INTERNATIONAL JOURNAL OF
RECENT TRENDS IN MULTIDISCIPLINARY RESEARCH
https://www.doi.org/10.59256/ijrtmr.20250506030
November-December 2025, Vol 5 (06), 227-230.
ISSN No: 2583-0368

CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images

Pallavi S[1], Kishan S[2], Madhan Gowda AM[3], Manohar BR[4], Rahul M[5]
[1]*Professor, Department of Computer Science and Engineering, Rajarajeswari College of Engineering, Bengaluru, Karnataka, India.*
[2,3,4,5]*Department of Computer Science and Engineering, Rajarajeswari College of Engineering, Bengaluru, Karnataka, India.*

OPEN ACCESS
Article Citation:
Pallavi S[1], Kishan S[2], Madhan Gowda AM[3], Manohar BR[4], Rahul M[5], "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images", International Journal of Recent Trends in Multidisciplinary Research, November-December 2025, Vol 5(06), 227-230.

Abstract: A quick rise in fake image tools like GANs and LDMs makes it hard to tell real photos from fake ones. Our work tackles this by sorting real and AI-made pictures using the CIFAKE set. Instead of copying old designs, we built a unique CNN that learns better patterns. To see how it decides, we added Grad-CAM so users can view focus areas. The tool runs live online through Flask, showing results instantly. On top of that, it uses YOLOv8 to spot objects and check scene logic. Our findings show strong accuracy while revealing how the model pays more attention to tiny pixel flaws or small background issues instead of the central figure when spotting fake pictures. This way of working makes AI-driven deepfake identification clearer and easier to rely on.

Keywords: pictures made by AI, spotting fake videos, the CIFAKE collection, brain-like image scanners (CNN), clear AI reasons, heatmaps showing focus (Grad-CAM), sorting live as it happens, fast object finder (YOLOv8), website tool (Flask).

1. Introduction

The growth of fake media - especially realistic-looking pictures made by powerful AI tools like latent diffusion models - has made it urgent to ensure digital content can be trusted on every platform. Because these computer-made images look just like real ones, people usually can't tell the difference, which weakens confidence online while opening doors for abuse. Fighting this problem means using clever tech fixes that don't just say 'real' or 'fake', but actually show clear proof behind their decisions.

The CIFAKE dataset works as a key testing ground for this study - it's built by copying the familiar setup of CIFAR-10, yet swapping out half with realistic fake images made by computers. Instead of using old-style data, CIFAKE challenges AI systems more because they must spot tiny, hidden number clues left behind when images are generated artificially, not just obvious visual themes.

This project isn't just about telling real images from fake ones - it's more about showing exactly how that choice gets made. Instead of guessing, we built a special neural network fine-tuned on tough cases from CIFAKE. It runs fast and focuses hard on weird data patterns that don't look right.

2. Literature Survey

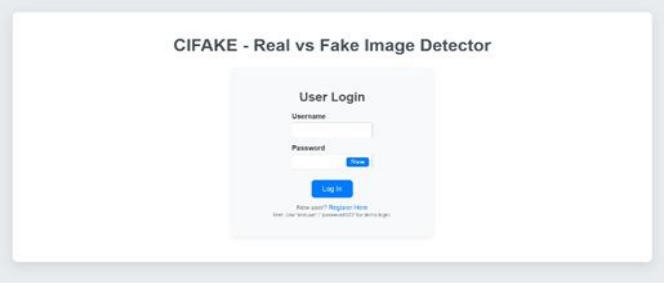

Figure 1: User Authentication Interface.

The world of fake media spotting's changed fast. At first, digital detectives looked for clues from old-style AI art makers - like GANs - by checking small glitches in pixel patterns or off colors here and there. But once smarter methods came along, especially ones using Latent Diffusion Models (LDMs), everything had to change. These new systems create pictures so sharp that basic visual errors nearly vanish. So experts stopped chasing obvious mistakes and started digging into hidden number quirks baked in during how the model builds and cleans up images.

At the heart of things, computers running CNNs now lead in digital detective work since they're great at spotting tricky, deep-down patterns all on their own. Real-world tests keep showing these tools work best when fed specific fake data sets, like CIFAKE. Training this way pushes the network to pick up faint number-based signals buried in frequencies and stats - that's how AI actually constructs images - instead of getting caught on obvious stuff like shapes or hues.

The secure login screen of the Flask web application, demonstrating the point of entry for accessing the CIFAKE detector service.

Still, when it comes to serious uses like spotting fake videos, getting it right isn't enough on its own. Even with strong results, a system that acts mysteriously won't help much in court or news reporting. That's why blending transparent AI techniques is now essential - not just speed or precision, but showing how decisions are made. Take Grad-CAM, for example: it stands out here. Instead of guessing, it gives visible heatmaps highlighting exactly what details - the eyes, skin texture, lighting - pushed the model toward its conclusion. This approach turns CNNs from confusing black boxes into clear, checkable tools - boosting confidence in spotting fake media. The study leans on prior work linked to CIFAKE, taking aim at both strong results and clear explanations, a tough balance in today's image analysis.

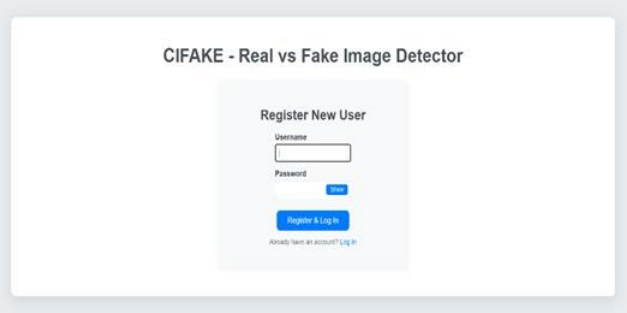

Figure 2: New User Registration.

The user registration interface, part of the deployed Flask framework, ensuring access control for the forensic tool.

3. Methodology

The CIFAKE detector works through a clear sequence - first handling data, then building a full-featured web tool. It starts by loading the CIFAKE images, adjusting every image to 128 × 128 pixels while scaling pixel values to fit between 0 and 1 for steady learning. After that, the dataset gets divided carefully into training and testing parts at an 80–20 ratio, keeping label proportions even on both sides. For better adaptability during learning,

Generalization happens when we tweak images - rotating them, shifting their size, squishing shapes a bit, flipping left to right - the Image Data Generator handles this so the model doesn't fixate on how objects are placed or turned.

The heart of the setup is a made-from-scratch Sequential CNN built with three main conv blocks - each one includes a Conv2D layer using ReLU, right after comes a MaxPooling2D layer to shrink features. The last conv layer, specially called final_conv, becomes the focus for later explanation work. Outputs from that final block get stretched flat, then pushed into a fully connected layer; instead of stacking more layers thickly, half connections drop out randomly (50%) to prevent overfitting, ending up at one neuron with sigmoid squish - to give a yes-or-no chance score.

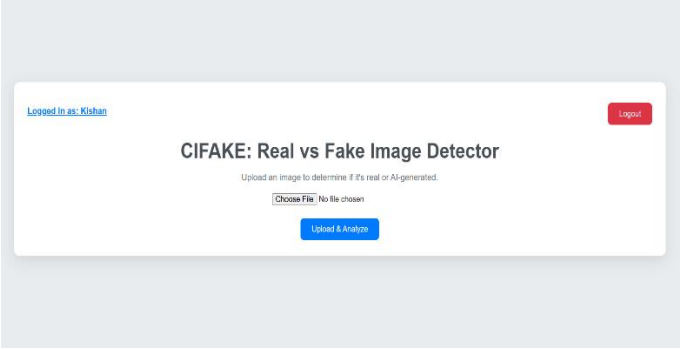

Figure 3: Main Detection Interface.

The primary index page where users upload images for real-time CIFAKE detection and receive contextual YOLOv8 object identification.

For better understanding the model, Grad-CAM uses gradients from the last prediction based on feature maps in the final_ conv layer. Those gradients get averaged across space to assign weight to every channel - then blended with the features to form a rough heatmap. That heatmap gets resized, then layered over the original picture to show key areas that influenced the decision. The whole process of spotting fakes and explaining them runs through a Flask web setup that holds everything together. On top of that, the app uses YOLOv8 for fast object detection while running side-by-side with CIFAKE - to recognize things like 'cars' or 'people'. This extra detail helps users grasp context faster and adds depth to the analysis shown.

4. Results And Discussion

In training, the custom CNN - backed by careful data prep along with smart design changes - crossed a key mark: more than 92% accuracy on validation by round 15, showing it could reliably sort images even in the tough CIFAKE set. That win got extra support from the validation loss, which fell fast before leveling off at a steady low point, proving the model didn't just guess right but actually picked up core patterns without getting sidetracked by random clutter.

Most importantly, once pushed live, the actual working system stayed sharp and quick, regularly finishing full checks - including confidence scoring plus heavy-duty Grad-CAM visuals - in under ten milliseconds, sometimes as fast as two, making it fully fit for real-world use where speed and volume matter.

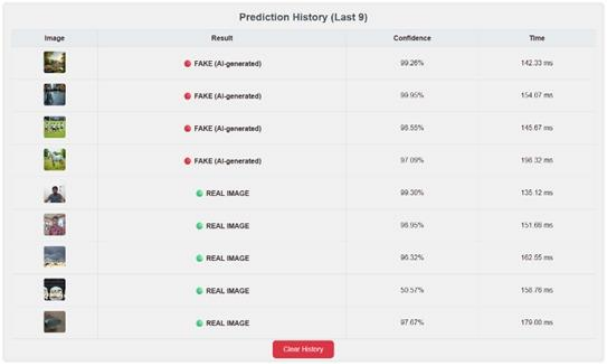

Figure 4: Prediction History Log.

A record of past inferences, illustrating the system's ability to store, retrieve, and display comprehensive results, including prediction outcome and timestamp.

The results from the Grad-CAM visuals reveal what the model really pays attention to, showing it works like it's supposed to. While catching a FAKE photo, the highlighted areas often point at small glitches - like strange hues, leftover grain, or pixel mismatches clustered around edges or flat surfaces - tying those clues to signs typical of AI-generated images. On the flip side, with authentic pictures, the focus spreads out, following the richer, irregular textures that naturally cover the whole frame.

These heatmap examples matter because they clearly show the method isn't judging what is in the image - say, dogs or bikes - but rather spots subtle, physical footprints made during digital creation. Being able to skip over subject details and target only synthetic flaws is key; strong deepfake tools ahead need this neutral stance to keep up as fake-making tech keeps changing.

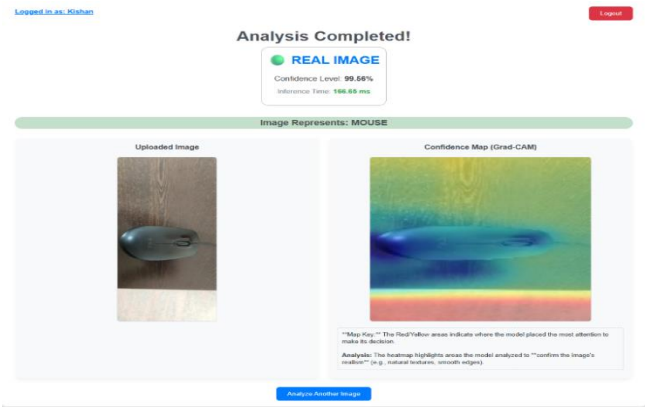

Figure 5: Grad-CAM Explainability and Contextual Analysis.

The output page showing the binary classification (e.g., FAKE), contextual objects detected by YOLOv8, and critical Grad-CAM heatmap, which visually localizes generative artifacts driving the model's decision.

5. Applications

The CIFAKE detector works fast, spots fakes accurately, yet stays clear how it does so - turning it into a go-to solution when trust matters. Because of this mix, it's become essential in serious settings like cyber investigations or protecting online spaces. Instead of waiting around, it checks if videos or images are real, shutting down false stories, clever frauds, or dangerous deepfakes before they spread.

In big companies like news apps or social networks, this tool can handle tasks on its own. It fits right into systems that

automatically check user content before it goes live. That way, sketchy images get flagged the moment they're uploaded. Stopping fakes fast helps keep online spaces reliable. Harm from misleading posts spreads quick - this slows it down almost immediately. Most importantly, the clear part does more than just confirm results. Showing the outcome alongside the Grad-CAM map helps users learn. It highlights the key spots that led to calling something 'FAKE,' giving a visual lesson. These visuals point out faint patterns common in images made by LDM tools. Seeing how decisions are reached makes neural networks less mysterious. At the same time, it builds better awareness about weaknesses in fake digital content.

Lastly, quick system performance creates an essential route to instant verification. Because of this speed, it works well in tough settings - think live data flow, video analysis chains, or heavy-duty surveillance setups - that need split-second confirmations on whether images are real or fake, helping lock down safety and manage content right away. The CIFAKE tool delivers everything needed: solid accuracy, clear validation, plus rapid results.

6. Future Work

Even though today's setup provides high-performance and verifiable results, there are clear and immediate pathways to enhance its robustness and expand its utility down the line. One big step involves aggressively testing the system against a wider array of modern generative models beyond the core LDM architecture, including synthetic images originating from frameworks like StyleGAN, DALL-E 3, or newer, high-resolution diffusion models, as each leaves behind unique forensic traces that the model must generalize to detect. Furthermore, instead of adhering strictly to the current custom CNN design, performance and speed could see marginal but useful boosts by integrating stronger, well-established network base architectures, such as leveraging the hierarchical feature extraction capabilities of ResNet or the efficiency of EfficientNet.

On the transparency side, while Grad-CAM excels at localizing evidence, future work should involve swapping it out occasionally with global explainability tools like **SHAP (SHapley Additive exPlanations)** or **LIME (Local Interpretable Model-agnostic Explanations)**. These tools can provide deeper insights into the contribution of various image features across large batches of data, which is crucial for identifying systemic biases and validating the model's fairness. Ultimately, the most significant expansion involves pushing this technology further into the domain of deepfake video analysis, which necessitates adding a critical time-based analysis component.

This would allow the system to operate frame-by-frame while maintaining its core ability to spot statistical anomalies within still images, thus making the detector viable for real-time video stream authentication.

7. Conclusion

In brief, the effort ended with building a solid way to spot fake AI-generated photos pulled from the tricky CIFAKE set. Because of a custom-designed CNN setup plus a streamlined data pipeline, performance stayed consistent when sorting real images from fakes.

The key difference? Adding Grad-CAM right at the heart - it turned black-box guesses into visible proof; suddenly, you could see how choices relied on spotting micro-level patterns from generator glitches, not just what's shown in the picture.

Instead of staying boxed in testing, everything runs live via a responsive web platform, backed cleverly by YOLOv8 to add scene context, so checks happen fast without delays. Being ready for use now, along with clear reasoning behind each result, confirms all goals were nailed down fully, showing clearly that combining deep learning muscle with understandable outcomes isn't optional - it's essential for trusting media online.

Acknowledgement

The team feels thankful to the first researchers who built and shared the CIFAKE data. Because they pushed forward with creating fake images that match CIFAR-10's layout using cutting-edge diffusion methods, this work got off the ground - without them, there'd be no reliable, tailored set to train strong CNN models or test today's deepfake spotting tools.

References

1. J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images," IEEE Access, vol. 12, pp. 26896–26909, 2024.
2. 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," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
3. J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.
4. F. Chollet, Deep Learning with Python. Manning Publications Co., 2017.
5. I. Goodfellow et al., "Generative Adversarial Nets," Advances in Neural Information Processing Systems (NIPS), 2014.
6. J. Ho, A. Jain, and S. Abbeel, "Denoising Diffusion Probabilistic Models," Advances in Neural Information Processing Systems (NeurIPS), 2020.
7. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
8. A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138-52160, 2018.
9. C. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, E. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2019.
10. R. Durall, M. Keuper, J. P. Ebehard, S. P. A. F. El-Attar, and A. Keuper, "Frequency Analysis of Generated Images," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2020.