Dataset Augmentation in Papyrology with Generative Models: A Study of Synthetic Ancient Greek Character Images

Matthew I. Swindall[1] , **Timothy Player**[2] , **Ben Keener**[3] , **Alex C. Williams**[8*] , **James H. Brusuelas**[3] , **Federica Nicolardi**[4] , **Marzia D'Angelo**[5] , **Claudio Vergara**[6] , **Michael McOsker**[7] , **John F. Wallin**[1]

[1]Middle Tennessee State University
[2]University of Tennessee, Knoxville
[3]University of Kentucky
[4]Università degli Studi di Napoli Federico II
[5]Istituto Papirologico 'G. Vitelli' (Università degli Studi di Firenze)
[6]Università di Pisa
[7]Institut für Altertumskunde Universität zu Köln
[8]Amazon

mis2n@mtmail.mtsu.edu, tplayer@vols.utk.edu, bdke225@uky.edu, acwio@amazon.com, james.brusuelas@uky.edu, federica.nicolardi@unina.it, marzia.dangelo@unifi.it, claudio.vergara@phd.unipi.it, mmcosker@uni-koeln.de, john.wallin@mtsu.edu

Abstract

Character recognition models rely substantially on image datasets that maintain a balance of class samples. However, achieving a balance of classes is particularly challenging for ancient manuscript contexts as character instances may be significantly limited. In this paper, we present findings from a study that assess the efficacy of using synthetically generated character instances to augment an existing dataset of ancient Greek character images for use in machine learning models. We complement our model exploration by engaging professional papyrologists to better understand the practical opportunities afforded by synthetic instances. Our results suggest that synthetic instances improve model performance for limited character classes, and may have unexplored effects on character classes more generally. We also find that trained papyrologists are unable to distinguish between synthetic and non-synthetic images and regard synthetic instances as valuable assets for professional and educational contexts. We conclude by discussing the practical implications of our research.

1 Introduction

Optical character recognition (OCR) is a mature field within deep learning and there are many OCR engines that yield impressive results for most modern documents. While deep learning models can classify characters with astounding accuracy, ancient handwritten manuscripts still pose many challenges. Existing OCR engines have been mostly trained on modern printed documents, and the variation in handwriting style for a given character can be rather extensive. The work in [Swindall *et al.*, 2021] sought to further research in this area by leveraging a crowdsourced dataset based on ancient Greek papyri to train machine learning models for the accurate classification of handwritten characters. These models challenged and outperformed pre-trained OCR platforms. However, due to the fragmentary nature of Greek papyri there was a notable imbalance in the dataset. Some characters were classified in large quantities, while others were not, whether due to difficulty in recognition by the crowdsourced transcribers and/or a lower preservation rate in the fragments themselves. Although these models performed with impressive accuracy, this imbalance in the dataset highlighted the need for further custom-trained models.

Machine learning datasets with class imbalances or insufficient data for training have been augmented by synthetic data to increase sample sizes. This method has been used in training facial recognition models and self-driving vehicles [Shrivastava *et al.*, 2017][Tremblay *et al.*, 2018]. For image datasets, Generative Adversarial Neural Networks (GANs) are a popular method for creating such data. Synthetic images have indeed become a method to challenge and improve machine learning datasets [Cronin *et al.*, 2020] [Frid-Adar *et al.*, 2018]. GANs thus offer a potential technique to further evaluate and improve existing models. Moreover, synthetic data has the potential to address issues in the scholarly, educational, and creative workflows of papyrologists themselves. For example, there is no resource that documents every stylistic variety of character shape present (or even now missing) in the surviving Greek papyri. GANs based on real-world examples can assist in and inspire the reconstruction of fragmented handwritten text.

We present findings from a two-part study to understand

*Work completed before joining Amazon.

the role synthetic characters can play in machine learning contexts and professional papyrology. To reduce class imbalance, we use PyTorch's StyleGAN2 to strategically increase the presence of character instances with limited sample sizes. We hypothesize that incorporating synthetic character images can enable models to not only classify character classes associated with these synthetic instances with higher accuracy, but also those for which synthetic instances have not been introduced. We begin by training a series of machine learning models (i.e., CNNs and ResNets) on AL-SYNTH, an augmented version of the AL-ALL dataset, and observe increases in per-character accuracy from 8% to 12%. We complement our model evaluation by engaging four expert papyrologists to examine the utility of synthetically produced character images in practice. We observe that expert papyrologists find significant value in synthetic character images as novel tools for manuscript reconstruction and educational assets. We conclude by discussing the relevance of our findings as they relate to synthetic instances and the professional study of ancient manuscripts.

2 Related Work

2.1 Machine Learning Image Datasets

Most machine learning datasets, such as MNIST [Deng, 2012] and the St. Gall database [Fischer *et al.*, 2011], consist of ideally cropped character images or custom made datasets. This is rarely the case for datasets derived from handwritten ancient manuscripts. Annotation of ancient manuscripts is tedious, time-consuming work, further limited by a shortage of experts trained to decipher them. Efforts to speed up the process through digitization and crowdsourcing can be costly and require years of effort. Machine learning datasets built through such efforts also tend to be extremely noisy [Swindall *et al.*, 2021]. Ancient manuscripts are often damaged and difficult to read even for experts. A persistent challenge is the difficulty in quantifying the ground-truth in labeling due to human annotation errors, especially in a crowdsourced setting.

2.2 Generating Synthetic Images with Machine Learning

Generative Adversarial Networks (GANs) have come to prominence in recent years because of their ability to produce synthetic data after being trained on a subset of real data. The process involves two neural network architectures, the generator and the discriminator. The discriminator is used to differentiate between the generated and real data. The generator network is designed to create an output by a randomized or latent representation of the real data as input. This is still an area of active research [Goodfellow *et al.*, 2014]. The website https://thispersondoesnotexist.com/ showcases an improvement on GAN architecture with StyleGAN2, as described in [Karras *et al.*, 2020]. This demonstrates how GANs have the potential to revolutionize machine learning applications for numerous use cases. In [Frid-Adar *et al.*, 2018] CNN performance increased roughly 10% when training data was augmented with synthetically generated images. In [Cronin *et al.*, 2020] we even find a style transfer GAN used to create a fully synthetic dataset of musculoskeletal ultrasound images. Additionally, creative uses of GAN architecture are being explored, such as the Creative Adversarial Network (CAN) in [Elgammal *et al.*, 2017] where GAN-like architecture is used to explore artistic styling.

Figure 1: An Oxyrhynchus papyrus fragment.

3 Domain Overview: Papyrology

Papyrology is a discipline that involves the conservation, editing, and interpretation of ancient texts written on papyrus. This field is critical to the study of the cultures of the ancient Mediterranean, from ancient Egypt to the Christian and Islamic periods.

3.1 Image Dataset: AL-ALL and AL-PUB

The Ancient Lives Project was a web-based crowdsourcing initiative that allowed volunteers to transcribe digital images of ancient papyrus fragments from 2011 until 2018 [Williams *et al.*, 2014]. This project resulted in millions of annotations from images of papyrus fragments, such as the one shown in Figure 1.

The annotation data from the Ancient Lives Project were compiled into consensus labels and pixel locations for each annotation. This consensus data was then used to create two crowdsourced datasets: AL-ALL, 399,421 images from unpublished and published papyri, and AL-PUB, 195,683 images from only published papyri. Both datasets consist of tightly cropped images of individual Greek characters from the annotated images of papyri, as shown in Figure 2. The datasets, however, are "noisy". Unlike a dataset such as MNIST [Deng, 2012], both were created from images containing holes, rips, missing segments, and faded ink, as shown in Figure 3. Accordingly, there is uncertainty in the ground truth for the character labels. For a given character, initiatives like the Ancient Lives Project must rely on untrained individuals to record both the correct classification and a reasonably accurate location within the images. Furthermore, these conditions that are conducive to "noise" in the data also produce sample bias. Due to physical damage, some characters simply appear in greater numbers than others, and some are transcribed less.

3.2 Image Dataset: AL-SYNTH

To address sample bias in the AL-ALL dataset we have created the AL-SYNTH dataset. AL-SYNTH is essentially AL-ALL, but augmented with an extra 904 images of Psi and 1201 images of Xi. This effectively doubles 2 of the 3 smallest sample sizes in the dataset. To test our hypothesis, these

Figure 2: Examples of each character in the AL-PUB dataset [Swindall *et al.*, 2021]

Figure 3: Examples of characters from damaged papyrus fragments.

characters were chosen since they both are indicative of the imbalance in the dataset and meet the threshold requirements of StyleGAN2, roughly 1000 to 2000 real images, to generate synthetic data.

4 Research Goal

In this paper, we aim to address two specific questions:

[RQ1] How do synthetic instances of Ancient Greek characters written on papyrus affect model performance?

[RQ2] How do synthetic instances of Ancient Greek characters written on papyrus affect professional practice?

We designed two separate studies to appropriately address each of our research questions. To address RQ1, we designed, implemented, and evaluated a series of machine learning models that are trained on an augmented version of the AL-ALL dataset. To address RQ2, we engaged a set of trained papyrologists in an experimental task that captured impressions of synthetic instances alongside their potential use in the broader practice of papyrology. We now describe the design and execution of these studies alongside their findings.

5 Study 1: Modeling with Synthetic Instances

Our approach for addressing RQ1 centered around three different types of machine learning models: (1) a GAN for generating synthetic data, (2) a traditional CNN, and (3) a convolution-based ResNet for classification. The two latter models (i.e., CNN and ResNet) were selected for use on the basis of being architecturally validated from prior models that were trained on the AL-PUB dataset [Swindall *et al.*, 2021]. Model weights for each of these models were recalculated through retraining over 75 epochs. Each model was trained 10 times, on the original AL-ALL dataset and on the new AL-SYNTH dataset containing additional synthetic images. Supplemental information regarding image generation and saved models are available at https://data.cs.mtsu.edu/al-pub/synth.html.

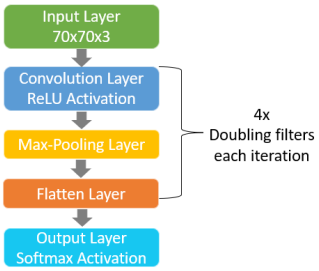


Figure 4: CNN Architecture

5.1 Model Type 1: Generative Model

The StyleGAN2 team provides a simple implementation via PyTorch available at https://github.com/lucidrains/stylegan2-pytorch. While training the GAN, multi-GPU training, data augmentation, and the addition of a single attention layer are utilized. The use of GPU training is included as a means to speed up the training process. The data augmentation option differentialy augments the images before the discriminator trains on them. This is a strategy for generating synthetic images when the training dataset is small, on the order of 1000 to 2000 samples. The attention option allows the addition of self-attention which can greatly improve results. As increasing attention increases training time, only 1 attention layer is added.

5.2 Model Types 2 & 3: Categorical Classification Models

The CNN model, described in Figure 4, is a very simple design utilizing 4 convolution layers, each followed by a max-pooling layer. Each convolution utilizes the ReLU activation function with a 3x3 kernel. The pooling layers have a 2x2 window size. The initial number of convolution filters is 96, which is doubled for subsequent convolution layers. The batch size was a constant 512 with a learning rate of 0.001. Sparse categorical cross-entropy, the standard loss function for multi-class classification is used along with the Adam optimizer. The output layer includes softmax activation for 24 classes.

The ResNet model, outlined in Figure 5, is a fairly standard residual model utilizing convolutional architecture. The early layers in the model consist of 2 convolution layers with ReLU activation, followed by a max-pooling layer and 18 residual blocks. The residual blocks consist of 2 cycles of convolution and batch normalization. After the residual block, the model finishes with a global average pooling layer, a single dense layer, a dropout layer, and finally a soft-max output layer for 24 classes.

5.3 Datasets

Machine learning results can be tricky to reproduce. Rather than rely on old results from [Swindall *et al.*, 2021], both the CNN and ResNet are trained from scratch utilizing the AL-ALL dataset and the new AL-SYNTH dataset. Both datasets are serialized with the pickle python library using pickle protocol 4. The data are loaded into each model, then randomly

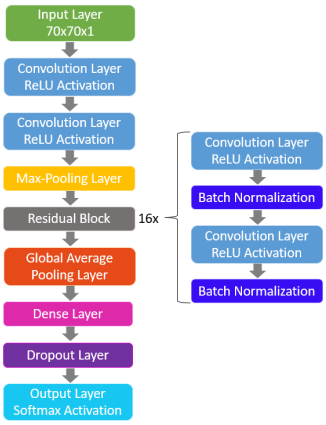
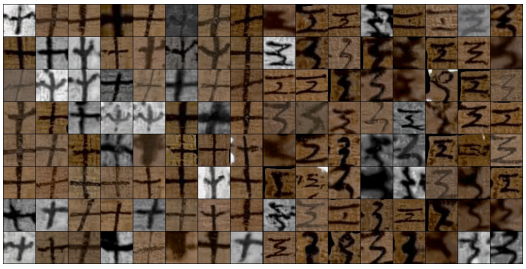


Figure 5: ResNet Architecture



(a) `Synthetic Psi` (b) `Synthetic Xi`

Figure 6: Images of synthesized Psi and Xi.

shuffled and sorted into training and validation subsets using scikit-learn's train_test_split() with the shuffle parameter set to True and an 80/20 training/validation split. The AL-ALL dataset was used instead of the publicly available AL-PUB because the sample sizes are much smaller in the public dataset; this would have dramatically degraded the StyleGAN2 results.

5.4 Generative Model Results

The smallest sample, the Greek character Sigma (Σ, σ), contains only 62 images, far less than the 1,000 to 2,000 images StyleGAN2's Pytorch implementation is designed for. To combat imbalance, we thus focus on the sample sizes for Psi (Ψ ψ) and Xi (Ξ ξ), 904 and 1201 respectively, which are better suited for StyleGAN2. To consider the effects of synthetic characters on the practice of papyrology, three of the most universally recognizable characters with larger sample sizes, Alpha, Delta, and Pi, were presented to experts. For each character, the GAN was trained for 40 iterations (each iteration produces an 8x8 grid of synthetic images). After generating the images, each was visually inspected and only reasonably well synthesized images were kept. StyleGAN2 produced amazingly realistic images. Figure 6 (a) and (b) show examples of synthetic Psi and Xi images, while Figure 7 shows synthetic Alpha, Delta, and Pi. We were able to double the sample size of Psi and Xi, and generate small samples, 64 each, of Alpha, Delta, and Pi for our experts to examine.

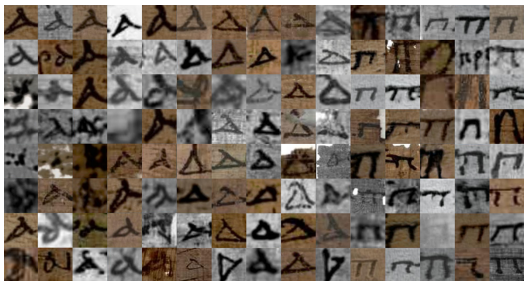


Figure 7: Images of Synthetic Alpha, Delta, & Pi

Psi(Ψ ψ)
The second smallest sample, at 904 images, is Psi. This is just shy of the 1000 data points StyleGAN2's implementation was designed for, but the results were excellent. By the 5th iteration, it was becoming difficult to distinguish some images from the real data. Sadly, many annotators labeled the christian cross symbol as Psi, since they are somewhat similar. The Psi sample is thus not truly representative of the Greek character. However, because so many are mislabeled, the accuracy for this sample is relatively high.

Xi(Ξ ξ)
StyleGAN2 produced high quality images that were challenging for trained individuals to visually distinguish from the original data. 1,201 Xi images were used to train the discriminator. Unlike the Psi sample, most of the Xi images are unambiguous and are clearly the correct character.

Additional Characters (A α, Δ δ, Π π)
Only 64 images were generated for Alpha, Delta, and Pi, since these characters were not being utilized to train any models. The sample sizes for theses characters are much larger than for Xi and Psi, and the per character accuracy was greater than for the smaller sample sizes. Alpha was the largest of the three samples at 42,538 images, followed by Pi at 17,112, and Delta at 11,716.

5.5 Categorical Classification Model Results

Figures 8 and 9 show the mean and confidence intervals for 10 runs of the CNN and ResNet models trained with the AL-ALL and AL-SYNTH datasets; the results are similar to a k-fold cross-validation.

CNN Results
As in [Swindall *et al.*, 2021], the CNN model did not perform as well as the ResNet, which is unsurprising as residual networks supplanted standard CNN architectures several years ago. Figure 8 shows that the training and validation accuracies for both CNNs are neck-and-neck with training accuracy above 95%. Similarly, the loss for both models is barely distinguishable, with considerable divergence early on. This, along with the dramatic gap between training and validation accuracy, likely suggests massive overfitting of the model, a problem inherent in a dataset with such skewed class imbalance. Generally, there appears to be little difference between models trained with the synthetic data and those with the original data.

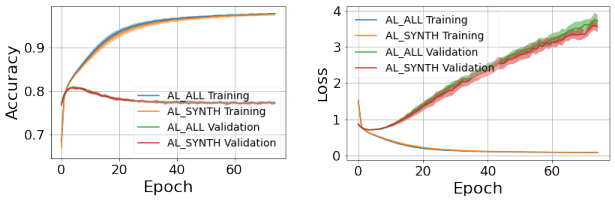


Figure 8: Accuracy & Loss for 10 CNN runs for each dataset.

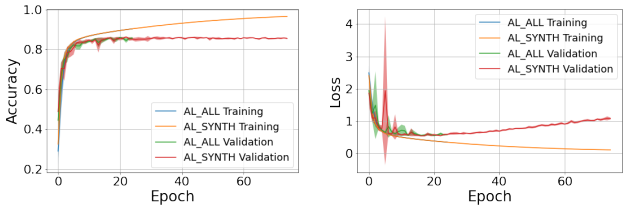


Figure 9: Accuracy & Loss for 10 ResNet runs for each dataset.

ResNet Results

As shown in Figures 8 and 9, the ResNet model outperforms the CNN and will thus be the focus when considering metrics such as accuracy, precision, and recall. The validation accuracy converges quite early on in the mid-to-upper 80's. Again we see evidence of over-fitting and little overall improvement when training with synthetic data. The validation loss for both models converge with training loss early, but steadily increases in later epochs. This model may benefit from additional strategies to combat the over-fitting which is likely due to class imbalance.

Precision & Recall

Accuracy alone is often not a clear indicator of model success. Precision and recall may be better indicators of how well the model generalizes in some cases. Table 1 shows the precision and recall for the CNN and ResNet, trained on the original AL-ALL and new AL-SYNTH datasets. Both metrics seem to fare worse for the CNN when trained with synthetic data. However, a considerable gain is evident when the ResNet is trained with synthetic data.

Per Character Accuracy

Table 2 shows the per-character accuracy for the Psi and Xi sub-samples. While accuracy varied slightly for all characters between models and training datasets, it is important to focus on the effects on individual character accuracy. A slight worsening in performance is seen from the CNN when trained with synthetic data. However, a considerable improvement is shown for the ResNet trained on AL-SYNTH. The accuracy for Psi increased by 12% while Xi's accuracy increased by about 8%. This result, combined with the lack of improvement of the models overall, suggests that while accuracy is increasing for the newly doubled samples, accuracy for other samples is likely being degraded. This warrants further study.

Metric	CNN	CNN-AL-SYNTH	ResNet	ResNet-AL-SYNTH
Precision	0.9307	0.9261	0.8548	0.9181
Recall	0.9141	0.9127	0.8135	0.8736

Table 1: Precision & Recall Comparison: AL-ALL & AL-SYNTH

Character	AL-ALL M1	AL-ALL M2	AL-SYNTH M1	AL-SYNTH M2
Psi(Ψ,ψ)	0.64	0.44	0.76	0.88
Xi(Ξ,ξ)	0.68	0.76	0.84	0.92

Table 2: Per Character Accuracy for Target Sub-samples. M1 Denotes model trained on AL-ALL. M2 Denotes model trained on AL-SYNTH

6 Study 2: Augmenting Creativity with Synthetic Instances

In order to address RQ2, we engaged four expert papyrologists with a survey that captured their professional impressions of the synthetically generated instances and their expert opinions about the usage of these instances in the broader practice of papyrology. In this section, we detail the methodology of our approach and discuss our findings.

6.1 Methodology: Web Survey

An important consideration for assessing the effect of synthetic instances in practice is understanding whether such instances can be identified. We therefore designed a web survey that asked domain experts (i.e., papyrologists) to complete two survey phases in support of comprehensively addressing RQ2:

[Phase 1] Label a set of character images as being machine-generated (i.e., synthetic) or having originated from a pre-existing digitization effort (i.e., non-synthetic).

[Phase 2] Answer a set of questions that inquire about the role of synthetic images in papyrological practice as a creative, a teacher, and a professional.

To eliminate any bias among respondents, we chose to limit our survey to individuals who had not seen our synthetic images and were not already aware of our generative efforts. We engaged a total of four experts as respondents and administered the survey experience via email.

Phase 1: Character Instance Labeling

In Phase 1, respondents were tasked with labeling character instances as "synthetic" or "non-synthetic". A set of 384 anonymized images was created, including 64 real images from AL-PUB and 64 synthetic images for each of the three widely recognizable Greek characters; Alpha (Aα), Delta (Δ δ), and Pi (Π π). A web-based interface was utilized for annotating each image as real or synthetic. Respondents are asked to drag-and-drop the anonymized images into the interface where each image can be labeled as 'r' for real or 's' for synthetic. Respondents can then download the combined results. These results are then compared to an anonymization key to determine each respondents accuracy.

Phase 2: Synthetic Instances and Practical Potential

In Phase 2, respondents were asked to answer four questions regarding the potential of synthetic instances as tools

Q1	*"How could you imagine such synthetic images, as a tool, affecting your practice as a creative in papyrology?"*
Q2	*"How could you imagine such synthetic images, as a tool, affecting your practice as a professional in papyrology?"*
Q3	*"How could you imagine such synthetic images, as a tool, affecting your practice as a teacher in papyrology?"*
Q4	*"Would you want to see this tool integrated into your existing systems for your profession?"*

Table 3: Four questions posed during Phase 2 of Study 2.

Expert	Precision	Recall	F1
1	0.54	0.36	0.43
2	0.62	0.32	0.42
3	0.53	0.22	0.32
4	0.62	0.62	0.62

Table 4: F1 score for each expert respondent in Survey 2.

for augmenting professional papyrological practice (see Table 3). In support of answering these questions, each expert was presented with three 8x8 grids of synthetically-generated instances of Alpha, Delta, and Pi images. We also inquired about respondents' age, gender, and professional expertise.

6.2 Findings

Phase 1. Synthetic Image Identification

The real and synthetic characters of anonymized images are compiled and compared to an anonymization key to determine survey respondent accuracy for all 384 images. The results suggest that it is very difficult for trained experts to distinguish between our real images and the newly synthesized images. The mean accuracy of the annotators is 55.14%. Annotator precision, recall, and F1 scores are detailed in Table 4.

Phase 2: Assessing Practical Potential

Thematic analysis was used to analyze qualitative data collected via the four questions in Table 3 [Braun and Clarke, 2012]. Respondents were 50% male, 50 % female, with a mean age of 30.75 years. All respondents agreed that synthetic instances can be valuable to a variety of professional contexts. The most common theme of responses centered around documentary reconstruction (i.e., reasoning about missing information in manuscripts) and educational usage (e.g., demonstrations of written characters). Reconstruction depends not simply on knowledge of ancient Greek, but also the likelihood that conjectured characters are palaeographically suitable to the manuscript. Respondents noted that synthetic characters could be used to fill the holes and gaps both to virtually reconstruct papyrus manuscripts and to verify the compatibility of characters conjectured with the remaining ink traces. Further machine learning applications based on synthetic data could also assist in this process, helping papyrologists estimate the number of missing characters based on style and shape and verify the compatibility of characters conjectured. In teaching Greek Palaeography, whether in a classroom or museum setting, synthetic characters could be used in palaeographic tables to explore the change and evolution of character shapes across a vast number of examples, such as examining the subtle differences in partially preserved but very similar characters, such as Delta, Alpha, and Lambda. Finally, responents noted that such instances may embody rare phenomena (e.g., characters with rare attributes) for further study and investigation. Overall, the respondents expressed interest in synthetic images being integrated into existing systems throughout papyrology (e.g., papyri.info [Ast and Bagnall, 2012]) for communal and practical use.

7 Conclusion

In this paper, we explored the utility of synthetically generated images of ancient Greek characters in the field of papyrology. Our examination of model performance suggests that augmenting a dataset of existing ancient Greek character images with synthetic instances can yield gains in performance. We observe that the 'simple' PyTorch implementation of StyleGAN2 produced realistic synthetic images of Alpha, Delta, Pi, Psi, and Xi when trained on sub-samples of AL-ALL. The produced images include 904 Psi, 1,201 Xi, and 64 each Alpha, Delta, and Pi. Both the CNN and ResNet showed insignificant changes in the overall accuracy and loss of the model when trained with the synthetic data. However, when trained with AL-SYNTH, the ResNet showed increases in accuracy of 8% and 12% respectively for the Xi and Psi characters.

Alongside our model evaluation, we find that domain experts see substantial utility in synthetic instances in a variety of professional contexts. Phase 1 of the survey demonstrated that experts find it difficult to distinguish between real images from AL-PUB and their synthetic counterparts. The mean respondent accuracy is 55.14% with F1 scores ranging from 0.32 to 0.62. In Phase 2 respondents suggest a wide range of uses for synthetic images in creative papyrology including museum exhibits, virtual document reconstruction, and teaching tools such as paleographic tables.

Taken collectively, our results introduce a variety of opportunities for future work at the intersection of machine learning, dataset augmentation, and the study of ancient manuscripts. The machine learning model results are suggestive that improving per-character accuracy via synthetically augmenting image datasets may have inverse effects on non-augmented samples. Exploration of GAN image synthesis with multiple datasets, similar to the approach taken in [Bowles *et al.*, 2018], may yield further understanding of the dynamics between overall accuracy and per-character accuracy. For papyrologists, generating rare characters and textural phenomena may have wide implications for teaching and analysis. Future work could focus on sub-samples of individual characters from AL-ALL that exhibit less-common attributes, which can then be used to generate additional examples with some level of variation. Our examination of synthetic instances and their utility specifically motivates a new pathway for visually reconstructing documents of significant deterioration or damage. We regard this pathway as such a valuable area for future work that we invite members of the machine learning and digital humanities communities to explore together.

References

[Ast and Bagnall, 2012] Rodney Ast and Roger Bagnall. Digital corpus of greek and latin literary papyri. https://caa.hcommons.org/deposits/item/hc:12447/, 2012. Accessed: 2022-01-15.

[Bowles *et al.*, 2018] Christopher Bowles, Liang Chen, Ricardo Guerrero, Paul Bentley, Roger Gunn, Alexander Hammers, David Alexander Dickie, Maria Valdés Hernández, Joanna Wardlaw, and Daniel Rueckert. Gan augmentation: Augmenting training data using generative adversarial networks. *arXiv preprint arXiv:1810.10863*, 2018.

[Braun and Clarke, 2012] Virginia Braun and Victoria Clarke. Thematic analysis. In *H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, K. J. Sher (Eds.), APA handbook of research methods in psychology, Vol. 2. Research designs: Quantitative, qualitative, neuropsychological, and biological*, pages 57–71. American Psychology Association, 2012.

[Cronin *et al.*, 2020] Neil J. Cronin, Taija Finni, and Olivier Seynnes. Using deep learning to generate synthetic b-mode musculoskeletal ultrasound images. *Computer Methods and Programs in Biomedicine*, 196:105583, 2020.

[Deng, 2012] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[Elgammal *et al.*, 2017] Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. Can: Creative adversarial networks, generating 'art' by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*, 2017.

[Fischer *et al.*, 2011] Andreas Fischer, Volkmar Frinken, Alicia Fornés, and Horst Bunke. Transcription alignment of latin manuscripts using hidden markov models. New York, NY, USA, 2011. Association for Computing Machinery.

[Frid-Adar *et al.*, 2018] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[Karras *et al.*, 2020] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119. IEEE, 2020.

[Shrivastava *et al.*, 2017] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.

[Swindall *et al.*, 2021] Matthew I. Swindall, Gregory Croisdale, Chase C. Hunter, Ben Keener, Alex C. Williams, James H. Brusuelas, Nita Krevans, Melissa Sellew, Lucy Fortson, and John F. Wallin. Exploring learning approaches for ancient greek character recognition with citizen science data. In *2021 17th International Conference on eScience (eScience)*, pages 128–137. IEEE, 2021.

[Tremblay *et al.*, 2018] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Proceedings of the ieee conference on computer vision and pattern recognition (cvpr) workshops. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2018.

[Williams *et al.*, 2014] Alex C. Williams, John F. Wallin, Haoyu Yu, Marco Perale, Hyrum D. Carroll, Anne-Francoise Lamblin, Lucy Fortson, Dirk Obbink, Chris J. Lintott, and James H. Brusuelas. A computational pipeline for crowdsourced transcriptions of ancient greek papyrus fragments. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 100–105. IEEE, 2014.