JOURNAL OF
EMERGING INVESTIGATORS

Article

Evaluating the effectiveness of machine learning models for detecting AI-generated art

Yunjiao Xiao[1], Jenny Zhao[2]
[1] Mission San Jose High School, Fremont, California
[2] IvyStreet, Fremont, California

SUMMARY

Breakthroughs in technology mean the creation of new programs that can produce results startlingly close to human creations. AI-generated art, for example, has garnered attention and debate by redefining artistic creations, originality, and the ethics of artificial intelligence. By training on existing artwork, these programs can generate images startlingly close to the intended art style and topic and can fool the ordinary person. In this paper, we examine whether the type and style of the input image will impact the accuracy of existing AI-detection machine learning models. The first part of the hypothesis is that the existing models are more accurate in classifying whether an image is AI or human-created if the image is centered around human subjects rather than the environment. The second part of the hypothesis is that the models are more accurate in classifying realistic images compared to images in animation style and traditional mediums. The data show that current models are slightly better at classifying images of the environment, with 84.7% accuracy, compared to images of human characters, with 81.9% accuracy, which doesn't support the first part of our hypothesis. However, most models have greater success with realism style than with traditional and animation styles, which supports the second part of our hypothesis. These results may serve as suggestions for further improvements in current models. With efficient machine learning models, artists and the general public can discern between AI and human-created art, which may help improve the regulation and usage of AI-generated art.

INTRODUCTION

As the global communication and information network gets more complex, the availability of massive amounts of data helps drive the development of new algorithms, which feeds into the growing field of machine learning (1). These rapid developments mean unimaginable possibilities and have impacted the art world lately (1). AI image generators typically generate visuals based on text prompts provided by a user (2). AI art is not new, but the accessibility of such programs (e.g. DALL-E 2 (3), DeepDream (4), Stable Diffusion (5), Midjourney (6)) has increased and it has been popularized by social media (1). One such program, Midjourney, is trained with all the data, text, and images it can pull from the internet and may be seen as a positive engine for creativity (7). The developments, however, have surpassed regulation and considerations of consequences. AI art generators learn to put together images by using existing artwork online, but artists do not have a way to take their work out of the training sets, and they are not compensated or credited (1). Convincing images that can trick the masses further complicate the question of ownership, raising concerns in many artists about their livelihoods (1). Furthermore, these generated images may produce bias if the dataset from which the AI learns is not a fair representation of the world, potentially resulting in the spread of misinformation and reinforcing stereotypes (1).

The process of training machine learning models involves feeding data to them. As such, the current AI art generators train on massive amounts of data to visualize a given prompt. With more training and data, these models can produce variations of the intended image in different styles (2). By that logic, one can also train the models to detect AI art from human-created art. Given a set of human-created artwork and AI counterparts, the model can learn to match images to each category, and thus predict future data.

Optic AI or Not analyzes media assets to find their source (8). The detector looks for known signs and characteristics of AI-generated images and compares the input to those signs and characteristics to make a classification (8). The model claims a 95% accuracy (8).

Hive's AI-Generated Media Recognition model is trained on a large dataset consisting of millions of AI-generated and human-created images across a broad range of categories, including photographs, digital and traditional art, and memes sourced from across the web (9). The model first conducts a binary classification to determine if an input is AI-generated, and then in the case of AI-generated, the model predicts the likely source engine that generated it (9).

Maybe's AI detector started as a repurposed model of an NSFW detection bot developed by a Reddit user (10). The model is trained with thousands of images on Reddit using Reddit Downloader, then labels the images as either "human" or "artificial" depending on which art subreddit the images are from (10). FastAI library assisted the training, but the initial results had many false positives and negatives (10). Image classification models on Hugging Face are used to construct the model and images before 2019 are used for the "human" category since that is when text-to-images became popularized (10). The final model used in the AI detector is a "SwinForImageClassification" model, which involves a Swin Transformer that constructs a hierarchical representation with small-sized patches and gradually combines neighboring patches in deeper Transformer layers (11). Maybe's AI detector produced false negatives corresponding to images generated with Stable Diffusion and DALLE-2, which most likely resulted due to a lack of representation in the training dataset (10).

DOI: https://doi.org/10.59720/23-132

Illuminarty combines various computer vision algorithms to generate an AI probability output for an image (12). Currently, the tool samples images from a few popular AI art generators such as Stable Diffusion, and Dall-E, and does not cover the diverse styles and tools of human artists (12). Known problems—such as failing on specific color schemes, high-resolution images, photographs, anime and game screenshots—associated with this model have been attributed to bias in the sampled dataset (12).

This study aims to test the accuracy of existing machine-learning models in detecting AI art on human characters versus the environment and in different art styles. Our hypothesis is that the existing models will be more accurate in detecting AI-generated images of human characters compared to detecting AI-generated images of the environment. We also hypothesize that the models will be more accurate in detecting realistic images compared to images in animation and traditional styles. We made this hypothesis considering the wide availability of photos of humans that exist online, which may allow for more training materials for the model to learn and thus make it easier to identify human subjects correctly. We tested this hypothesis by gathering sample images that are representative of each category (character realism, character animation, character traditional, environment realism, environment animation, and environment traditional) and feeding them to four online models (AI or Not, Hive Moderation, AI detector by Maybe, and Illuminarty). We selected these models because they are open to the public and easily accessible, but they also train on datasets of various sizes and sources. Contrary to the first part of the hypothesis, we found that the models are slightly more accurate at classifying images of the environment rather than images of human characters. The majority of the models, however, do give more accurate results when the image is in a realistic style rather than animation or traditional, abstract styles.

RESULTS

A total of 36 images (3 AI and 3 human-created for each category) were entered into each of the models: AI or Not (8), Hive Moderation (13), Maybe's AI detector (14), and

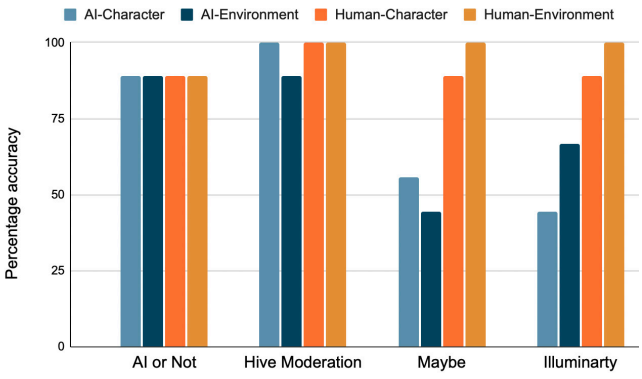

Figure 1: Each model's accuracy in detecting AI-generated and human-generated images of human characters and environment. The classification accuracy of AI detectors, AI or Not, Hive Moderation, Maybe, and Illuminarty, in all four image categories. Each image category is fed to AI detectors in equal amounts, and the accuracy of the tests are totaled up.

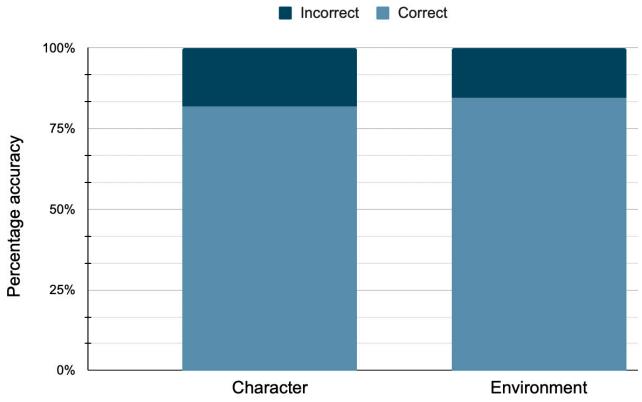

Figure 2: The combined accuracy of all the models in detecting whether an image is AI-generated or human-generated. The classification accuracy of AI detectors for images of human characters and environment. An equal sized sample of both subjects is fed to AI detectors, with the accuracy noted. P = 0.655, so the result is not significant at $p < 0.05$.

Illuminarty (15). The overall accuracy of each model is as follows: 88.89% for AI or Not, 97.22% for Hive Moderation, 72.22% for Maybe, and 75.0% for Illuminarty. In terms of detecting whether a human character or environment image is AI-generated, the Hive Moderation model has the highest overall accuracy with a 100% accuracy rate in detecting AI-generated human character images, human-created human character images, and human-created environment images, and an 88.89% accuracy rate if the image is AI-generated environment images **(Figure 1)**. AI or Not has a consistent accuracy rate of 88.89% for classifying images of human characters and environments **(Figure 1)**. Maybe's AI detector can correctly classify AI-generated human characters 55.56% of the time, AI-generated environments 44.44% of the time, human-created human characters 88.89% of the time, and human-created environments 100% of the time **(Figure 1)**. Illuminarty can correctly classify AI-generated human characters 44.44% of the time, AI-generated environments 66.67% of the time, human-generated human characters 88.89% of the time, and human-created environments 100% of the time **(Figure 1)**.

The combined result of all four models shows that they can accurately detect whether an image is AI-generated or not 84.7% of the time for images of environments, which is higher than 81.9% for images of human characters, although this difference was not significant in a chi-squared test **(Figure 2)**. Regarding the art and image style, all models except for Illuminarty have a higher accuracy rate in the realism style than traditional **(Figure 3)**. Trends in data show that all four models falsely marked AI images of environments that simulate traditional art media, such as watercolor and abstract oil paintings, as human-created. Models accurately categorize human-generated images more than AI-generated images. Paintings in the cubism style are also an area of low accuracy **(Table 1)**.

To determine if these results are significant, we used a chi-square test. The result is not significant at $p < 0.05$. This result means that we fail to reject the null hypothesis, which states that there is no correlation between the type of subject (or art style) and AI's classification accuracy.

DOI: https://doi.org/10.59720/23-132

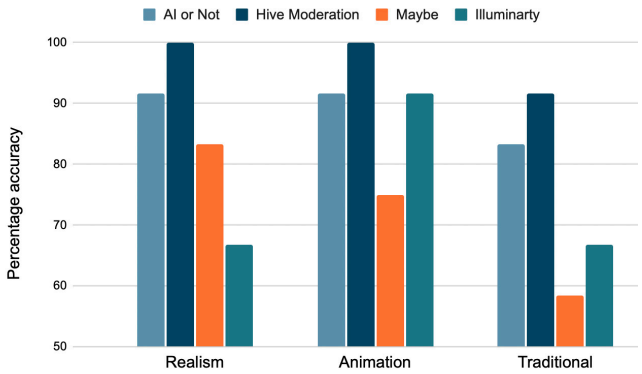

Figure 3: Influence of the art/image style on each model's accuracy. The classification accuracy of AI detectors for images in realism, animation, and traditional styles. An equal sized sample of all three styles is fed to AI detectors, with the accuracy noted. P = 0.142, so the result is not significant at $p < 0.05$.

DISCUSSION

In this study, we found that AI detection models are slightly more accurate at detecting the authenticity of an image if the image is of environments rather than human characters **(Figure 2)**. Certain models, such as Maybe and Illuminarty, are better at correctly classifying human-generated images compared to AI-generated images, but all models struggled with images in abstract painting styles, such as impressionism and cubism **(Figure 3)**. The data show a disparity across the models depending on the type of input image, but the overall accuracy suggests that these models can be acceptable options to combat AI art concerns. Hive Moderation showed the highest overall accuracy, followed by AI or Not, then Illuminarty, and lastly Maybe. This difference may be due to the size of the training data. Hive Moderation processes billions of API calls a month and partners with many social platforms, which contributes to a greater testing and training dataset (13). Maybe, on the other hand, is a personal project trained on images from Reddit, so its data sources are likely more limited (10).

This study has several limitations that influenced data collection. By the nature of categorizing images into different subjects and styles, random sampling cannot be fully ensured. The sample size for this study was limited and there may be other AI-detection models that could yield different results but were excluded from the study due to accessibility reasons. Future research can explore different styles and types of images and artwork. Therefore, this study serves as a rough generalization of the styles surveyed and provides a basic evaluation that indicates future areas of improvement.

The size and variety of training datasets directly determine the accuracy of machine learning models. Of the models used in this study, most were more accurate in classifying human-generated images than AI-generated images. This may be due to a training dataset focusing more on human-generated images than AI-generated images or because the AI images in the training set were much less developed than the current ones. To make sure that progress in AI image generators is not overlooked, developers should constantly improve detectors with new data in online AI libraries. AI or Not and Hive Moderation's greater accuracy for AI-generated images than Maybe's AI detector and Illuminarty may also be a training dataset issue.

The low accuracy for abstract art in traditional mediums indicates another area for development. A noteworthy machine learning model is the Art Recognition algorithm that helps fight art forgery and has correctly detected several paintings masquerading as works of famous artists working in traditional mediums (16). The basis of the algorithm is a deep convolutional neural network trained to detect characteristics specific to an artist from a collection of original works (16). The training materials are good-quality photos of original artworks, and several patches are generated per image to augment the data set. While the Art Recognition algorithm claimed great success with detecting authentic paintings, it is unclear whether or not the algorithm works just as well if the input image is AI-generated, since their training data consists of art by famous human artists. However, well-defined structure and brushstrokes are likely the reason for the algorithm's better performance with impressionists (16). Thus, the models in this study may be falsely flagging traditional paintings simply due to a lack of training with such data. The models haven't fully learned the brushstrokes associated with AI paintings, resulting in overgeneralizing painting textures as human-generated. Accuracy may be improved by feeding traditional paintings to the models' training set and also feeding AI paintings that try to simulate traditional art.

In general, future experiments can explore using a greater set of data from various sources and styles. To a

Subgroups	*AI or Not*	*Hive Moderation*	*Maybe*	*Illuminarty*
AI-character realism	100%	100%	100%	0%
AI-character animation	66.67% (Stable Diffusion)	100%	33.33% (anime and Stable Diffusion)	100%
AI-character traditional	100%	100%	33.33% (watercolor and cubism)	33.33% (watercolor and cubism)
AI-environment realism	100%	100%	33.33%	66.67%
AI-environment animation	100%	100%	66.67%	100%
AI-environment traditional	66.67% (impressionist)	66.67% (impressionist)	33.33% (watercolor and comic)	33.33% (watercolor and impressionist)
Human-character realism	100%	100%	100%	100%
Human-character animation	100%	100%	100%	66.67%
Human-character traditional	66.67% (cubism)	100%	66.67% (cubism)	100%
Human-environment realism	66.67%	100%	100%	100%
Human-environment animation	100%	100%	100%	100%
Human-environment traditional	100%	100%	100%	100%
Total accuracy	88.89%	97.22%	72.22%	75.0%

Table 1: Collected Data on the Percentage of Accurate Detection. Labels: AI: AI-generated, Human: human-generated; character: images of human character, environment: images of the environment; realism & animation & traditional: the art style/type. The colored highlights indicate that the model made an incorrect classification, and the labels in parentheses are the sub-groups that the model predicted incorrectly.

DOI: https://doi.org/10.59720/23-132

	Human Character	*Environment*
Successes	59	61
Failures	13	11

Table 2: Chi-square test for images of human character versus environment. Performing the test results in a chi-square statistic of 0.2. The p-value is 0.654721. The result is not significant at $p < 0.05$.

human eye, the inconsistencies in some of the AI-generated human character images are clear: extra limbs, nonsensical lighting, melting faces and fingers, asymmetric eyes, extra muscles, blurry clothing lines, etc. In the dataset for this research, these characteristics are especially prominent in the AI-generated images of human characters. For future developments, models can learn to recognize and focus on important features such as faces and hands to detect any inconsistencies.

Other factors not explicitly tested in this study influence the model's classification. Maybe's AI art detector has been noted to produce false positives with images involving screenshots, memes, or other 'captioned' images; cropping and adding filters to an image will also influence the result. This may be due to the removal of inconsistencies near the edges of an AI-generated image or DALLE-2 and Stable Diffusion's watermark, which the model may have picked up on (10). The influence of filters complicates the relationship between artists and technology. It is possible that the models are picking up the inconsistent blurred texture or diffused bright colors introduced by filters to human-generated art. To solve these issues, train the models on images of various sizes and with filters. Another possible idea is for models to analyze an image in small pieces and as a whole. Just like how Turnitin's AI detector splits the text into smaller segments so models can analyze images in small chunks and assign a score of AI-generated possibility for each chunk (17). Most of the pieces used in this study are Western so future research could test pieces from other cultures.

While models may learn to recognize patterns and trends in AI art, it is still difficult to define art. For example, plenty of human artists employ inconsistent lighting and anatomy in their artworks. This suggests that even with massive amounts of training, models can still falsely mark human-generated art as AI. It is not clear why the models have an easier time classifying human-generated images compared to AI-generated images, but it's likely due to the fast evolution of AI-generated images which makes it hard to get a representative training dataset for the models.

MATERIALS AND METHODS

Data Sampling

To quantify the data, images were split into a few major categories, with human character and environment as the main subject, in realism (images and illustration in photorealistic style), traditional (photos of traditional paintings in mediums like watercolor and oil painting), and iconic animation (Japanese anime, Disney, etc) styles/types. This amounted to a total of six categories (character realism, character animation, character traditional, environment realism, environment animation, and environment traditional). Each of these categories had two subgroups: human and AI. To ensure a comprehensive review, three samples were gathered for each of these subgroups to achieve a sum of 36 pieces. The categories were chosen for their prevalence in modern media as well as the high rate of modeling in AI programs.

	Realism	*Animation*	*Traditional*
Successes	41	43	36
Failures	7	5	12

Table 3: Chi-square test for different styles. Performing the test results in a chi-square statistic of 3.9. The p-value is 0.142274. The result is not significant at $p < 0.05$.

AI-generated images of human characters were mainly sourced from Lexica, an online library of AI images, as it contains a variety of styles (18). AI-generated images of the environment were sourced from both Lexica and Krea as well as a few from Deviant Art and PaintBot to ensure a variety of styles (19; 20; 22). These libraries mostly feature work generated by Stable Diffusion; pieces generated by other AI tools were used less due to a lack of free libraries online. Human-generated images were gathered from a greater variety of sources: realistic photos of human characters and environments were mostly from Unsplash except for one picture being a charcoal portrait; animation and traditional style artworks were sourced from anime, Disney, past famous artists, current online artists, and movie screenshots. Traditional images, in particular, included impressionism pieces, watercolors, cubism pieces, and comics. Animation images included screenshots from anime, movies, and digital rendering in the animation styles. Each image is only in one sub-group. All sampled images are accessible to the public online and are free to use.

Testing

After image gathering, all images were labeled and sorted into their respective categories. Images were fed into the models one by one, and the results were recorded. The percentage of correct detection is recorded for each category and this is repeated for all four models. The models used were: Aiornot.com (8), Hivemoderation.com (13), AI Image Detector by umm-maybe (14), and App.illuminarty.ai (15).

For Maybe's AI detector and Illuminarty, the results were shown through a percentage of human vs. artificial. If the result is 50-50, it's counted as wrong. Otherwise, the higher percentage determines the final decision, regardless of the closeness between the two percentages.

Counting the number of successes and failures for both images of human character and environment (**Table 2**) and performing the chi-square test produces the chi-square number of 0.2. Using 1 degree of freedom and an alpha of 0.05, the p-value becomes 0.655 which is higher than 0.05. Performing the test for the three styles (**Table 3**) results in a chi-square statistic of 3.9. Using two degrees of freedom and 0.05 alpha gives a p-value of 0.142, which is higher than 0.05.

ACKNOWLEDGMENTS

Author Yunjiao Xiao would like to thank her parents, Qiang Wen and Yan Xiao, for their encouragement and support.

Received: June 19, 2023
Accepted: October 18, 2023
Published: January 5, 2024

DOI: https://doi.org/10.59720/23-132

REFERENCES

1. Greg. "What is AI Art? How Does It Work And Is It Really Art?" *ExpressVPN*, 3 March 2023, www.expressvpn.com/blog/what-is-ai-art/.
2. Vaus, Jacob. "How Does AI-Generated Art Work?" *Built In*, 28 Dec. 2022, builtin.com/artificial-intelligence/how-does-AI-generated-art-work.
3. "Dall·E 2." *DALL·E 2*, openai.com/dall-e-2. Accessed Jun. 9, 2023.
4. "Trending Dreams: Deep Dream Generator." *Trending Dreams | Deep Dream Generator*, deepdreamgenerator.com. Accessed Jun. 9, 2023.
5. "Stable Diffusion Online." *stable diffusion web*, stablediffusionweb.com. Accessed Jun. 9, 2023.
6. "Midjourney." *Midjourney*, midjourney.com/home/. Accessed Jun. 9, 2023.
7. Vincent, James. "An Interview with Midjourney Founder David Holz." *The Verge*, 2 Aug. 2022, www.theverge.com/2022/8/2/23287173/ai-image-generation-art-midjourney-multiverse-interview-david-holz.
8. "Optic AI or Not." *AI or Not*, www.aiornot.com. Accessed Jun. 9, 2023.
9. Hive. "Detect and Moderate AI-Generated Artwork Using Hive's New API." *Hive*, 7 Apr. 2023, thehive.ai/blog/detect-and-moderate-ai-generated-artwork-using-hives-new-classification-model#anchor3.
10. Maybe, Matthew. "Can an AI Learn to Identify 'AI Art'?" *Medium*, 22 May 2023, medium.com/@matthewmaybe/can-an-ai-learn-to-identify-ai-art-545d9d6af226.
11. Liu, Ze., et al. "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows." *IEEE/CVF International Conference on Computer Vision (ICCV)*, 17 Aug. 2021, https://doi.org/10.48550/arXiv.2103.14030.
12. "Big Spring Update." *Illuminarty*, illuminarty.ai/en/posts/big-spring-update.html. Accessed 20 Aug. 2023.
13. "AI-Generated Content Detection." *Hive Moderation*, hivemoderation.com/ai-generated-content-detection. Accessed Jun. 9, 2023.
14. "Maybe's AI Art Detector." *Huggingface.co*, huggingface.co/spaces/umm-maybe/AI-image-detector. Accessed Jun. 9, 2023.
15. "Illuminarty." *Illuminarty.ai*, App.illuminarty.ai. Accessed Jun. 9, 2023.
16. Bailey, Jason. "Can AI Art Authentication Put an End to Art Forgery?" *Artnome*, 12 Sept. 2019, www.artnome.com/news/2019/9/12/can-ai-art-authentication-put-an-end-to-art-forgery.
17. "AI Writing Detection Capabilities." *Turnitin*, www.turnitin.com/products/features/ai-writing-detection/faq.
18. "Lexica." *Lexica*, lexica.art. Accessed May. 1, 2023.
19. "KREA." *KREA*, search.krea.ai. Accessed May. 1, 2023.
20. "Home." *Deviant Art*, deviantart.com. Accessed May. 1, 2023.
21. "PAINTBOT." *PAINTBOT*, paintbotapp.com. Accessed May. 1, 2023.