International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 8 Issue 5, Sep-Oct 2024 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Detection of AI-Generated Images

Misal Thakre[1], Satyam Kadu[2], Ronit Bera[3], Rohit Atre[4], Prof. Shreya Bhanse[5]

[1,2,3,4]School of Science, G H Raisoni University, Amravati, Maharashtra, India
[5]Assistant Professor, G H Raisoni University, Amravati, Maharashtra, India

ABSTRACT

Generative AI has gained enormous interest nowadays due to new applications like ChatGPT, DALL E, Stable Diffusion, and Deepfake. In particular, DALL E, Stable Diffusion, and others (Adobe Firefly, ImagineArt, etc.) can create images from a text prompt and are even able to create photorealistic images. Due to this fact, intense research has been performed to create new image forensics applications able to distinguish between real captured images and videos and artificial ones. Detecting forgeries made with Deepfake is one of the most researched issues. This paper is about another kind of forgery detection. The purpose of this research is to detect photorealistic AI-created images versus real photos coming from a physical camera. Id est, making a binary decision over an image, asking whether it is artificially or naturally created. Artificial images do not need to try to represent any real object, person, or place. For this purpose, techniques that perform a pixel-level feature extraction are used. The first one is Photo Response Non-Uniformity (PRNU). PRNU is a special noise due to imperfections on the camera sensor that is used for source camera identification. The underlying idea is that AI images will have a different PRNU pattern. The second one is error level analysis (ELA). This is another type of feature extraction traditionally used for detecting image editing. ELA is being used nowadays by photographers for the manual detection of AI-created images. Both kinds of features are used to train convolutional neural networks to differentiate between AI images and real photographs. Good results are obtained, achieving accuracy rates of over 95%. Both extraction methods are carefully assessed by computing precision/recall and F1 -score measurements. The proliferation of AI-generated images, often referred to as deepfakes or synthetic media, has revolutionized how digital content is created, consumed, and shared. While these advancements offer immense creative potential, they also pose significant challenges, particularly in terms of authenticity and misinformation. This paper explores the growing need for AI-generated image detection, especially in social media applications, and delves into the methods used for distinguishing between human-made and AI-generated content. It outlines the technical challenges, the potential societal impact, and strategies for implementing robust detection mechanisms in social media platforms.

How to cite this paper: Misal Thakre | Satyam Kadu | Ronit Bera | Rohit Atre | Prof. Shreya Bhanse "Detection of AI-Generated Images" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-8 | Issue-5, October 2024, pp.805-810, URL: www.ijtsrd.com/papers/ijtsrd69453.pdf

IJTSRD69453

KEYWORDS: *Computer Vision, Object Detection, Image Classification*

I. INTRODUCTION

Nowadays, generative artificial intelligence is one of the top themes of computer engineering research. The emergence of transformers [1] as a key tool for generating content has opened a world of new applications where automated systems can create productions that, until now, were exclusive to human authorship. Transformers were first used for automated translation systems, where a first processing stage (the encoder) transforms the input text into a numerical representation of text meaning; then a second stage (the decoder) converts (like an inverse transform) those intermediate data into text in another language [2]. Besides neural machine translators, other impressive applications have arisen. Famously, ChatGPT is a conversational engine created with a decoder transformer [3]. Using transformers, models for translating regular text into images have also been developed. The most known and documented examples of these last ones are DALL E [4,5] and Stable Diffusion [6]. But other examples have quickly been released, like OpenArt [7], ImagineArt Adobe Firefly and many others. These artificial image generators have reached the point where they may create photorealistic images that can make humans hesitate on whether a particular image is coming from a camera or is an artificial creation. As an example, in Figure 1, three AI-created images are presented. They were created by three different engines: DALL E 2, Stable Diffusion, and OpenArt (after testing many applications, these three were found the most appropriate for photorealistic images; other models are good at producing drawings or illustrations and not so much at imitating real photographs). The prompt was the same for the three images: “realistic photo, a portrait of a dog in a library, Sigma 85 mm f/1.4”. Note that details about the lens were added (85 mm focal lens, f/1.4 numeric aperture); this is a common trick used for getting more realistic results. In the same figure, we also present three real photographs that will be processed later. The purpose of this work is to make a binary decision between two options: AI image (fully created AI image) and real image. Sensors 2023, 23, x FOR PEER REVIEW 2 of 15 photorealistic images that can make humans hesitate on whether a particular image is coming from a camera or is an artificial creation.

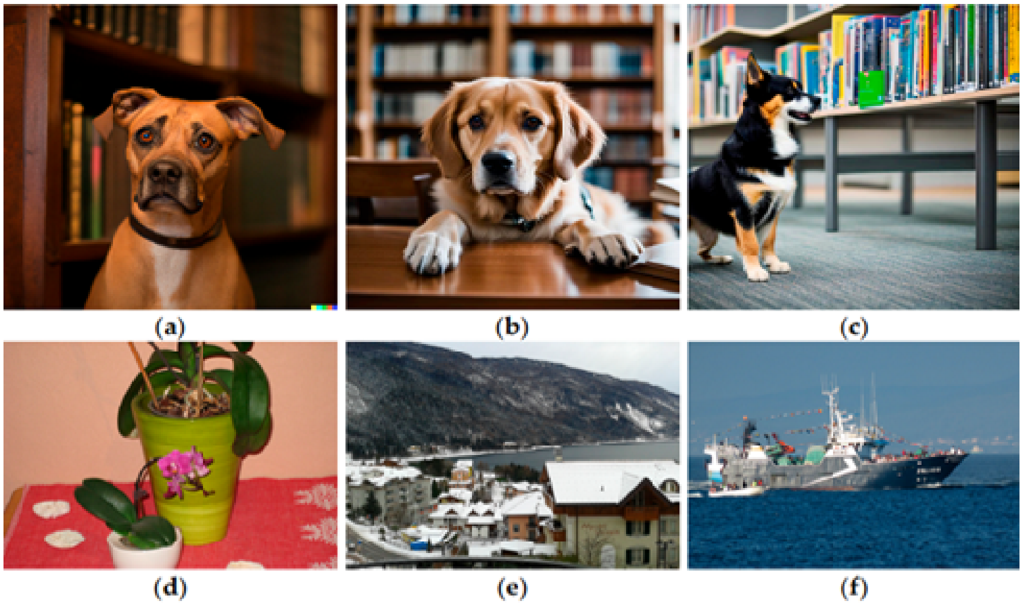

Figure 1. (a) DALL E 2 image, (b) Stable Diffusion image, and (c) OpenArt image. (d–f) Real photos

Another impressive AI application is Deepfake . Deepfake can create photos and videos mixing plausible information from previous photos and/or videos. For example, creating a video of a person mixing the body of one given individual and the face of another one. The potential danger of this technology being used for fraud or other illegal purposes (defamation, pornography, etc.) has sparked much research in the field of detecting Deepfake image creation . For example, in, Rössler et al. start by creating a large dataset of fake videos. In authors exploit what is, perhaps, the most intuitive method: finding image artifacts that can reveal synthetic content. In, a system called “FakeCatcher” is described; this system generated) images, which is another type of problem as they deal with images that were created with intensive human intervention. For a similar need, Google has recently announced a new tool called SynthID which adds an invisible watermark to AI-generated images so that they can be identified. Note that this will identify AI images only if the creation engine watermarks them. Because of the need to classify whole images with no assumption about image content, the system was designed based on methods from other image forensics applications.

The main idea is to extract some relevant t information from images before applying a convolutional neural network. Convolutional neural networks (CNNs) are very useful in distinguishing between classes that are visually different for humans, like digit classification distinguishing objects relevant for making driving decisions in real traffic, and many other similar applications Nevertheless, in this case, classes are not visually different, and that suggests that the direct application of CNNs could not be very useful (besides the experience from the Deepfake case). For this reason, pixel-wise feature extraction was used. This means using processing stages that convert images into other images with the same size (it converts each pixel to a new pixel) but containing a reduced amount of information that should be relevant to the particular problem of distinguishing AI images.

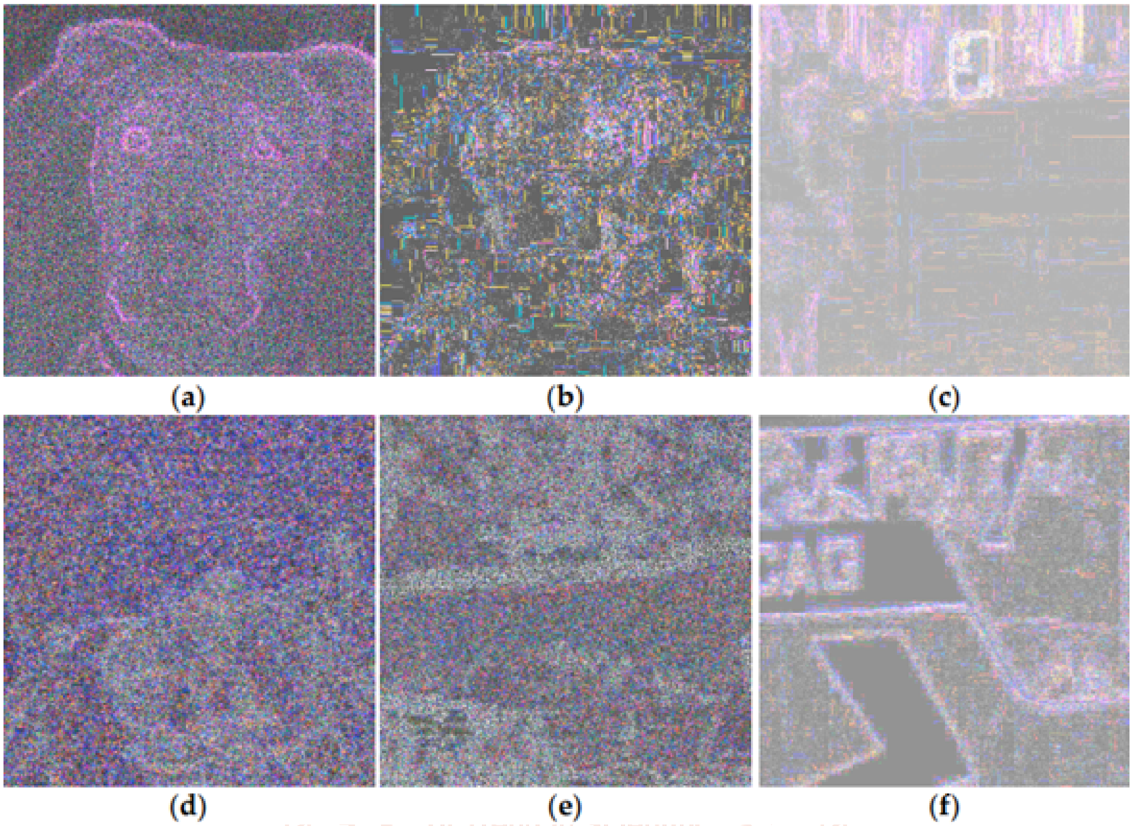

Fig 2. Classification of Images

II. BACKGROUND

A. AI-Generated Image Techniques

The advancement of AI-generated images can largely be attributed to the development of Generative Adversarial Networks (GANs). GANs consist of two neural networks: a generator that creates synthetic data and a discriminator that evaluates whether the data is real or AI-generated. Over time, the generator improves its ability to create realistic images, resulting in AI-generated visuals that are often indistinguishable from real photographs.

In addition to GANs, models like DALL-E and Stable Diffusion have further refined image generation, using text inputs to create complex visuals. These models have been used for everything from creating art to generating images for advertisements, sparking both enthusiasm and concerns about their potential misuse.

B. The Rise of Deepfakes

Deepfakes are AI-generated media, often created using GANs, that replace the likeness of a person in a video or image with someone else. They can be used for comedic or creative purposes but have also been weaponized to create non-consensual imagery, fake news, or other forms of misinformation. The rise of deepfakes has triggered the need for more sophisticated detection methods to prevent malicious use.

C. Challenges in Detection

Detecting AI-generated images is increasingly difficult as AI models become more advanced. While early AI-generated images often contained visible flaws, modern systems can create visuals with highly realistic details such as lighting, texture, and facial expressions. This poses a significant challenge for detection algorithms, which must keep pace with rapidly evolving AI techniques. Moreover, AI-generated content can be modified to evade detection, making it necessary for detection systems to be continuously updated.

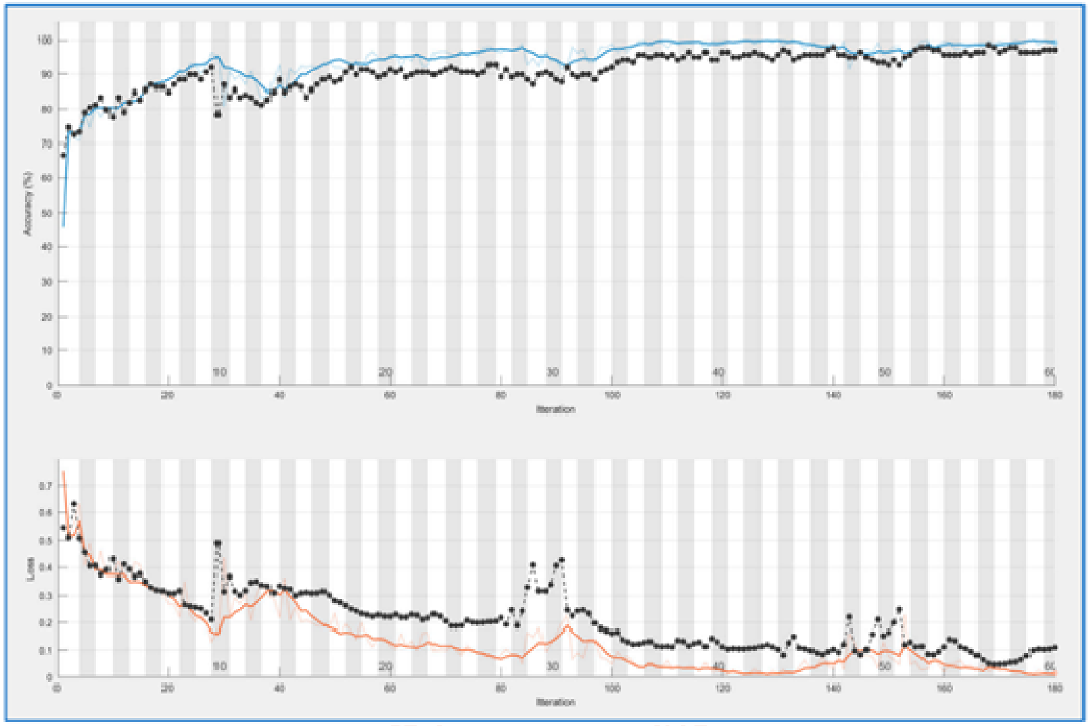

Fig 3. Challenges in detection

III. TECHNIQUES FOR AI-GENERATED IMAGE DETECTION

A. Deep Learning-Based Detection

Deep learning models trained to recognize patterns unique to AI-generated images have emerged as a leading solution. These models analyze pixel-level anomalies or inconsistencies in texture, lighting, or edge detection that are more common in synthetic images than in real ones. Convolutional Neural Networks (CNNs) are commonly used for this purpose due to their ability to extract complex features from images.

CNN-based detectors are trained on large datasets containing both real and AI-generated images, allowing them to learn the subtle differences between the two. However, as AI-generated images become more sophisticated, deep learning models must be frequently retrained on new datasets to maintain accuracy.

B. Analysis

Another method for detecting AI-generated images involves analyzing the metadata associated with the image file. AI-generated images often lack specific metadata or contain metadata that is inconsistent with that of genuine photos, such as information about the camera or software used. Detection algorithms can flag these inconsistencies as potential indicators of synthetic content.

However, this approach has limitations, as metadata can be easily altered or removed, making it less reliable as a standalone detection method.

C. Frequency Domain Analysis

The frequency domain refers to the representation of an image in terms of its frequency components, typically obtained through Fourier or wavelet transforms. AI-generated images often contain frequency artifacts or irregularities that are not present in naturally occurring images. By analyzing the frequency domain, detection systems can identify patterns that distinguish synthetic images from real ones.

This technique is particularly useful for detecting fine-grained differences in textures and shading, but it can struggle with more complex images or images that have been post-processed to remove artifacts.

D. Blockchain and Digital Watermarking

One proposed solution to the problem of AI-generated image detection is the use of blockchain technology or digital watermarking to verify the authenticity of images. Blockchain could be used to track the provenance of an image from its original creation, ensuring that any modifications or AI alterations are recorded. Digital watermarking involves embedding invisible information into an image that can be detected to verify its authenticity.

While promising, these techniques require widespread adoption and cooperation from content creators and platforms to be effective. Additionally, they do not address the detection of AI-generated images that have not been watermarked or tracked via blockchain.

IV. IMPLEMENTATION IN SOCIAL MEDIA PLATFORMS

A. User Experience And Integration

The successful implementation of AI-generated image detection in social media applications requires seamless integration into the platform's existing infrastructure. Users should be able to trust that the platform is filtering out or flagging suspicious content without compromising their experience. Features such as real-time content analysis, automated flagging, and transparency reports on detected content should be made available to users and moderators alike.

Social media platforms can also give users tools to verify images themselves, such as "Verify Image" buttons or notifications that inform users when a piece of content has been flagged as potentially AI-generated.

B. Ethical Considerations

While AI-generated image detection is necessary to prevent misinformation, there are ethical concerns that must be addressed. Algorithms must be transparent in how they detect AI-generated content to avoid bias or censorship, and users should be informed about why certain images are flagged. Additionally, false positives could lead to real, user-generated content being unfairly flagged or removed, necessitating a balanced approach that includes human oversight.

C. Legal and Regulatory Framework

The rise of AI-generated images has prompted discussions around the need for new regulations to govern their use, particularly in social media. Governments are exploring the idea of mandatory disclosure for AI-generated content, which would require platforms to label synthetic media. By integrating detection systems, social media companies can stay ahead of regulatory changes while promoting responsible content sharing.

D. Future Directions

As AI-generated images become more difficult to detect, new approaches will be needed to keep detection systems effective. Hybrid methods that combine deep learning, frequency analysis, and metadata analysis may offer the most robust solutions. Research into adversarial machine learning, where detection systems are trained to combat AI systems designed to evade them, will also be critical to staying ahead of emerging threats.

Moreover, the development of decentralized verification systems, such as blockchain-based content tracking, could offer a long-term solution to the challenges posed by AI-generated media.

V. CONCLUSION

The rise of AI-generated images presents both opportunities and challenges for social media platforms. While these technologies enable creative expression and innovation, they also pose significant risks in terms of misinformation and trust. Implementing AI-generated image detection systems is crucial for ensuring the authenticity of content on social media platforms. By using advanced detection techniques such as deep learning, frequency analysis, and metadata examination, platforms can better protect users from deceptive content and maintain the integrity of their ecosystems.

As this field continues to evolve, ongoing research and collaboration between developers, regulators, and platform operators will be essential to creating a safer and more trustworthy online environment.

This paper could serve as a strong foundation for your research or product development. Let me know if you'd like to focus on any specific areas for further exploration or discuss potential implementation strategies for your application.

VI. REFERENCES

[1] Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2022), "An Analytical Perspective on Various Deep Learning Techniques for Deepfake Detection", *1st International Conference on Artificial Intelligence and Big Data Analytics (ICAIBDA),* 10th & 11th June 2022, 2456-3463, Volume 7, PP. 25-30

[2] Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2022), "Revealing and Classification of Deepfakes Videos Images using a Customize Convolution Neural Network Model", *International Conference on Machine Learning and Data Engineering (ICMLDE)*, 7th & 8th September 2022, 2636-2652, Volume 218, PP. 2636-2652, https://doi.org/10.1016/j.procs.2023.01.237

[3] Usha Kosarkar, Gopal Sakarkar (2023), "Unmasking Deep Fakes: Advancements, Challenges, and Ethical Considerations", *4th International Conference on Electrical and Electronics Engineering (ICEEE)*,19th & 20th August 2023, 978-981-99-8661-3, Volume 1115, PP. 249-262, https://doi.org/10.1007/978-981-99-8661-3_19

[4] Usha Kosarkar, Gopal Sakarkar, Shilpa Gedam (2021), "Deepfakes, a threat to society", *International Journal of Scientific Research in Science and Technology (IJSRST)*, 13th October 2021, 2395-602X, Volume 9, Issue 6, PP. 1132-1140, https://ijsrst.com/IJSRST219682

[5] Usha Kosarkar, Prachi Sasankar(2021), " A study for Face Recognition using techniques PCA and KNN", Journal of Computer Engineering (IOSR-JCE), 2278-0661,PP 2-5,

[6] Usha Kosarkar, Gopal Sakarkar (2024), "Design an efficient VARMA LSTM GRU model for identification of deep-fake images via dynamic window-based spatio-temporal analysis", Journal of Multimedia Tools and Applications, 1380-7501, https://doi.org/10.1007/s11042-024-19220-w

[7] Usha Kosarkar, Dipali Bhende, "Employing Artificial Intelligence Techniques in Mental Health Diagnostic Expert System", International Journal of Computer Engineering (IOSR-JCE),2278-0661, PP-40-45, https://www.iosrjournals.org/iosrjce/papers/conf.15013/Volume%202/9.%204045.pdf?id=7557