Image Classification: AI-Generated and Real Images

Sanya Madan
2021561
CSAI'25
sanya21561@iiitd.ac.in

Parisha Agrawal
2021270
CSAI'25
parisha21270@iiitd.ac.in

Brinda Muralie
2021140
ECE'25
brinda21140@iiitd.ac.in

Abstract

Differentiating between real and AI-generated images has become a complex task for humans due to the remarkable progress in AI image generation using General Adversarial Networks (GANs) and Stable Diffusion Models. Misleading images that convincingly imitate reality can cause significant harm like spreading fake news, damaging reputations, and manipulating public opinion through media. The judicial system also often relies on visual evidence. Identification and mitigation of fake images is crucial to ensure the authenticity of visual content.

1. Introduction

This project aims to develop robust classification methods for distinguishing between real and AI-generated images. Using datasets such as CIFAKE[2] which includes both types of images, extracting discriminative features, and evaluating various models, the project seeks to address the challenge of identifying increasingly realistic AI-generated images, with potential applications in image regulation both online and in real world.

2. Literature Survey

[1] **CIFAKE: Advancing AI-Generated Image Recognition**[2] : The paper *CIFAKE: Advancing AI-Generated Image Recognition*[2] by Jordan J. Bird and Ahmad Lotfi, focuses on the critical necessity of data reliability and authentication in the era of AI-generated synthetic images. The authors propose a method to enhance our ability to recognize these images through computer vision.

The study involves the creation of a synthetic dataset, CIFAKE[2], which mirrors the ten classes of the CIFAR-10[1] dataset. The dataset is generated using latent diffusion, providing a contrasting set of images for comparison to real photographs. They used prompt modifiers to generate 6,000 images in each class.

The authors then propose the use of a Convolutional Neural Network (CNN) to classify the images into two categories: Real or Fake. After hyperparameter tuning and training of 36 individual network topologies, the optimal approach could correctly classify the images with an accuracy of 92.98%.

To further understand the classification process, the study implements explainable AI via Gradient Class Activation Mapping. This reveals that the model focuses on small visual imperfections in the background of the images for classification, rather than the actual image itself.

The CIFAKE[2] dataset is made publicly available to the research community for future work. This study provides a significant contribution to advancing AI-generated image recognition.

[2] **GenImage: Advancing AI-Generated Fake Image Detection**[3]:

The paper *GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image*[3] addresses the growing concern about the spread of misinformation due to the extraordinary ability of generative models to generate photographic images. The authors introduce the GenImage dataset.

The GenImage dataset has three main advantages:

1. Plenty of Images: It includes over one million pairs of AI-generated fake images and collected real images.
2. Rich Image Content: It encompasses a broad range of image classes.
3. State-of-the-art Generators: The dataset synthesizes images with advanced diffusion models and GANs.

The authors conducted a comprehensive analysis of the dataset and proposed two tasks for evaluating the detection method in resembling real-world scenarios.

The first task was the cross-generator image classification task which measured the performance of a detector trained on one generator when tested on others. The second task was the degraded image classification task which assessed the capability of detectors in handling degraded images such as low-resolution, blurred, and compressed images.

With the GenImage dataset, researchers can effectively expedite the development and evaluation of superior AI-generated image detectors in comparison to prevailing methodologies.

[3] **Generalizable Synthetic Image Detection via Language-guided Contrastive Learning**[4]:

The paper *Generalizable Synthetic Image Detection via Language-guided Contrastive Learning*[4] by Haiwei Wu, Jiantao Zhou, and Shile Zhang, addresses the challenge of detecting AI-generated synthetic images. The authors propose a novel method for synthetic image detection using language-guided contrastive learning.

The rapid development of synthetic image generating models has led to the creation of highly realistic AI-generated images. However, these images can be used maliciously, such as in the spread of fake news or creation of fake profiles. Existing forensic algorithms for detecting synthetic images have limitations, particularly in their generalization capability.

The authors proposed a new approach to this problem. They augmented training images with carefully-designed textual labels, enabling the use of joint image-text contrastive learning for forensic feature extraction. They also reformulated the synthetic image detection problem as an identification problem, which is a significant upgrade from traditional classification-based approaches.

Their proposed model, LanguAge-guided SynThEsis Detection (LASTED), demonstrates improved generalizability to unseen image generation models and delivered promising performance that exceeds state-of-the-art competitors by +22.66% accuracy and +15.24% AUC. The authors have made the code available for further research.

3. Dataset

In this section, we provide a comprehensive overview of the dataset used for our project and detail the preprocessing techniques applied to prepare the data for model training.

3.1. Dataset Description

The dataset utilized in this study is known as CIFAKE[2], which consists of two primary classes: real and fake images. CIFAKE[2] contains a total of 120,000 images, evenly distributed between synthetically-generated and real images. Key characteristics of the dataset are as follows:

- **Classes:** Both the classes in CIFAKE[2], real and fake, are divided into ten classes, each representing distinct objects or entities. These classes include: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck, with 6,000 images per class.
- **Data Sources:** The real images used in the dataset were extracted from the CIFAR-10 dataset introduced by Krizhevsky and Hinton[1]. The synthetic images, on the other hand, were generated using Hugging Face's Stable Diffusion Model Version 1.4. To enhance diversity within the synthetic dataset, different prompt modifiers were applied for each class label.
- **Data Split:** To facilitate training and evaluation of machine learning models, the dataset was partitioned into 100,000 training images (50,000 per class) and 20,000 testing images (10,000 per class). All images were represented in RGB format and resized to a uniform resolution of 32 × 32 pixels.
- **Scaling:** It is noteworthy that no additional scaling was applied to the images during preprocessing since all images were already resized to the desired 32 × 32 pixel dimensions during the data collection process.
- **Visualisation:** t-SNE.

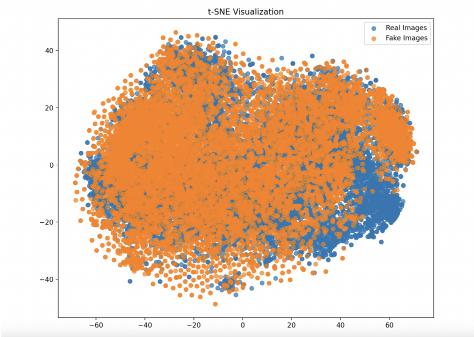

Figure 1. t-SNE Visualisation

3.2. Data Preprocessing Techniques

After loading the data from the provided directory, and segregating images into their respective classes ('REAL' or 'FAKE'), The following steps were undertaken to preprocess the dataset in preparation for model training:

1. **Image Resizing:** To ensure uniformity in image dimensions, each image was resized to 32 × 32 pixels using the OpenCV library (For precaution as in data provided each image was 32 × 32 pixels only).

2. **Label Encoding:** Class labels were encoded to numeric values using the LabelEncoder from scikit-learn. This transformation is essential for machine learning model compatibility.

3. **No Outliers:** Using data visualisation and boxplots, we analysed that there are no outliers.

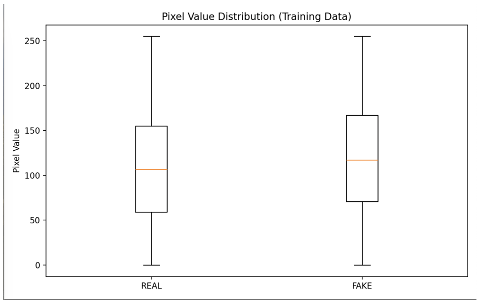

Figure 2. Box Plot

4. **Dimension Reduction using PCA:** Using PCA for dimension reduction is not useful here as the dataset has already provided the optimal reduced dimension. We had analysed this using data visualisation, TSNE as well as training model on different PCA values. Here is a table showing the number of features and the accuracy.

Number of components	Accuracy
10	0.71375
50	0.795
100	0.81175
500	0.81125
1000	0.81325
1024 (32 × 32)	0.8135

By following these preprocessing techniques, we ensured that the dataset was suitably prepared for training and evaluation of machine learning models.

4. Methodology

- **Logistic Regression** : Logistic regression a fundamental machine learning algorithm used for binary classification. We performed binary classification (Real vs. AI generated images) using logistic regression implemented in PyTorch and scikit-learn. It includes data preprocessing, model creation and training, and evaluation of both PyTorch and scikit-learn logistic regression models. We achieved 0.6793 Validation Accuracy and Test accuracy of 0.67725.

Value	Validation	Test
Accuracy	0.6793	0.67725
Precision	0.66386	0.6926
Recall	0.7218	0.6374
F1-Score	0.6916	0.6638
Specificity	0.6370	0.6374
Confusion Matrix:	[[6393 3642] [2772 7193]]	[[6374 3626] [2829 7171]]
False Positive Rate	0.3629	0.3626

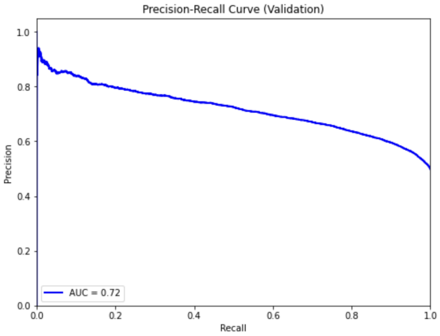

Figure 3. Precision Recall Curve Logistic Regression

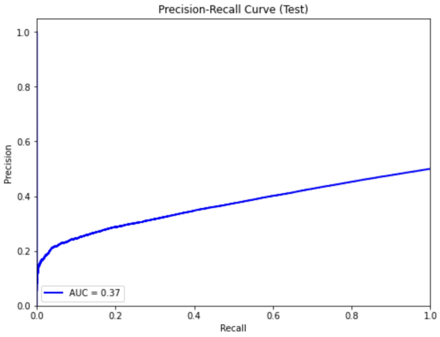

Figure 4. Precision Recall Curve Test Logistic Regression

- **Naïve Bayes Classifier** : Naive Bayes is a probabilistic machine learning algorithm based on Bayes theorem while assuming features to be independent of each other. We performed Multinomial Naïve Bayes classifier on image data for binary classification (Real vs. AI generated) using scikit-learn. It includes data loading, preprocessing, model training, and evaluation using scikit-learn. We achieved 0.5893 Validation Accuracy and Test accuracy of 0.59275.

Value	**Validation**	**Test**
Accuracy	0.5893	0.59275
Precision	0.59412	0.59570
Recall	0.57269	0.5773
F1-Score	0.58321	0.5863
Specificity	0.60602	0.6082
Confusion Matrix:	[[6039 3926] [4288 5747]]	[[6082 3918] [4227 5773]]
False Positive Rate	0.3939	0.3918

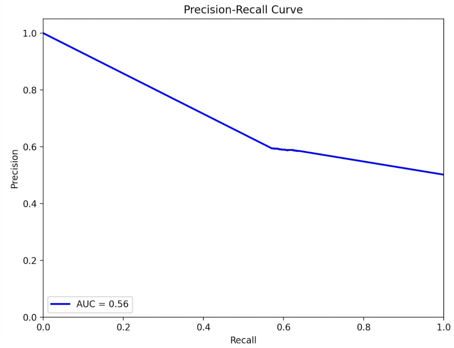

Figure 5. Precision Recall Curve Naïve Bayes

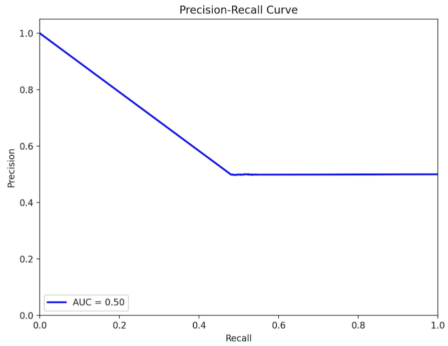

Figure 6. Precision Recall Curve Test Naïve Bayes

- **Decision Tree Classifier** : Decision Tree is a supervised machine learning algorithm, it is a tree-like model represting decisions and consequences in hierarcial manner. We performed Decision Tree classifier on image data for binary classification (Real vs. AI generated) using scikit-learn. It includes data loading, preprocessing, model training, and evaluation using scikit-learn. We trained a Decision Tree classifier for image classification (Real vs. AI generated) and achieved 0.69345 Validation Accuracy and Test accuracy of 0.69765.

Value	**Validation**	**Test**
Accuracy	0.6869	0.69745
Precision	0.68457	0.6983
Recall	0.6972	0.6953
F1-Score	0.6908	0.6967
Specificity	0.6764	0.6996
Confusion Matrix:	[[6741 3224] [3038 6997]]	[[6996 3004] [3047 6953]]
False Positive Rate	0.3235	0.3004

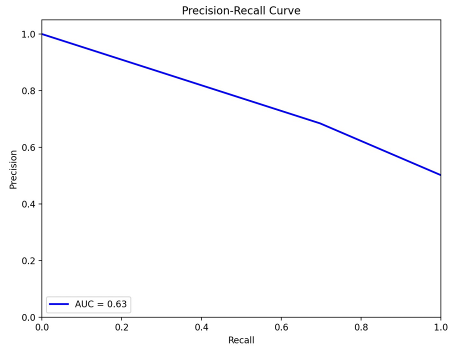

Figure 7. Precision Recall Curve Decision Tree

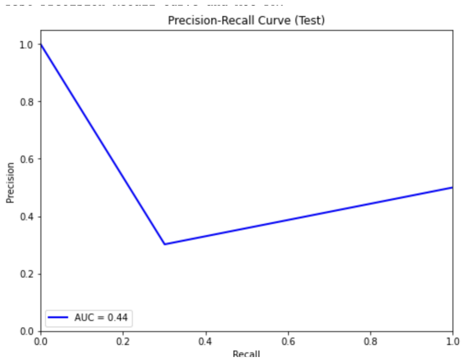

Figure 8. Precision Recall Curve Test Decision Tree

- **Random Forest Classifier** : Random Forest is a machine learning for regressiona nd classification tasks, based on the concept of decision tree. It includes data loading, preprocessing, model training, and evaluation using scikit-learn. Trained a Random Forest classifier for image classification (Real vs. AI generated) and achieved 0.8266 Validation Accuracy and Test accuracy of 0.8307.

Value	Validation	Test
Accuracy	0.8272	0.82925
Precision	0.8498	0.80630
Recall	0.7963	0.8667
F1-Score	0.8222	0.8354
Specificity	0.8583	0.8667
Confusion Matrix:	[[8553 1412] [2044 7991]]	[[8667 1333] [2082 7918]]
False Positive Rate	0.1416	0.1333

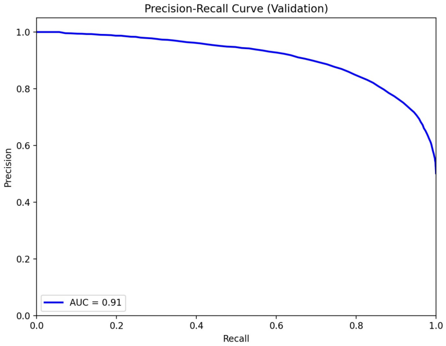

Figure 9. Precision Recall Curve Random Forest

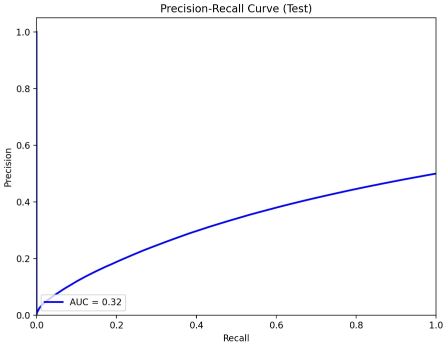

Figure 10. Precision Recall Curve Test Random Forest

5. Results and analysis

In the precision-recall curves (Figure 3-10), we observed that

In our evaluation of four machine learning models-Logistic Regression, Naïve Bayes, Decision Trees, and Random Forests-we observed varying levels of performance.

Random Forests achieved the highest test accuracy, reaching 83.07%, making it the top-performing model in our study. This can be due to its ability to handle complex image data, reduce overfitting, and capture intricate patterns.

On the other hand, Naïve Bayes exhibited the poorest performance, with a test accuracy of 59.275%. This outcome is mainly due to its simplistic assumption of feature independence, which is inadequate for image data characterized by pixel correlations and complex spatial relationships.

Logistic Regression showed moderate performance with a test accuracy of 67.725%. However, its linear nature limits its capability to capture intricate patterns and relationships present in image datasets.

Decision Trees performed reasonably well, achieving a test accuracy of 69.765%. These models are adept at capturing non-linear relationships but are susceptible to overfitting, contributing to their performance.

Model	Validation Accuracy	Test Accuracy
Logistic Regression	0.6793	0.67725
Naïve Bayes	0.5893	0.59275
Decision Tree	0.69345	0.69765
Random Forest	0.8266	0.8307

6. Conclusions

In conclusion, this project addresses the critical challenge of distinguishing real images from AI-generated ones, which have the potential to spread misinformation and manipulate public opinion. By utilizing the CIFAKE[2] dataset and employing various machine learning models, we have achieved promising results.

However, it's essential to acknowledge the project's limitations. Due to the novelty of topic, we couldn't find any suitable papers whose models that we could implement on our own. Since it is an image classification problem, we couldn't meet the hardware requirements of some of the pre-existing models and datasets (for example, the dataset introduced in GenImage was 16 GBs, and our personal systems could not accommodate that). These include potential dataset diversity issues, the Naïve Bayes model's independence assumption, and the absence of deep learning models like CNNs. The project mainly relies on accuracy as an evaluation metric, which may not fully capture model performance.

Despite these limitations, this project represents a valuable step in tackling the challenges posed by AI-generated images. Future work could focus on optimization and the incorporation of advanced techniques such as SVMs and neural networks. Overall, this project underscores the ongoing need for research to ensure the authenticity of visual content in the digital era.

7. References

1. Will Cukierski. (2013). CIFAR-10 - Object Recognition in Images. Kaggle. https://kaggle.com/competitions/cifar-10

2. Bird, J.J., Lotfi, A. (2023). CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images. arXiv preprint arXiv:2303.14126.

3. M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, and Y. Wang, “Genimage: A million-scale benchmark for detecting ai-generated image,” arXiv preprint arXiv:2306.08571, 2023.

4. Haiwei Wu, Jiantao Zhou, Shile Zhang, ”Generalizable Synthetic Image Detection via Language-guided Contrastive Learning”, arXiv preprint arXiv:2305.13800, 2023.

5. Krizhevsky, A. (2009) Learning Multiple Layers of Features from Tiny Images. Technical Report TR-2009, University of Toronto, Toronto. https://www.cs.toronto.edu/ kriz/learning-features-2009-TR.pdf