Research Article

AI Image Detection through Human-Centered Training Methods

Eva Samuel[1], Sahil Dev[2]

[1]Northview High School, 10625 Parsons Road, Johns Creek, Georgia, 30097, United States;
[2]Cornell University, 300 Day Hall Ithaca, New York, 14853, United States

ABSTRACT

As AI-generated images become increasingly prevalent in digital media, the ability to distinguish between real and manipulated content is essential for combating misinformation. Our study investigates whether targeted training can significantly improve individuals' ability to detect AI-generated images. Somoray and Miller (2023) found that deepfake detection accuracy remained low, averaging around 55%, regardless of whether participants were given a list of detection strategies [1]. Our study builds on this by implementing a structured training program, which includes video demonstrations and interactive practice with feedback. We investigate whether detection accuracy improves after participants view videos explaining how to identify deepfakes for AI-image identification instead of just a list of strategies. The experiment began with a pre-test to assess participants' baseline ability to distinguish between real and AI-generated images. Next, two-thirds of the participants received targeted training on identifying inconsistencies, while one-third served as a control group with no training. Finally, we administered a post-test to measure any improvements in their detection skills after training. Demographic and experiential factors such as age, sleep, AI experience, and screen time did not significantly impact detection accuracy. A paired t-test was performed to evaluate the impact of training on detection accuracy, and the results show a statistically significant improvement in detection accuracy post-training ($p=0.009$). A statistically significant positive correlation was found between the time spent analyzing images and detection accuracy ($p < 0.0001$), indicating that more thorough analysis improves performance.

Keywords: AI-generated images; Deepfake detection; Computer vision; Machine learning; Misinformation; Human-AI interaction; Image classification; Training algorithms

INTRODUCTION

AI-generated images have increasingly appeared in various domains, such as entertainment, advertising, and digital media (2). Advancements in artificial intelligence, particularly in generative adversarial networks (GANs), diffusion, and transformer models, have enabled machines to generate high-quality images that are increasingly indistinguishable from images created by humans (3). The rapid advancements of AI-generated images pose significant challenges in various domains, from art forgery to widespread misinformation. The goal of our study is to evaluate whether a structured training program can significantly improve individuals'

Corresponding author: Eva Samuel, E-mail: eva10samuel@gmail.com.

Received June 18, 2025**; Accepted** August 10, 2025
https://doi.org/10.70251/HYJR2348.34281289

ability to detect AI-generated images. We hypothesize that participants who undergo targeted training will demonstrate a measurable increase in detection accuracy compared to those who receive no training. By providing video demonstrations, we aim to investigate whether a more immersive and guided approach to AI image identification leads to improved results. This study aims to provide insights into the effectiveness of such training programs in enhancing the public's ability to identify AI-generated content.

LITERATURE REVIEW

One focus is on the technical development of AI algorithms to improve the realism, diversity, and efficiency of image generation (4). Cave *et al*. [2019] and Brundage *et al*. [2020] identify the legal and ethical implications of using AI-generated content, including issues related to property rights, privacy concerns, and the potential for misuse or manipulation (4, 5). For example, Johnson *et al*. [2020] discussed the broader implications of AI-generated images for national security. It highlights the potential for misuse in misinformation and cyber attacks. The improvement of deepfakes raises concerns about their potential misuse, particularly for national security. They report existing research on deepfakes and their threats to national security (6).

While research has focused on the technical advancements and applications of AI-generated images, our study builds on how humans can be trained to detect AI-generated portraits. By restricting our investigation to portrait photographs, we can better understand the specific visual cues and inconsistencies unique to AI-generated images. Ha *et al*. [2024] found that non-artist participants had an accuracy rate of approximately 63% when attempting to distinguish between AI-generated and human-created art. In contrast, trained professional artists performed slightly better, reaching around 72% accuracy. Supervised classifiers, like the Hive model, achieved an accuracy rate as high as 85% in distinguishing AI-generated images from human-made ones, surpassing even the expert human performance (7). Lu *et al*. [2023] provided an understanding of the differences between humans and AI models when evaluating AI-generated images. With the rise of sophisticated AI image generation techniques, there's a growing concern about the spread of deepfakes and other manipulated media. The study uses a large dataset (Fake2M) containing real and AI-generated images. Humans performed worse than random guessing (50%), with only a 38.7% success rate in differentiating real from AI-generated images. Factors like age, gender, and experience with AI-generated content did not significantly affect human performance. The best AI detection model achieved a 13% error rate (8).

The study by Bhatt and Varghese (2022) provides a global perspective on human detection of AI-generated media, revealing low accuracy rates across various countries. The authors highlight that cultural differences and differing levels of exposure to digital media contribute to the discrepancies in detection performance. Despite variations, participants consistently struggled to correctly identify AI-generated audio, images, and text. Most notably, participants tended to classify AI-generated media as human-made. German participants performed better in detecting AI-generated audio, likely due to the lower quality of AI-generated samples in the German language. This highlights a potential need for more unbiased testing data in this area of research (9). A seminal work investigated the confidence levels of individuals in detecting deepfakes, revealing that people often overestimate their abilities, leading to susceptibility to deception. Offering financial incentives did not significantly improve detection accuracy. People showed a bias towards mistaking deepfakes for real videos. The study suggests that people tend to take videos at face value unless they find clear-cut evidence of it being fake and have lots of overconfidence in their abilities. This makes them very susceptible to deepfakes (10).

These findings highlight a key limitation in human perception: individuals are easily fooled by sophisticated deepfakes. The study underscores the need for more research into the cognitive processes behind this inability. Groh *et al*. [2021] compared the effectiveness of different detection methods (7). They concluded that machine-informed human crowds outperform both standalone human and machine detectors. Humans and machines make different types of errors. Humans were generally more successful at identifying deepfakes based on facial inconsistencies, but they struggled particularly with deepfakes that disrupted facial features such as symmetry, reflections, or blending textures. On the other hand, machine models excelled in detecting pixel-level anomalies, such as noise or texture irregularities, that humans may not notice (11).

Given the challenges associated with detecting AI-generated images, it is paramount to try to improve detection accuracy. However, existing approaches have failed to generate improvements in detection accuracy. Notably, Somoray and Miller [2023] attempted to improve

detection accuracy by providing viewers with a list of detection strategies and compared their performance on identifying manipulated videos against their baseline performance (13). Before receiving any strategies, their accuracy was around 60.7%. After being given detection strategies, their accuracy slightly improved to 62.1%. Participants who received strategies tended to overestimate their detection abilities. They were not able to find a statistically significant correlation between receiving detection strategies and higher accuracy ($p > 0.05$) (1). Thus, there is still a gap in effective training to improve AI-generated image detection accuracy.

METHODS AND MATERIALS

In the experiment, we test the participants' abilities to identify AI-generated and authentic images before any training using a test made with Google Forms. The pre-test comprised 20 images, nine real and 11 AI-generated, all randomly chosen. We source images from publicly available datasets such as the FFHQ (Flickr-Faces-HQ) (12) and generate them using advanced AI algorithms like StyleGAN2 and take them from online communities such as reddit.com/r/midjourney (13). Appendix Figure 1 shows the fourth image on the pre-test, an example that was commonly identified correctly as real. Participants were instructed to classify each image as 'AI-generated' or 'Real.' The other questions on the test were demographic questions, such as their screen time, familiarity with AI, and age range. We record detection accuracy, confidence in answers, and specific reasons that caused subjects to choose the answer they chose.

This study was conducted as part of an educational outreach initiative and was determined to involve no more than minimal risk to participants. As such, formal IRB or ethics board approval was not sought. While the study was not anonymous, no sensitive personal identifiers (e.g., names, contact information) were collected. Participants' responses were associated only with non-sensitive metadata such as age, AI experience, and screen time to allow for exploratory subgroup analysis. Participation was entirely voluntary, and all participants, and, where applicable, their legal guardians, provided electronic written consent prior to participation.

Our experiment tested 39 individuals. We randomly assign participants to one of two groups. Twenty-eight individuals received training on identification strategies for AI-generated images, and 11 individuals acted as the control group, getting no training and taking the test with their baseline knowledge. Participants ranged in age from 14 to 54 years. Familiarity with AI ranged from none to high (hands-on work with AI for several years).

Participants were randomly assigned to either the control group ($n = 11$) or the training group ($n = 28$). The group sizes were not equal by design. We expected greater variability in how participants responded to the training, and thus allocated more participants to that group to better capture a range of outcomes. Because of this, we chose to have a larger training group to better capture that range of possible outcomes.

We identify two main categories of inconsistencies for determining whether an image is AI or real. The first category, facial feature inconsistencies, includes unnatural proportions, asymmetrical features, and irregularities in skin texture. The second category, general image anomalies, includes inconsistencies in lighting, shadow placement, and the interaction of subjects with their backgrounds. AI systems frequently struggle to replicate realistic lighting conditions, leading to noticeable discrepancies that can be critical for identification.

We then train our participants with a series of three narrated videos, all around two minutes in length. The videos were posted on YouTube and sent to the participants. These videos could be paused or replayed at any time. The first explores the background of AI-generated images in depth. The second explains facial feature inconsistencies with AI-generated images. The third describes general image inconsistencies.

The second video reviews facial feature inconsistencies. It gives instruction on recognizing unnatural proportions and asymmetrical features, common indicators of AI-generated images, and training on identifying irregular skin texture and inconsistencies in details such as pores and wrinkles, which AI models often struggle to replicate accurately. It identifies common issues in AI-generated images, such as unnatural reflections in the eyes, inconsistent shading, and misaligned features around the mouth. It also taught participants to spot anomalies in hair strands and background elements, which may appear blurred or artificially blended in AI-generated images.

The third video discusses general image inconsistencies. It provides insight into identifying inconsistencies in lighting and shadow placement, which can indicate an AI-generated image due to the difficulty of accurately replicating natural lighting conditions. The video teaches participants to observe the overall context of the image, including background elements and interactions between the subject and their surroundings,

to identify discrepancies. The videos emphasize the importance of consistent texture and detail across the entire image, as AI-generated images often exhibit variations in these areas.

Following the training, participants were administered for a post-test similar in structure to the pre-test. This post-test included 20 images, with 9 AI-generated and 11 genuine human portraits. Images in the post-test were different from those used in the pre-test.

RESULTS

The primary metric for evaluation was the accuracy of participants' classifications, defined as the age of correct identifications of AI-generated and genuine images. The baseline score is 50% by random guessing. To analyze the data, we compared pre-test and post-test accuracy within the training group and also compared results to a control group that received no training. Figure 1 presents the pre- and post-training accuracy per-participant in the treatment group. Appendix Figure 2 presents the distribution of the frequency of certain detection accuracies for the pre-test, with a mean of approximately 49.64% and a median of 50%, close to the 50% baseline for random guessing.

Figure 2 shows the distributions of the pre- and post-training accuracy in the treatment group. A paired t-test was performed to evaluate the impact of training on detection accuracy, and the results show a statistically significant improvement in detection accuracy post-training ($p=0.009$). The probability of the observed improvement being due to random chance is less than 1%.

Similarly, Figure 3 shows the pre- and post-training accuracy per-participant in the control group, while Figure 4 shows the distributions of the pre- and post-training accuracy in the control group. We performed the same paired t-test on the control group and found no significant difference in scores between the pre-test and post-test ($p=0.341$). This means that the improvement

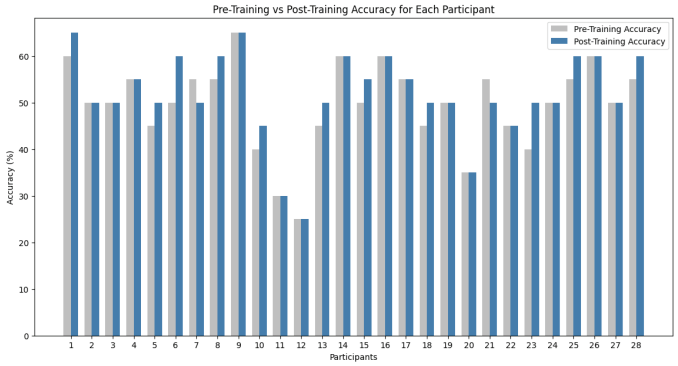


Figure 1. Pre- and post-training classification accuracy for each participant in the treatment group.

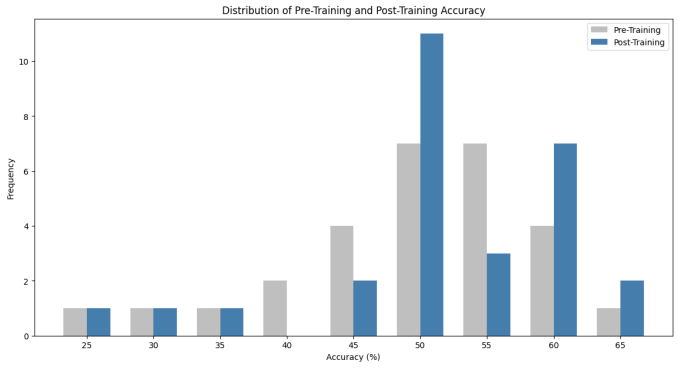


Figure 2. Distribution of pre- and post-training classification accuracy in the treatment group.

Figure 3. Pre- and post-training classification accuracy for each participant in the control group.

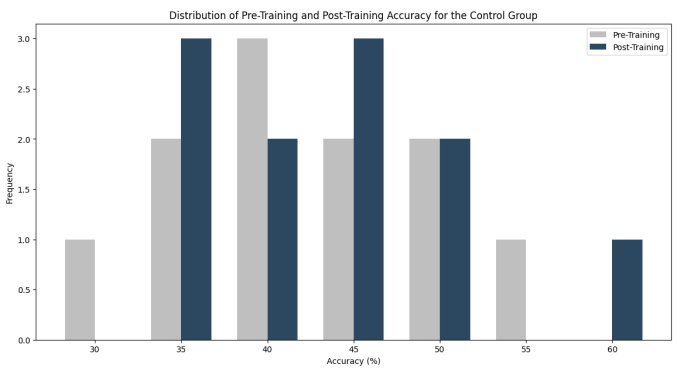


Figure 4. Distribution of pre- and post-test accuracy in the control group.

shown in the experimental group is unlikely to be due to factors other than the training.

After establishing the significant effect of video training on participant accuracy, we then examined whether other factors were correlated with improved performance. Due to our limited age data, we cannot find a significant correlation between participants' age and detection accuracy ($p=0.271$). However, we did find a strong positive correlation between the time taken to analyze an image and detection accuracy. A Pearson correlation analysis was conducted to examine the correlation between the time taken to analyze an image and detection accuracy. The analysis produced a correlation coefficient of $r=0.743$ ($p < 0.0001$). This strong positive correlation allows us to reject the null hypothesis (H_0), showing that longer analysis times are associated with higher detection accuracy. The correlation of time taken on the test with detection accuracy is shown below in Figure 5. Appendix Figure 3 presents the distribution of the time taken to identify all of the images, with a mean of approximately 25.75 minutes and a median of 27.0 minutes.

We found no significant correlation between sleep on the previous night and test performance ($p=0.552$) (see Figure 6).

We also hypothesized that experience with AI might correlate with test performance. Participants self-reported a low, medium, or high level of experience (see Figure 7). We used ANOVA to compare detection accuracy across different levels of AI experience. We were unable to show a statistically significant correlation between AI experience and detection accuracy ($p=0.150$).

We hypothesized that a higher daily screen time correlates with a higher amount of time spent on each test, due to the screentime causing a lack of focus. The relationship between daily screen time and the time taken to identify deepfakes was assessed (see Figure 8). We were unable to show a statistically significant correlation between screen time and time taken to identify AI-images ($p=0.163$).

The results indicate that training significantly improves participants' ability to detect AI-generated images. A strong positive correlation exists between the time taken to analyze images and detection accuracy. However, other factors, such as age, sleep, experience with AI, and familiarity with subjects, did not significantly impact detection accuracy. These findings emphasize the importance of targeted training for enhancing detection skills while identifying areas for further research.

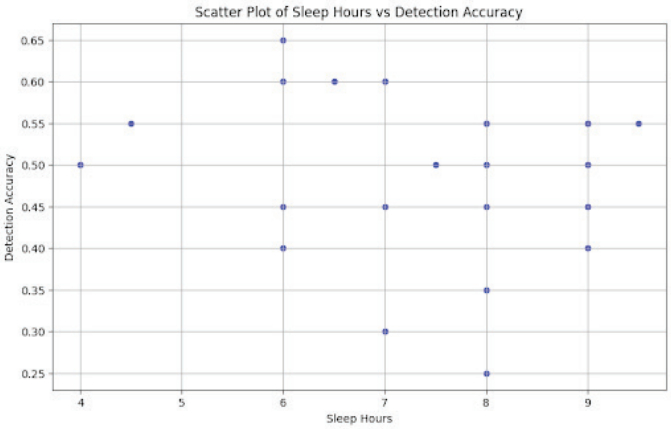


Figure 6. Relationship between hours of sleep the previous night and detection accuracy.

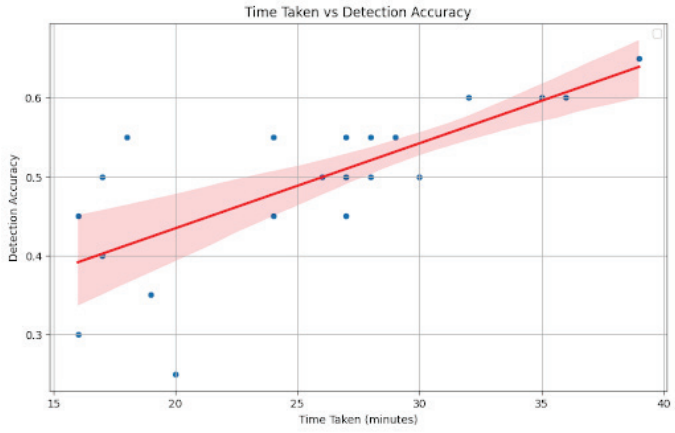


Figure 5. Correlation between time taken per image and detection accuracy.

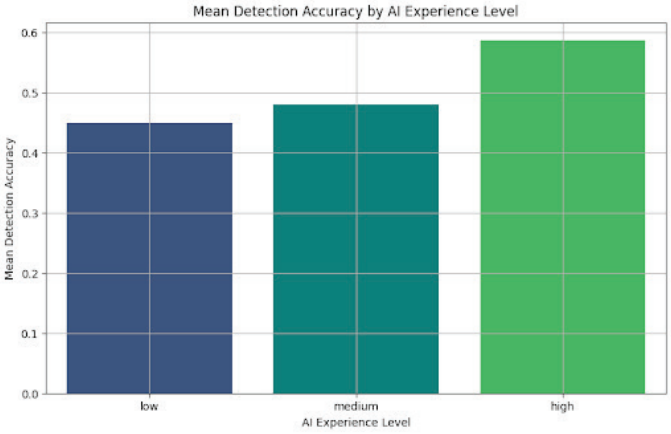


Figure 7. Detection accuracy across self-reported levels of AI experience.

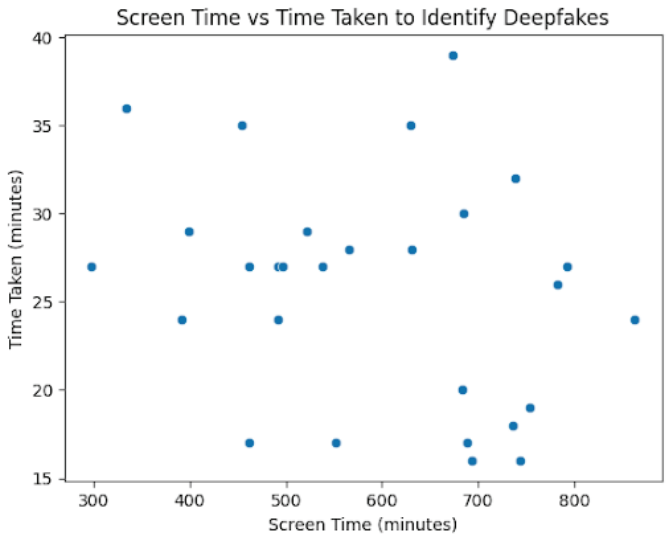


Figure 8. Correlation between daily screen time and average time taken per image.

DISCUSSION

By equipping participants with specific strategies for identifying AI-generated content, we observe a significant increase in their ability to discern between real and synthetic images. The significant improvement in detection accuracy following the training program can be attributed to several key factors related to the videos and interactive components of the training. First, the videos provided participants with clear, visual examples of what to look for when identifying AI-generated images. By highlighting subtle inconsistencies, such as unnatural lighting, irregular shadows, and distorted facial features, participants were able to internalize these specific strategies and apply them more effectively during the post-test. The ability to see these patterns in real-time likely helped participants better recognize them in unfamiliar images.

Moreover, the training emphasized not only what to look for, but why these visual markers are common in AI-generated images. Brief explanations of how generative models work, such as the way GANs or diffusion models can struggle with spatial coherence, helped participants understand the underlying reasons behind these telltale signs. This deeper understanding likely enhanced their ability to generalize their knowledge to a wider range of image types.

Overall, the combination of clear visual instruction, interactivity, and foundational knowledge created a powerful learning experience. These findings underscore the importance of accessible, targeted training programs in enhancing digital literacy and preparing individuals to navigate a media landscape increasingly influenced by synthetic content.

The correlation between the time spent analyzing the images and detection accuracy also shows an important aspect of visual analysis. The strong positive relationship indicates that participants who invest more time examining images are likely to better distinguish between AI-generated and genuine content. Longer examination periods should be encouraged, as they allow individuals to identify the more subtle discrepancies.

We could not identify significant relationships for several factors, including age, sleep, experience with AI, and familiarity with subjects. Due to the lack of well-distributed data, we could not establish a statistically significant correlation across all age groups. The lack of correlation between sleep and detection accuracy suggests that sleep loss may not significantly impact cognitive performance in this context. Further research is necessary to explore the nuances of these relationships.

The lack of significant differences in detection accuracy across varying levels of AI experience ($p=0.150$), suggests that while no definitive conclusion can be drawn from this data, there may be a trend toward improved detection accuracy with greater AI experience. The p-value indicates that the result is not statistically significant, but it is still worth noting that with a larger dataset, this trend could potentially reach statistical significance. Future research with a larger sample size is needed to provide more conclusive insights into the relationship between AI experience and detection accuracy.

FUTURE WORKS AND LIMITATIONS

Future work may include investigating different training methods, such as integrating interactive learning tools that may enhance engagement and retention of detection strategies.

Another area for future research is exploring background factors' role in detection accuracy. Studies could examine how different backgrounds, such as varying levels of digital literacy or familiarity with specific image types, influence individuals' ability to detect AI-generated content.

Given the rapid advancements in AI technology, future work should also focus on the effectiveness of detection methods as AI-generated images become

increasingly realistic.

One limitation is that as AI images improve, it might eventually be impossible to tell them apart from real ones. More research is needed to support policy and technology solutions that ensure transparency in AI-generated content.

Our small sample size does not represent the broader population of Americans.

A larger, more diverse sample could show statistically significant trends in how various demographic factors perform and provide a better understanding of how various demographic factors influence detection accuracy. Future studies should include participants from varied backgrounds, ensuring a more representative sample that can provide insights applicable to a broader audience.

Variations in testing environments and individual participant characteristics may also influence results. Future studies should consider standardizing testing conditions and incorporating measures to ensure the test environment is the same for all participants.

CONCLUSION

We investigated the effectiveness of training on the detection accuracy of AI-generated images, focusing on identifying strategies that can enhance individuals' ability to discern between real and synthetic content. The findings demonstrate that targeted training significantly improves detection accuracy, highlighting the importance of educational interventions in enhancing digital literacy in the age of AI.

The study revealed a positive correlation between the training program and detection accuracy, suggesting that individuals who receive specific training on identifying AI-generated images are better prepared to identify AI images. The training group showed more variability in results than the control group, which aligns with our initial assumption that the effects of training would differ across participants. This justified our decision to assign more participants to the training group in order to better capture a range of learning outcomes.

As AI technology advances, there is a need for ongoing research and the development of effective training programs to equip individuals with the skills necessary to navigate a landscape increasingly populated by synthetic content. Once AI gets good enough, training programs alone may not be sufficient to help distinguish between reality and AI-generated content. Using AI tools themselves may become the only viable solution for identifying AI-generated content in the future.

In conclusion, this study demonstrates the effectiveness of training in enhancing the detection accuracy of AI-generated images and highlights the importance of time spent analyzing these images. While several demographic and experiential factors did not significantly impact the study, the results suggest that targeted interventions focused on teaching specific detection strategies can significantly improve performance. As AI technology continues to evolve, ongoing research will be essential to develop effective detection methods and training programs that adapt to the changing landscape of digital content.

DECLARATION OF CONFLICT OF INTERESTS

The authors declare that there are no conflicts of interest regarding the publication of this article.

APPENDIX

Links to training videos:

- What is an AI-generated image? (https://www.youtube.com/watch?v=Kl3suwWtOQc)
- Identifying AI-generated Images based on Facial Features (https://www.youtube.com/watch?v=RLi05IvnPJQ)
- Identifying AI generated images with General Inconsistencies (https://www.youtube.com/watch?v=BcXGv3X-sdk)

REFERENCES

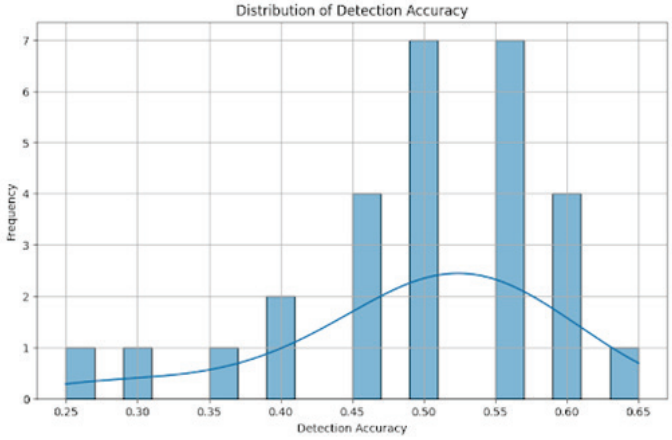
1. Bhatt S and Varghese A. A representative study on human detection of artificially generated media across countries. *Int J Artif Intell*. 2022; 39 (4): 234–256.
2. Brundage M, Avin S, Wang J, Belfield H, *et al*, Toward trustworthy AI development: Mechanisms for supporting verifiable claims. *Harvard Data Sci Rev*. 2020; 2 (1): Available from: https://doi.org/10.1162/99608f92.85f9fd23 (accessed on 2025-3-22)
3. Cave S, Óh Éigeartaigh SS, and Whittlestone J. Transparency and the use of machine learning in the public sphere. *Philos Trans R Soc A Math Phys Eng Sci*. 2019; 377 (2142): 20180142. Available from: https://doi.org/10.1098/rsta.2018.0142 (accessed on 2025-2-27)
4. Reddit Community. MidJourney Images. Available from: https://www.reddit.com/r/midjourney (accessed on 2025-4-12)
5. Forbes Technology Council. AI utilization in the media and entertainment world. Available from: https://www.forbes.

com/councils/forbestechcouncil/2024/05/07/ai-utilization-in-the-media-and-entertainment-world/ (accessed on 2025-4-01)

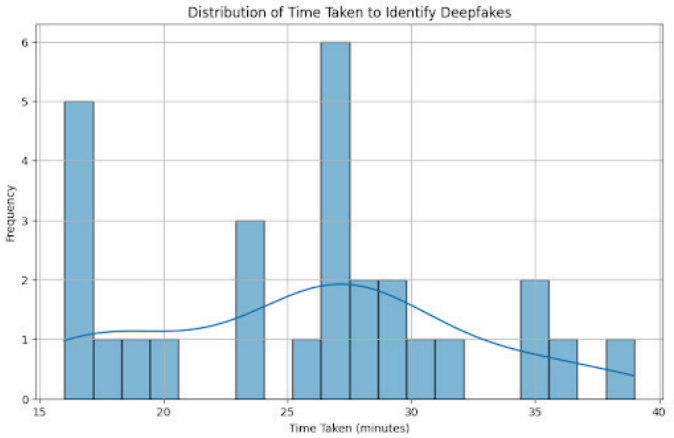
6. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, *et al*. Generative adversarial nets. *Adv Neural Inf Process Syst*. 2014; 27: 2672–2680. Available from: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf (accessed on 2025-3-09)
7. Groh A, Yang W, Wu L, Shah A, and Seltzer M. Deepfake Detection by Human Crowds, Machines, and Machine-Informed Crowds. *arXiv preprint*, 2021. arXiv:2105.06496. Available from: https://arxiv.org/abs/2105.06496 (accessed on 2025-1-30). https://doi.org/10.1073/pnas.2110013119
8. Ha AYJ, Passananti J, Bhaskar R, Shan S, *et al*. Organic or diffused: Can we distinguish human art from AI-generated images? In: *Proc. ACM Conf. on Computer and Communications Security (CCS)*, Salt Lake City, UT, 2024. https://doi.org/10.1145/3658644.3670306
9. Johnson A and Smith B. Deep fakes and national security. *J Strateg Stud*. 2020; 43 (6): 820–846.
10. Köbis NC, Dolezalová K, and Soraperra I. Fooled Twice: People Cannot Detect Deepfakes But Think They Can. *J Media Psychol*. 2021; 33: 105–117. https://doi.org/10.1016/j.isci.2021.103364, https://doi.org/10.2139/ssrn.3832978
11. Lu Z, Wang L, Chen J, and Zhang T. Seeing is not always believing: Benchmarking Human and Model Perception of AI-Generated Images. *arXiv preprint*, 2023. arXiv:2304.13023. Available from: https://arxiv.org/abs/2304.13023 (accessed on 2025-2-14)
12. NVIDIA Research. Flickr-Faces-HQ Dataset (FFHQ). Available from: https://github.com/NVlabs/ffhq-dataset (accessed on 2025-3-04)
13. Somoray K and Miller T. Providing Detection Strategies to Improve Human Detection of Deepfakes: An Experimental Study. *J Appl Psychol*. 2023; 42: 153–168. Available from: https://doi.org/10.1016/j.jap.2023.02.004 (accessed on 2025-4-16)

APPENDIX FIGURES

Appendix Figure 1. This is an example of an image that was commonly identified incorrectly as real.



Appendix Figure 2. Distribution of the frequency of certain detection accuracies for the pre-test.



Appendix Figure 3. Distribution of the time taken to identify all of the images.