Improving Fake Image Detection with the Power of Transfer Learning and CNN

Juhi Singh[1], Aditi Singh[2], Raghuraj Singh[3], and Millan Saxena[4]
[1]*Corresponding author, Harcourt Butler Technical University, Kanpur-208002*
[2]*Contributing author, Harcourt Butler Technical University, Kanpur-208002*
[3]*Contributing author, Harcourt Butler Technical University, Kanpur-208002*
[4]*Contributing author, Founder, E-Cosys Consultancy Pvt. Ltd., New Delhi, 110077, Delhi*

Abstract: **Artificial Intelligence (AI) generated fake Images are becoming more common day by day and it is very hard to differentiate them from the real ones. In this study, an efficient and reliable technique to detect AI generated images has been proposed by using a strong pre-trained model and fine-tuning it as per the required domain. We apply a deep learning based approach to help improve the detection of these AI generated synthetic images by using transfer learning with the well-known VGG16 model originally trained on the CIFAR-10 dataset [7]. To boost the performance of the technique, it is enhanced by adding extra layers, dropout, and batch normalization. The improved version of VGG16 achieved an impressive 95.90% validation accuracy, along with high precision and recall.**

Index Terms: **AI-generated images, CNN-based image classification, deep learning, transfer learning, VGG16 model.**

I. INTRODUCTION

AI has transformed the way we create and interact with digital content. Generative models like GANs and Diffusion Models have reached a point where they can create images so realistic that they often appear indistinguishable from actual photographs. GANs achieve this through a dynamic interplay between two neural networks, a generator and a discriminator, resulting in highly life like images. Diffusion Models, on the other hand, start with random noise and refine it step by step to produce images with remarkable detail and realism. While these advancements have opened up exciting possibilities in the fields like entertainment, design, and media, they also raise serious concerns about misinformation, fraud, and digital security. Spotting AI-generated images has become a pressing challenge as synthetic visuals grow increasingly life like and harder to distinguish from the real ones.

As AI-generated images continue to evolve, the need for reliable detection methods grows. This research aims to explore the latest advancements in deep learning, dataset development, and AI-generated image detection techniques. By studying state-of-the art models and large-scale datasets, we hope to contribute to the ongoing efforts to safeguard digital authenticity in an era where seeing is no longer believing.

II. LITERATURE REVIEW

Researchers have been working tirelessly to develop effective detection models. One study [1] looks into using deep learning models like ResNet, VGG16, and CNNs to detect fake images, with Error Level Analysis (ELA) helping to prepare the images before training. Their work found that ResNet delivered the best results, with a validation accuracy of 93%, making it the most reliable model for real-world applications. Similarly, another study [2] introduces the CIFAKE dataset, which replicates real-world image categories using AI-generated data. Their model, using CNN and explainable AI techniques, achieved 92.98% accuracy and revealed that subtle visual imperfections in backgrounds, rather than main objects, are key indicators of synthetic images.

One of the major challenges in this field is the lack of large-scale datasets that cover different types of AI-generated images. To address this, researchers created GenImage [3], a massive dataset containing over a million real and fake image pairs generated using advanced GANs and Diffusion Models. This dataset is designed to help train and test detectors in more diverse and realistic scenarios. AI-generated images aren't just a problem in media but the fraudulent use of deep fakes in financial transactions has also become a serious threat. A study [4] proposes using GAN-based models to detect

manipulated images in online payments, achieving a detection rate of over 95% and significantly improving security in digital transactions.

A study [5] tested nine well-known detection models on a new dataset called Chameleon and found that most of them failed to differentiate between real images from AI-generated ones. To improve detection accuracy, they introduced AIDE, a hybrid AI detector that combines high-level semantic analysis with low-level artifact detection. Another large-scale dataset, WildFake [6] was designed to test model generalization by including images from a variety of generative techniques, ensuring that AI detectors can work across different types of synthetic images.

However, even with these advancements, many existing detection models struggle when faced with highly sophisticated AI-generated images and there is a strong need to explore this area and develop further advanced methods and models to distinguish between the AI generated fake images and the real images. A details summarisation of the work done by identifying the dataset used, methodology applied, main findings, limitations of the work and the research gaps is shown in the Table 1,

Table 1: Summary of the literature review

Research Paper Name	Dataset Used	Methodology	Main Finding	Limitation	Research Gap
WildFake: A Large-scale Challenging Dataset for AI Generated Images Detection	It contains more than 3.7 million images- about 1 million real and 2.7 million generated by AI	Used different AI tools to gather fake images, sorted them for easy analysis, and included real ones to keep it balanced	WildFake is a big and varied fake image dataset, organized in a smart way to help study how well detection models handle different types of AI generated images	Not mentioned	Not mentioned
A Sanity Check For AI Generated Image Detection	Chameleon	Created Chameleon and AIDE to make AI image detection smarter by using more diverse data and deeper image understanding	Even the best AI detectors, like the AIDE model, work well on standard tests but still have a hard time spotting super realistic fakes from the Chameleon dataset	Detecting AI generated images is still a big challenge, especially with realistic datasets like Chameleon, where even the best models struggle	Most benchmarks are too simple, and models trained on one kind of AI often fail to spot others
Detection of AI Deepfake and Fraud in Online Payments Using GAN Based Models	5,000 real online payment images from sources like Kaggle and Alipay, and 5,000 AI generated ones created with models like StyleGAN	Used a GAN model where one part created fake payment images and the other learned to spot them	The GAN based model was highly accurate - over 95% - and performed well across all metrics, showing it's effective at catching deepfake fraud in online payments	Fraud detection still falls short - it's not flexible enough and often can't keep up in real-time situations	Handling complex cases, like improving generalization, using multimodal data, and testing in real time systems
Fake vs Real Image Detection Using Deep Learning Algorithm	CASIA	Used CNN, ResNet, and VGG16 models along with ELA to spot fake images	ResNet came out on top with 95% training and 93% validation accuracy, while CNN and VGG16 also performed well but with slightly lower results	The study was limited by using only a few models, simple data cleaning, and a prototype that could be more user-friendly	Explore more models, fine tune the data processing, and make the prototype more user-friendly
GenImage: A Million-Scale Benchmark for Detecting AI Generated Image	GenImage	Using ImageNet classes and fake images generated by 8 advanced AI models	GenImage includes over a million fake and real image pairs and tests how well detectors work, showing that many models still struggle when faced with different types of AI generated images	Not mentioned	Detectors still struggle with GAN and diffusion images because there's not enough diverse data to train them well
CIFAKE: Image Classification and Explainable Identification of AI Generated Synthetic Images	CIFAR-10	Applied Grad CAM to understand how the model made its decisions.	The study created fake images, and a CNN detected them with 93% accuracy by focusing on background flaws	Not mentioned	Not mentioned

III. METHODOLOGY

The proposed technique to detect AI generated images uses a strong pre-trained model after fine tuning. Deep learning and transfer learning based approach has been used to improve the detection of AI generated synthetic images with the well-known VGG16 model originally trained on the CIFAR-10 dataset [7]. Performance of the technique is further enhanced by adding extra layers, dropout, and batch normalization. Methodology, to work with the dataset, build and train the model, and measure the performance is discussed below.

A. Dataset

CIFAKE [7] dataset from Kaggle, which includes a total of 120,000 images out of which half are real photos taken from the popular CIFAR-10 dataset [7] and the other half are AI-generated images using modern image generation techniques (Latent Diffusion Models) has been used.

- Real Images: These images are taken from the CIFAR-10 dataset [7] and show everyday objects like animals, vehicles, and scenery.
- Fake Images: These images, as shown in Fig. 1 were generated by AI and made to look like real images, even though they were created by a computer.

The dataset is neatly organized into two folders; one for real and one for fake images. This data was split into training and testing sets, making sure that each set has a good mix of both real and fake images.

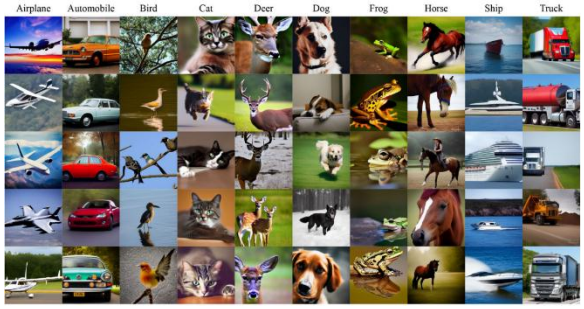

Fig. 1: Randomly selected AI-generated image from our dataset

B. Preparing the Data

Before training, the model, the images have been pre-processed by following the steps given below to get the images ready for the model.

- Resizing: All images were resized to 32x32 pixels to keep things simple and training fast.
- Batching: Images were grouped into batches of 500 for the efficient training.
- Shuffling: The training images were shuffled so that the model doesn't learn from any fixed order.
- Normalization: Pixel values were scaled to a range between 0 and 1 to help the model learn more effectively.

GPU was used for training, to significantly reduce the time required to train the model on this large dataset.

C. Model Development

Instead of starting from the scratch, a powerful pre-trained model called VGG16 has been used. This model already trained for recognizing features in the images. The process of model development is described below.

Base Model (VGG16): The top (final classification) layer of the original VGG16 model has been removed and only that part which extracts the image features has been kept. This base model has been trained on a large dataset to understand visual patterns.

Adding New Layers: To make the training more stable, a batch normalization layer has been added. After that two dense (fully connected) layers with 256 and 64 neurons, respectively have been added. These layers help the model learn deeper patterns specific to real and fake images. Regularization and dropout has also been used to reduce over-fitting so that the model doesn't just memorize the training images. Finally, a single output layer with a sigmoid function, which gives a score between 0 and 1 signifying how likely the image is real or fake has been added. The model architecture based on VGG16 with fully connected layers has been shown in Fig. 2.

D. Model Training

The Adamax optimizer which is a type of Adam optimizer has been used to train the model using the binary cross entropy loss function. This function is ideal for yes/no classification problems. Early stopping, which stops training if the model does not improve any further, helping us avoid over-fitting has also been used. This saves time and gives the best version of the model. To measure the performance following metrics have been used.

Accuracy: How often the model gets it right.

Precision: How good it is at correctly identifying fake images.

Recall: How good it is at not missing any fake images.

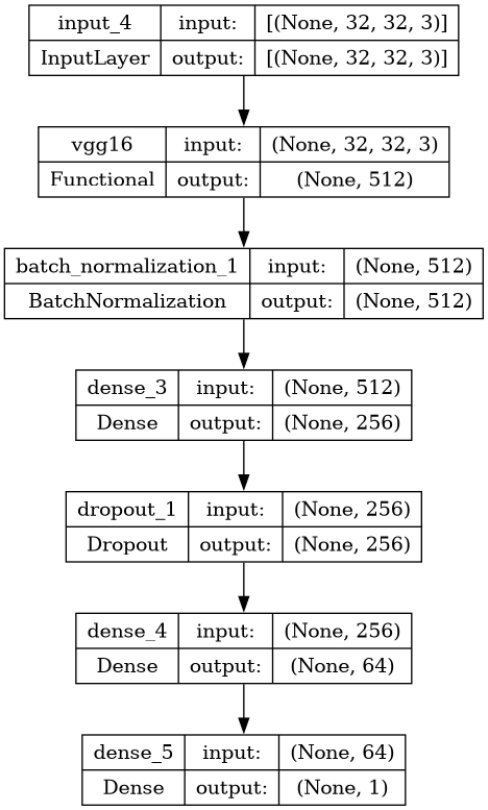

Fig. 2: Model architecture based on VGG16 with fully connected layers.

E. Model Evaluation

After training, the model is tested using the separate test dataset. During evaluation we have looked at:

Validation Loss: How much error the model made on the test set.

Validation Accuracy: The percentage of correctly identified images.

Precision and Recall: To check how well the model identifies fake vs. real images.

Graphs were also plotted to show how accuracy, precision, and loss changed during training.

IV. RESULTS AND DISCUSSION

A fine-tuned VGG16 model to differentiate between real and AI-generated images, leveraging its deep learning capabilities to enhance classification accuracy was used for the proposed model development. Model is trained on the CIFAKE dataset. Fig. 3 shows that both training and validation losses decrease steadily over various epochs, showing effective model learning. The validation loss is just 0.1407, indicating that the model didn't just memorize the training data but learned meaningful patterns.

Accuracy, precision and recall of the proposed model over various epochs has been shown in the Fig. 4, Fig. 5 and Fig. 6 respectively. The model is able to correctly classify the images with 95.90% accuracy, 96.98% precision and 94.76% recall indicating that the model accurately detects most of the fake images while keeping false alarms low. Overall, the model is very promising in spotting the AI-generated images and could be a helpful tool in maintaining the authenticity of visual content.

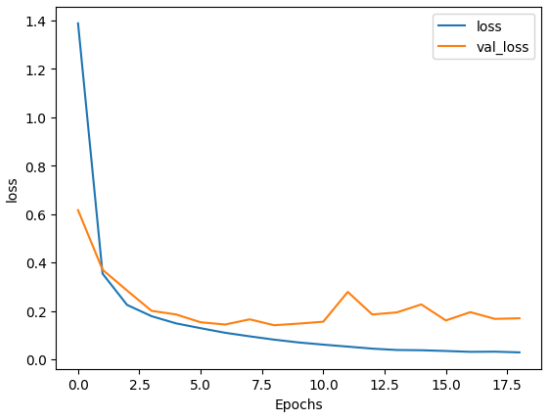

Fig. 3: Training and Validation Losses over various epochs

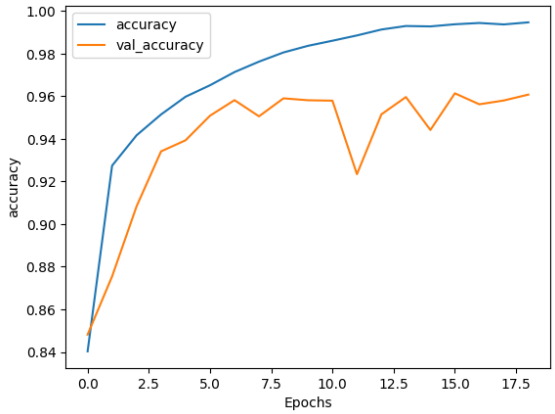

Fig.4: Training and Validation Accuracy Over Epochs

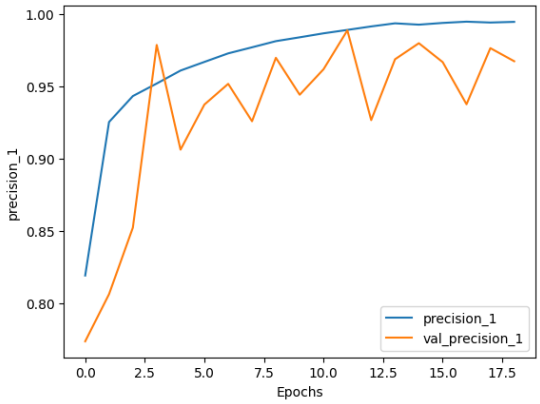

Fig.5: Precision vs. Epochs

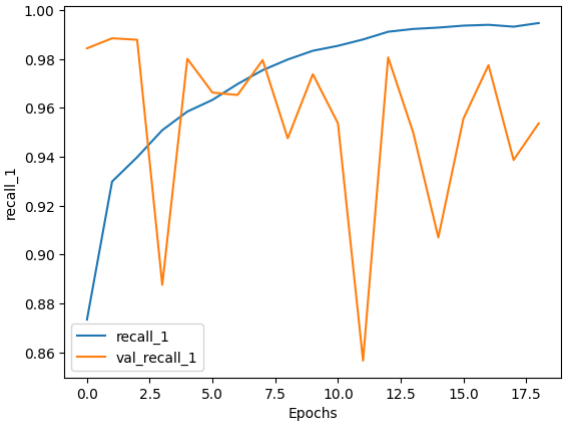

Fig.6: Recall Performance Over Epochs

V. CONCLUSIONS AND FUTURE SCOPE

As AI continues to evolve, the ability to create highly realistic fake images has become easier and more convincing. This brings a real concern about misinformation, fraud, and trust in digital content. As fake image generation gets more advanced, our detection methods need to be more accurate and reliable. While there's been consistent progress but the challenge isn't still over. It is observed that large, diverse datasets and explainable AI techniques help develop models which learn better and give insight into the decision making process. In this work, a model has been proposed which detects fake images with high accuracy, precision and recall by exploring deep learning model VGG16.

Moving forward, there is a lot of potential to improve fake image detection by using smarter and more efficient CNN models. The newer CNNs can deliver better accuracy while being fast, lightweight and perfect for real-time applications. With a bit of fine-tuning, data augmentation, and smart training strategies, key metrics like precision, recall, and F1-score can be improved further. Also, combining

different models or using ensemble techniques could make the system more reliable and adaptable, especially as fake image generation keeps getting more advanced. In a world where seeing is no longer always believing, building trust through technology is more important than ever, we need smarter, more adaptable tools that can not only detect fake content accurately but also explain their reasoning.

REFERENCES

[1]. Fatoni, Tri Basuki Kurniawan, DeshintaArrova Dewi, Mohd Zaki Zakaria, Abdul Muniif Mohd Muhayeddin, (2025), "Fake vs Real Image Detection Using Deep Learning Algorithm", Applied Data Sciences, 6(1), 366–376, bright-journal.org, (1), https://doi.org/10.47738/jads.v6i1.490.

[2]. Jordan J. Bird and Ahmad Lotfi, (2024), "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images", IEEE Access, 12, ieee.org, (128), https://doi.org/10.1109/ACCESS.2024.3356122 .

[3]. Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, Yunhe Wang, (2023), "GenImage: A Million Scale Benchmark for Detecting AI-Generated Image", neurips.cc, (131), 37th Conf. Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

[4]. Zong Ke, ShichengZhou,Yining Zhou, Chia Hong Chang, Rong Zhang, (2025), "Detection of AI Deepfake and Fraud in Online Payments Using GAN-Based Models ", arxiv.og, (15), IEEE at 2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE 2025).

[5]. Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, Weidi Xie, (2025), "A Sanity Check For AI-Generated Image Detection", arxiv.og, (19), Published as a conference paper at ICLR 2025.

[6]. Yan Hong, Jianming Feng, Haoxing Chen, Jun Lan, Huijia Zhu, Weiqiang Wang, Jianfu Zhang, (2024), "WildFake: A Large-scale Challenging Dataset for AI-Generated Images Detection", arxiv.org, (10).

[7]. A. Krizhevsky, G. Hinton, (2009). Learning multiple layers of features from tiny images. J.J. Bird, and A. Lotfi, 2024. CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. IEEE Access. Real images are from Krizhevsky Hinton (2009), fake images are from Bird Lotfi (2024).