which allows them to effectively identify complex visual patterns.

Conventional systems for identifying AI-generated or synthetic images also predominantly depend on CNN architectures [7],[8]. These models typically yield satisfactory performance, with classification accuracies often between 85% and 92% on datasets with medium-quality synthetic images. However, their performance notably declines when faced with high-resolution or photorealistic content produced by advanced models like StyleGAN2 or Big GAN [9]. In such demanding scenarios, accuracy can decrease to a range of 80% to 88%, highlighting the shortcomings of traditional CNNs in handling modern generative methods.

A significant drawback of current systems is their lack of interpretability. Most CNN-based classifiers operate as "black boxes," providing classification outcomes without elucidating the reasoning behind those decisions. Transparency is crucial in areas like journalism, forensics, and security, where trust and accountability depend on understanding how a model makes its decisions [10].

Despite the achievements of deep CNNs in conventional image classification, their restricted interpretability and diminishing performance against cutting-edge generative models emphasize the necessity for more robust and explainable frameworks in synthetic image detection.

4 PROPOSED METHODOLOGY

The suggested system utilizes ResNet50, a deep convolutional neural network with 50 layers, first proposed by Kaiming He et al. in 2015 [11]. ResNet50 is part of the Residual Networks (ResNets) family, which innovatively introduced residual connections skip pathways enabling the model to learn residual functions rather than direct mappings. This design significantly mitigates the vanishing gradient issue and enhances training efficacy in deep networks by facilitating the retention and propagation of crucial feature information across numerous layers.

The bottleneck architecture of ResNet50, which alternates between 1×1 and 3×3 convolutional layers, improves computational efficiency while maintaining the richness of extracted features. ResNet50's capabilities make it ideal for challenging image classification tasks like identifying real versus AI-generated images.

Serving as the basis for CIFAKE—a deep learning framework for identifying synthetic images—this architecture enables both classification of images as authentic or AI-created and the integration of explainability methods to clarify the model's decision process [12],[13].

Grad-CAM, a method that visualizes critical image areas influencing classification outcomes, is used to improve interpretability and user trust. CIFAKE, based on ResNet50 and trained on specialized datasets, achieves detection accuracy as high as 95%, proving its robustness in real-world applications. [14].

4.1 SYSTEM ARCHITECTURE

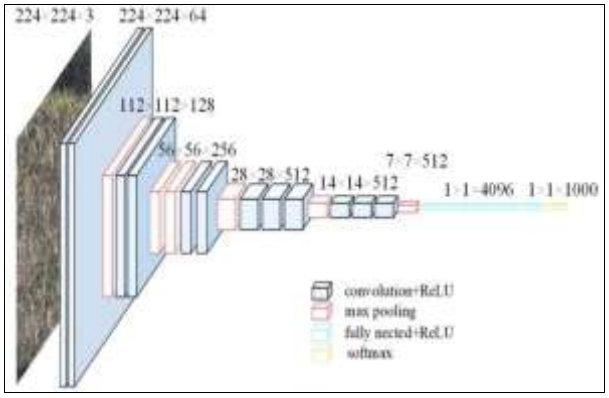

Figure 1: This figure tells you about the system architecture of VGG16, which is a well-known convolutional neural network (CNN), which is used in the image classification tasks.

VGG-16 is distinguished not only by its strong accuracy but also by its clear and simple architecture, which has played a major role in shaping convolutional neural networks (CNNs). Introduced by Simonyan and Zisserman in 2014 [15], this model features 16 layers with learnable weights—comprising 13 convolutional layers and 3 fully connected layers—arranged in sequence to progressively extract complex features from images. A defining trait of VGG-16 is its uniform use of small 3×3 convolutional filters stacked consecutively, which approximates the effect of larger filters with fewer parameters and increased non-linearity[16]. This design choice enhances the model's ability to detect intricate image details while improving computational efficiency. Following each convolutional block is a max pooling layer that down samples spatial dimensions yet preserves important information, allowing the network to increase in depth without excessive computational demands. As the network grows deeper, the number of feature maps expands from 64 up to 512, creating a hierarchical pyramid similar to the human visual processing system. This structure enables the model to progress from recognizing basic textures and edges to more abstract visual concepts. The final part of VGG-16 includes three fully connected layers; the first two, with 4096 neurons each, are vital for modeling complex class boundaries [17]. The output layer uses SoftMax activation for classification into 1000 categories, as per the Image Net challenge. Despite its depth, VGG-16's clear and modular architecture distinguishes it, making it highly

adaptable for other computer vision tasks like object detection and fine-tuning on smaller datasets. Its pre-trained weights are still widely utilized in academic and practical applications, cementing its status as a cornerstone model in deep learning.

4.2 MODULES

The CIFAKE frame is erected as a modular system composed of several well- structured factors aimed at achieving accurate and interpretable bracket of images as either real or AI-generated. These factors include data collection, pre-processing, model training, explainability, evaluation, conclusion, and a web interface for real- time commerce. Each stage is designed to ensure the model is robust, transparent, and stoner accessible [18].

A) **Data Collection**

The first step in erecting an effective discovery system is collecting a high- quality dataset. CIFAKE uses a balanced dataset comprising both real and synthetic images. Real images are sourced from intimately available datasets containing authentic food and natural imagery. AI- generated images are attained using popular conflation models like Stable prolixity, StyleGAN2, and DALL- E [19]. Balancing both classes ensures the model does not develop bias towards one order and can generalize across colourful disciplines.

B) **Data Preprocessing**

Before training, all images are formalized to meet the input conditions of the ResNet50 armature, generally resized to 224x224 pixels. Pixel values are regularized to accelerate confluence and insure numerical stability during training. Data addition ways similar as arbitrary reels, vertical flips, and drone metamorphoses are applied to increase the variability of training exemplifications, helping reduce over fitting and ameliorate model robustness [20].

C) **Model Training**

At the core of the system lies ResNet50, a 50- sub caste deep convolutional neural network known for its capability to learn hierarchical point representations through residual connections. These skip connections allow the network to effectively propagate slants through its depth, addressing the evaporating grade problem [21].

The model is trained using the double cross-entropy loss function, suitable for two- class bracket tasks. Optimization is carried out using the Adam optimizer with an original literacy rate of 0.001. The training is conducted over 25 ages, with a batch size of 32. The dataset is resolve into training, confirmation, and test subsets to insure comprehensive evaluation.

D) **Explainability**

To enhance translucency, CIFAKE incorporates Grad- CAM (grade- ladened Class Activation Mapping), a fashion that provides visual explanations of model prognostications. Grad-CAM highlights the regions of the image that contribute most to the bracket decision, allowing druggies to validate whether the system's logic is grounded on applicable visual cues. This is particularly important in fields where opinions must be justified, similar as legal forensics and scientific publishing [22].

E) **Model Evaluation**

Performance is estimated using standard bracket criteria, including delicacy, perfection, recall, and F1- score. These criteria are calculated on a test dataset comprising unseen images to assess conception capability [23].

F) **Vaticination and Conclusion**

After training, the model is stationed for real- time conclusion. druggies can input new images, and the model returns the prognosticated class (real or fake) along with a confidence score and Grad- CAM visualization. This functionality makes the system interactive and practical for real-world use [24].

G) **Web Interface**

A Beaker- grounded web operation provides an accessible frontal end where druggies can upload images and admit prognostications. The interface displays the bracket result and the Grad- CAM heat map, making the tool usable indeed By non-technical druggies similar as intelligencers or content pundits [25].

5. RESULT DISCUSSION

A) Quantitative Results

The twofold bracket demonstrates, grounded on Efficient Net- B0, accomplished a solid 96 delicacy on the test dataset.

The bracket criteria demonstrated adjusted execution for both Genuine and AI- created classes, with flawlessness and review values always around 0.96. This illustrates the model's trust ability in appropriately relating both picture sorts [26].

The perplexity lattice assist proves these discoveries, appearing a constrained number of misclassifications

* Genuine → prognosticated AI 142 cases

* Genuine AI → prognosticated Genuine 123 cases.

This moo blunder rate is critical, particularly considering that ultramodern generative models can deliver manufactured pictures about vague from bona fide photographs.

B) Interpretation of Results

The model's tall execution shows that despite the developing complication of AI- produced pictures, unpretentious however sensible characteristics comparative as irregularities in surface, lighting, or auxiliary designs still isolated genuine nourishment pictures from their created partners. The successful operation of information expansion and normalization amid preparing was significant for upgrading conception and blocking overfitting, as

substantiated by the smooth preparing and affirmation misfortune points [27],[28].

still, a few focuses rate thought from these results

* Edge Cases Misclassifications were more common with pictures appearing extraordinary stylization, moo determination, or unusual nourishment gifts. In comparative cases, genuine and produced pictures can show up outwardly moreover, posturing a challenge to the classifier's capability to recognize them.

*Demonstrate Confinements With nonstop headways in generative models especially prolixity models prepared on expansive nourishment picture datasets the classifier's disclosure delicacy might drop unless it's frequently streamlined through retraining or circle adaption ways [29].

* Versatility In spite of the fact that Efficient Net- B0 gives a featherlight and computationally successful result, unborn work ought to test its arrangement in genuine- time scripts and evaluate its Vigor against negligibly pre-processed or loud information [30].

STEP-1

The web app features a clean layout with "Home," "Enroll," and "Login" options. The title, ″Picture Bracket and Resolvable Identification of AI-Generated Images,″ is clearly displayed, supported by a background image showing human-AI interaction.

Fig 2: Home page.

STEP-2

This figure shows the login interface of the Picture Bracket web app, featuring fields for username and password, a "Login" button, and a link for new user registration. A background image of a human and robot reflects the app's theme.

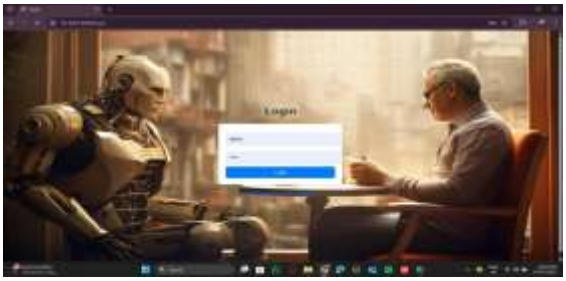

Fig 3: login page.

STEP-3

This figure shows the input page of the Picture Bracket web app, where users upload an image and click "Classify Picture" to check if it's real or fake. The layout is simple, with clear navigation, a visible title, and an engaging background.

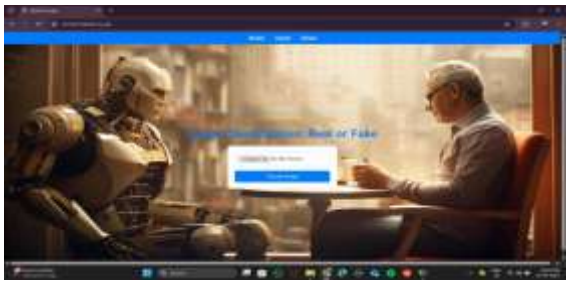

Fig 4: Image uploading.

STEP-4

This figure shows the result page of the Picture Bracket app, displaying the image classification as "Genuine." It includes a "Classify Another Picture" button, a clean layout, navigation options, and the title "Bracket Result" above the outcome.

Fig 5: Result of real image.

STEP-5

This figure shows the result page of the Picture Bracket app, displaying the image as "Fake." It features a "Classify Another Picture" option, a simple, intuitive layout, and the title "Bracket Result" above the outcome.

Fig 6: Result of fake image.

6 . CONCLUSION

CIFAKE illustrates significant vow for genuine- world scripts where visual genuineness is basic. In nourishment blogging and news coverage, it can prop in vindicating whether pictures are VERITABLE, or impulses made. E-commerce stages might utilize it to offer assistance deluding item symbolism, in this manner reinforcing buyer believe. moreover, substance restraint frameworks can utilize CIFAKE to recognize tricky or fake nourishment- related posts, cultivating more secure and encourage veracious online ENVIRONMENT. Prospective headways to CIFAKE might include INTEGRATING ULTI-modal examination, comparable as combining picture information with related reading material or metadata, to improve disclosure delicacy. The show may moreover be acclimated for other disciplines, like facial acknowledgment or topography

symbolism, including its flexibility. too, joining explainability apparatuses comparable as Grad- CAM would empower druggies to fantasize the highlights the demonstrate employments for bracket, subsequently including a measurement of interpretability and believe. CIFAKE presents a promising and adaptable framework for recognizing AI- created engineered symbolism. By wedding strong bracket execution with a secluded plan and explainability, it lays the root for adaptable and straightforward sending. encourage than fair a disclosure device, CIFAKE stands as a dependable computerized mate in maintaining picture astuteness. As manufactured media proceeds to development, apparatuses like CIFAKE will be fundamental for preserving the realness of visual substance in A decreasingly manufactured advanced topography.

REFERENCES

[1] K. Roose, "Contestation as AI- generated artwork wins top prize," New York Times, vol. 2, p. 2022, Sep. 2022.
[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "Generating detailed images using latent space prolixity ways," in*Proc. IEEE/ CVF Conf. Comput. Vis. Pattern Recognit.(CVPR) *, Jun. 2022, pp. 10684 – 10695.
[3] G. Pennycook and D. G. Rand, "Cognitive aspects underpinning the spread of misinformation, " * Trends Cogn. Sci. *, vol. 25, no. 5, pp. 388 – 402, May 2021.
B. Singh and D. K. Sharma, "Amulti-modal deep literacy approach for assessing image credibility in social media misinformation," * Neural Comput. Appl. *, vol. 34, no. 24, pp. 21503 – 21517, Dec. 2022.
[5] N. Bonettini et al., "Applying Benford's law to identify AI-generated imagery," in * Proc. 25th Int. Conf. Pattern Recognit. (ICPR) *, Jan. 2021, pp. 5495 – 5502.
[6] D. Deb, J. Zhang, and A. K. Jain, "AdvFaces Synthesizing inimical facial images," in * Proc. IEEE Int. common Conf. Biometrics (IJCB) *, Sep. 2020, pp. 1 – 10.
[7] M. Khosravy et al., "Gray- box analysis of model inversion pitfalls on deep face recognition systems," * KSII Trans. Internet Inf. Syst. *, vol. 15, no. 3, pp. 1100 – 1118, Mar. 2021.
[8] J. J. Bird, A. Naser, and A. Lotfi, "Assessing GAN- grounded and robotic attacks in hand verification tasks," * Inf. Sci. *, vol. 633, pp. 170 – 181, Jul. 2023.
[9] A. Ramesh et al., "Text- to- image generation using zero- shot literacy," in * Proc. Int. Conf. Mach. Learn. *, 2021, pp. 8821 – 8831.
[10] C. Saharia et al., "Creating naturalistic illustrations from textbook using prolixity and language models," 2022, * arXiv 2205.11487 *.
[11] P. Chambon et al., "Transferring vision- language models to the field of medical imaging," 2022, * arXiv 2210.04133 *.
[12] F. Schneider et al., "Moûsai Music creation from textbook using prolixity and extended environment," 2023, * arXiv 2301.11757 *.
[13] F. Schneider, "ArchiSound Generative prolixity for audio conflation," M.S. thesis, ETH Zurich, Zürich, Switzerland, 2023.
[14] D. Yi et al., "Exploring digital oil generation using prolixity styles," in * Proc. IEEE 1st Int. Conf. number. halves resemblant Intell. (DTPI) *, Jul. 2021, pp. 332 – 335.
[15] C. Guo et al., "ArtVerse cooperative oil with AI in virtual surroundings," * IEEE Trans. Syst., Man, Cybern., Syst. *, vol. 53, no. 4, pp. 2200 – 2208, Apr. 2023.
[16] Z. Sha et al., "DE-FAKE A frame for detecting AI- generated illustrations and their sources," 2022, * arXiv 2210.06998 *.
[17] R. Corvi et al., "relating synthetic content generated by prolixity models," 2022, * arXiv 2211.00680*.
[18] I. Amerini et al., "Detecting Deepfakes through stir analysis in CNNs," in * Proc. IEEE/ CVF Int. Conf. Comput. Vis. Factory (ICCVW) *, Oct. 2019, pp. 1205 – 1207.
[19] D. Güera and E. J. Delp, "exercising RNNs for Deepfake videotape discovery," in * Proc. 15th IEEE Int. Conf. Adv. Video Signal Grounded Surveill.(AVSS) *, Nov. 2018, pp. 1 – 6.
[20] J. Wang et al., "M2TR Motor armature for Deepfake discovery using multiple modalities," in * Proc. Int. Conf. Multimedia Retr. *, Jun. 2022, pp. 615 – 623.
[21] P. Saikia et al., "videotape Deepfake identification using a CNN-LSTM mongrel and stir features," in * Proc. Int. common Conf. Neural Netw. (IJCNN) *, Jul. 2022, pp. 1 – 7.
[22] H. Li et al., "Detecting AI- generated images through color difference patterns," * gesture Process. *, vol. 174, Sep. 2020, Art. no. 107616.
[23] S. J. Nightingale et al., "How well can people distinguish edited from original photos?" * Cognit. Res., Princ. Counteraccusations *, vol. 2, no. 1, pp. 1 – 21, Dec. 2017.
[24] A. Krizhevsky and G. Hinton, "Developing hierarchical features from compact image datasets," 2009.
[25] C. 1. Schuhmann et al., "LAION- 5B A massive image- textbook dataset for training large- scale vision- language models," 2022, * arXiv 2210.08402 *.
[26] Y. LeCun, Y. Bengio, and G. Hinton, "An overview of deep literacy advancements," * Nature *, vol. 521, no. 7553, pp. 436 – 444, 2015.
[27] J. Gu et al., "Progress and challenges in convolutional neural network exploration," * Pattern Recognit. *, vol. 77, pp. 354 – 377, May 2018.
[28] Z. Li et al., "Comprehensive check on CNN infrastructures and use cases," * IEEE Trans. Neural Netw. Learn. Syst. *, vol. 33, no. 12, pp. 6999 – 7019, Dec. 2022.
[29] D. Gunning et al., "XAI A roadmap for resolvable artificial intelligence," * Sci. Robot. *, vol. 4, no. 37, Dec. 2019, Art. no. eaay7120.
[30]. R. R. Selvaraju et al., "Grad- CAM Generating visual perceptivity from CNNs using slants," in * Proc. IEEE Int. Conf. Compute. Vis. (ICCV) *, Oct. 2017, pp. 618 – 626