A Machine Learning Approach to Unmasking Art Forgery: Detection and Analysis

Kutay Ölmez
TED University
Department of Computer Engineering
kutay.olmez@tedu.edu.tr

***Abstract*—As the capabilities of generative models continue to evolve, the distinction between the media produced with artificial intelligence and the real media is becoming increasingly difficult. This phenomenon raises significant concerns regarding image forgery and copyright infringement. This paper addresses the issue on automatic detecting the AI generated art-work paintings. In this study, I aim to detect and analyse AI-generated images using a machine learning model trained on a dataset of more than 12,000 real and generated images in two different categories. The outcomes of this study intended to contribute to the development of tools and strategies for maintaining the integrity of artistic expression and protecting the rights of content creators in the digital era.**

I. Introduction

Significant advances in deep learning are evident in many fields today. Generative artificial intelligence (AI) also provides various media with these developments. These days, the quality of generative media is increasing day by day and it is becoming difficult to distinguish them from real media. This situation raises a concern over image forgery and copyright infringement. This is because the most successful generative models are trained with image datasets, which generally contain millions of images, the vast majority of which are copyrighted. Trained models can successfully recreate an ordinary painting in the style of a famous painter as if it were a work of art or they can convert the style of an artwork painting to a different painting style successfully. Due to the high quality and high accuracy of these paintings, it has become very difficult to manually recognize whether they were created by an artist or generated by AI. Therefore, it is necessary to develop innovative solutions to protect the originality and intellectual property of artistic works.

In response to these challenges, this paper addresses the development of a robust detection system for artwork images generated by AI. Leveraging the DeepfakeArt Challenge dataset [1], which has more than 32,000 image data in 4 categories and consists of both AI-generated images and real images, our approach uses machine learning techniques to automatically identify and analyze these digitally generated artworks. The cornerstone of our methodology is the use of the ResNet50 model as a feature extractor. In this way, the mastery of capturing the intricate details and nuanced features in the images provided by the convolutional neural network (CNN) is aimed to provide a solid basis for distinguishing the nuanced features that distinguish human-created works of art and works produced by AI. Building upon the feature extraction, a Logistic Regression classifier was used to distinguish patterns and make informed predictions.

The subsequent sections will delve into the literature review, methodology, results, and conclusions derived from this study, shedding light on the promising strides made in safeguarding the integrity of artistic expression in the face of evolving AI technologies.

II. Literature Review

A large high-resolution dataset containing real and fake artwork images was initially needed to detect AI-generated artwork images. For this reason, several datasets were identified for further research. These are DeepfakeArt Challenge [1], ArtiFact [3], WikiArt [4], Art by AI [5], and Art500K [6]. Among these, many issues such as the dataset containing both real and fake quality artwork images, having an organized structure, actively receiving updates, the diversity in the style of the images created by AI, different models, and different

techniques were considered and finally it was decided to use the Deepfake Art Challenge Dataset. Then, in the literature review, studies for the detection of AI-generated images were reviewed. In addition, techniques and models such as GAN, StyleGAN, ProGAN, ResNet50, ResNet101, Stable Diffusion 2, adversarial data poisoning attacks, and cutmix, which are used when generating an image and detecting images generated by artificial intelligence, were researched.

After this review, the work of Stef Herregods et al [8] was taken as a basis for the development of this project. In their project based on a research for the detection of images generated with various GAN derivatives [2], they worked on AI-generated art image detection with the project they developed with the ResNet50 NoDown model. In consideration of these researches, we decided to use the ResNet50 model as a feature extractor in our own project and to the classification of these extracted features with Logistic Regression.

III. Methodology

In order to develop the project more easily and with solid foundations, we proceeded in 3 stages. These stages are, first collecting and organizing the dataset, then extracting the features of the images in this dataset, and finally developing a model that will successfully classify these extracted features.

A. Data Collection

After conducting several researches, we decided to use the DeepFakeArt Challenge Dataset [1] as the dataset of our project, which provided the necessary features for us. This dataset had more than 32,000 data from 4 different categories. This dataset was even divided into test and train sets. However, it was decided whether the categories ”Data Poisoning” and ”Cutmix” were necessary for our project, and only two categories named ”Inpainting” and ”Style Transfer” would be used. Therefore, these two categories contained in the dataset were removed and they were also removed from the test and train sets. Thus, around 12,000 data remained in the dataset. After this, an arrangement was made in the file organization of the data set, and the file path of the images in the train and test sets was also changed according to the new file organization. The ”Inpainting” category consists of artwork images created with the Stable Diffusion 2 model. The total number of data is 5036 in this category. While the ”Style Transfer” category contains images of various styles, it also contains copies created by AI in different styles from real images using ControlNet. This category consists of 3074 entries.

B. Feature Extraction

After collecting data for the project, the ResNet50 model was used to extract nuanced features of human-created artwork images and AI-generated images. After the images were resized to the appropriate shape, all images in the dataset were passed through the ResNet50 model, in which classification layers were removed, and the features of each image were extracted. In Figure 1, you can observe the feature map of one of the images whose features were extracted. At this stage, unlike the studies we relied on, we performed a resize operation on the images and used ResNet50 not as a classifier but as a feature extractor.

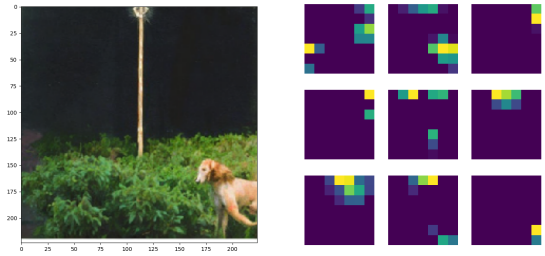

Fig. 1. Feature Map of AI-Generated Art Image

C. Model Training & Evaluation

Based on feature extraction, a Logistic Regression classifier was used to distinguish patterns and make informed predictions. This model was trained in 3 different ways on a dataset consisting of 2 categories. These are observing the results by training the categories separately and observing the results by training the 2 categories together. On each way, the data was scaled and then model training was carried out. The results obtained in these ways are the outputs of the performance of the model on the test set. At this stage of the project, hyperparameter tuning was not performed.

IV. Results

When the results obtained by classification using logistic regression after the feature extraction process performed with ResNet50 were observed, firstly, the accuracy rate of the model obtained for the inpaintings category was recorded as 0.7871. According to the classification report in Table I, this

model demonstrates a balanced performance for both real and generated classes, with similar precision, recall, and f1-score values.

TABLE I
INPAINTING CATEGORY CLASSIFICATION REPORT

	Precision	Recall	F1-Score	Support
Real	0.79	0.78	0.79	1221
Generated	0.79	0.79	0.79	1221
Accuracy		0.7871		

The second model obtained for the Style transfer category has a high accuracy rate of 0.9757 and high classification performance. The high precision, recall, and f1-score values values for both real and generated classes are remarkable. This model's results are presented in Table II.

TABLE II
STYLE TRANSFER CATEGORY CLASSIFICATION REPORT

	Precision	Recall	F1-Score	Support
Real	0.98	0.97	0.98	658
Generated	0.97	0.98	0.98	658
Accuracy		0.9757		

Thirdly, the accuracy rate of the combined model including inpaintings and style transfer categories was obtained as 0.8348. it was observed that the model performed well for both classes. Performance metrics of the model are presented in Table III.

TABLE III
COMBINED CLASSIFICATION REPORT

	Precision	Recall	F1-Score	Support
Real	0.84	0.82	0.83	1879
Generated	0.83	0.85	0.84	1879
Accuracy		0.8348		

V. DISCUSSION

When we look at the results of this study, which will contribute to the literature on art forgery detection by developing a model that can differentiate artworks generated by AI from real paintings, it has been observed that although the model we trained only with the Inpainting category yields satisfactory results. However, its performance is comparatively lower than the other models. The Combined category performed well and had the second-highest accuracy score. The model that classifies the Style Transfer category performed a much better performance than the other two models. In general, the Style Transfer model exhibited the highest performance, while the other two models also achieved balanced and satisfactory results.

This study, with potential future enhancements, could reach a higher level, serving as both a reference point for future research and an actively utilized tool for the detection of AI-generated artwork images. To exemplify potential improvements, such as determining the layers that will give the best performance on the feature extraction model and developing the model that consists only of these layers. In addition, the dataset can be diversified, experiments can be performed with different classification models, and hyperparameter tuning can be applied to the classification model.

VI. CONCLUSION

In conclusion, the success of the developed model in distinguishing between real and AI-generated artworks demonstrates the potential of machine learning against challenges related to authentication and copyright protection of artworks. Although it provides satisfactory results for future research, the current approach needs to be improved or different approaches need to be experimented and the resources used need to be increased for this study to be ready for the real-world scenario.

REFERENCES

[1] Aboutalebi, Hossein, et al. "DeepfakeArt Challenge: A Benchmark Dataset for Generative AI Art Forgery and Data Poisoning Detection." arXiv.Org, 11 June 2023, URL arxiv.org/abs/2306.01272.

[2] Gragnaniello, Diego, et al. "Are Gan Generated Images Easy to Detect? A Critical Analysis of the State-of-the-Art." arXiv.Org, 6 Apr. 2021, URL arxiv.org/abs/2104.02617.

[3] Rahman, Md Awsafur, et al. "Artifact: A Large-Scale Dataset with Artificial and Factual Images for Generalizable and Robust Synthetic Image Detection." arXiv.Org, 24 Feb. 2023, URL arxiv.org/abs/2302.11970.

[4] "Visual Art Encyclopedia." URL www.wikiart.org. Accessed 2023.

[5] "Art by AI – How to Create the Dataset." Kaggle, 23 June 2019, URL www.kaggle.com/code/vbookshelf/art-by-ai-how-to-create-the-dataset.

[6] Mao, Hui, et al. "ART500K Dataset." Art500k Dataset, The Hong Kong University of Science and Technology, URL deepart.hkust.edu.hk/ART500K/art500k.html. Accessed 2023.

[7] Herregods, Stef, et al. "Evaluation of a GAN Generated Image Detector (ResNet50 NoDown) on Stable Diffusion Generated Images." GitHub, 2023, URL github.com/StefHerregods/datathon-AI-art/tree/main/GAN_Detection.