ECE324 - Final Report

Classifying Artwork: AI vs Human

4.15.2023

—

Rainey Fu (1006949929), Humzah Khan (1007236215), Tabitha Kim (1006788736)
University of Toronto

Abstract

The integration of artificial intelligence (AI) into the art world has significantly transformed the creative landscape, introducing novel forms of artistic expression while raising crucial ethical implications. This report presents an AI model developed to classify art as human-made or AI-generated, achieving an overall accuracy rate of over 90%. We also introduce a novel prompt filtering technique using NLP tools and K-means clustering to greatly improve performance. By distinguishing between human and AI-generated artwork, the model aims to address pressing ethical concerns such as transparency, authorship, art valuation, market impact, and intellectual property and copyright issues. However, the model has limitations and does not fully address the complexities and evolving nature of these concerns.

Transparency and authorship are essential for maintaining the integrity of the art world. Accurate attribution of creative works is vital for the recognition and credibility of both artists and AI developers. Our AI model plays a role in promoting transparency by distinguishing between human-made and AI-generated artwork, thus reducing the chances of false attributions. However, the model does not explicitly address concerns around the potential loss of human creativity or the implications of AI-generated art on the broader concept of artistic authorship as a whole.

Potential future ethical considerations include public perception and stigmatization of AI-generated art, which may lead to a dichotomy between human and AI creations, affecting the market value and reputation of artists working with AI. Our model's ability to classify art could inadvertently contribute to this divide, necessitating careful consideration of how to maintain a balanced view of AI-generated art. Furthermore, unintended consequences on emerging artists arise as our AI model might unintentionally favor established artists, while less-known creators might be misclassified as AI-generated. Transparency in the classification process is essential to prevent biases and protect opportunities for emerging talents.

Moreover, AI accountability and liability pose challenges in attributing responsibility when the AI model's misclassification leads to economic or reputational damages. Developing legal frameworks and policies that address accountability and liability in art classification is crucial to maintaining trust in the system. Additionally, the dependence on AI and loss of human expertise threatens the preservation of valuable human skills, knowledge, and intuition in the art world as a whole as AI technology becomes increasingly integrated. Balancing the benefits of AI classification with the need to value human expertise may demand a combination of expert and AI collaboration.

Our source code can be found at the following GitHub Repo.

Table of Contents

1.0 Introduction

The art industry has long been a realm of creativity, expression, and innovation. However, in recent years, artificial intelligence (AI) has begun to make its mark on this domain, transforming artistic practices and giving rise to novel forms of expression. As AI-generated art takes the art world by storm, it not only captivates audiences but also raises complex ethical questions that demand thorough examination and responsible solutions. In this report, we unveil an innovative AI model designed to classify art as human-made or AI-generated with a remarkable accuracy of 90%, addressing these ethical challenges head-on while staying true to our responsibilities as ethical engineers.

This report is structured into six main sections, providing a comprehensive overview of our groundbreaking AI model and its potential impact on the art world. In the Prior Work section, we delve into the key influences and existing research in the field of AI-generated art and classification models. Drawing inspiration from the pioneering works of Corvi et al. (2022) and Fu et al. (2002), we set the stage for our own work. We also compare the benefits and tradeoffs of using ResNet-50 and ResNet-101, two state-of-the-art architectures that contribute to the model's efficacy. The Data Collection and Handling section outlines our meticulous approach to gathering and managing data, ensuring a robust foundation for training and testing our model. Next, we explore the Data Cleaning and Handling section, where we detail the crucial preprocessing and filtering steps that guarantee the reliability and suitability of our dataset.

In the Implementing and Training the Models section, we illuminate the technical aspects of our AI model, discussing its implementation, training process, and performance evaluation. Here, we highlight the model's achievements, limitations, and the transformative role of the novel prompt filtering algorithm. Lastly, the Ethical Implications section delves into the complex ethical landscape surrounding AI-generated art. We address existing concerns, such as transparency and authorship, art valuation and market impact, intellectual property, and copyright, while also discussing potential future ethical considerations that may arise as AI continues to influence the art world.

Upon completing this report, readers will gain a thorough understanding of our AI model's capabilities, the ethical implications arising from its use in the art world, and the importance of ongoing collaboration between artists, AI developers, legal experts, and other stakeholders. It is through this interdisciplinary partnership that we can ensure a responsible and innovative integration of AI technology in the field of art, continuing to captivate and inspire audiences for generations to come.

2.0 Prior Work

The progress of AI-generated content, particularly diffusion models, has necessitated the development of innovative methods for distinguishing between genuine and synthetic images. In our project, we have built upon the findings of previous studies in the field of AI-generated content recognition. Our research focuses specifically on the works of Corvi et al. (2022) and Fu et al. (2002).

Corvi et al. (2022) investigated the problem of distinguishing between synthetic images made by diffusion models and pristine images [8]. They investigated the spectral characteristics of various diffusion models, including Stable Diffusion, Latent Diffusion, and GLIDE. The researchers detected notable peaks in the spectra produced by these common diffusion models by examining noise residuals and performing the Fourier transform. In contrast, they discovered that models like DALLE2 and ADM produced weaker spectra, which presented a greater challenge in detecting AI-generated content.

The findings of Corvi et al.'s study had a significant influence on our project. We expanded on their work by employing a pre-trained ResNet-101 model rather than the ResNet-50 model they used, to improve performance. Furthermore, their observation that traditional tactics, such as using residuals as input and radical augmentation, only delivered minor advantages inspired our approach to data processing and model training. Furthermore, the trend they discovered about the weaker spectra produced by the DALLE2 and ADM models also was visible in our results. We aimed to develop a more robust and accurate system for detecting synthetic images in the art domain by connecting our work with the insights provided by Corvi et al.'s study.

Fu et al. (2002)'s "Detecting GAN-generated face images via hybrid texture and sensor noise-based features" was another notable paper that influenced our research. This work used texture and sensor noise to recognize GAN-generated face images [9]. While this study focuses on GAN-generated images rather than diffusion models, it nevertheless provides useful context for our work because it tackles the broader problem of distinguishing between genuine and synthetic images.

Our project, classifying AI-generated and human-crafted artwork, drew inspiration from and expanded on these related prior works to maximize our performance.

2.1 ResNet-50 vs. ResNet-101: Key Differences and Motivations

ResNet-50 and ResNet-101 are two widely used deep residual network architectures, with the primary difference lying in the number of layers they possess. ResNet-50 consists of 50 layers, while ResNet-101 contains 101 layers [10]. The increased depth of ResNet-101 allows the network to learn more complex features and achieve higher accuracy in some tasks, such as image classification [11]. However, the additional layers also lead to increased computational requirements and training time, which may pose challenges when working with large datasets or under resource constraints [12].

In our study, we implemented ResNet-101 to investigate their performance in classifying art as human-made or AI-generated. Our motivation for trying ResNet-101 was to explore the potential improvements in model accuracy and performance stemming from its increased depth and capacity to learn more intricate features. Although the tradeoffs in computational requirements and training time were considered, the pursuit of enhanced accuracy justified our exploration of ResNet-101 as a viable alternative to ResNet-50 as was used in the work by Corvi et al.

3.0 Data Collection and Handling

3.1 Data Collection

Before training our model, we need to first collect our dataset. Thankfully, due to the nature of diffusion models, data gathering is relatively easy. We were able to find large datasets with images for each of our classification labels: Midjourney [13], Stable Diffusion [14], Dalle 2 [15], and authentic [16]. Overall, we have over 2 billion different images. However, we have neither the time nor resources to train with such a large dataset. Our dataset is a subset comprised of 20,000 images approximately evenly split between the four labels. For future work, we plan on collecting data from a larger variety of sources, especially authentic images, to improve the generalizability of our data.

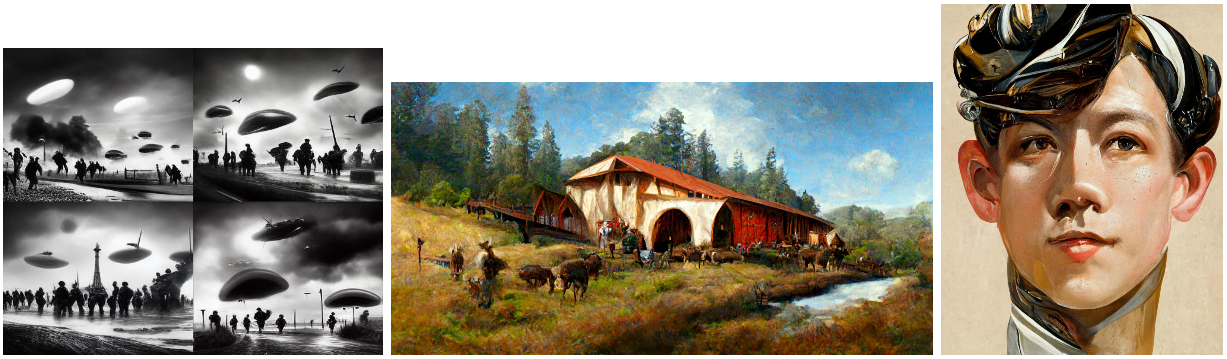

Figure 1: Images from Midjourney Dataset

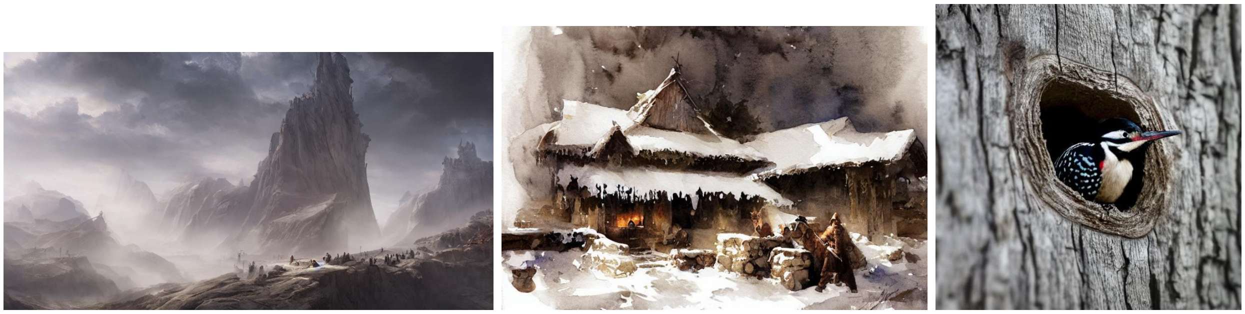

Figure 2: Images from Stable Diffusion Dataset

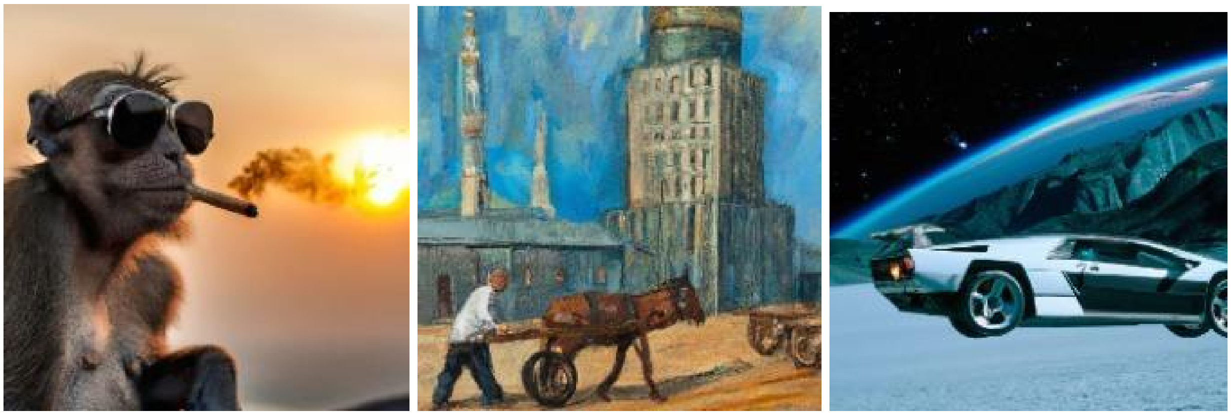

Figure 3: Images from Dalle 2 Dataset

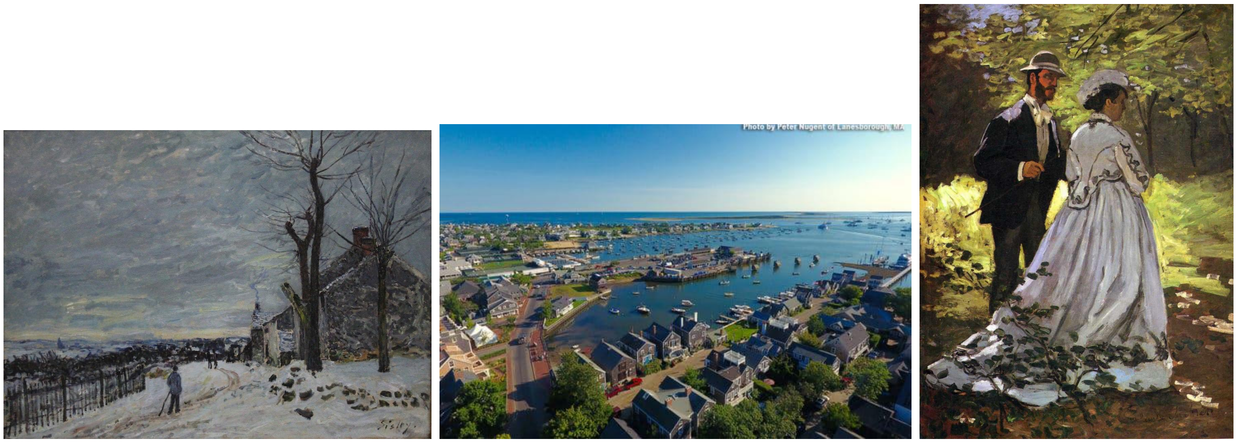

Figure 4: Images from Authentic Dataset

3.2 Data Cleaning and Handling

We also need to do some data cleaning so our data is in the correct format. For authentic and Midjourney images, we have URLs to the image. To handle this, we make a request to the URL and attempt to save it. If successful, we save the image to the hard drive. Similarly, for Stable Diffusion and Dalle 2, we have the images already so we simply load them into memory.

Additionally, one of the issues we faced was RAM usage. Google Colab has a limit of 12 GB of ram. Since we have a large dataset, we are unable to load all of our training data at once. To optimize our memory usage, we first download the image and save it to the hard drive with a unique id. Thus, instead of saving the image itself, we can save the path to the image as a string and load the image when we actually use it. We define a custom ImageDataset object to act as a torch.utils.data.Dataset that also handles loading the image from a string.

To improve the generalizability of our dataset, we use a variety of different data augmentations: random horizontal flip (p=0.5), random rotation (15 degrees), random crop (scale = (0.8, 0.1), ratio = (0.75, 1.333)), and color jitter (brightness=0.2, contrast=0.2, saturation=0.2, hue=0.1). For future work, we plan on tuning the intensity of the augmentations and also adding Gaussian noise.

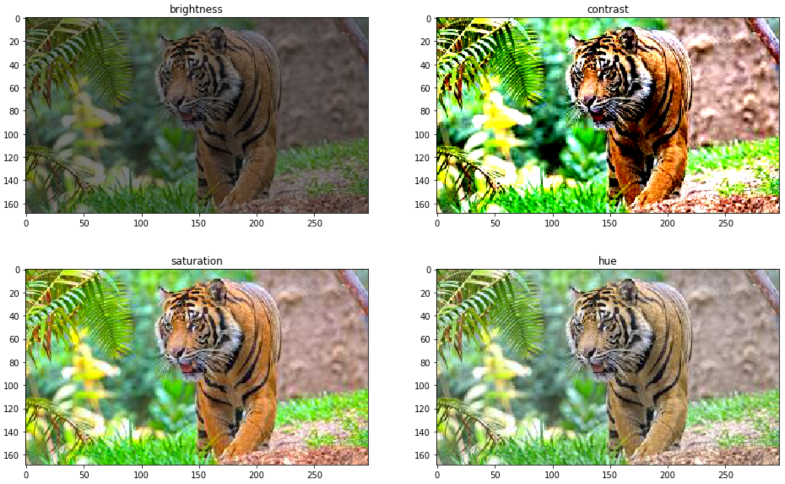

Figure 5: Data Augmentation Examples

4.0 Implementing and Training the Models

After loading the images, we now have everything necessary to begin training. We define our training set as 90% of our total dataset. This leaves us with more than 2000 images in our test set, more than enough to be statistically significant. Before training, we randomly order our training data. We use a pre-trained version of ResNet-101 and replace the last layer with a four-output fully connected linear layer. This will now output a 4-dimensional vector that represents the probabilities of each label. For future work, we plan on including an initial layer for residual extraction and having no down-sampling in the first layer.

We also have various hyperparameters that can be fine-tuned. Specifically, we are using a batch size of 64, 20 epochs, an initial learning rate of 0.01, and stochastic gradient descent. We can further tune the exact values of the batch size, epochs, and learning rate. We can also explore using a different optimizer such as ADAM. For our learning rate, we adaptively lower the learning rate by a couple of magnitudes when the performance of the model plateaus. We also utilize early stopping, to stop the training process before the model overfits to the training data.

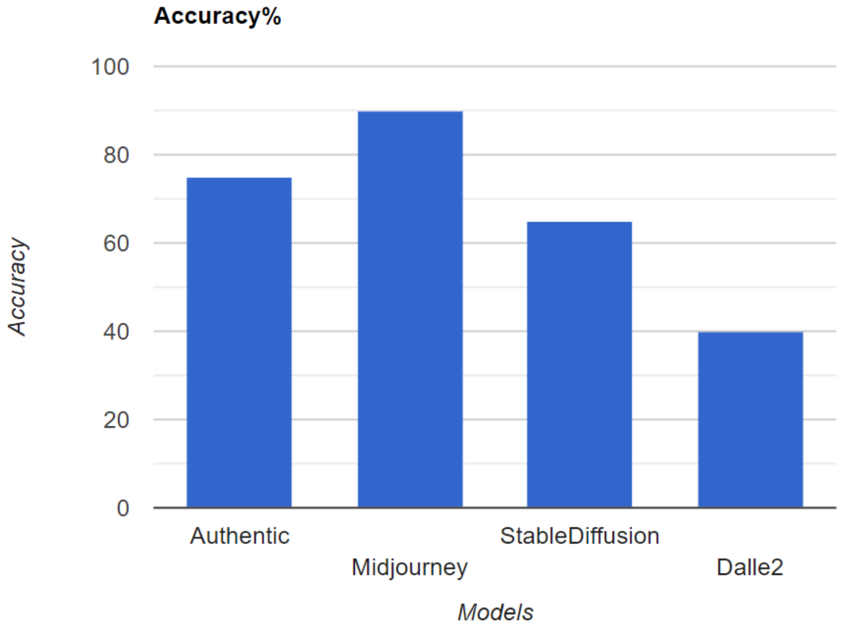

Figure 6: ResNet-101 Different Model Performance

As you can see in the figure above, the accuracy for Dalle 2 is the lowest. This corroborates previous work [8] that says Dalle 2 leaves less significant traces compared to the other models. Since there are 4 different outputs, the accuracy should be at least 25%.

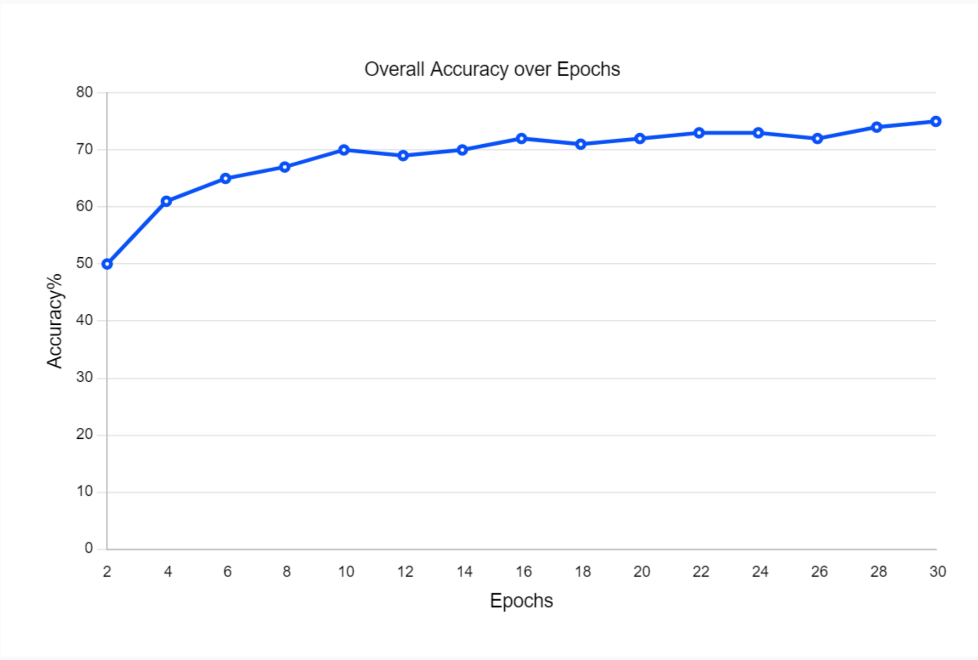

Figure 7: ResNet-101 Accuracy over Epochs

Figure 7 shows our model's overall accuracy over epochs. The benefits from adaptively changing the learning rate can be seen at epochs 12, 18, and 26 where small improvements in accuracy can be seen shortly after. Our accuracy is relatively good. This is probably because we are using ResNet-101 compared to previous works which used ResNet-50. We wanted to expand on previous work and further improve the performance. However, ResNet-101 is more computationally expensive and uses more RAM. This makes the training process longer and also allows us to load fewer images into RAM at once.

Next, we want to explore how the training data can impact the performance of our model. Particularly, we are concerned about biases in the data. Since the users of diffusion models are not representative of the entire population, they may generate art that only fits their interests. For example, they may be interested in anime and generate artwork in a similar style. This is a problem for us because we do not want our model to classify the artwork based on its objects, but rather on the stylistic differences between diffusion model art and authentic art.

To handle this, we focus on authentic images and images generated by Midjourney since we have the corresponding prompts. We will now introduce the novel process we used to filter our prompts. We use Term Frequency-Inverse Document Frequency to convert our prompts into a matrix of features. We ignore common English stop words like "the" and "a". Then, we use K-Means to group our prompts into 100 different clusters based on their similarity of TF-IDF features. We do this for authentic and Midjourney images so we have 2 sets of clusters, one for each dataset. For each cluster in a dataset, we take the cosine similarity between the average of all feature vectors and clusters in the opposing dataset. If the similarity score meets a certain threshold, then we use the data. Otherwise, we ignore it. This allows us to categorize our data and only use data that is substantial and common across our authentic images and Midjourney images. Note, the initial number of clusters and the threshold are hyperparameters whose specific values were chosen after rigorous tuning.

If we take a look at the clusters and the prompts belonging to each cluster, we can see that this is working as intended. The majority of our clusters are related to different artwork styles. For example, "ultraviolet fire male Angel in armour full shot by Annie Leibovitz" and "ultraviolet fire male Angel with wings in armour full shot by Annie Leibovitz" belong to the same cluster. Similar trends can be seen for each cluster.

However, since we are now handling the prompts for each image. We can only use 1000 images for Midjourney and authentic images for a total of 2000 images. We follow a similar training process as previously described. We replace the last layer with a two-output fully connected linear layer. As such, the accuracy is expected to be at least 50%.
By using this data and tuning our hyperparameters, we get the following ResNet-101 performance:

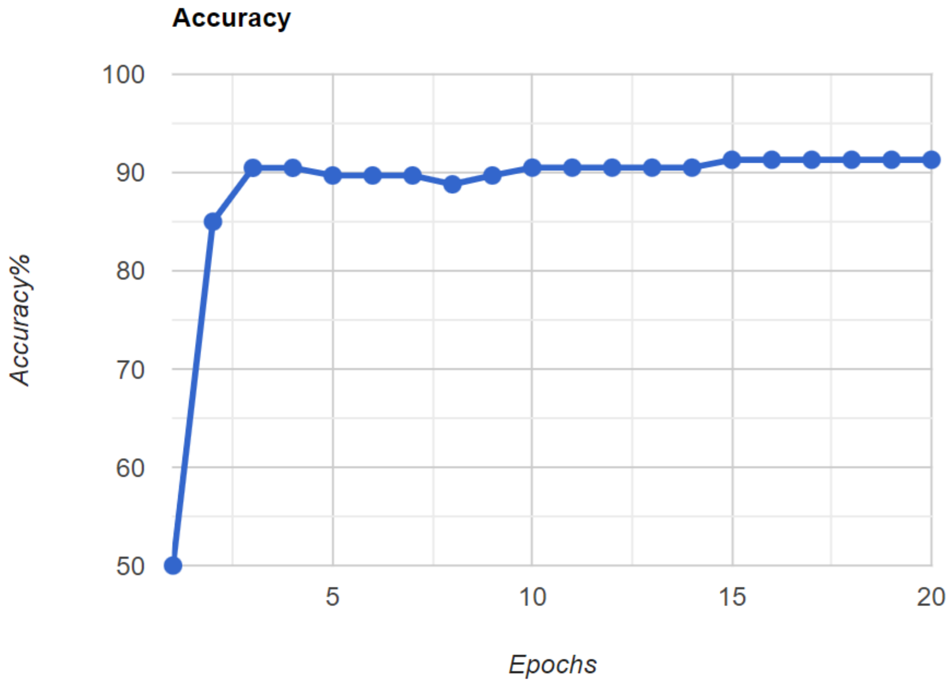

Figure 8: ResNet-101 Accuracy Over Epochs with Filtered Prompts

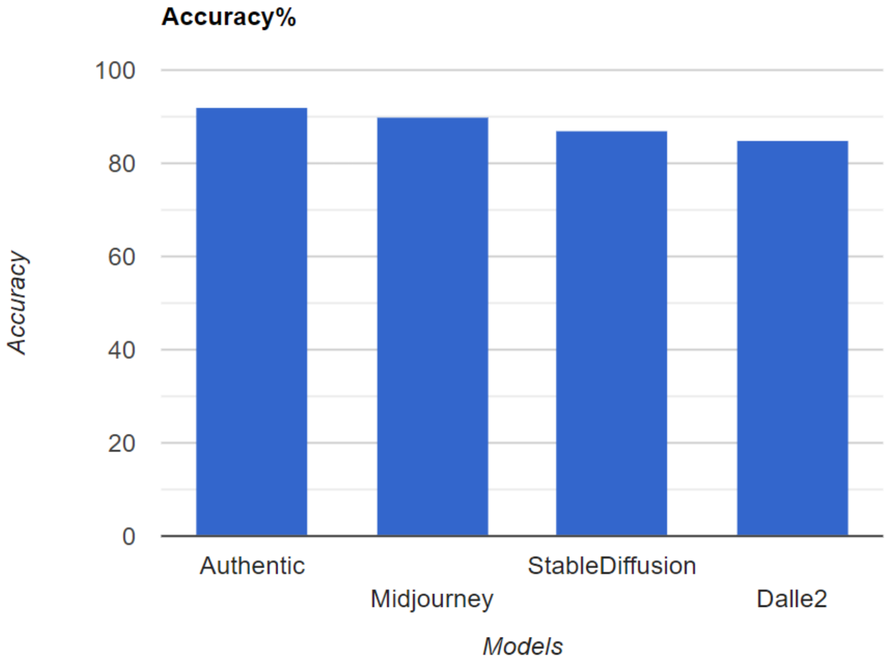

Figure 9: ResNet-101 Different Model Performance with Prompt Filtering

As seen in Figure 9, our model performs well in classifying artwork generated by all models. Again, we note that Dalle 2 has the lowest accuracy. Training a model using images generated by Dalle 2 would be an interesting investigation. However, our dataset does not contain prompts for Dalle 2 images. We will leave this for future work.

To further explore if the performance can be improved and to address some of our previous issues around computing resources, we use the same process but on a ResNet-50 model instead. This will allow us to use more training data as ResNet-50 requires less RAM. Furthermore, we also utilize a Google Colab Pro subscription to further increase our computing resources. This allows us to double our training data size to 4000 images.

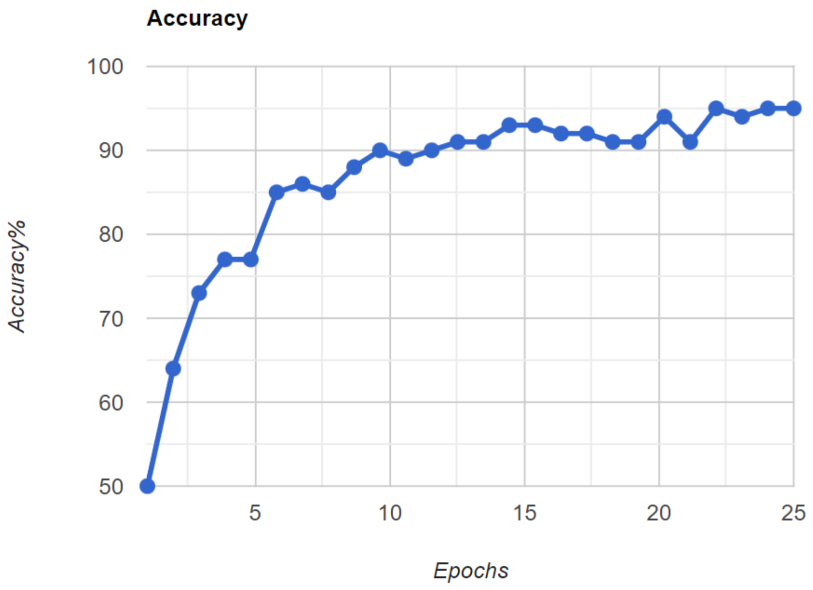

Figure 10: ResNet-50 Performance over Epochs with Prompt Filtering

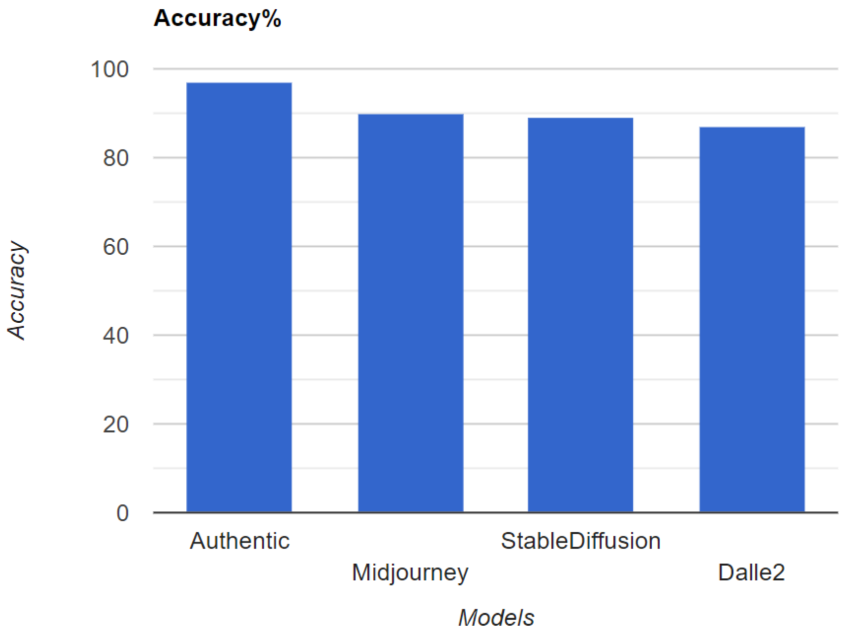

Figure 11: ResNet-50 Performance over Epochs with Prompt Filtering

As seen in Figure 5 and Figure 6 above, the performance of our ResNet-50 model is slightly better than ResNet-101. This can likely be explained by the fact that our dataset for the ResNet-50 model is over 2 times larger since ResNet-50 uses fewer computing resources than ResNet-101. Overall, our best performing model is ResNet-50 with prompt filtering.

5.0 Unique Contributions and Future Work

5.1 Unique Contributions

Our team has made substantial advancements in the classification of art as human-made or AI-generated, building upon the foundational works of Corvi et al. (2022) and Fu et al. (2002). Our unique contributions include the curation of a diverse and extensive dataset, which exposes our model to various artistic styles and techniques, thereby improving its generalization and classification capabilities. We have also developed robust data cleaning processes and employed data augmentation techniques to enhance the model's performance and ability to recognize subtle features in the artworks. Furthermore, our novel prompt filtering algorithm greatly increases the model's accuracy and efficiency by focusing on the most relevant features of the input artwork. Additionally, our implementation of ResNet-101 and its success in comparison with the ResNet-50 architecture has provided valuable insights into the trade-offs between model complexity, computational requirements, and classification accuracy. Collectively, these contributions have resulted in an AI model with an overall accuracy rate of over 90%, demonstrating significant improvements in classifying art as human-made or AI-generated.

5.2 Future Work

Our team has made significant progress in the classification of art as human-made or AI-generated, but there remains room for growth and further exploration. One potential avenue for improvement is experimenting with additional deep learning architectures like DenseNet, EfficientNet, or Transformer-based models, which may enhance our model's performance. As we expand and diversify our training dataset, we can expect to see further improvements in our model's generalization and classification capabilities. In parallel, investigating advanced feature extraction techniques, such as attention mechanisms or unsupervised learning methods, can help our model better capture the nuances of various artistic styles.

Enhancing model interpretability and explainability is crucial for increasing transparency and trust in our model's outcomes, allowing stakeholders to understand and rely on its decision-making process. Broadening the model's scope to analyze and classify various forms of art, such as sculpture, digital media, or performance art, can extend its applicability and relevance in the evolving art world. A real-time classification system can also be developed to cater to applications like live art authentication or monitoring AI-generated content on social media platforms.

In summary, by addressing these potential areas of improvement and building upon our current achievements, we can continue to refine our model and contribute positively to the intersection of AI and art, ensuring a responsible and innovative future in this rapidly evolving domain.

6.0 Ethical Implications

6.1 Addressing Existing Issues

Artificial intelligence has transformed the creative landscape, bringing forth novel forms of artistic expression and raising important ethical implications that warrant thorough discussion. As AI-generated art gains prominence, it is essential to address issues like transparency and authorship, art valuation and market impact, and intellectual property and copyright to ensure a responsible and fair evolution of the art world [1][2][3][4][5]. Our AI model, with an overall accuracy rate of over 90% in classifying art as human-made or AI-generated, contributes to addressing these concerns. We recognize that the complexities of these issues warrant ongoing analysis and refinement and that our model is only a first step in addressing these evolving issues.

The first ethical implication, transparency, and authorship are crucial for maintaining the integrity of the art world. Accurate attribution of creative works is vital for the recognition and credibility of both artists and AI developers. Our AI model plays a role in promoting transparency by distinguishing between human-made and AI-generated artwork, thus reducing the chances of false attributions such as when a famous Instagram photographer was caught using AI, or a genuine artist was falsely accused of using AI [6][7]. If buyers and art collectors are unable to determine whether or not a piece was generated by a human or a machine, the art industry as a whole will lack credibility and transparency. It's important to note that our model only partially addresses this issue of authorship, as the 90% accuracy rate indicates that some instances of misattribution will likely still occur. Furthermore, the model does not explicitly address concerns about the potential loss of human creativity or the implications of AI-generated art on the broader concept of artistic authorship as a whole.

The impact of AI-generated art on valuation and market dynamics is another significant ethical concern. By differentiating between human-created and AI-generated artwork, our AI model helps establish a more accurate appraisal and valuation system for both types of art. This fosters a fair market for traditional artists and AI-generated art enthusiasts alike, ensuring that the value of human creativity is not diminished while maintaining the potential for a distinct market purely for AI art enthusiasts, which at times may be valued at close to half a million dollars [17][18]. However, the model does not fully account for the constantly evolving nature of the art market, nor does it directly address potential inequalities in access to AI technology and resources among artists.

Intellectual property and copyright issues are increasingly important in the context of AI-generated art. Currently, the AI art giants Stable Diffusion and Midjourney AI are facing an infamous Copyright Lawsuit that could set the precedent for future laws regarding the operations of the AI art market [5]. Our AI model aids in clarifying ownership by identifying the source of an artwork, thus making it easier to resolve potential disputes and adapt legal frameworks to the evolving artistic landscape. Although the model contributes to addressing these concerns, it does not offer a comprehensive solution to the intricate legal and ethical challenges surrounding ownership and rights in the age of AI-generated art, nor does it address the concerns of training data using other artists' work [19][20].

In summary, our AI model is a valuable tool for addressing the ethical implications of AI-generated art. However, it is important to acknowledge the complex and evolving nature of these concerns and the limitations of the model in fully addressing them. Collaboration between artists, AI developers, and legal experts will be crucial for navigating this intersection and ensuring responsible innovation in the field of AI and art.

6.2 Potential Future Ethical Consideration

As we continue to explore the ethical landscape of using AI to classify art as human-made or AI-generated, it is crucial to consider the multifaceted nature of the challenges that arise from this technology. The integration of AI into the art world brings forth a myriad of complex ethical implications, which require a nuanced and balanced approach to fully comprehend their impact.

Firstly, the public perception and stigmatization of AI-generated art may lead to a dichotomy between human and AI creations, potentially affecting the market value and reputation of artists working with AI. Our model's ability to classify art could inadvertently contribute to this divide, necessitating careful consideration of how to maintain a balanced view of AI-generated art. Secondly, the unintended consequences on emerging artists arise as our AI model might unintentionally favor established artists, while less-known creators might be misclassified as AI-generated [7]. Transparency in the classification process is essential to prevent biases and protect opportunities for emerging talents.

Moreover, AI art and cultural appropriation raise concerns about the ethical use of cultural elements in AI-generated art. Our model's capacity to recognize and distinguish between various artistic styles might inadvertently contribute to the commodification and misrepresentation of cultural heritage [21][22]. It is imperative to establish guidelines that ensure the respectful and responsible use of cultural, racial, and gender-based elements [23]. Additionally, AI accountability and liability pose challenges in attributing responsibility when the AI model's misclassification leads to economic or reputational damages as they have been found to do in false accusations of student homework assignments, and the thoroughly researched case of false positive results from rare disease tests [24][25][26][27]. Developing legal frameworks and policies that address accountability and liability in art classification is crucial to maintaining trust in the system.

Furthermore, the dependence on AI and loss of human expertise threatens the preservation of valuable human skills, knowledge, and intuition in the art world as a whole as AI technology becomes increasingly integrated. Balancing the benefits of AI classification with the need to value human expertise may demand a combination of expert and AI collaboration [28]. Finally, AI transparency and explainability must be addressed, as the model's classification process should be transparent and easily understandable, allowing stakeholders to grasp the rationale behind its decisions and maintain trust in the system [29][30].

In conclusion, these newly arising ethical implications must be weighed against the benefits our AI model brings in addressing transparency and authorship, art valuation and market impact, and intellectual property and copyright issues. As ethical engineers, we must be aware of the potential pitfalls and embrace the complexity of these issues to ensure that the AI model we develop is ethically sound and contributes positively to the art world.

7.0 Conclusion

In this report, we have presented an AI model that classifies art as human-made or AI-generated with an overall accuracy rate of over 90%. Our work also proposes a novel prompt filtering algorithm that greatly improved our model's performance. Additionally, we also demonstrate some of the benefits and tradeoffs of using ResNet-50 or ResNet-101. Our classification model is a significant advancement in the field, as it addresses some of the ethical concerns arising from the integration of AI into the art world, such as transparency and authorship, art valuation and market impact, and intellectual property and copyright issues. While our model has made strides in addressing these concerns, it is important to acknowledge the limitations of the model and the complexities of the ethical implications that continue to evolve.

Furthermore, as AI-generated art becomes increasingly prevalent, potential future ethical considerations, such as public perception and stigmatization, cultural appropriation, AI accountability and liability, the preservation of human expertise, and AI transparency and explainability must be taken into account. These considerations highlight the need for collaboration among artists, AI developers, legal experts, and other stakeholders to ensure responsible innovation in the field of AI and art.

In conclusion, our AI model represents a valuable tool for addressing the ethical implications of AI-generated art. As ethical engineers, we must embrace the complexity of these issues, striving to develop AI models that are ethically sound and contribute positively to the art world. By fostering open dialogue and interdisciplinary collaboration, we can navigate the intersection of AI and art in a manner that is both responsible and innovative.

Bibliography

[1] Dignum, Virginia. “The Art of AI - Accountability, Responsibility, Transparency.” *Medium*, Medium, 4 Mar. 2018, https://medium.com/@virginiadignum/the-art-of-ai-accountability-responsibility-transparency-48666ec92ea5.

[2] “AI Creating 'Art' Is an Ethical and Copyright Nightmare.” *Kotaku*, 25 Aug. 2022, https://kotaku.com/ai-art-dall-e-midjourney-stable-diffusion-copyright-1849388060.

[3] Amelia, Brandt. “The Ethical Dilemma of AI Art Generation.” *Medium*, ILLUMINATION'S MIRROR, 23 Jan. 2023, https://medium.com/illuminations-mirror/the-ethical-dilemma-of-ai-art-generation-1a25d314903f.

[4] Ajao, Esther. “Implications of AI Art Lawsuits for Copyright Laws: TechTarget.” *Enterprise AI*, TechTarget, 1 Feb. 2023, https://www.techtarget.com/searchenterpriseai/news/365530156/Implications-of-AI-art-lawsuits-for-copyright-laws.

[5] Vincent, James. “AI Art Tools Stable Diffusion and Midjourney Targeted with Copyright Lawsuit.” *The Verge*, The Verge, 16 Jan. 2023, https://www.theverge.com/2023/1/16/23557098/generative-ai-art-copyright-legal-lawsuit-stable-diffusion-midjourney-deviantart.

[6] Stokel-Walker, Chris. “A Huge Subreddit Suspended a User for Posting AI Art, but the Work Is 100% Human-Made.” BuzzFeed News, BuzzFeed News, 6 Jan. 2023, https://www.buzzfeednews.com/article/chrisstokelwalker/art-subreddit-illustrator-ai-art-controversy.

[7] Stewart, Jessica. “Popular Instagram Photographer Confesses That His Work Is AI-Generated.” *My Modern Met*, 23 Feb. 2023, https://mymodernmet.com/joe-avery-ai-deception-instagram/.

[8] T. Fu, M. Xia, and G. Yang, “Detecting GAN-generated face images via hybrid texture and sensor noise based features,” Multimedia tools and applications, vol. 81, no. 18, pp. 26345–26359, 2022, doi: 10.1007/s11042-022-12661-1.

[9] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," 2022, doi: 10.48550/arxiv.2211.00680.

[10] He, K., Zhang, X., Ren, S., & Sun, J. "Deep Residual Learning for Image Recognition.", 2016, https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf.

[11] Han, D., Liu, Q., & Fan, W. "A new image classification method using CNN transfer learning and web data augmentation.", 2017, https://www.sciencedirect.com/science/article/abs/pii/S0957417417307844.

[12] Canziani, A., Paszke, A., & Culurciello, E. "An analysis of deep neural network models for practical applications.", 2016, https://arxiv.org/abs/1605.07678.

[13] "Midjourney User Prompts & Generated Images (250k)." *Kaggle*, https://www.kaggle.com/datasets/da9b9ba35ffbd86a5f97ccd068d3c74f5742cfe5f34f6aaf1f0f458d7694f55e?resource=download.

[14] "DiffusionDB." *DiffusionDB*, https://poloclub.github.io/diffusiondb/.

[15] superpotato9. "Dalle Recognition Dataset." *Kaggle*, 20 Jan. 2023, https://www.kaggle.com/datasets/superpotato9/dalle-recognition-dataset?select=real.

[16] "Laion." *LAION*, https://laion.ai/.

[17] "Portrait by AI Program Sells for $432,000." *BBC News*, BBC, 25 Oct. 2018, https://www.bbc.com/news/technology-45980863.

[18] Gaskin, Sam. "When Art Created by Artificial Intelligence Sells, Who Gets Paid?" *Artsy*, 17 Sept. 2018, https://www.artsy.net/article/artsy-editorial-art-created-artificial-intelligence-sells-paid.

[19] Metz, Rachel. "These Artists Found out Their Work Was Used to Train AI. Now They're Furious | CNN Business." *CNN*, Cable News Network, 21 Oct. 2022, https://www.cnn.com/2022/10/21/tech/artists-ai-images/index.html.

[20] "Why Those AI-Generated Portraits All over Social Media Have Artists on Edge | CBC Radio." *CBCnews*, CBC/Radio Canada, 12 Dec. 2022, https://www.cbc.ca/radio/asithappens/artificial-intelligence-ai-art-ethics-greg-rutkowski-1.6679466.

[21] "Times AI Created Artwork Based on Stereotypes." *HerZindagi English*, HerZindagi, 5 Jan. 2023, https://www.herzindagi.com/society-culture/artificial-intelligence-ai-artwork-based-on-stereotypes-article-218643.

[22] Dhanesha, Neel. "AI Art Looks Way Too European." *Vox*, Vox, 19 Oct. 2022, https://www.vox.com/recode/23405149/ai-art-dall-e-colonialism-artificial-intelligence.

[23] Buolamwini, Joy. "Artificial Intelligence Has a Racial and Gender Bias Problem." *Time*, Time, 7 Feb. 2019, https://time.com/5520558/artificial-intelligence-racial-gender-bias/.

[24] Jimenez, Kayla. "Professors Are Using CHATGPT Detector Tools to Accuse Students of Cheating. but What If the Software Is Wrong?" *USA Today*, Gannett Satellite Information Network, 13 Apr. 2023, https://www.usatoday.com/story/news/education/2023/04/12/how-ai-detection-tool-spawned-false-cheating-case-uc-davis/11600777002/.

[25] Akle, Sebastian, et al. "Mitigating False-Positive Associations in Rare Disease Gene Discovery." *Human Mutation*, U.S. National Library of Medicine, Oct. 2015, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4576452/.

[26] "Large-Scale Study Finds Genetic Testing Technology Falsely Detects Very Rare Variants." *ScienceDaily*, ScienceDaily, 15 Feb. 2021, https://www.sciencedaily.com/releases/2021/02/210215211036.htm.

[27] Kumar, Harshit. "Technical Fridays." *False Positive Paradox*, https://kharshit.github.io/blog/2018/10/12/false-positive-paradox.

[28] *Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making*, https://dl.acm.org/doi/fullHtml/10.1145/3491102.3517732.

[29] Grennan, Liz, et al. "Why Businesses Need Explainable AI-and How to Deliver It." *McKinsey & Company*, McKinsey & Company, 29 Sept. 2022, https://www.mckinsey.com/capabilities/quantumblack/our-insights/why-businesses-need-explainable-ai-and-how-to-deliver-it.

[30] Yalçın, Orhan G. "5 Significant Reasons Why Explainable AI Is an Existential Need for Humanity." *Medium*, Towards Data Science, 12 June 2022, https://towardsdatascience.com/5-significant-reasons-why-explainable-ai-is-an-existential-need-for-humanity-abe57ced4541.