

Advancing Digital Papyrology: Machine Learning and Blockchain Tools for Modernizing the Study of Ancient Greek Manuscripts

by

Matthew I. Swindall



A dissertation submitted in partial fulfilment of the requirements of the degree of
Doctor of Philosophy in Computational Science

Middle Tennessee State University
April, 2024

Dissertation Committee:
Dr. John F. Wallin, Chair
Dr. James H. Brusuelas
Dr. Joshua L. Phillips
Dr. Qiang Wu

ABSTRACT

The study and preservation of ancient Greek papyri poses unique challenges due to the degraded and fragmented state of these highly damaged, ancient manuscripts. While digital imaging has aided in documenting these texts, manual transcription by experts remains a formidable bottleneck. In my work I present novel machine learning and blockchain-based approaches designed to accelerate and streamline the transcription and archiving of papyrological manuscripts. Key to this work is the creation of crowdsourced datasets of ancient Greek character images, annotated through a crowdsourcing initiative, that enable the training of deep learning models for character detection, segmentation, and recognition. Other contributions include augmenting datasets with synthetically generated characters to reduce sampling bias, and techniques for identifying annotation uncertainty via ensemble modeling to improve classification accuracy. The models and algorithms created in these works form the core of a pipeline that combines human oversight with automated processes for diplomatic transcription of papyrus fragments. Current and future work, based on these contributions, includes advances in optical manuscript dating and novel approaches to character spotting. To support collaborative scholarship within the field of papyrology, a blockchain framework utilizing smart contracts and decentralized storage is proposed for managing versions of transcribed texts. Implemented as a prototype, this framework demonstrates feasibility and potential benefits over traditional editorial workflows. Collectively, the methods developed aim to provide an AI-assisted platform tailored for papyrologists and other humanities researchers. By uniting machine learning, human computation, and distributed ledger technologies, this interdisciplinary research proposes a modernized paradigm for studying the ancient world through its surviving manuscripts.

ACKNOWLEDGEMENTS

I would like to express my appreciation for previous efforts, without which my work would not have been possible. First, I want to acknowledge the Ancient Lives Project and the volunteer, citizen scientists for their efforts in annotating hundreds of thousands of characters in thousands of documents. I want to thank the amazing research team who have guided and inspired my research. Specifically I want to thank Dr. John F. Wallin, Dr. James H. Brusuelas, and Dr. Alex C. Williams, for their mentorship and guidance. I also want to thank Dr. Graham West, Dr. Kritagya Upadhyay, Ben Keener, Timothy Player, Gregory Croisdale, and Chase C. Hunter for encouraging me to expand my horizons and to become a better researcher. Additional thanks go to other amazing collaborators I have worked with: Nita Krevans, Melissa Sellew, Lucy Fortson, Federica Nicolardi, Marzia D Angelo, Claudio Vergara, and Michael McOsker. I would also like to express my appreciation to the dissertation committee members, including Dr. Joshua L. Phillips and Dr. Qiang Wu, for your thoughtful analysis and feedback during the dissertation process. And finally, I want to thank my amazing partner in life, Dejan Rauzanovic. I can't imagine making it this far without you. Thank you.

Table of Contents

List of Figures	ix
------------------------	-----------

List of Tables	x
-----------------------	----------

I Crowdsourced Greek Character Datasets and Optical Character Recognition **1**

1 Introduction **2**

1.1	Research Overview	2
1.2	Publications	5
1.2.1	IEEE eScience 2021	5
1.2.2	International Joint Conference on Artificial Intelligence 2022 (IJCAI)	6
1.2.3	Conference on Human Computation and Crowdsourcing Doctoral Consortium 2022 (HCOMP)	6
1.2.4	Journal of Data Mining and Digital Humanities 2024 (JDMDH)	6
1.2.5	ACM Conference on Intelligent User Interfaces 2024 (IUI)	7
1.2.6	ACM SIGMIS Computers and People Research 2024	7

2 Exploring Learning Approaches for Ancient Greek Character Recognition with Citizen Science Data **8**

2.1	Introduction	8
2.2	Related Literature	10
2.2.1	Handwritten Character Recognition	11
2.2.2	Crowdsourcing and Noise Labelers	11
2.2.3	Crowdsourcing Manuscript Transcription	11
2.2.4	Contribution	12
2.3	Ancient Lives	13
2.3.1	Task Interface	13
2.3.2	Processing Pipeline & Data Quality	14
2.3.3	Cropping Algorithm and Consensus Label	15
2.3.4	The Ancient Lives Dataset	15
2.4	Methods	16
2.4.1	Baseline: Tesseract	16
2.4.2	Learning Approaches	17
2.4.3	Hyperparameter Tuning: Procedure	21
2.4.4	Reduction of Explored Values	21
2.4.5	Hyperparameter Tuning: Results	22
2.5	Results	24
2.5.1	Tesseract	24
2.5.2	CNN-BASE, CNN-XGB, and RESNET	24
2.5.3	Auditing Misclassifications with RESNET and CNN-BASE	26
2.6	Theia: A Web UI for Character Classification	30

2.6.1	User Experience and Web Interface	30
2.7	Discussion	31
2.8	Summary	32
3	Dataset Augmentation in Papyrology with Generative Models: A Study of Synthetic Ancient Greek Character Images	34
3.1	Introduction	34
3.2	Related Work	35
3.2.1	Generating Synthetic Images with Machine Learning	35
3.3	Domain Overview: Papyrology	36
3.3.1	Image Dataset: AL-ALL and AL-PUB	36
3.3.2	Image Dataset: AL-SYNTH	37
3.4	Research Goal	37
3.5	Study 1: Modeling with Synthetic Instances	38
3.5.1	Model Type 1: Generative Model	38
3.5.2	Model Types 2 & 3: Categorical Classification Models	39
3.5.3	Datasets	39
3.5.4	Generative Model Results	39
3.5.5	Categorical Classification Model Results	42
3.6	Study 2: Augmenting Creativity with Synthetic Instances	46
3.6.1	Methodology: Web Survey	46
3.6.2	Findings	48
3.7	Summary	50
II	Improving Crowdsourcing and Developing Tools for Digital Papyrology	52
4	Crowdsourcing Image Datasets	53
4.1	Motivation	53
4.2	Pipeline for Crowdsourcing Image Datasets	54
5	Incorporating Crowdsourced Annotator Distributions into Ensemble Modeling to Improve Classification Trustworthiness for Ancient Greek Papyri	56
5.1	Introduction	56
5.1.1	Sources of Noise in the Dataset	57
5.1.2	Normalized Distribution of Annotations (NDA)	58
5.2	Methods	61
5.2.1	Deep Learning Architecture	61
5.2.2	Ensemble Model: Stacked Generalization and K-Nearest Neighbors	64
5.2.3	Evaluation Metrics	64
5.2.4	Measuring Classification Uncertainty with Shannon Entropy	65
5.3	Numerical Experiments	65
5.3.1	Experimental setup	66
5.3.2	Results	66

5.3.3	Entropy Analysis	72
5.4	Summary	77
6	Smart Digital Edition Management: A Blockchain Framework for Papyrology	79
6.1	Introduction	79
6.1.1	Ancient Texts and Papyrology	79
6.1.2	Current Practices in Editing and Peer Review for Digital Papyrology	81
6.1.3	Papyrological Editions	82
6.1.4	Motivation	83
6.1.5	Why Blockchain?	85
6.1.6	Our Contributions	85
6.1.7	Outline	86
6.2	Modern Digital Papyrology and Digital Edition Management	86
6.2.1	Literature Survey	86
6.2.2	Existing Systems	87
6.3	Methodology	88
6.3.1	System Architecture	88
6.3.2	User Interactions	89
6.3.3	Dataset	95
6.3.4	Data Storage: IPFS and Blockchain	96
6.3.5	Smart Contract	97
6.4	Results	97
6.4.1	IPFS Results	98
6.4.2	Testnet Results (Sepolia)	99
6.5	Limitations and Future Work	101
6.6	Summary	102
III	Past, Meet Future: Software Tools for Digital Papyrology	104
7	Towards a Platform for AI-Assisted Papyrology	105
7.1	Introduction	105
7.2	Our AI/ML Tools	105
7.3	Handwritten Text Recognition	106
7.4	Character Spotting	107
7.5	Text Line Segmentation	110
7.6	Manuscript Dating	110
7.7	An AI-Driven Platform for Papyrology	112
8	Future Work	114
9	Conclusion	117
	References	120

List of Figures

Figure 1	Arthur Hunt and Bernard Grenfell, Oxford papyrologists who excavated the Oxyrhynchus papyri.	2
Figure 2	Papyrus fragment Z_POxy.v0068.n4652.a.01_138225. This image, including ink barely clinging to papyrus, is one of the worst examples of the poor condition of many manuscripts in the Oxyrhynchus collection.	3
Figure 3	An example from the Oxyrhynchus papyri collection.	9
Figure 4	A visual representation of our contributions and how they build on prior work.	9
Figure 5	The Ancient Lives interface.	12
Figure 6	CNN Architecture	19
Figure 7	ResNet Architecture	19
Figure 8	Visualizations of the grid search for the <i>CNN-BASE</i> , <i>CNN-XGB</i> , and <i>RESNET</i> models over 75 epochs using <i>AL-TUNE</i> . In contrast to the other models, <i>CNN-XGB</i> produces a smoother curve due to its decision tree design. Best models in blue.	20
Figure 9	Accuracy and loss of the <i>CNN-BASE</i> , <i>CNN-XGB</i> , and <i>RESNET</i> models over 75 epochs on the <i>AL-ALL</i> dataset.	23
Figure 10	Images that were misclassified by <i>RESNET</i> and <i>CNN-BASE</i> with confidence scores and consensus labels in red.	25
Figure 11	Blurriness and confidence for all misidentified Alpha images from <i>RESNET</i> and <i>CNN-BASE</i> in the <i>AL-ALL</i> dataset.	27
Figure 12	An overview of Theia’s interface.	29
Figure 13	Examples of each character in the AL-PUB dataset	37
Figure 14	Examples of characters from damaged papyrus fragments.	37
Figure 15	Images of synthesized Psi and Xi.	40
Figure 16	Images of Synthetic Alpha, Delta, & Pi	41
Figure 17	Accuracy & Loss for 10 CNN runs for each dataset.	43
Figure 18	Accuracy & Loss for 10 ResNet runs for each dataset.	44
Figure 19	Normalized Distribution of Annotations (NDA) for the image Shown in Figure 20.	60
Figure 20	Example of the character Gamma ($\Gamma\gamma$) from AL-PUB. Though the consensus label for this character is Gamma, the ground-truth is unclear due to the degradation exhibited in the source manuscript.	60
Figure 21	Plots of the fraction of correct classifications versus entropy of the output distribution for each model. Note that the KLD model is NOT better than the other models. Rather, it has consistently higher entropy values than the other models.	68
Figure 22	Per character precision and recall for each of the three models. Note that the KNN ensemble model is consistently higher than the individual models for most characters.	70
Figure 23	Precision and recall of all three models versus per character sample size. Notice the drop in recall for characters with smaller sample sizes	71

Figure 24	Histogram of entropies for the NDAs and the three models' output distributions for both the entire data set and solely the character Alpha ($A\alpha$). We split the histograms into two sets: 1) images whose model classification agreed with the human consensus and 2) those which disagreed. . . .	74
Figure 25	Two-dimensional histograms of the three models' output entropies versus the number of human annotations. The left column contains all characters and the right column contains only Alpha ($A\alpha$). We further split the plots into two portions: 1) images whose model classification agreed with the human consensus and 2) those which disagreed. Notice how the mean entropy value tends to decrease for large annotator counts for the correctly classified images.	75
Figure 26	Per character precision and recall of the SVM model which predicts model classification accuracy based on the entropy of Softmax outputs.	77
Figure 27	Papyrus Fragment from the Oxyrhynchus Collection. Note the extensive damage. Such fragments are exceedingly difficult to edit. In this fragment we see missing papyrus segments as well as missing, torn, and faded ink. . . .	80
Figure 28	Due to damage and the unique features of ancient Greek papyri, even the simplest sentence can become challenging to decipher. The phrase, "I baked a delicious cake" becomes unrecognizable under conditions common for ancient Greek papyri.	80
Figure 29	The basic standard for a digital edition of an ancient Greek papyrus is an articulated text with a critical apparatus. This is sufficient for many papyrus texts. Yet, for new or unknown literary and sub-literary works, more components are required for an edition to be considered critical. These include: Paleographic Apparatus, Critical Apparatus, Testimonia, and Text Translations.	83
Figure 30	The few existing systems for managing editions of papyrological transcriptions utilize centralized, database architectures that offer little flexibility or transparency, and may be vulnerable to data loss. A blockchain and smart contract enabled system can offer better security, transparency and data provenance.	84
Figure 31	The architecture consists of 4 layers. The top layer, a user interface, enables the submission of digital editions. In the second layer, a smart contract automatically manages submission and retrieval of digital editions. The third layer consists of the blockchain itself, where the smart contract is deployed and where the location(s) of all associated editions are stored. The final layer is the storage layer, which allows for flexibility of storage options including cloud storage, local server, or decentralized storage options such as IPFS.	88
Figure 32	The proposed architecture includes a user interface and a smart contract deployed on the blockchain that tracks all editions and stores URL or IPFS hash addresses for the associated data on the blockchain.	91

Figure 33	This diagram illustrates the process of submitting an original edition. In this example, the edition data is stored on IPFS while it's hashed address is stored on the blockchain. This information is available to other users for review and critique.	91
Figure 34	This diagram illustrates the process of submitting an expanded edition. The original edition is retrieved from IPFS using the URL that was stored on the blockchain. Expanded editions are then stored on IPFS and their URL stored on the blockchain.	92
Figure 35	An illustration of the process of submitting an original edition. The user interacts only with the user interface. All blockchain transactions and data transmissions are handled in the background via APIs.	92
Figure 36	An illustration of the process for retrieving an original edition for review, then submitting an expanded edition. While the underlying connections between files is complex, blockchain can manage the storage and retrieval of data files without a complex database infrastructure.	93
Figure 37	An illustration of the process of retrieving all editions associated with a given smart contract.	94
Figure 38	Image of Homer's Odyssey on papyrus from the Schoyen Collection. During experimentation, this image was a stand-in for the the original manuscript and was stored on IPFS.	95
Figure 39	During experimentation, individual text and image files were transmitted to IPFS (left side). Data stored on the blockchain ledger (right side) consist of the IPFS addresses to original edition, expanded editions, and the associated image file. In a real-world system, it may be advantageous to transmit a single file to IPFS rather than multiple. XML and JSON files allow for mixed data types and unstructured data in a single document. Additional notes may also be included both on IPFS and the blockchain.	96
Figure 40	The smart contract contains functions that facilitate storage and retrieval of original and expanded editions. In our experiments, we utilized three functions which allowed for the submission of an original edition, submission of expanded editions, and a function for retrieving all associated editions.	97
Figure 41	The Moralis API requires all data files to be encoded as Base64. Here we see a histogram of the Base64 encoding times for expanded edition data files.	98
Figure 42	Here we see the recorded times for data file transmission to IPFS. These transmission times, while greater than those for Base64 encoding, are relatively small and unlikely to be a burden for users.	99
Figure 43	Here we see transaction times for transmission of IPFS URLs to the smart contract on Sepolia. While these times are typical of blockchain transaction times, they are slow compared to a database system. This may be mitigated by using a local blockchain.	100

Figure 44	Here we see transaction costs (in ETH) for each of 3 transaction types performed on the Sepolia Testnet. While these costs are very small, it is unlikely that any implementation of the proposed system will pass on such costs to users.	101
Figure 45	Accuracy & Loss for our ResNet trained on AL-ALL version 2	106
Figure 46	Training Results for transfer learning with YOLOv5 trained with images of Greek characters on Papyrus. The metrics suggest a quality model with mAP_0.5 nearing 0.8, and validation box loss below 0.04.	108
Figure 47	Transfer learning with YOLOv5 trains the model to locate Greek characters. Here we see bounding boxes around detected characters in an image from the validation set. Each box is accompanied by a model confidence value.	109
Figure 48	Character Spotting. YOLOv5 was trained to locate handwritten Greek characters on papyrus (bottom left), but does a surprisingly good job at locating characters in other scripts including Bangla (top left), Korean (top right), and Hebrew (bottom right).	109
Figure 49	Visualization of line-segmentation algorithm results. Blue circles denote the bounding box radius from the new character segmentation model, while the red lines connect all characters in a given row of text.	110
Figure 50	Here we see a confusion matrix for early test of our manuscript dating model. For images containing greater than 25 characters, we see validation accuracy of over 72%.	112
Figure 51	Character Location User Interface. The newly trained character segmentation model locates characters in the image. Users can then edit the located characters and add additional characters the model may have missed.	113
Figure 52	An AI-assisted transcription of Greek text on papyrus. The color-coding (right) denotes line association. Each character is accompanied by a classification probability.	114
Figure 53	During this interaction with Claude 3, I gave the model an image of a fragment for the Oxyrhynchus collection to read. The model recognized that the image contained a manuscript written on papyrus, but mistook the Greek characters for Hebrew or Aramaic. The model was unable to read the text and suggested that an expert's help is needed to read the fragment.	116

List of Tables

Table 1	Counts for each letter in the Ancient Lives dataset.	15
Table 2	Accuracy for each model during hypertuning.	22
Table 3	An overview of the explored hyperparameter space alongside the best-performing values for each model.	23
Table 4	Average validation accuracy for all three models with both datasets across a 10-fold cross-validation procedure.	25
Table 5	Precision & Recall Comparison: AL-ALL & AL-SYNTH	45
Table 6	Per Character Accuracy for Target Sub-samples. M1 Denotes model trained on AL-ALL. M2 Denotes model trained on AL-SYNTH	46
Table 7	Four questions posed during Phase 2 of Study 2.	48
Table 8	F1 score for each expert respondent in Survey 2.	49
Table 9	Number of images correctly/incorrectly classified by the CXE- and KLD-ResNets. Both models agreed with consensus annotation for 353,549 images, and both disagreed for 10,871 images.	62
Table 10	Overall and per-character precision and recall (with respect to the human consensus) for the CXE, KLD, and KNN ensemble models.	67
Table 11	Contribution Comparison: The body of existing works related to our proposed system is limited, as our approach appears to be truly novel, covering all necessary and relevant aspects.	87
Table 12	Counts for each letter in AL-ALL version 2. This new version of the dataset contains 419,439 Greek character images.	107

Part I

**Crowdsourced Greek Character Datasets
and Optical Character Recognition**



Figure 1: Arthur Hunt and Bernard Grenfell, Oxford papyrologists who excavated the Oxyrhynchus papyri.

1 Introduction

1.1 Research Overview

Papyrology, the conservation, editing, and interpretation of ancient Greek texts written on papyrus, is a delicate and time-consuming discipline. This field is critical to the study of the cultures of the ancient Mediterranean, from Ancient Egypt to the Christian and Islamic periods. In the late 19th and early 20th century, two Oxford papyrologists, Arthur Hunt and Bernard Grenfell, Figure 1, excavated an astounding quantity of manuscripts from rubbish dumps found at the edge of the ancient Egyptian city of Oxyrhynchus [13]. These manuscripts consist of papyrus fragments written in a number of languages including Coptic, Aramaic, and Syriac. The vast majority, however, were written in Greek. Among the fragments were civil records such as tax documents, legal codes, historical texts, and religious works. Some of the oldest existing copies of the New Testament of the Bible, and portions of Euclid’s “the Elements” were among the fragments unearthed at Oxyrhynchus. Hundreds of thousands of papyrus fragments were collected, the vast majority of which are currently stored at the Bodleian Art Library of Oxford University.



Figure 2: Papyrus fragment Z_POxy.v0068.n4652.a.01_138225. This image, including ink barely clinging to papyrus, is one of the worst examples of the poor condition of many manuscripts in the Oxyrhynchus collection.

Over 100 years later, only a small fraction of these papyrus fragments have been transcribed. While the dry desert climate aided in the preservation of the Oxyrhynchus papyri, the fragments exhibit high levels of degradation as they were, after all, rubbish. The fragments exhibit extensive tears, holes, and faded ink. In some cases, virtually nothing but the ink remains (Figure 2). There are only a modest number of experts trained to decipher these texts, and the fragments themselves are very fragile. To aid in the conservation and transcription of the Oxyrhynchus and other papyrus collections, the Imaging Papyri Project at Oxford University created digital images of numerous papyrus fragments. Many of the images of the Oxyrhynchus fragments are available online at <http://www.papyrology.ox.ac.uk/POxy/>.

With the astonishing recent advances in A.I. and machine learning, particularly models for Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR), one would expect that automating the transcription of ancient documents from their digital images would be trivial. This couldn't be further from the truth. The central dogma of handwritten character recognition (also commonly referred to as handwritten text recognition) remains inextricably linked to optical character recognition methods for print media. Alongside their reliance on proprietary data and lack of open-access software, the applicability of these optical character recognition methods to handwritten characters from low-quality documents remains limited. Additionally, most machine learning image datasets, such as MNIST [25] and the St. Gall database [28], consist of ideally

cropped character images or custom-made datasets annotated through costly services such as Amazon’s Mechanical Turk [55] or through years of expert transcription work. The alternative approach is crowdsourced datasets, another venture requiring years of effort. Crowdsourcing opens up transcription to mostly untrained citizen scientists, which leads to a number of complications which make this approach challenging for image datasets. Regardless of the source of transcription data, most state-of-the-art systems, Tesseract [76] for example, perform poorly for manuscripts from the ancient world [81].

Novel approaches will be required to aid papyrologist and other humanities professionals and educators in the important work of documenting ancient, historical texts. My efforts are designed to expand development of a pipeline for creating A.I. enabled tools to aid papyrologist in the diplomatic transcription process utilizing crowdsourced annotations of ancient Greek papyrus fragments. This pipeline will be discussed in more detail in Section 4.

One of the greatest challenges in this process has been data quality. Crowdsourced annotations performed by citizen scientists can be difficult to model and, at times, untrustworthy. This greatly complicates deriving ground-truth labels. The damaged nature of the source material adds another dimension of difficulty. In many cases, even trained experts may disagree on the transcription of individual manuscripts, as ink degradation or missing papyrus make many of the Greek characters’ identities ambiguous. The digital images were created using various methods, and are of an unknown number of resolutions, complicating methods used to locate and extract characters.

The Oxyrhynchus papyri are literally *trash*, and the data derived is indicative of that. When work began on this project, it seemed unlikely that many of the approaches would bear fruit, but surprisingly, the apparent flaws in the data may have actually been advantages (Section 2.4.2). The following sections consist of excerpts from my publications, detailing the efforts of myself and my fellow researchers toward those goals. The data gathered in previous works have allowed for the creation of impressive deep learning models for character segmentation and character classification for papyrology. When combined with a line-segmentation algorithm, these models have made automatic transcription of Greek papyrus fragment images feasible, and these models are currently

being developed into software tools for papyrology. Additionally, analysis of the data, from raw annotations to model inference probabilities, has revealed new insights that may lead to improved methods for crowdsourcing datasets using entropic measures of uncertainty for neural network predictions and crowdsourced annotations.

1.2 Publications

Prior work in the digital humanities was key to my research, especially the Ancient Lives project, a <https://www.zooniverse.org/> project, which generated millions of crowdsourced annotations for 12,070 images of papyrus fragments between 2011 and 2018. The consensus annotations developed from the Ancient Lives project include centroid locations and character identifications [94]. These consensus annotations and the associated fragment images are at the core of my research and was a critical first step of the larger project.

This dissertation is comprised of modified excerpts from the publications associated with this work in the sections 2 through 7, followed by a concluding section detailing the aggregate results of the research as well as current and future work (Section 9).

1.2.1 IEEE eScience 2021

My first publication, **Exploring Learning Approaches for Ancient Greek Character Recognition with Citizen Science Data**, was accepted to the 17th IEEE eScience conference in Innsbruck, Austria in 2021. I am the lead author on this publication which is highlighted in Section 2. This article features a large-scale image dataset, AL_ALL, created using the annotations from the Ancient Lives Project. Numerous deep learning models were trained on the dataset demonstrating the efficacy of the dataset for deep learning applications and producing highly performant classification models.

1.2.2 International Joint Conference on Artificial Intelligence 2022 (IJCAI)

My second publication, **Dataset Augmentation in Papyrology with Generative Models: A Study of Synthetic Ancient Greek Character Images**, was accepted to the 31st International Joint Conference on Artificial Intelligence in Vienna, Austria in 2022. I am the lead author on this publication which is highlighted in Section 3. This article explores the use of generative models trained on images from AL_ALL. Synthetic instances of several Greek characters on papyrus were generated. These synthetic characters were used to explore their utility in the professional practice of papyrology, and to augment smaller samples within the dataset to reduce sampling bias for two deep learning architectures.

1.2.3 Conference on Human Computation and Crowdsourcing Doctoral Consortium 2022 (HCOMP)

My third publication, **Crowdsourcing Image Datasets: An Examination of Ground-Truth in Labeling, Text Segmentation, & Sampling Bias**, was accepted to the Tenth Annual Conference on Human Computation and Crowdsourcing Doctoral Consortium, held virtually in 2022. I am the lead author on this publication which is highlighted in Section 4. This article explores the methodology used to create the AL_ALL dataset and outlines a pipeline, from crowdsourced annotations, to deep learning models used for character recognition in images of highly damaged, ancient manuscripts. The pipeline is likely generalizable and can be utilized to create character image datasets for other language scripts.

1.2.4 Journal of Data Mining and Digital Humanities 2024 (JDMDH)

My fourth publication, **Incorporating Crowdsourced Annotator Distributions into Ensemble Modeling to Improve Classification Trustworthiness for Ancient Greek Papyri**, was accepted to the Journal of Data Mining and Digital Humanities, Special Edition on Historical Documents and Text Recognition in 2024. I am the co-lead author on this publication with Dr. Graham West, which is highlighted in Section 5. In this article we explore a novel approach to ensemble modeling

for character classification models. Two nearly identical ResNet models are trained on AL_ALL. One of these models uses typical deep learning methodologies, while the other model utilizes a normalized distribution of the crowdsourced annotation as labels, along with an atypical loss function, Kullback-Leibler Divergence. These two models are then combined using k-nearest neighbor clustering to achieve higher accuracy than either component model. In addition, we explore the use of Shannon entropy for measuring classification uncertainty in each model, as well as the original annotator distribution.

1.2.5 ACM Conference on Intelligent User Interfaces 2024 (IUI)

My fifth publication, **Towards a Platform for AI-Assisted Papyrology**, was accepted to the 29th Annual ACM Conference on Intelligent User Interfaces, “Past Meets Future” Workshop Proceedings in 2024. I am the lead author on this publication which is highlighted in Section 7. In this article, we discuss the state of digital and deep learning tools for the practice of papyrology, highlight our contributions to the field, and propose an approachable user interface that combines our suite of tools into a comprehensive platform for digital papyrology. Additionally we discuss current and future work related to the aforementioned platform for papyrology.

1.2.6 ACM SIGMIS Computers and People Research 2024

My sixth publication, **Smart Digital Edition Management: A Blockchain Framework for Papyrology**, has been accepted for publication in the ACM SIGMIS Computers and People Research conference to be held May 29 through June 1, 2024 in Murfreesboro, TN. I am the lead author on this publication which is highlighted in Section 6. In this article we propose a novel framework for management of digital papyrology editions from numerous authors that employs blockchain technology. This decentralized, distributed ledger approach side-steps the challenges of complex database storage methodologies and enable automated storage and retrieval of digital editions while ensuring data security and provenance.

2 Exploring Learning Approaches for Ancient Greek Character Recognition with Citizen Science Data ¹

2.1 Introduction

Labeled datasets of handwritten digits and characters, such as MNIST, have been critical in advancing the field of machine learning over the past three decades [25, 49, 48]. Within the past few years, studies have continued to acknowledge the relevance of such datasets with particular interests in extending them (e.g. from digits to letters [21]). When data is unavailable, character recognition researchers are generally required to create their own datasets [64], which can often be costly in terms of time, effort, and money. Today, the number of openly available datasets remains significantly limited in quantity despite their growing demand, particularly in cultural heritage contexts [26, 69].

Over the past decade, one cost-effective and increasingly common method of data collection is crowdsourcing, in which a gold standard label is generated by multiple annotators [31]. Despite being generally cost-effective, the use of multiple annotators often results in a set of noisy labels that are non-uniform and maintain some level of disagreement. Such noise has been shown to heavily influence a dataset’s utility in machine learning contexts [62, 67, 98], and research has therefore given significant attention to engineering techniques to mitigate noise by various statistical measures [6, 82, 92]. Other approaches have simply thrown out the noisy labels altogether [99]. Modern datasets generally fail to reflect the reality that crowdsourced data is not only imperfect, but so large in magnitude that identifying annotator errors is both challenging and time-consuming in task settings in which a ground truth label may be ambiguous. One such example is transcribing the deteriorated papyrus manuscript in Figure 3.

In this section, we compare and contrast the effectiveness of a state-of-the-art optical character

¹Modified from the original article published in the 17th IEEE eScience Conference, 2021. Authored by Matthew I. Swindall, Gregory Croisdale, Chase C. Hunter, Ben Keener, Alex C. Williams, James H. Brusuelas, Nita Krevans, Melissa Sellev, Lucy Fortson, John F. Wallin



Figure 3: An example from the Oxyrhynchus papyri collection.

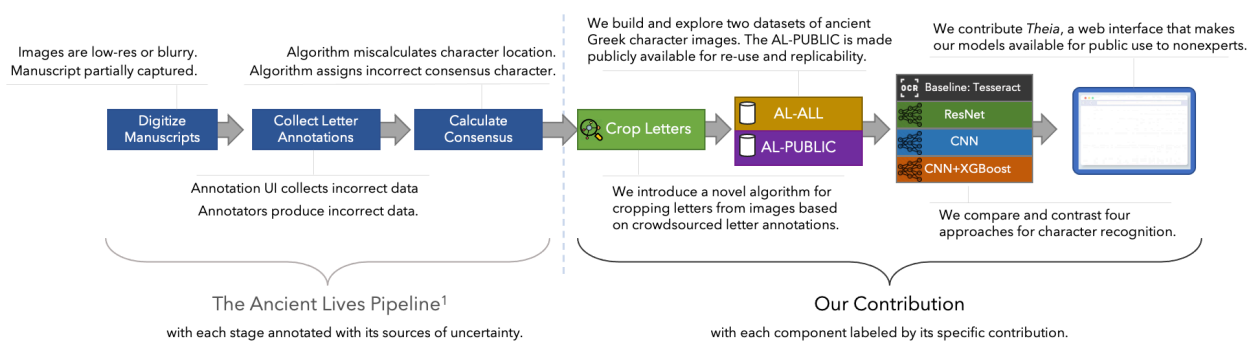


Figure 4: A visual representation of our contributions and how they build on prior work.

recognition tool (i.e., Tesseract) to a set of novel machine learning approaches that share the task of classifying handwritten character images. Our approaches are fueled by the *Ancient Lives* dataset, a collection of digitized images of handwritten ancient Greek characters that are the product of the *Ancient Lives Project*. We first present a novel cropping algorithm that isolates and extracts each character in each manuscript into independent and labeled image files. Using the *Ancient Lives* dataset as source material, we establish two ancient Greek character image datasets: (1) *AL-ALL* and (2) *AL-PUB*. We use these two datasets to train three unique machine learning models and compare their effectiveness against the Tesseract OCR tool. We find that all three model approaches perform more effectively than Tesseract, with our best-performing model achieving an accuracy of 92.73%. Following an analysis of our results, we conduct an audit of misclassifications across our model and introduce *Theia*, a web interface that allows non-experts (e.g., in the humanities) to utilize our classification models without the need for expertise. We conclude the section with a discussion on our findings and their implications for research at the intersection of machine learning, citizen science, noisy labeling, and handwritten character recognition in cultural heritage contexts.

2.2 Related Literature

In this section we will discuss previous and related work. Past efforts to develop computational models for OCR and HTR for Greek papyri have largely been impeded by a lack of available data. The data we will discuss throughout comes from a unique, crowdsourcing initiative which yields complicated data that requires an interdisciplinary team to navigate. Thus, it is salient at this stage to point out an important oversight in the development of *AL-ALL*. The original version of *AL-ALL* contains the character Sigma (Σ, σ), rather than the Lunate Sigma (C, ς) that is typically found on papyri from this period. This oversight resulted in lower accuracy for the models produced from the original dataset. This is discussed in further detail in Section 7.3.

2.2.1 Handwritten Character Recognition

For more than three decades, advances in handwritten character recognition have relied substantially on publicly available datasets. Amid the range of datasets MNIST, a dataset of 70,000 images of handwritten digits written by high school students and employees of the United State Census Bureau. The dataset has served as a focal point of application, education, and extension in machine learning [21]. Following the widespread usage of MNIST, a myriad of character recognition datasets have been collected and made publicly available, such as those that focus on handwriting in multiple languages [24, 46, 51], symbols [40, 84], or noisy handwriting [42].

2.2.2 Crowdsourcing and Noise Labelers

Crowdsourcing is an exceptionally popular technique for building datasets both in and beyond computer vision. As a by-product of engaging labelers of varying backgrounds, expertise, and personal characteristics, crowdsourcing research is often concerned with techniques for managing noisy labels [62, 67, 98]. Controlled studies have, for example, focused substantially on the development of characterizing the reliability of annotators, particularly in cultural heritage projects where finances (i.e., for labeling) are often very limited [92]. To mitigate the concerns around cost and reliability, humanities-oriented transcription projects, such as Old Weather [8] and Operation War Diary [35], generally rely on volunteers who are intrinsically motivated to participate in labeling as opposed to other types of labelers who are extrinsically motivated (e.g. Amazon Mechanical Turk workers) [55]. Despite the sustained use of crowdsourcing for data collection in manuscript contexts, handwritten character recognition datasets remain relatively limited within the purview of ancient manuscripts, alongside the digital humanities at large.

2.2.3 Crowdsourcing Manuscript Transcription

Enlisting the help of the public, either through monetary or voluntary means, has been a growing topic of interest for the digital humanities in recent years. Many such projects are geared toward the production of digital transcriptions that are of interest both to archival repositories and to scholars

working on critical editions. Transcribe Bentham [19], for example, is a long-running and thriving project that enlists the help of the public to transcribe the journal pages of Jeremy Bentham. Beyond the task of document transcription, newer projects, such as Ancient Lives and Scribes of the Cairo Geniza, employ their crowdsourced transcriptions to identify or simply contextualize unstudied manuscripts [93], which may require specialized computational pipelines for data processing and preparation [94]. The popularity of this kind of research is evidenced by the rise of platforms dedicated to hosting projects for archives, libraries, and museums since 2010, most notably the *Zooniverse* citizen science platform [75], the *From the Page* platform [63], and the *SciFabric / Pybossa* platform [9]. While crowdsourced manuscript transcription efforts and initiatives have risen in popularity, the public availability and re-use of crowdsourced transcription data from these projects has yet to become the norm.

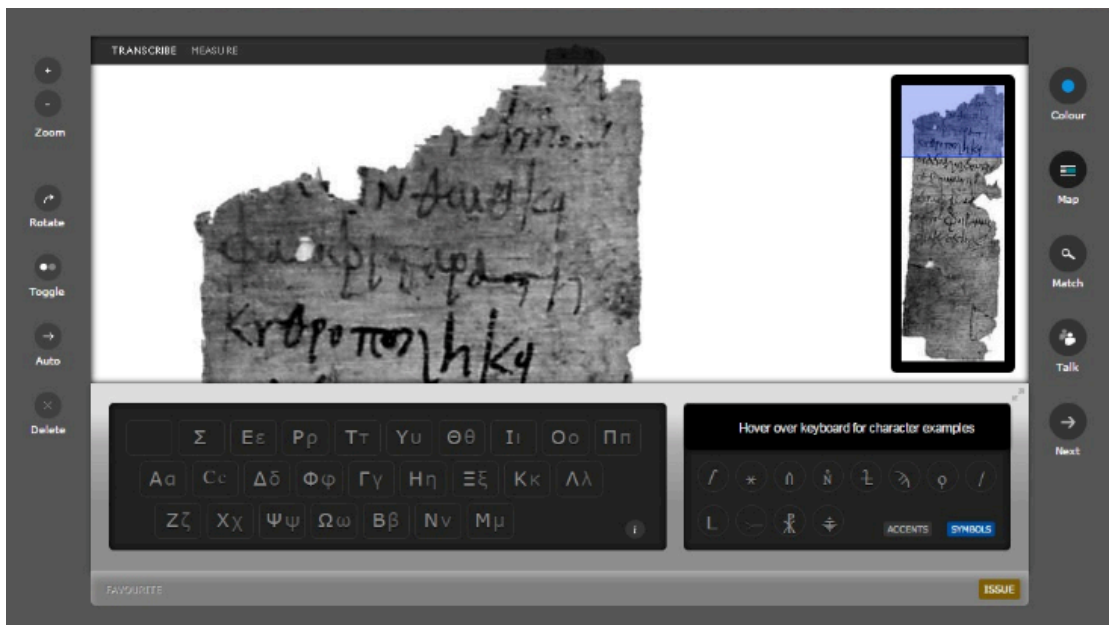


Figure 5: The Ancient Lives interface.

2.2.4 Contribution

In this work, we make several contributions, each of which expands on the prior literature in a unique way. First, we contribute a novel cropping algorithm that extracts individual character images from manuscript images using crowdsourced data. Second, we introduce two datasets of

ancient Greek character images, one of which is made publicly available for re-use and replicability among the digital humanities, computer vision, and machine learning communities. Third, we design, implement, and optimize three machine learning modelling approaches that are trained with state-of-the-art machine learning toolkits. Alongside these models, we contribute an evaluation of these models, focusing specifically on their ability to classify ancient Greek characters in comparison to a state-of-the-art OCR tool. Our final contribution is *Theia*, a web interface tool that allows non-experts to utilize our machine learning models with their own character images. An overview of our contributions is shown in Figure 4.

2.3 Ancient Lives

Ancient Lives was a web-based citizen science project that was launched in June 2011 in coordination with the Zooniverse [75] and concluded in June 2018. Through the internet, the project enlisted the help of volunteers from across the world to transcribe deteriorated ancient Greek papyrus fragments (*i.e.*, remnants of a larger manuscript). All 12,070 papyri fragments that were transcribed via the Ancient Lives system belonged to the family of the Oxyrhynchus papyri, an established collection of ancient Greek papyrus manuscripts discovered in the ancient Egyptian city of Oxyrhynchus [11, 17, 12].

2.3.1 Task Interface

Counter to most citizen science projects for transcription, Ancient Lives’s task interface treats a transcription event as an object detection process. Users are asked to transcribe each fragment by finding one letter at a time. At the beginning of a task, users are presented with a papyrus fragment along with a virtual keyboard that allows the user to customize their transcription experience (*e.g.*, change annotation color). Annotations can be added to the interface simply by clicking on the image. An annotation’s position can be updated at any time by clicking and dragging the UI element to a new location. Once an annotation is created, a user can assign a letter to the annotation, or update the annotation’s letter, by clicking on the appropriate letter on the virtual keyboard. When

a user hovered over any key on the virtual keyboard, two example images of the Greek letter or symbol were shown on the top-left panel of the keyboard. Lastly, users are given a mini-map to show the field-of-view of the papyrus fragment being viewed. Users can change their field-of-view of the image by clicking and dragging on either the mini-map or the image itself. The interface enforced no constraints about how (e.g. in what order) letters or symbols should be annotated. The interface is shown in Figure 5.

2.3.2 Processing Pipeline & Data Quality

To organize and collate the wealth of annotation data, an existing computational pipeline developed by Williams et al. [94] was leveraged. The pipeline implements several algorithms that facilitate the processes of (1) aggregating letter annotations into consensus annotations and (2) creating “chains” of letter annotations to create strings of text.

In the creation of the dataset at hand, the team only makes use of (1) as the research is limited to individual characters. The quality of the data produced through the pipeline’s procedure has been vetted [94], and the Ancient Lives data itself has been used toward several other contexts (e.g., deteriorated manuscript identification [93]).

Character	Count	Character	Count
Alpha (A, α)	42,538	Nu (N, ν)	44,896
Beta (B, β)	2,534	Xi (Ξ, ξ)	1,201
Gamma (Γ, γ)	6,907	Omicron (O, o)	46,334
Delta (Δ, δ)	11,716	Pi (Π, π)	17,112
Epsilon (E, ϵ)	31,581	Rho (P, ρ)	20,448
Zeta (Z, ζ)	1,425	Sigma (Σ, σ)	62
Eta (H, η)	15,062	Tau (T, τ)	32,034
Theta (Θ, θ)	7,575	Upsilon (Y, υ)	15,762
Iota (I, ι)	25,593	Phi (Φ, ϕ)	6,063
Kappa (K, κ)	17,932	Chi (X, χ)	9,155
Lambda (Λ, λ)	13,253	Psi (Ψ, ψ)	904
Mu (M, μ)	13,225	Omega (Ω, ω)	16,043

Table 1: Counts for each letter in the Ancient Lives dataset.

2.3.3 Cropping Algorithm and Consensus Label

The *Ancient Lives* web interface was used to collect annotations on manuscript images that included one or more characters. To generate a dataset of characters, we designed a cropping algorithm that was applied to each manuscript image to extract individual character images. Using the the coordinate information of each annotation, the distance to the nearest adjacent character (δ) is calculated. Each image was then cropped by $1.1(\frac{\delta}{2})$ pixels including a 10% buffer, from the indicated location of the character along the vertical and horizontal dimensions. All cropped images were automatically resized to 70 x 70 pixels. To limit the number of extraction errors (e.g. extracting markings), a filtering criterion was applied in which images that had fewer than three annotators in agreement were removed. As the focus is on alphabetic characters, symbols and miscellaneous markings were excluded. Target labels for each image were chosen by taking the consensus of annotators’ labels (i.e., majority vote). Table 1 shows the dataset’s character distribution.

2.3.4 The Ancient Lives Dataset

We used the cropping algorithm to produce two versions of the *Ancient Lives* dataset for training machine learning models:

- *AL-PUB*: Includes 195,683 labeled character images from 5,043 published manuscript im-

ages that were used in the Ancient Lives web interface.

- *AL-ALL*: Includes 399,330 labeled character images from 12,070 published and unpublished manuscript images that were used in the Ancient Lives web interface. This includes all images from the *AL-PUB* dataset.

Our motivation to create two versions of the *Ancient Lives* dataset was driven by one issue: scholars working on unpublished and unidentified manuscripts. In order not to reveal manuscript data still under papyrological research, the AL-PUB dataset stems from previously published material in The Oxyrhynchus Papyri Series. Nevertheless, each image in both dataset versions *ideally* contains one tightly cropped Greek character, stored in JPEG format, and can be of a range of several resolutions. All 24 Greek alphabet characters are represented in the dataset. Images are sorted into sub-directories for each character and follows a file-naming convention that enumerates its associated labels.

2.4 Methods

The goal of our research is to compare and contrast the performance of learning approaches trained on the Ancient Lives dataset to state-of-the-art optical character recognition methods. Here, we describe the various modelling approaches we explored alongside the OCR tool used as a baseline measure.

2.4.1 Baseline: Tesseract

Tesseract [76] is a state-of-the-art optical recognition software tool for extracting text from images. Prior examinations of Tesseract have demonstrated its effectiveness for character recognition for extracting text across a multitude of languages [77], including ancient Greek texts [29, 91]. Further, comparative studies suggest that the effectiveness of Tesseract varies between contexts (e.g., license plates in greyscale vs. in color) in contrast to proprietary OCR alternatives (e.g. Transym) [66]. Based on the wealth of prior research reinforcing its utility, we used the Tesseract engine

designed for ancient Greek² as our baseline measure. We ran Tesseract on individual character page segmentation mode with a whitelist consisting solely of the characters in the dataset. Despite being on single character mode, Tesseract may return multiple characters. To remedy this without penalizing Tesseract, we considered any transcription with the target character present to be accurate.

2.4.2 Learning Approaches

We designed and implemented three unique learning approaches using Tensorflow and Keras that stem from the broader family of Convolutional Neural Networks (CNNs). The decision to employ CNNs as an alternative method to optical character recognition tools was motivated by the wealth of literature that has demonstrated the success of CNNs as tools for unique handwritten character recognition scenarios [20, 36, 87]. During early experimentation, it was observed that our models performed worse when using certain techniques to avoid overfitting. Specifically, the use of image data augmentation significantly degraded the accuracy of our models. We believe this may be due to the noise inherent in the dataset which may act as a form of built-in data augmentation. We have yet to explore this possibility in depth, but chose to continue our experiments without using further augmentation. Our explored learning approaches include:

1. Standard CNN (*CNN-BASE*): A standard CNN with a configurable architecture of convolution layers (e.g., that control image tensor dimensionality) and max-pooling layers. This base architecture is both highly configurable and highly performant for character recognition [20]. This architecture, highlighted in Figure 6, includes 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

²<https://ancientgreekocr.org/>

softmax activation for 24 classes.

2. CNN + XGBoost (*CNN-XGB*): XGBoost is a machine learning library which applies gradient boosted trees in classifying data. When layered with a standard CNN, XGBoost is able to classify characters with more depth and greater accuracy using the extracted features from the CNN. The concept is modified from ConvXGB [85], with some modifications to hyperparameters tuned.
3. ResNet Model (*RESNET*): This model is loosely based on the Residual Learning framework first introduced in 2015 [37]. The framework, illustrated in Figure 7 is a modification of the typical CNN used in image recognition, where a residual learning component is added to combat problems when increasing depth. This in turn allows the model to remain time efficient and accurate at above average depth, resulting in more detail captured from our character. 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 softmax output layer for 24 classes.

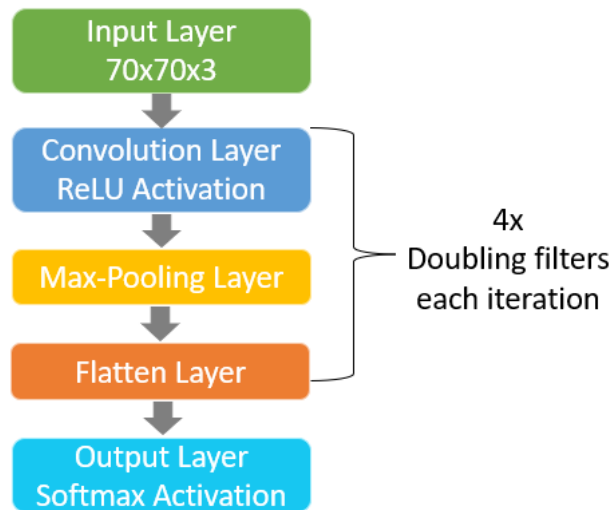


Figure 6: CNN Architecture

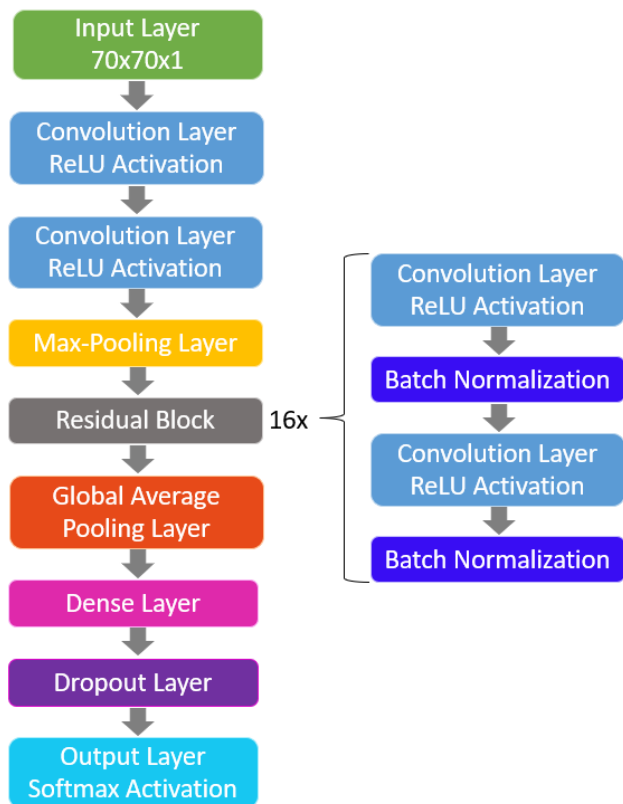
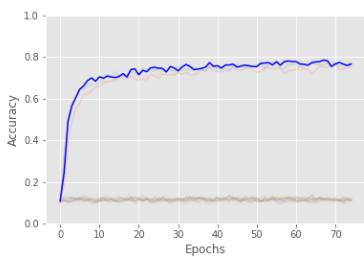
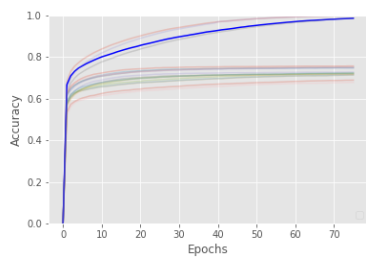


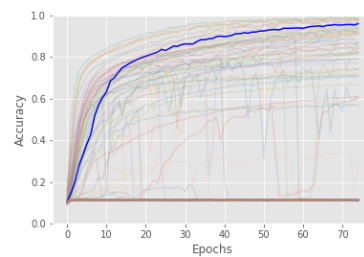
Figure 7: ResNet Architecture



(a) *CNN-BASE*



(b) *CNN-XGB*



(c) *RESNET*

Figure 8: Visualizations of the grid search for the *CNN-BASE*, *CNN-XGB*, and *RESNET* models over 75 epochs using *AL-TUNE*. In contrast to the other models, *CNN-XGB* produces a smoother curve due to its decision tree design. Best models in blue.

2.4.3 Hyperparameter Tuning: Procedure

We optimized our machine learning models by conducting both a coarse grid search and fine grid search on the standard list of configurable hyperparameters for convolutional neural network models [5]. Our initial plan of execution involved the exploration of multiple values for all hyperparameters across each model. However, preliminary runs of the hypertuning process using the *AL-PUB* and the *AL-ALL* datasets were estimated to require a compute time of upward of a year to execute to completion. To reduce the amount of time required to conduct an adequate grid search, we conducted a hypertuning process that utilized a reduced dataset of character images alongside a smaller number of explored parameters.

2.4.3.1 AL-TUNE: A Reduced Dataset of Character Images

We randomly sampled 39,924 (10%) character images from the *AL-ALL* dataset to create a derivative dataset, which we henceforth refer to as *AL-TUNE*. After the sampling process had completed, we confirmed that character distributions between the *AL-ALL* and *AL-TUNE* datasets were relatively similar, suggesting that the reduced dataset was statistically representative of the larger dataset. While representation was generally maintained, several characters (i.e., Sigma, Psi) were not well-represented in the reduced *AL-TUNE* dataset as a by-product of having significantly lower representation in the larger *AL-ALL* dataset.

2.4.4 Reduction of Explored Values

A preliminary grid search using the *AL-TUNE* dataset with all three model types revealed that several hyperparameters (i.e., activation function, convolutional kernel size, max pooling kernel size, and momentum) had minimal effect on the outcome of validation accuracy across searches. We therefore reduced the range of explored values for these hyperparameters by setting them to static values as shown in Table 3. The range of explored values for other hyperparameters was guided by tool documentation and prior studies that suggest appropriate values for configuring CNNs [5]. The complete list of explored hyperparameter values, alongside best-performing value

for each hyperparameter for each model, are shown in Table 3. The procedure was executed on a five-node GPU cluster with Intel i9 9820x processors and dual NVIDIA RTX 2080TI GPUs. All runs were conducted on individual cluster nodes, and each model took approximately 48 hours to complete.

2.4.5 Hyperparameter Tuning: Results

Figure 8 shows the validation accuracy for all hyperparameter configurations for each model across 75 training epochs. Throughout the grid search procedure, we observed significant variance in accuracy across the various model combinations. We specifically observed that, when the learning rate of a CNN was greater than 0.001, model accuracy tended to stagnate around 10%. Additionally, the *CNN-XGB* model’s accuracy across 75 epochs tended to be positively correlated with the maximum tree depth. We chose the best models from this hyperparameter tuning procedure by averaging the 90th percentile of the accuracies that were generated by each epoch, which are shown in Table 2. Following the conclusion of the hyperparameter tuning procedure, the best-performing model configurations were statically implemented for each model and subsequently executed on the same cluster system using both the *AL-ALL* and *AL-PUB* datasets.

Table 2: Accuracy for each model during hypertuning.

Model	Val. Accuracy
CNN-BASE	86.7%
CNN-XGB	84.4%
RESNET	80.4%.

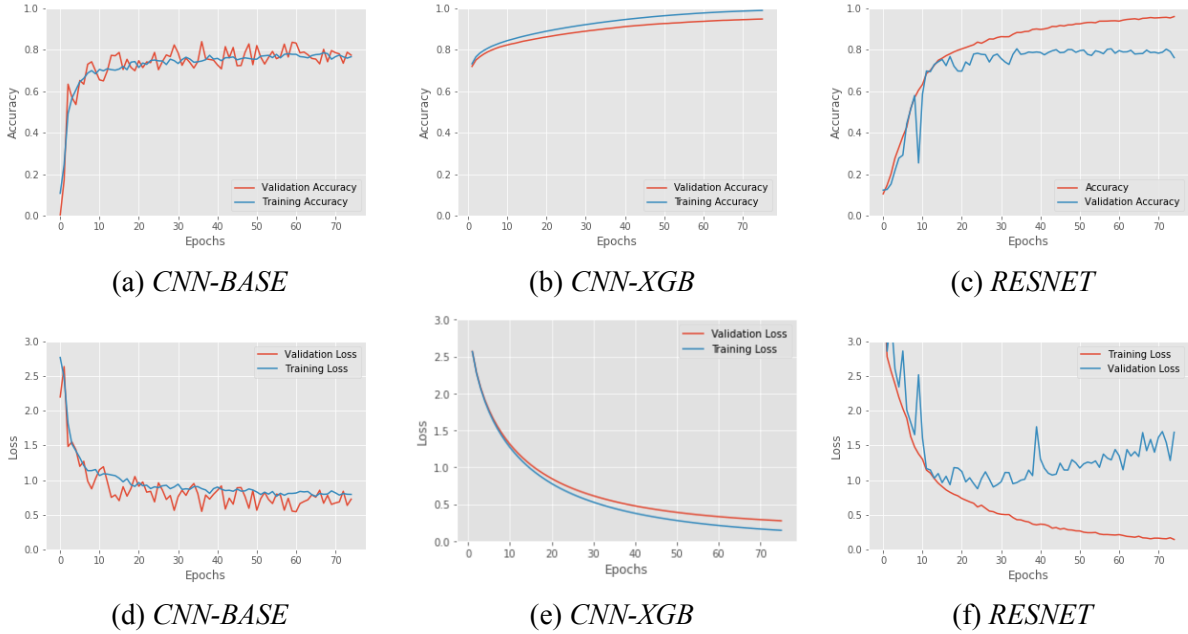


Figure 9: Accuracy and loss of the *CNN-BASE*, *CNN-XGB*, and *RESNET* models over 75 epochs on the *AL-ALL* dataset.

	Hyperparameter	Explored Values	Best
CNN-BASE	Learning Rate	0.001, 0.01, 0.03, 0.05	0.001
	Optimizer	Adam, RMSprop	Adam
	Number of Filters	16, 32, 64, 96	96
	Activation Func.	RELU	RELU
	Convolutional Kernel	3x3	3x3
	Max Pooling Kernel	2x2	2x2
CNN-XGB	Learning Rate (CNN)	0.001, 0.01, 0.1	0.001
	Number of Filters (CNN)	8, 16, 32, 64	16
	Eta (XGB Learning Rate)	0.01, 0.1, 0.2, 0.3, 0.4	0.1
	Max_Tree_Depth	5, 10, 30, 50, 100, 200	100
	Min_Child_Weight	2, 4, 6, 8, 10	10
RESNET	Learning Rate	0.001, 0.01, 0.03, 0.05	0.001
	Num. of Hidden Layers	2, 4, 8, 16, 32, 50	32
	Optimizer	Adam, RMSprop	RMSprop
	Number of Filters	16, 32, 64, 96	96
	Activation Function	RELU	RELU
	Kernel Size	3x3	3x3
	Momentum	0.9	0.9

Table 3: An overview of the explored hyperparameter space alongside the best-performing values for each model.

2.5 Results

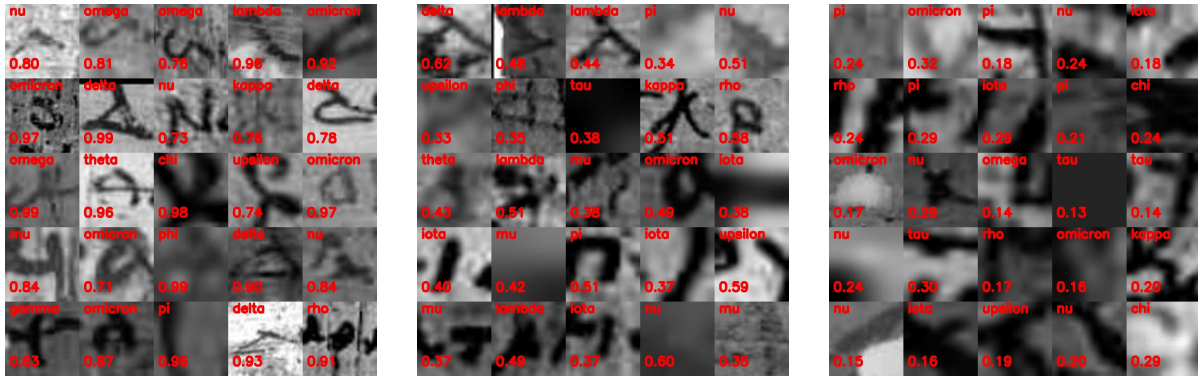
In this section, we discuss the the results of all four explored approaches for both the *AL-ALL* dataset and the *AL-PUB* dataset. As Tesseract uses a pre-trained model, we discuss its accuracy both in aggregate and at the letter level. In contrast to Tesseract’s discussion of results, we employed a standard k-fold cross validation procedure with each of the three learning approaches. For each of these approaches, we discuss validation accuracy and loss and reported the observation accuracy. We conclude this section by revisiting instances of misclassifications to better understand both why and how our explored modeling approaches fail.

2.5.1 Tesseract

Overall, the pretrained Tesseract Ancient Greek model performed significantly worse than any of the newly trained model alternatives. Tesseract correctly produced a text transcription (i.e., that included the expected Greek character) in 45,988 (11.15%) of the character images in the *AL-ALL* dataset. Across this same dataset, Tesseract failed to recognize any text whatsoever in 133,950 character images (33.54%). Mirroring the tool’s performance on *AL-ALL* dataset, Tesseract correctly produced a text transcription in 20,027 (10.23%) of the character images in the *AL-PUB* dataset while failing to recognize any text at all in 67,186 character images (34.33%). In examining the recognition effectiveness of individual letter images in the *AL-ALL* dataset, we observe that Tesseract was most effective at recognizing Epsilon images, classifying 4,576 of the 31,581 Epsilon images (14.49%) correctly. In contrast, we find that Beta images were the most frequently failure case for Tesseract, classifying only 68 of the 2534 Beta images (2.68%) correctly.

2.5.2 CNN-BASE, CNN-XGB, and RESNET

Table 4 shows the average validation accuracy for each of the three model types for the standard k-fold cross-validation procedure over 10 iterations. Figure 9 shows accuracy and loss for model training across each of the three model types.



(a) 66% or Greater Confidence. (b) 33-66% Confidence. (c) 0-33% Confidence.

Figure 10: Images that were misclassified by *RESNET* and *CNN-BASE* with confidence scores and consensus labels in red.

In general, all three modeling approaches significantly outperformed Tesseract OCR. Among the three learning models, the *RESNET* model achieved the best performing accuracy for both the *AL-ALL* and *AL-PUB* datasets, achieving an accuracy of 92.73% and 92.57% respectively. In contrast, the *CNN-XGB* exhibited significant variance between datasets, achieving an accuracy of 80.32% on the *AL-PUB* dataset and a substantially higher accuracy of 90.81% on the *AL-ALL* dataset. The *CNN-BASE* model averaged a similarly lower accuracy of 80.24% on the *AL-PUB* dataset while achieving a lower average of accuracy of 80.79% on the *AL-ALL* dataset.

The *RESNET* model’s sustained accuracy across datasets suggests that it is the most reliable model for use in practice. However, all three approaches demonstrate a significant level of practical utility as the smallest average of accuracy among our models is 80.32%, which indicates – in the worst case – 38,510 of the *AL-PUB*’s 195,683 character images were on average incorrectly classified.

Table 4: Average validation accuracy for all three models with both datasets across a 10-fold cross-validation procedure.

Model	AL-PUB	AL-ALL
CNN-BASE	82.24% ($\sigma=0.03$)	80.79% ($\sigma=0.01$)
CNN-XGB	80.32% ($\sigma=0.63$)	90.81% ($\sigma=1.21$)
RESNET	92.57% ($\sigma=4.23$)	92.73% ($\sigma=3.44$)

2.5.3 Auditing Misclassifications with RESNET and CNN-BASE

To better understand the shortcomings and failures of our learning approaches, we conducted a high-level examination of the errors encountered in our best-performing model: the *RESNET* model. In assessing the model’s misclassifications, we observe that the *RESNET* model incorrectly classified a total of 29,533 character images (7.4%) out of the total 399,330 character images in the *AL-ALL* dataset. In comparison, the *CNN-BASE* model incorrectly classified 53,427 character images in the same dataset (13.1%). A total of 17,780 character images were misclassified by both of these models. Figure 10 shows a collage of examples across these models with confidence scores and consensus labels in red. To better understand these misclassifications, we examined the characters that were misclassified by both the *RESNET* model and by the *CNN-BASE* model. We direct our investigation toward the 1,405 misclassified character images that were labeled as the Greek letter Alpha (α) according to our labeling pipeline shown in Figure 4. Restricting our sample to confidence levels greater than 33% (i.e., Figures 10a and 10b), we observe that misclassified character images can be categorized into two categories: (1) images that were misclassified by the volunteers or (2) images that are blank or unreadable. Based on a small sample of 100 characters in this low confidence group that were labeled as Alpha, but classified as another Greek character, 97 of the testing set images were found to be mislabeled. In other words, the *RESNET* model missed only three of the characters that were actually Alpha. In manually reviewing this small sample, we find the *RESNET* model to be more effective at classifying image examples of high ambiguity (e.g., cursive characters) than our research team members who lack formal training in Greek paleography.

In contrast, the vast majority of misclassified character images with confidence levels less than 33% (i.e., Figure 10c) in our sample were blurred, if not simply impossible to read by our research team’s qualified experts. As shown in Figure 11, we visually observe that images with higher levels of blurriness tend to have lower confidence scores in the *RESNET* model, suggesting that future experiments may benefit significantly from an image dataset thresholded on image blurriness. This observation suggests that the original digitization process, the cropping algorithm, or the consensus

algorithm that was employed to create the dataset encountered a failure. In other words, our observation suggests that it was not the *RESNET* and *CNN* models that encountered a misclassification error, but the prior pipeline stage that was used to produce the dataset.

Based on this preliminary analysis, we estimate that at least 80% of the misclassifications in the *RESNET* model can be attributed to labeling errors, stemming from any one of the pre-processing stages (e.g., cropping, consensus assignment, or imperfect annotations) shown in Figure 4. Based on this brief manual examination of this data, only about 1% Alpha character images were incorrectly classified by the *RESNET* model. There were 1,125 Alpha character images that were misclassified by the *RESNET* model, but were not misclassified by the *CNN* model. When we examine these characters, we find about 20% of the characters with confidence greater than 66%

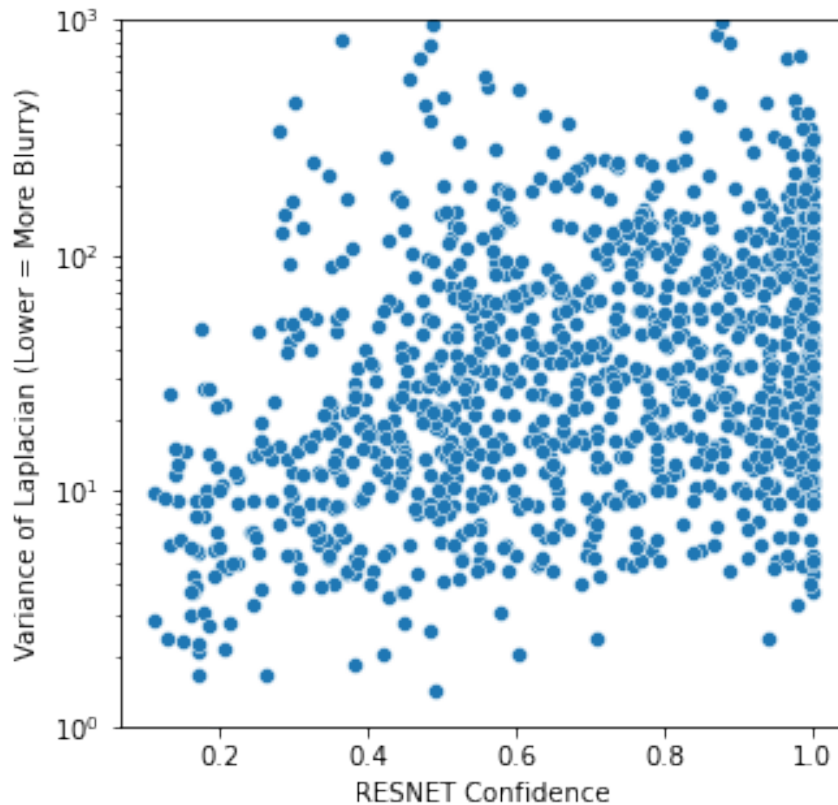
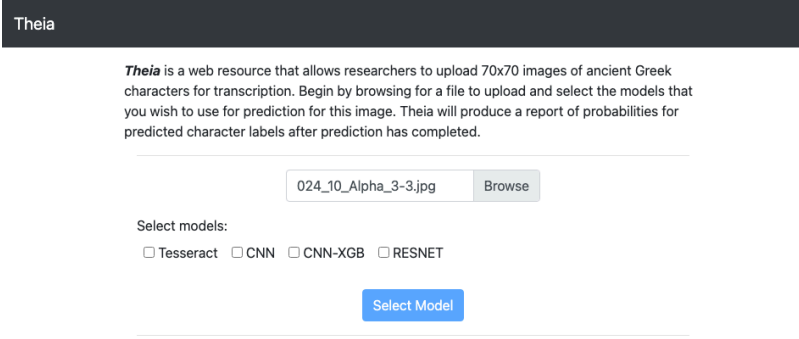


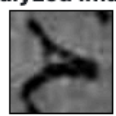
Figure 11: Blurriness and confidence for all misidentified Alpha images from *RESNET* and *CNN-BASE* in the *AL-ALL* dataset.

were incorrectly classified. For the misclassifications situated between 33% and 66% range of confidence, about 5% of the actual Alpha characters were incorrectly classified.



(a) The base interface in which users upload a character image and make a selection for approaches to character recognition.

Analyzed Image:



	Tesseract	CNN
Alpha	0%	91.258%
Beta	0%	0%
Chi	0%	0%
Delta	0%	0%
Epsilon	0%	0%
Eta	0%	0%
Gamma	0%	0%
Iota	0%	0%
Kappa	0%	0%
Lambda	0%	8.741%
Mu	0%	0%
Nu	0%	0%
Omega	0%	0%
Omicron	0%	0%
Phi	0%	0%
Pi	100%	0%
Psi	0%	0%
Rho	0%	0%
Sigma	0%	0%
Tau	0%	0%
Theta	0%	0%
Upsilon	0%	0%
Xi	0%	0%
Zeta	0%	0%
Best	Pi	Alpha

(b) An example of Theia's classification output.

Figure 12: An overview of Theia's interface.

2.6 Theia: A Web UI for Character Classification

Our findings suggest that each of our three machine learning models substantially outperform the Tesseract OCR engine. Here, we introduce *Theia*, a static web interface that allows members of the research community to easily and intuitively reproduce analysis on our approaches. The interface is specifically designed for nonexperts (e.g., in the humanities) who lack the necessary technical experience to train and evaluate machine learning approaches. *Theia* was built using the Tensorflow.js and Tesseract.js libraries, both of which use client-side resources (i.e., GPU using WebGL compute shaders) to load learning modules and perform model-related computation. *Theia*'s front-end centric design facilitates an affordable hosting model for the academic and research community by requiring only a standard HTTP server (i.e., for statically hosted content) to function. Further, Tensorflow.js maintains embedded support for importing re-usable learning models such that future research teams can engineer their own models (e.g., in other language formats), which could be later incorporated into and hosted within the *Theia* system. The *Theia* system can be found on the project webpage³.

2.6.1 User Experience and Web Interface

Theia aims to enable nonexperts with the ability to identify which of our trained models are most appropriate for their own project. *Theia* could, for example, allow a team of archaeologists to explore the applicability of our models to a dataset of digitized ancient Greek manuscripts that are the product of a digitization initiative that used a unique digitization technique (e.g., that is dissimilar to the technique used to digitize the Oxyrhynchus papyri studied in this paper). Users can use *Theia* by uploading an image of appropriate size (70x70), selecting the desired models to use for classification (e.g. Tesseract, *CNN-BASE*, *CNN-XGB*, or *RESNET*), and clicking the “Transcribe” button to send the image to the models for classification. After the models have completed their classification procedure, the image will be rendered to the user alongside a list of possible character labels and their associated classification probabilities. Note that the Tesseract model may return

³<https://utk-pairs.github.io/theia/>

multiple characters, so there may appear to be more than one “best” transcription. By engaging in this process iteratively, users are capable of determining which model performs best under their specific circumstances. An overview of Theia’s user experience is shown in Figure 12.

2.7 Discussion

Our work demonstrates that learning methods can be usefully applied to the task of identifying handwritten characters in digitized images of ancient manuscripts. In this study’s context, we designed, implemented, optimized, and explored the classification effectiveness of three unique approaches for identifying ancient Greek characters in manuscripts from the Oxyrhynchus papyri collection. Our comparative analysis with the Tesseract OCR engine suggests that each of our trained models were significantly more accurate in classifying all possible characters images across the ancient Greek alphabet.

An important consideration for interpreting our findings is the fundamental basis in which the Tesseract engine itself was trained. Like our three learning approaches, Tesseract utilizes an embedded learning approach (i.e., LSTM) to facilitate its optical character recognition [15]. Our decision to use Tesseract as a baseline measure of performance was motivated by its widespread utility as an out-of-the-box tool for effective character recognition. A follow-up study could explore training Tesseract with the *AL-ALL* and *AL-PUB* datasets to create a custom new Tesseract OCR model, which could also be made available via Theia. Such a study would help draw finer conclusions about the performance trade-offs between the approaches evaluated in our work.

Alongside our comparative analysis, our research introduces new insights into the applicability of our learning approaches to scenarios that involve imperfect data. A wealth of research in crowdsourcing and machine learning research have given ample attention to the development of computational techniques for identifying reliable annotators, filtering out unreliable sources of data, and more generally, improving the quality of crowdsourced data [23, 41, 92]. In conjunction with new aggregation frameworks for learning from imperfect annotation data [68], we find evidence that suggests that our modelling approaches are surprisingly resilient to imperfect data. Specifically,

our audit of misclassifications in Section 2.5.3 highlights that our learning approaches encounter misclassification errors with character images that may have been labeled by human annotators incorrectly, algorithmically assigned an incorrect consensus letter, or simply algorithmically cropped in an imperfect fashion. Further, we find that many misclassifications may stem from the dataset’s initial digitization as described in Figure 4. Our audit introduces a new frontier for further exploring the effect of these “error-induced” misclassifications and how their exclusion from the *AL-ALL* and *AL-PUB* datasets may affect model performance.

Lastly, our findings set a compelling benchmark for future research at the intersection of character recognition, citizen science, and machine learning for cultural heritage contexts. Historically, research initiatives for manuscript transcription retain their data as a by-product of platform policy agreements. By making the *AL-PUB* dataset publicly available and introducing *Theia* into the research community, we believe that we’ve taken the necessary steps toward establishing a change of attitude in sharing data, models, and resources among the research communities that work at this intersection. Researchers can, for example, extend *Theia* to datasets of alternative languages of interest and incorporate new learning models for widespread use among the community. With the *AL-PUB* dataset, machine learning researchers now have a readily-available alternative to the MNIST dataset [25] that has been exhaustively used for nearly three decades of research. In general, these contributions facilitate the re-use, replication, and extension of our research in the interest of generating new advances for computing and the humanities alike.

2.8 Summary

In this section, we explored the performance of state-of-the-art optical character recognition tools for print and learning models engineered with state-of-the-art machine learning toolkits trained on handwritten inputs. Using Tesseract OCR as a baseline, we build, optimize, and evaluate three types of convolutional neural networks that are trained on the *AL-ALL* and *AL-PUB*⁴ datasets, a collection of images of handwritten ancient Greek characters that were labeled by volunteers through the

⁴<https://data.cs.mtsu.edu/al-pub>

Ancient Lives online citizen science project. We find our best-performing machine learning model to be 92.57% accurate compared to Tesseract OCR's 11.15%. Each of our models could have benefited from early stopping, though the degradation in later epochs was minimal for most variations during cross-validation. Following our analysis, we present a brief examination of our models' shortcomings, introduce the publicly-available *AL-PUB* dataset, and, describe *Theia*, a web-based tool that democratizes our machine learning models for public use. We conclude by discussing the promise of our findings for advancing research at the intersection of machine learning, manuscript transcription, and the digital humanities.

3 Dataset Augmentation in Papyrology with Generative Models: A Study of Synthetic Ancient Greek Character Images⁵

3.1 Introduction

In the previous section, the AL-All and AL-PUB datasets have been shown to be valuable tools for digital papyrology. 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 [74][86]. 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 [22] [30]. 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.

In this section we present findings from a two-part study to understand the role synthetic characters can play in machine learning contexts and professional papyrology. To reduce class imbalance,

⁵Modified from the original article published at The 31st International Joint Conference on Artificial Intelligence 2022 (IJCAI-ECAI). Authored by Matthew I. Swindall, Timothy Player, Ben Keener, Alex C. Williams, James H. Brusuelas, Federica Nicolardi, Marzia D'Angelo, Claudio Vergara, Michael McOsker, John F. Wallin

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.

3.2 Related Work

3.2.1 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 [33]. The website <https://thispersondoesnotexist.com/> showcases an improvement on GAN architecture with StyleGAN2, as described in [43]. This demonstrates how GANs have the potential to revolutionize machine learning applications for numerous use cases. In [30] CNN performance increased roughly 10% when training data was augmented with synthetically generated images. In [22] 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 [27] where GAN-like architecture is

used to explore artistic styling.

3.3 Domain Overview: Papyrology

3.3.1 Image Dataset: AL-ALL and AL-PUB

AL-ALL and AL-PUB are “noisy” image datasets. The highly damaged nature of the source papyri, as shown in Figure 14, is the single greatest source of noise in the data. Due to physical damage, some characters simply appear in greater numbers than others, and some are transcribed less. The datasets and the various sources of noise are detailed further in Section 5.1.1,



Figure 13: Examples of each character in the AL-PUB dataset [81]



Figure 14: Examples of characters from damaged papyrus fragments.

3.3.2 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 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.

3.4 Research Goal

In this section, 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.

3.5 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, as discussed in Section 2 [81]. 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>.

3.5.1 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 differentially 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.

3.5.2 Model Types 2 & 3: Categorical Classification Models

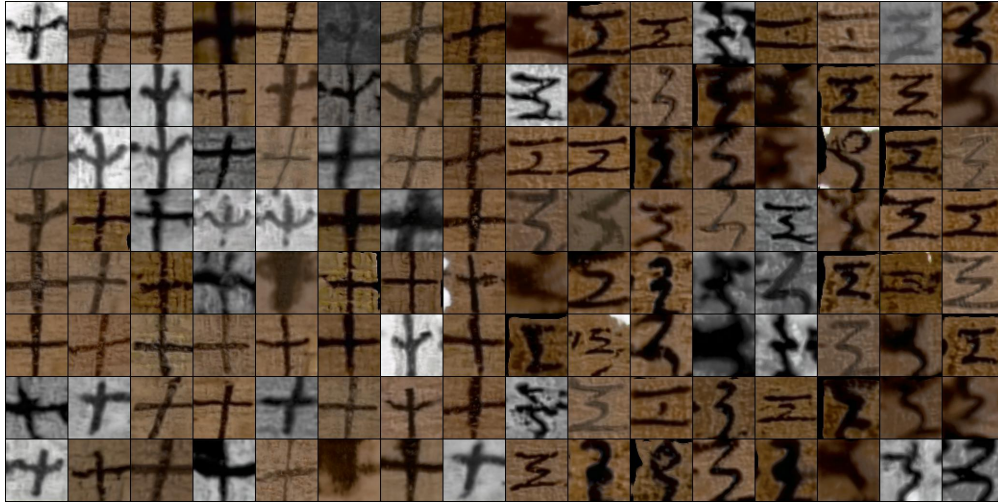
The CNN model, illustrated in Figure 6, is a very typical CNN model. The details of this model, including network layers, activation functions, and hyperparameters is discussed thoroughly in Section 2.4.2. The more complex ResNet model, illustrated in Figure 7, is also a fairly typical implementation. This models full details are outline in Section 2.4.2 as well.

3.5.3 Datasets

Machine learning results can be tricky to reproduce. Rather than rely on old results from [81], 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 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.

3.5.4 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 sythetic 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



(a) Synthetic Psi

(b) Synthetic Xi

Figure 15: Images of synthesized Psi and Xi.

(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 15 (a) and (b) show examples of synthetic Psi and Xi images, while Figure 16 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.

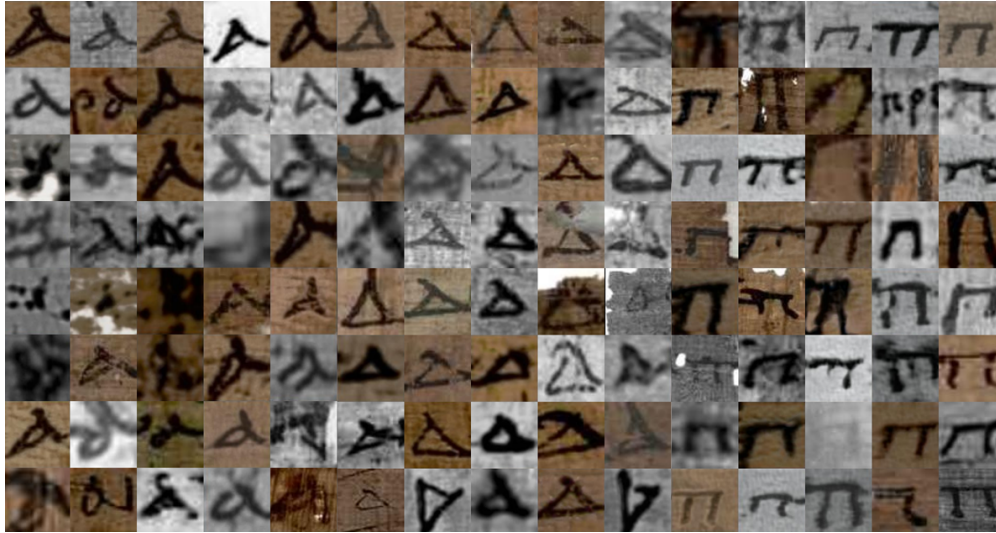


Figure 16: Images of Synthetic Alpha, Delta, & Pi

3.5.4.1 Psi($\Psi \psi$)

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.

3.5.4.2 Xi($\Xi \xi$)

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.

3.5.4.3 Additional Characters ($\text{A } \alpha, \Delta \delta, \Pi \pi$)

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 these 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.

3.5.5 Categorical Classification Model Results

Figures 17 and 18 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.

3.5.5.1 CNN Results

As in Section 2 [81], 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 17 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.

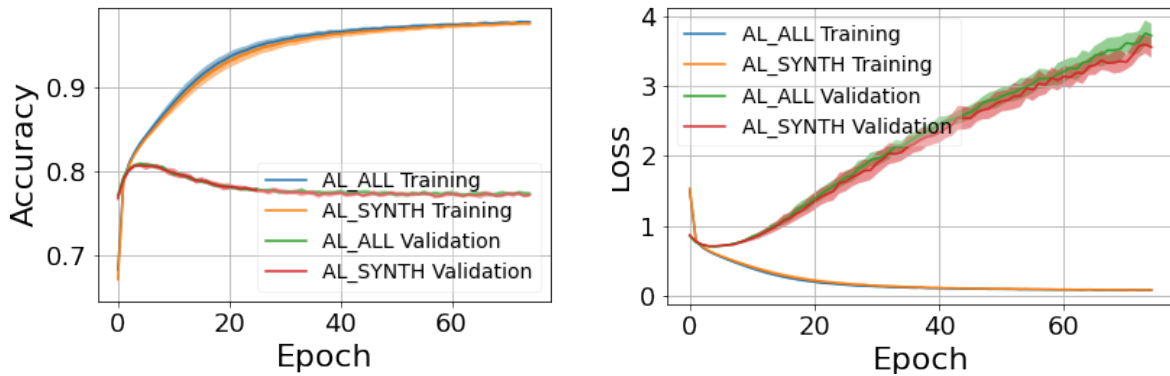


Figure 17: Accuracy & Loss for 10 CNN runs for each dataset.

3.5.5.2 ResNet Results

As shown in Figures 17 and 18, 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.

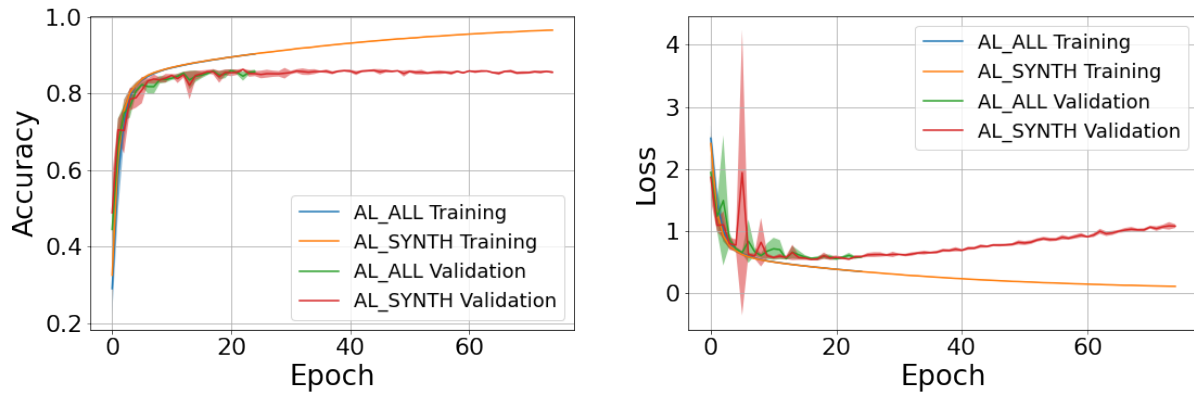


Figure 18: Accuracy & Loss for 10 ResNet runs for each dataset.

3.5.5.3 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 5 shows the precision and recall for the CNN and ResNet, trained on the original AL-ALL and new AL-SYNT 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.

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 5: Precision & Recall Comparison: AL-ALL & AL-SYNTH

3.5.5.4 Per Character Accuracy

Table 6 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.

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

3.6 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.

3.6.1 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.

3.6.1.1 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 ($A\alpha$), Delta ($\Delta \delta$), and Pi ($\Pi \pi$). 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.

3.6.1.2 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 for augmenting professional papyrological practice (see Table 7). 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.

- 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 7: Four questions posed during Phase 2 of Study 2.

3.6.2 Findings

3.6.2.1 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 8.

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 8: F1 score for each expert respondent in Survey 2.

3.6.2.2 Phase 2: Assessing Practical Potential

Thematic analysis was used to analyze qualitative data collected via the four questions in Table 7 [14]. 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, respondents 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 [4]) for communal and practical use.

3.7 Summary

In this section, 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 improvements in the classification of under-represented samples, such as Psi and Xi. We do not, however, see any meaningful improvement in overall model accuracy for either architectures. 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. As with the models discussed in Section 2, our models in this section would have benefited from early stopping, but degradation in later epochs was relatively minimal.

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 [10], 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.

Part II

**Improving Crowdsourcing and Developing
Tools for Digital Papyrology**

4 Crowdsourcing Image Datasets⁶

4.1 Motivation

As discussed in earlier sections, character and text recognition are mature fields. However, ancient documents are a vastly different quality of data source compared to those used to train OCR engines. These manuscripts are challenging if not impossible for state-of-the-art systems to decipher. It is of immense importance to the study of ancient texts, and the history they contain, to create systems capable of rapidly transcribing extensively damaged manuscripts. Crowdsourcing annotations of ancient documents is an important step in developing such systems, but including non-expert annotations into the process introduces a great deal of uncertainty in ground-truth labeling. Volunteer annotators regularly make incorrect judgments and are often in disagreement with each other. Qualifying and quantifying the trustworthiness of volunteer annotations is a necessary next step toward practical use of crowdsourced datasets. In addition to ground-truth uncertainty, crowdsourced datasets contain additional noise including sampling bias. Some characters are less common than others in a given language, or in a given collection of documents. For the rarest characters, class imbalance ensures models yield low accuracy for such samples. Automatic text extraction is yet another challenge for ancient documents, as text recognition engines are easily confused by the damage in these manuscripts. Solving the problems of ground-truth uncertainty, sampling bias, and text extraction for ancient manuscripts would make it possible to design a pipeline for developing interactive agents capable of assisting experts in transcribing these documents and dramatically reduce the time necessary to transcribe entire collections. An additional challenge lies in managing the various digital editions of studied papyri. Current systems for digital edition management rely on centralized and cumbersome databases that lack the flexibility to manage the complicated associations between multiple editions from multiple authors for a single manuscript.

⁶Modified from the original article published in The Tenth Annual AAAI Conference on Human Computation and Crowdsourcing 2022 (HCOMP), Doctoral Consortium. Authored by Matthew I. Swindall

4.2 Pipeline for Crowdsourcing Image Datasets

Building on the work discussed in Part I, we propose a complete pipeline for crowdsourcing image datasets for the development of AI/ML tools to aid in the practice of digital papyrology. This pipeline consist of the following components:

1. **Crowdsourcing annotations (preexisting work):** The starting point for this larger project is the collecting of millions of annotations of Greek characters on ancient, damaged papyrus fragments through the crowdsourcing initiative called The Ancient Lives Project, which was discussed in more detail in Section 1.
2. **Creating consensus annotations (preexisting work):** These annotations were combined to produce consensus annotations (locations and classification) for hundreds of thousands of characters. This is also discussed further in Section 1.
3. **Creating a dataset of individually cropped Greek characters on papyrus:** These consensus annotations were then used to crop out individual character from the fragment images to create the AL-ALL and AL-PUB datasets which were detailed in Section 2.
4. **Train neural networks on the dataset for classification tasks:** Also a major component of my first paper (Section 2), was the design, training, and analysis of multiple deep learning models, trained on the new image dataset for Greek character classification.
5. **Train neural networks to locate Greek characters in images:** More recently, an additional model was created using transfer learning and the object detection model YOLO which locates Greek characters in digital images. This ongoing development is discussed in Section 7.4.
6. **Utilize these neural networks to create tools to aid papyrologists:** Additional models and computational tools for papyrology have been developed, building on the core models and dataset including synthetically generated Greek characters (Section 3), ensemble model

approaches for quantifying classification uncertainty (Section 5), an image-based manuscript dating model, line-finding algorithm (Section 7.5), and a software prototype for interactive, automated transcriptions (Section 7.7).

7. **Develop tools for automatic dating of manuscripts:** Deciphering the time-frame in which a manuscript was written is a lengthy and arduous process. Direct dating techniques such as radio-carbon dating can be prohibitively expensive and would not be practical for large manuscript collections. As such, papyrologist and paleographers employ indirect dating mechanisms. One such method involves detailed analysis of individual character shapes. In Section 8, we detail our novel approach to ensemble modeling for manuscript dating based on character shapes.
8. **Create an digital edition management framework for creating, editing, and storing digital editions:** The tools we have developed have limited efficacy as stand-alone software implementations. To enable the intended end users, a larger ecosystem is required, one with a front-end user interface for editing transcriptions, and a blockchain and smart contract enabled back-end for managing storage and retrieval of digital editions (Section 6.3.5)

A pipeline such as this can be easily modified for similar datasets collected via crowdsourcing or through professional services where multiple annotators collaboratively label data. Other dataset produced from Zooniverse collaborations may have a similar annotation structure that would lend itself well to our approach. Applying our pipeline to another dataset, such as Galaxy Zoo (<https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/>) would be a challenging test of our proposed methodologies. In the remainder of Part II and Part III, I will discuss the developments born of our pipeline in the form of published articles, current experiments, and future work.

5 Incorporating Crowdsourced Annotator Distributions into Ensemble Modeling to Improve Classification Trustworthiness for Ancient Greek Papyri ⁷

5.1 Introduction

One of the greatest challenges in the domain of machine learning is the acquisition of useful datasets. Three main-stream approaches include:

- Professional annotation services
- Synthetic dataset generation
- Crowdsourcing initiatives

Professional annotation services such as Amazon Mechanical Turk outsource labeling to human annotators who are unlikely to be domain experts, i.e. [79]. The approach can also be expensive and take a long time depending on the size of the dataset. While Generative Adversarial Networks (GANs) can produce astonishingly realistic synthetic image datasets, as demonstrated in Section 3 [80], it is impractical to verify that every feature and label are correct for large scale projects. Crowdsourcing initiatives face some of the same challenges as the first approach, but with some additional ones. Due to the number of annotators, crowdsourcing makes it more practical to have multiple annotators review and annotate the same image. This can be a helpful mechanism for managing annotation errors, but it also leads to an inherent uncertainty in labeling. If multiple annotators disagree about the class to which an image belongs, what should its label be? Regardless of the approach used, there is usually a lack of certainty, i.e. “noise” in the ground-truth of the produced dataset.

⁷Modified from the original article published in The Journal of Data Mining and Digital Humanities, Special Edition on Historical Documents and Text Recognition, 2024. Authored by Graham West, Matthew I. Swindall, Ben Keener, Timothy Player, Alex C. Williams, James H. Brusuelas, John F. Wallin

In the field of papyrology, these challenges are compounded by the lack of quality image datasets and a shortage of experts. Efforts to produce a crowdsourced image dataset from such papyri, such as AL-ALL and AL-PUB have shown that, even though the ground-truth in labeling may be questionable, the utilization of annotator consensus can produce datasets that can be accurately modeled using deep learning approaches.

To build on these efforts, we propose a stacked generalization method comprised of two ResNets trained on the AL-ALL dataset with identical architectures but different loss functions and labeling schemes. It is important to clarify that the focus of this work, as well as our contribution, is the incorporation of this novel labeling scheme, not ensemble approaches in general. We thoroughly discuss this scheme, called Normalized Distribution of Annotations (NDA), in later sections. We then explore the utility of the Shannon entropy of the various models' output probability distributions as a measure of classification uncertainty. Our results demonstrate that the proposed ensemble method yields greater accuracy than either of the individual models for this dataset. It also highlights the ability of the Shannon entropy to quantify the uncertainties of both the model output distributions as well as the distribution of crowdsourced annotations.

5.1.1 Sources of Noise in the Dataset

As AL-ALL and AL-PUB were derived from heavily damaged papyrus fragments annotated by untrained citizen scientists, various sources of noise have been introduced into the dataset. The most prominent sources of noise can be split into three main categories that will be discussed below.

5.1.1.1 Class Imbalance

There is significant class imbalance inherent to the datasets due to multiple factors. The primary source of class imbalance is the lack of instances of certain characters in the source manuscripts themselves. This imbalance is compounded by the tendency of human annotators to “chase characters” i.e., to preferentially annotate specific characters. While most characters have many thousands of samples (see Table 1), there are several which have a relatively small amount. The most signif-

icant of these imbalanced outliers is Sigma ($\Sigma \sigma$), with only 62 samples. As discussed in earlier sections, Sigma is atypical in ancient papyri. Due to its small sample size, Sigma is the most challenging character to classify accurately and our models perform very poorly on this character, but the Lunate Sigma is much more common. We discuss the Lunate Sigma issue further in Section 7.3

5.1.1.2 Image Noise

As shown in Figure 14, many images within AL-ALL and AL-PUB contain illegible or incomplete characters, or potentially contain multiple characters. A further source of noise is the cropping algorithm used to create images of the individual characters. The cropping bounds are determined by approximating a manuscript's mean inter-character distance which is calculated from the character locations as determined by the annotators. Since characters on a manuscript are rarely evenly-spaced and the annotator locations are imperfect, this approach can lead to image crops which are zoomed too far out (including multiple characters) or too far in (eliminating portions of the character of interest).

5.1.1.3 Annotator Noise

In addition to the uncertainty in character locations, annotators supply an additional source of noise since they, at times, classify characters incorrectly and regularly disagree with other annotators. As most annotators are not trained professionals, they may easily confuse characters with similar shapes such as Alpha, Delta, and Lambda (A, Δ, Λ). It is also possible for annotators to confuse a sequence of characters for a single character due to the connection (or near-connection) of their ligatures. For example: confusing two side-by-side Tau's ($\tau\tau$) for Pi (π).

5.1.2 Normalized Distribution of Annotations (NDA)

The amalgamation of all annotations for a single character in AL-ALL and AL-PUB can be treated as a sparse distribution over the set of 24 character classes. When normalized, this distribution

is similar to a probability distribution. We illustrate this Normalized Distribution of Annotations (NDA) in Equation 1. If we let $M = 24$ classes and x_i be the number of human annotations for class i for some image (i.e., the values of the non-normalized distribution of annotations), then we can represent NDA of the image by x'_i where

$$x'_i = \frac{x_i}{\sum_{j=1}^M x_j} \quad (1)$$

Figure 19 shows a single example of this NDA. This distribution is an excellent illustration of the annotator noise inherent in the datasets as the ground-truth is not crystal-clear. The character image corresponding to this NDA is shown in Figure 20 with a consensus label of Gamma (Γ). However, a large percentage of annotators (roughly 33%) believed the character to be a Psi (Ψ). The extant damage in the papyrus fragment makes it challenging even for expert papyrologists to agree on the correct classification for this image. The NDA is reminiscent of, yet very different from, a non-continuous version of a Softmax output distribution in neural networks. In the Methods section, we make use of the NDA in the KLD loss function of our second ResNet.

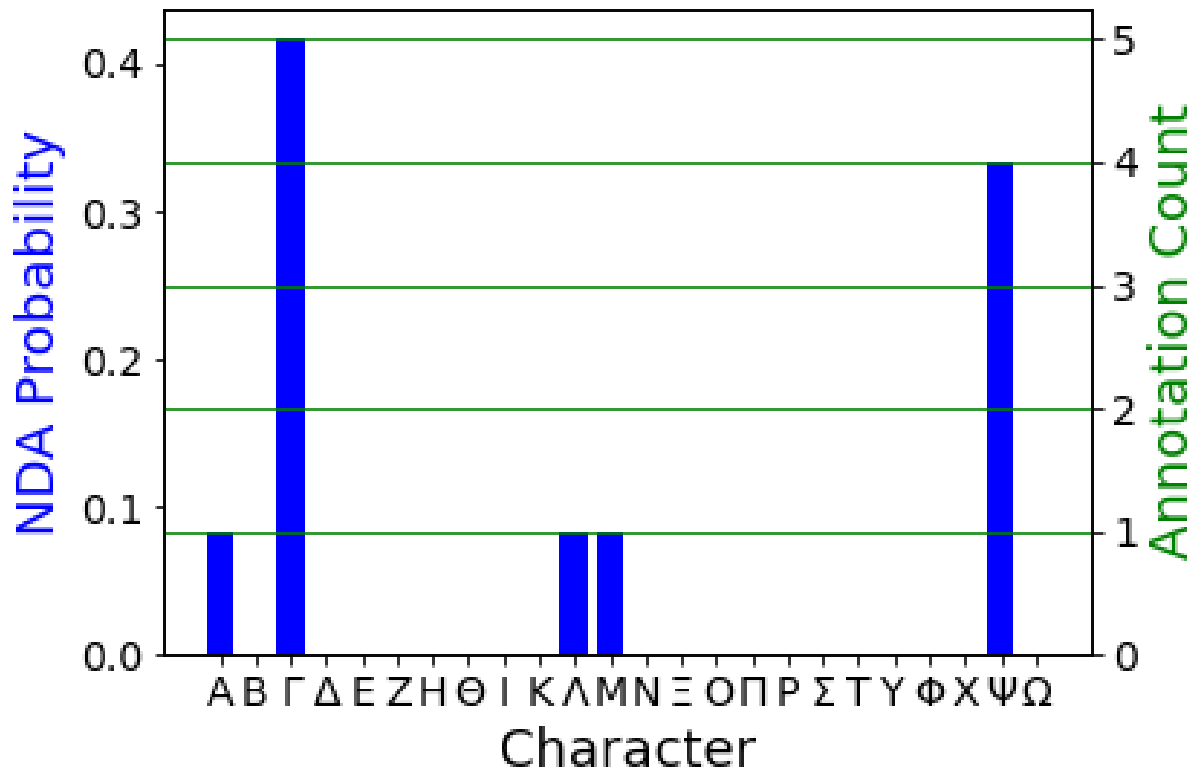


Figure 19: Normalized Distribution of Annotations (NDA) for the image Shown in Figure 20.

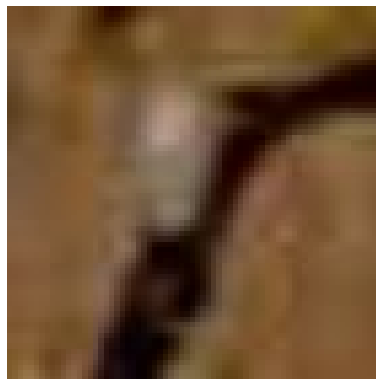


Figure 20: Example of the character Gamma ($\Gamma\gamma$) from AL-PUB. Though the consensus label for this character is Gamma, the ground-truth is unclear due to the degradation exhibited in the source manuscript.

5.2 Methods

In this section, we discuss the various methods used in this study. We describe the architecture of the ResNets, their labeling schemes, and their loss functions. We then describe our approach for performing stacked generalization to combine the ResNets’ outputs to create input features for a k -nearest neighbors model. We discuss the evaluation metrics which were used to monitor the training progress of the ResNets. Finally, we discuss the use of Shannon entropy as a means to measure classification uncertainty.

An important semantic distinction must be addressed before further discussion. As the ground-truth can often be ambiguous for crowdsourced datasets, it is necessary to assume that the labels (derived from consensus of human annotations) are the ground-truth. Thus, when we indicate that a particular image was “correctly classified” by a given model, we are stating that the model prediction agrees with the annotation consensus. Conversely, when we state an image is “incorrectly classified,” we are indicating that the model disagrees with the consensus.

5.2.1 Deep Learning Architecture

The core of the proposed ensemble model consists of two, nearly identical architectures. The ResNet, outlined in Figure 7 and based on the model discussed in Sections 2 and 3, is a fairly standard convolutional residual network, As in [38], The early layers in the model consist of two convolution layers with ReLU activation, followed by a max-pooling layer and 16 residual blocks. The residual blocks consist of two 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 an output layer. These models differ only in the class labels and loss functions. The architecture is based on the models utilized in [81]. For both models, the input features are identical: the raw images. The first model, CXE-ResNet, takes standard, single value numerical class labels while the second model, KLD-ResNet, takes the NDA from the Ancient Lives Project data as labels.

		KLD	
		Correct	Incorrect
CXE	Correct	353549	16407
	Incorrect	18594	10871

Table 9: Number of images correctly/incorrectly classified by the CXE- and KLD- ResNets. Both models agreed with consensus annotation for 353,549 images, and both disagreed for 10,871 images.

The ResNets, written in Tensorflow 2.3 with GPU acceleration, were trained on a Linux system consisting of 2 Intel Xeon E5-2687W Processors and 2 NVIDIA GeForce RTX 2080 Ti CUDA cores. In the interest of limiting the differences between the two ResNets, the same randomization seed was used for both models, though this likely had minimal effect on training or inference. Recent work, such as [59], suggests that GPU non-determinism dominates the effects of seed randomization.

5.2.1.1 Loss functions

Our ResNets used two different loss functions, respectively: sparse categorical cross-entropy (CXE) and Kullback-Leibler divergence (KLD). Both of these functions are based on the Shannon entropy H of a probability distribution p [18] (see Equation 2).

$$H(p) = \mathbf{E}[-\log(p)] = - \sum_{i=1}^M p_i \log(p_i) \quad (2)$$

Here, the distribution $p = (p_1, \dots, p_M)$ is defined over M classes. The entropy function (which takes the entire probability distribution as input) returns the amount of “uncertainty” present in the distribution. Thus, for distributions which are very narrowly peaked, the entropy will be low, while for distributions which are more spread out and uniform, the entropy will be high. The full range of values of the Shannon entropy is from 0 to $\log(M)$.

The first of our loss functions—sparse categorical cross entropy—can be defined in terms of

the Shannon entropy of p with respect to q as in Equation 3

$$H_q(p) = \mathbf{E}_q[-\log(p_i)] = - \sum_{i=1}^M q_i \log(p_i) \quad (3)$$

where p is the model’s output distribution and q the “target” distribution. Since we are using *sparse* CXE, this means that the q_i ’s are effectively one-hot encoded, where the only $q_i = 1$ is that of the consensus label. In order to calculate the CXE for the entire data set, we simply perform the sum in Equation 4

$$\text{CXE} = \sum_{j=1}^N -\log(p_{consensus}^j) \quad (4)$$

where $p_{consensus}^j$ is the model output distribution’s probability of the consensus class for image j (summed over N images). Notice that all references to the q_i are gone since the effective one-hot encoding renders them irrelevant.

The second loss function—Kullback-Leibler divergence [72]—is defined in Equation 5.

$$D_{KL}(p||q) = H_q(p) - H(p) \quad (5)$$

This function is simply the difference between the model output, p , with respect to the target distribution, q , and the Shannon entropy of the model output, p . A major difference between this function and the sparse CXE is that the target distribution is no longer one-hot encoded. Instead, we use the NDA. Also, as with the CXE, one calculates the KLD for the entire dataset by summing $D_{KL}(p||q)$ for the individual images.

The main motivation for using the NDA instead of one-hot encoding on the consensus is to be able to utilize more information in the dataset which was being discarded by the one-hot encoding. Although this does introduce annotator disagreement into the loss function, it also introduces information about the similarity of different characters. This will allow for more avenues for future analysis on character misclassifications which are due to the characters having similar shapes. Furthermore, as will be shown in the next section, adoption of the NDA also causes the KLD-ResNet

outputs to tend to have higher entropy than those of the CXE-ResNet.

5.2.2 Ensemble Model: Stacked Generalization and K-Nearest Neighbors

As will be shown below, though both ResNets performed well, there is a sizable subset of the data which one network classified correctly while the other was incorrect. Because of this disagreement between the ResNets, it is possible to construct an ensemble model which uses the two models’ outputs to achieve an accuracy greater than that of the two individual networks, as informed by [32]. We employ the technique of “stacked generalization”, outlined in [97], whereby the output distributions of the two ResNets are used as inputs to a new model. Using two ResNets with different loss functions decreases the amount of correlation between the models, giving us a more robust ensemble model. We use a k -nearest neighbors model (hereafter KNN) with $k = 50$. We create the inputs for the KNN in the following manner. Let C_j^i, K_j^i be the probability (taken from the CXE- and KLD-ResNet outputs, respectively) that the i -th image belongs to the j -th class. We can then create a new set of input data X for the KNN model via Equation 6.

$$\begin{aligned} X_{1,\dots,M}^{1,\dots,N} &= C_{1,\dots,M}^{1,\dots,N} \\ X_{M+1,\dots,2M}^{1,\dots,N} &= K_{1,\dots,M}^{1,\dots,N} \end{aligned} \tag{6}$$

Note that the lower index of X takes on $2M$ values since there are two models each with $M = 24$ classes whose output probabilities must be incorporated.

Similar to the ResNets, the KNN also produces an output distribution of classification probabilities which is based on the fraction of neighbors of different classes. This allows us to perform the same entropic analyses on all three models.

5.2.3 Evaluation Metrics

While the CXE-ResNet utilizes the standard accuracy metric for evaluation, the approach used by the KLD-ResNet is atypical of deep learning approaches. This is by nature of the labeling scheme used. Thus, Mean-Absolute-Error (MAE) was utilized as the primary metric. This posed

a challenge when comparing the results from training the two models. The most straight-forward way of comparing accuracy for our models is to run inference on the entire dataset, including both training and validation data, and calculating the accuracy of model predictions. For consistency, we continue this approach in analysis of the ensemble model.

5.2.4 Measuring Classification Uncertainty with Shannon Entropy

Techniques exist to minimize and measure classification uncertainty using varied approaches. As [61] explains, combining multiple annotations can often reduce misclassifications. Other approaches, such as [39], seek to utilize the magnitude of Softmax maxima as a misclassification indicator. Yet another approach, in [34], sought to make use of Softmax distributions which yielded a 4% decrease in misclassifications. We take an approach that has much in common with past efforts and utilizes the Shannon entropy of the ResNets' output probabilities, as in [96]. This approach is well suited for our work because of the probabilistic nature of the NDA, which is atypical of labeling schemes, allowing us to compare the uncertainties of the human annotations and ResNet output distributions.

5.3 Numerical Experiments

We now move on to discuss the different numerical experiments performed on the data. We will begin with a discussion of the experimental setup and then proceed to compare the results of the different models tested. We will also provide some analysis of these results based on the Shannon entropy of the models' output distribution. For brevity, we only share here a limited number of plots consisting of the results for the entire dataset and results for the single character Alpha ($A\alpha$). Hundreds of plots were produced showing results for all classes. These plots, the ResNet models, and a demonstrational Jupyter notebook are available in an online repository available at <https://github.com/mis2n/JDMDH-NDA-Results> .

5.3.1 Experimental setup

For measuring the accuracy of our models, we make use of the precision and recall statistics, given in Equation 7

$$\begin{aligned} \text{Precision} &= TP / (TP + FP) \\ \text{Recall} &= TP / (TP + FN) \end{aligned} \tag{7}$$

where TP is the number of true positives, FP is the number of true negatives, and FN is the number of false negatives.

5.3.2 Results

Table 9 shows the number of correct and incorrect classifications made by the CXE- and KLD-ResNet models. While both models correctly classify approximately 93% of the images (see below for a more thorough look at model accuracy), the off-diagonal elements in the table show that there is a sizeable fraction of images which one model classified correctly while the other classified incorrectly. As stated earlier, stacked generalization can provide a boost in accuracy in such a case, where there is enough disagreement between two well-performing models that the best of both can be incorporated.

Table 10 and Figure 22 contain the precision and recall for the CXE-ResNets, KLD-ResNets, and KNN ensemble models per character and for the entire data set. Overall, the models achieved accuracies of 92.6%, 93.2%, and 95.6%, respectively.

Note that for the majority of characters and models, the precision and recall are >0.9 , with the KNN tending to have the highest values. Cases with lower values for these statistics tend to be characters with fewer sample images (see Figure 23). The most notable of these would be Sigma ($\Sigma\sigma$), for which there are only 62 samples (compared to $>42,000$ for Alpha ($A\alpha$)). While the two ResNets had precisions of 0.500 and 0.857 for Sigma ($\Sigma\sigma$), respectively, the recalls were <0.1 for both. Even more significant is the fact that the KNN achieved both precision and recall of 0.0.

	CXE		KLD		KNN	
	pre.	rec.	pre.	rec.	pre.	rec.
Total	0.928	0.926	0.932	0.932	0.956	0.956
<i>Aα</i>	0.935	0.932	0.916	0.956	0.954	0.962
<i>Bβ</i>	0.957	0.812	0.928	0.888	0.952	0.927
<i>Γγ</i>	0.910	0.764	0.951	0.767	0.942	0.841
<i>Δδ</i>	0.928	0.936	0.955	0.900	0.958	0.950
<i>Eε</i>	0.964	0.931	0.943	0.950	0.964	0.968
<i>Zζ</i>	0.957	0.733	0.953	0.787	0.957	0.872
<i>Hη</i>	0.918	0.934	0.962	0.913	0.956	0.958
<i>Θθ</i>	0.936	0.894	0.952	0.866	0.958	0.934
<i>Iι</i>	0.891	0.890	0.910	0.868	0.924	0.927
<i>Kκ</i>	0.959	0.951	0.965	0.958	0.978	0.971
<i>Λλ</i>	0.917	0.841	0.940	0.834	0.943	0.913
<i>Mμ</i>	0.887	0.939	0.939	0.918	0.953	0.951
<i>Nν</i>	0.938	0.972	0.951	0.975	0.974	0.978
<i>Ξξ</i>	0.926	0.796	0.842	0.866	0.909	0.891
<i>Oο</i>	0.939	0.952	0.925	0.952	0.956	0.969
<i>Ππ</i>	0.770	0.956	0.943	0.895	0.942	0.947
<i>Pρ</i>	0.953	0.902	0.884	0.946	0.955	0.953
<i>Σσ</i>	0.500	0.016	0.857	0.097	0.000	0.000
<i>Tτ</i>	0.959	0.897	0.917	0.954	0.957	0.958
<i>Υυ</i>	0.915	0.911	0.900	0.922	0.930	0.949
<i>Φφ</i>	0.960	0.925	0.962	0.931	0.972	0.967
<i>Xχ</i>	0.952	0.962	0.955	0.951	0.972	0.974
<i>Ψψ</i>	0.916	0.820	0.930	0.740	0.938	0.855
<i>Ωω</i>	0.919	0.966	0.937	0.974	0.972	0.977

Table 10: Overall and per-character precision and recall (with respect to the human consensus) for the CXE, KLD, and KNN ensemble models.

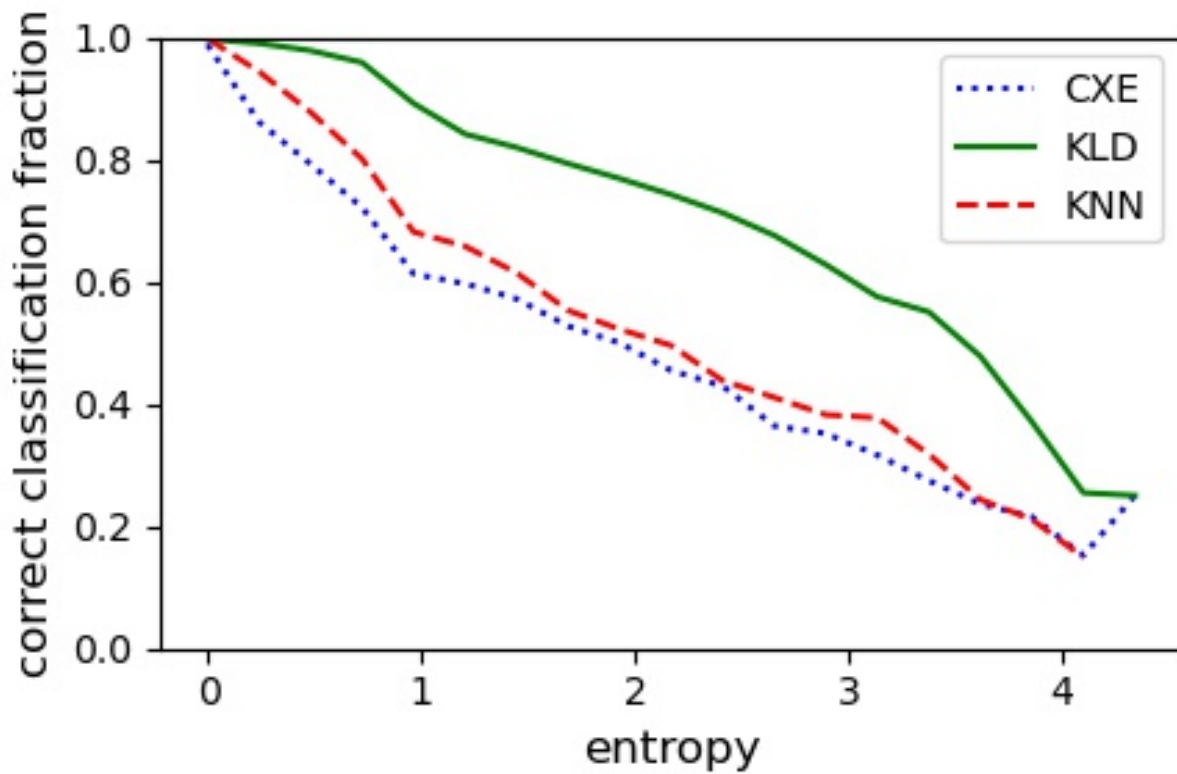


Figure 21: Plots of the fraction of correct classifications versus entropy of the output distribution for each model. Note that the KLD model is **NOT** better than the other models. Rather, it has consistently higher entropy values than the other models.

This is reasonable since the number of neighbors (50) is near the number of samples (62), making the chance of a majority of Sigma ($\Sigma\sigma$) neighbors quite low.

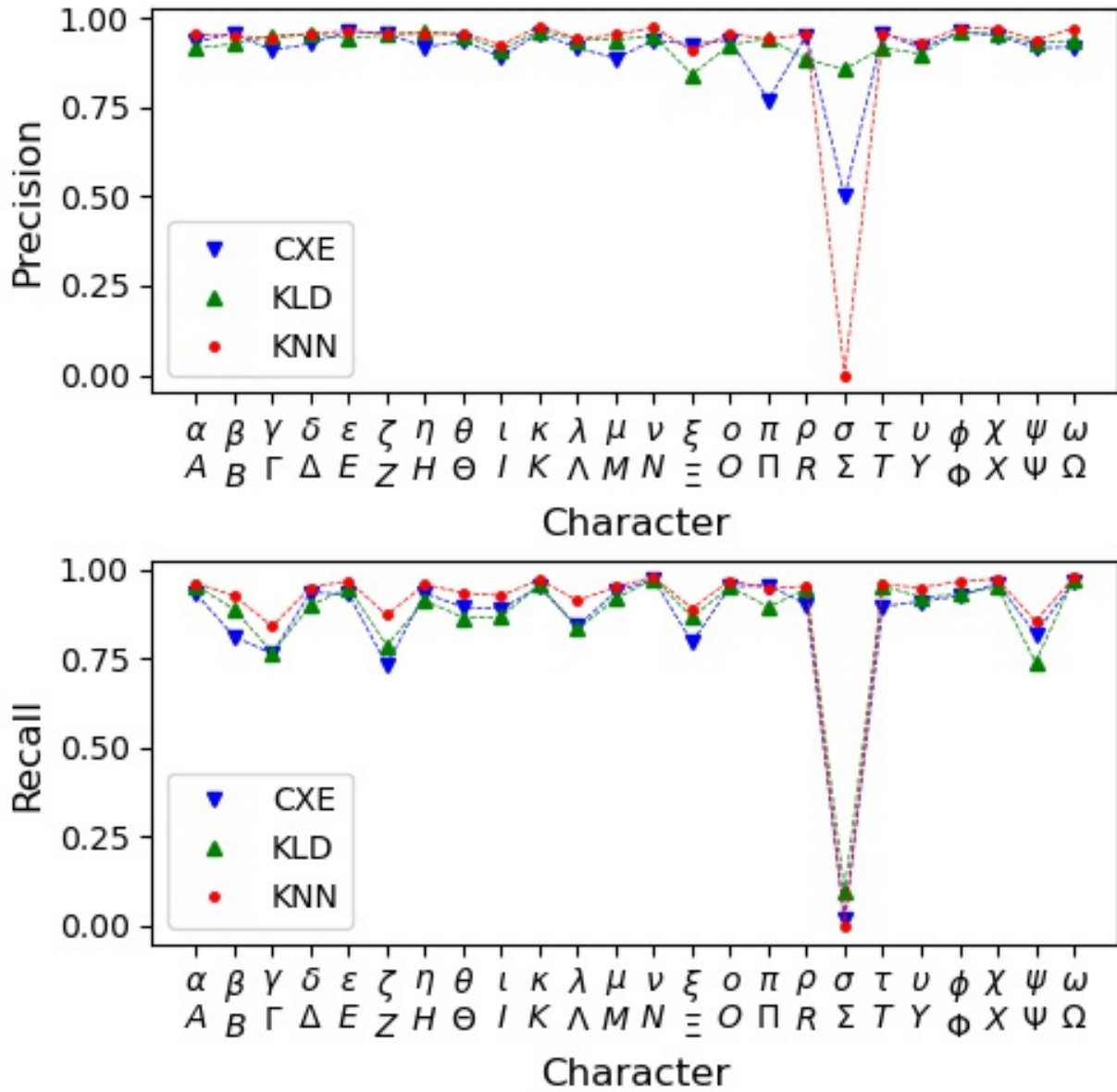


Figure 22: Per character precision and recall for each of the three models. Note that the KNN ensemble model is consistently higher than the individual models for most characters.

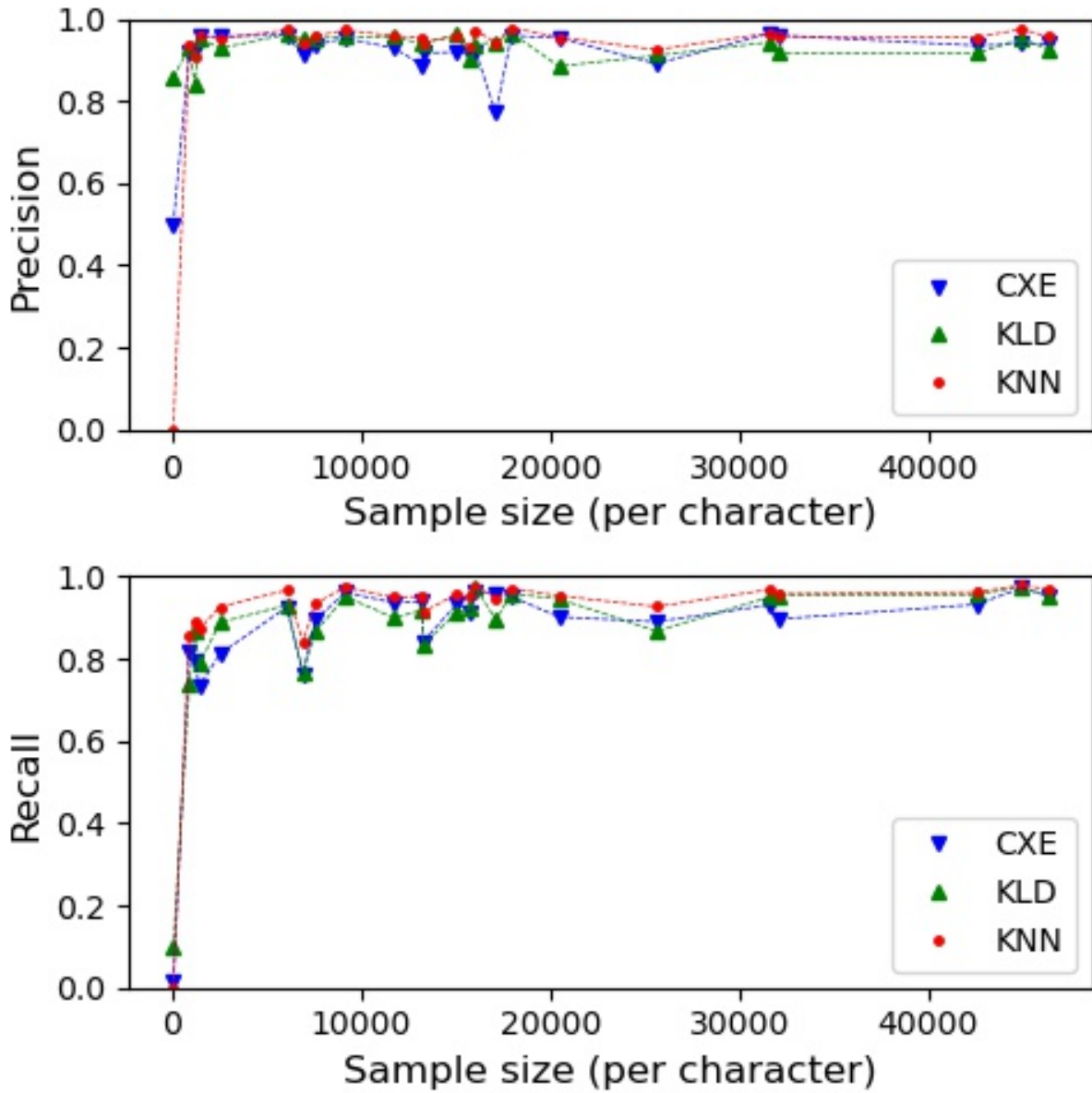


Figure 23: Precision and recall of all three models versus per character sample size. Notice the drop in recall for characters with smaller sample sizes

5.3.3 Entropy Analysis

We also performed entropic analyses on the NDA, ResNets, and ensemble models' output distributions. Figure 24 shows histograms for the entropy of each of these distributions for both the entire data set and for just the character Alpha. The plots are split into two, with one plot containing only those images which were correctly classified by the respective model and the other plot containing those which were misclassified.

There are several interesting points revealed by these plots. First—and perhaps most readily noticed—is the fact that the majority of images have a low entropy in their respective output distributions, regardless of which model was used. For the NDA and correctly classified model histograms, the first bin has the greatest number of images by far. This means that for most images which are classified correctly, the model tends to be quite certain in its classification. For the misclassified model histograms, however, this is not the case (with the exception that the CXE-ResNet histograms in the first bin is approximately equal to the fourth). In fact, the tails of these histograms tend to contain the majority of samples. This means that for images which are classified incorrectly, the models tended to have less certainty in classification. Because of this discrepancy between the entropies of correct and incorrect classifications, entropy can serve not only as a good measure of classification uncertainty, but also as a predictor of classification accuracy (we discuss this further below). Images which have lower entropy will tend to be classified correctly while images with higher entropy will tend to be classified incorrectly. Figure 21 shows the fraction of correctly classified images versus entropy for the three models, illustrating the previous point well. Note, however, that the plot should not be interpreted as the KLD-ResNet significantly outperforming the other models. Rather, by nature of its labeling scheme, its output distributions have a high entropy on average.

Second, note the difference in Figure 24 between the three models' entropy profiles. The KLD-ResNet tends to have higher entropy values than both the CXE-ResNet and ensemble model. The discrepancy in entropy between the CXE-ResNet and KLD-ResNet is reasonable, since the CXE-ResNet only took into account the consensus label while the KLD-ResNet used the entire NDA.

Additionally, it is reasonable for the ensemble entropies to be quite small since a KNN model's output distribution is merely derived from the fractions of neighbors of each class. Due to the clustering of the CXE-ResNet and KLD-ResNet output distributions (the input features for the ensemble model), the majority of samples of a given character are surrounded in feature space by other samples of the same character, inducing a low entropy in the KNN's output distribution.

Figure 25 shows two-dimensional histograms of the entropy of the models' output distributions for each image against the number of human annotations each image received. The most notable feature of these plots is related to the difference in entropy profiles between the correctly classified and misclassified images. For the correctly classified images, we see that the range and mean of entropy values decreases as the annotation count increases. On the other hand, for the misclassified images, we see that the range and mean of entropy values does not tend to decrease (or even vary strongly) as annotator count increases. The CXE and KNN plots appear to show a very slight instance of this trend while the KLD plot seems to show the opposite (with the mean entropy increasing as annotation count increases).

These trends are reasonable when considered in conjunction with the law of large numbers. One might imagine in the limit of an infinite number of annotations, each image would have an NDA which converges to some distribution, say, P . Consequently, the entropy $H(P)$ would also converge. Also, as discussed above, images which are classified correctly tend to have low entropies. Therefore, as the annotation count for a given image (which would be correctly classified) increases, the entropy will tend to converge to a small value (though likely non-zero). Concerning plots of the misclassified data, we do not see a consistent trend as we did with the correctly classified data. For the KLD-ResNet plots, we see that the entropy tends to take on higher values at higher annotation counts. For the CXE-ResNet and KNN plots, it is difficult to discern a clear trend.

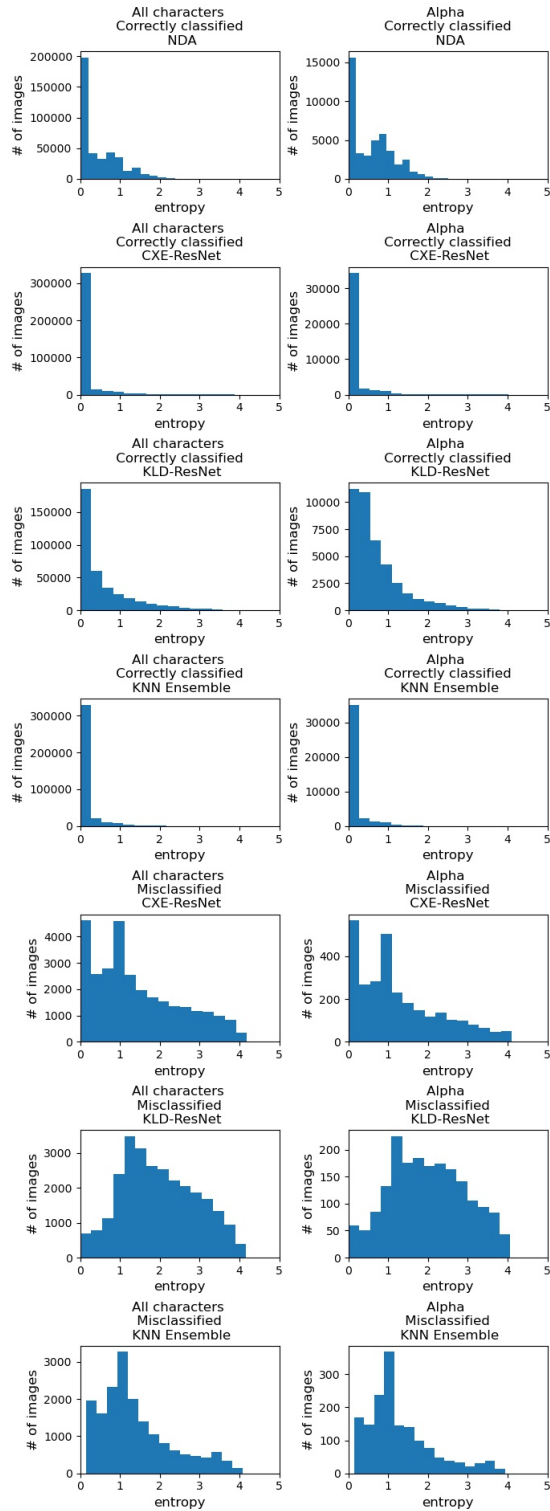


Figure 24: Histogram of entropies for the NDAs and the three models' output distributions for both the entire data set and solely the character Alpha ($A\alpha$). We split the histograms into two sets: 1) images whose model classification agreed with the human consensus and 2) those which disagreed.

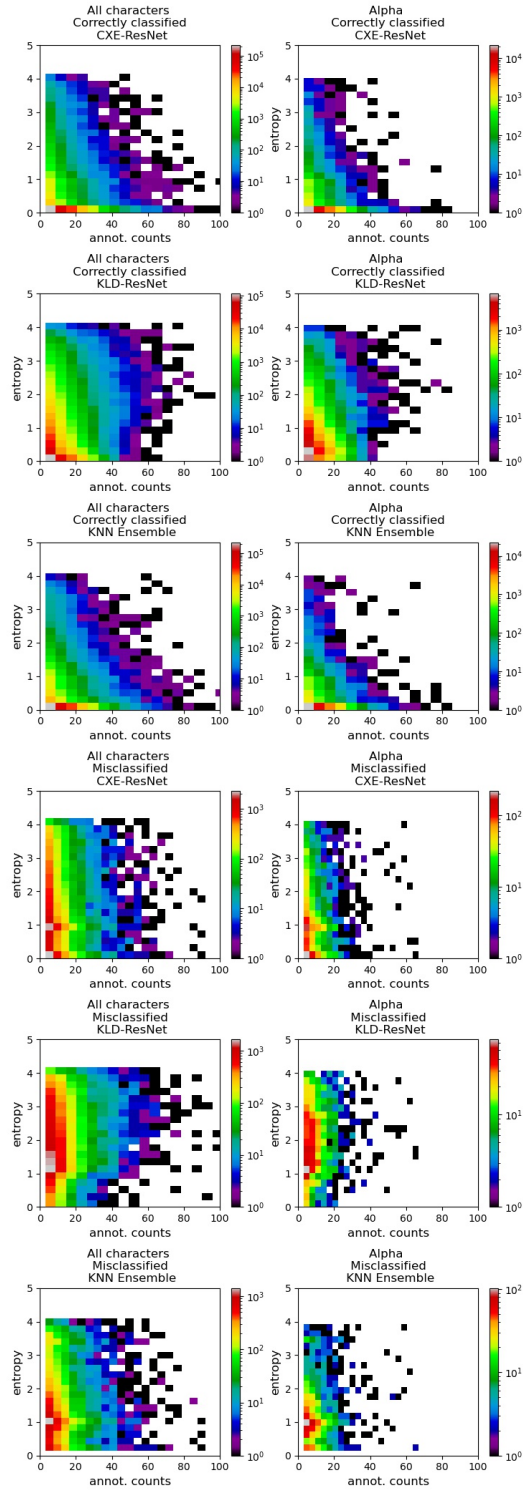


Figure 25: Two-dimensional histograms of the three models' output entropies versus the number of human annotations. The left column contains all characters and the right column contains only Alpha (A_α). We further split the plots into two portions: 1) images whose model classification agreed with the human consensus and 2) those which disagreed. Notice how the mean entropy value tends to decrease for large annotator counts for the correctly classified images.

5.3.3.1 Entropy as a predictor for classification accuracy

As mentioned above, the entropy of a given image's output distribution (regardless of which model is used) is effective for predicting whether that image will be classified correctly by the model. To illustrate this, we trained a Support Vector Machine (SVM) to take the entropy of a model's output distribution for a given image as a single input feature and predict whether the model which produced this distribution would correctly classify the respective image. Thus, there are only two classes: 1) images correctly classified by the model or 2) images misclassified by the model.

We trained a separate SVM for each character, since each character has a different entropy profile and a different number of samples. We also used an 80/20 train/test split for each character. Additionally, due to the massive imbalance in the number of correctly and incorrectly classified characters, we balanced these data by choosing an equal number of correctly and incorrectly classified samples. Figure 26 depicts the results of this experiment, providing precision and recall statistics for both classes for all models and characters. It should be noted that it is unnecessary to display the precision *and* recall for both classes of a classification problem since there is a dependency relation between these statistics, but both are displayed for clarity.

For the CXE-ResNet, precision and recall values for both classes and all characters (other than Sigma) were between 0.486 and 1.000. For the KLD-ResNet, they were between 0.523 and 0.937. For the KNN, they were between 0.636 and 0.975. It should be noted that these SVMs tend to be biased toward predicting that the ResNets and ensemble models will correctly classify images. When examining their confusion matrices, there was a consistently larger number of false positives (where the SVM incorrectly predicted that the ResNets or ensemble models' classification would be correct) than false negatives. Also, it can be seen from Figure 26 that there is a consistent trend across the different characters where both the recall of the incorrectly classified images (bottom left) and the precision of the correctly classified images (top right) are lower than their converse values.

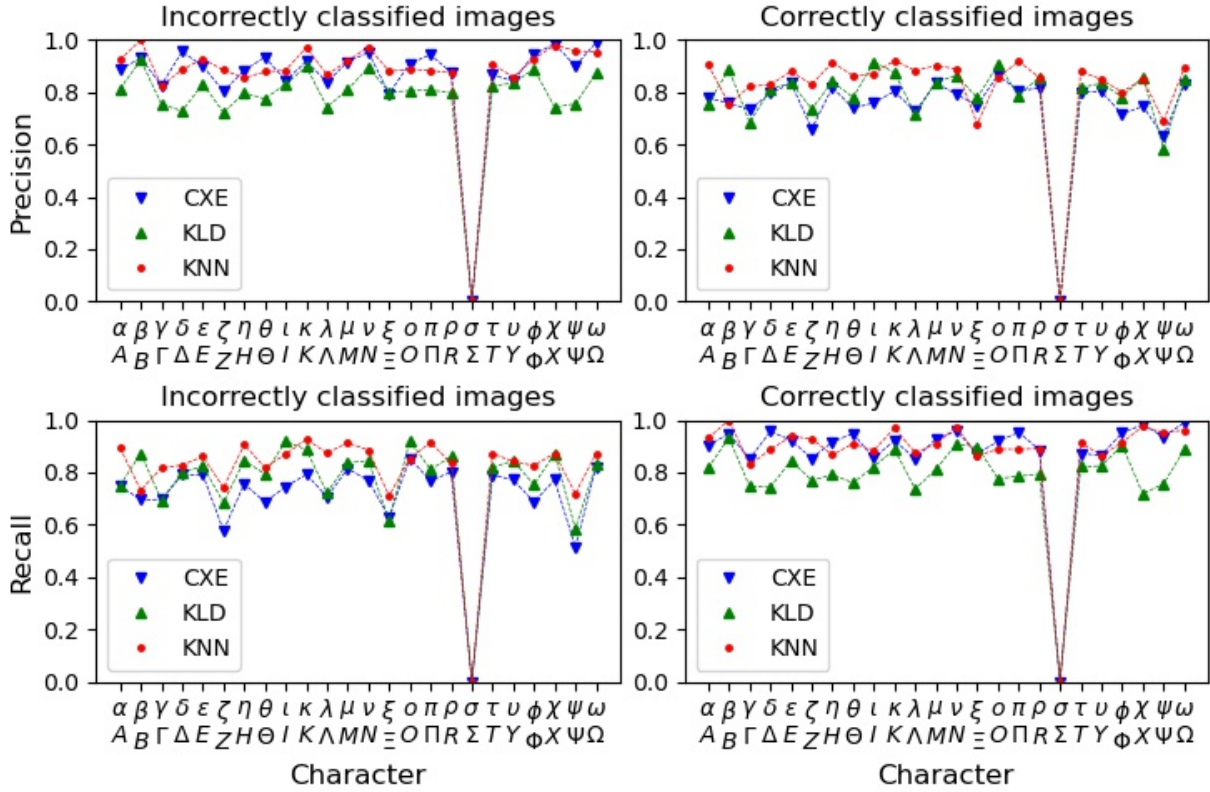


Figure 26: Per character precision and recall of the SVM model which predicts model classification accuracy based on the entropy of Softmax outputs.

5.4 Summary

In this section, we have demonstrated that the multi-valued NDA labeling scheme used in our KLD model performs comparably to the standard, single-value labeling used in the CXE model. The CXE and KLD models largely agree with one another, but not always. This ensures that the ensemble model can utilize the differences between the component models effectively. We have also shown that by performing stacked generalization, with our pair of well-performing ResNet models, we are able to increase our prediction accuracy to a level beyond that achieved by either of the original models. We have also shown that the Shannon entropy of our models' output distributions is both an excellent measure of classification uncertainty and a predictor of classification accuracy for our data, with misclassified images tending to have a higher entropy than those correctly classified. Although our techniques have only been applied to a single crowdsourced dataset (AL-ALL),

we believe it may be a useful approach for other datasets to reduce the effects of ground-truth uncertainty in labeling. The amalgamation of these results yield models with improved classification trustworthiness, as well as the means to quantify the trustworthiness of individual classification predictions based on the annotation distributions (NDA).

In the future, this work may allow for the removal of ambiguous and mislabeled images from the AL-ALL dataset (as well as other crowdsourced datasets) by manually inspecting images whose NDA shows high entropy. There are a significant number of images in the dataset which have both a high NDA entropy and a high annotation count. Images such as these are often unintelligible, so removing them from the dataset may be beneficial when training models. Conversely, there is a possibility that such noise may be useful in avoiding model-overfitting, similar to the effects of image augmentation. Also, there are many images which are clearly intelligible but have been misclassified by the human annotators. These may be identified by using the ensemble model's results as the ground-truth which the human annotators must predict. Images where the humans disagree with the model when the model is clearly correct (determined by visual inspection of the image) can either be corrected or removed.

6 Smart Digital Edition Management: A Blockchain Framework for Papyrology⁸

6.1 Introduction

6.1.1 Ancient Texts and Papyrology

As discussed in previous sections, the study and editing of ancient Greek papyri is a challenging discipline. Aside from the damaged nature of these ancient texts, editing ancient Greek papyri involves additional challenges. These ancient texts were written without word division and often without any punctuation. Ancient Greek, specifically manuscripts written before the medieval period, also lacks the modern notion of upper and lower case, thus utilizing one constant set of characters. Sentences are not always simple to decipher, even for experts, and the text is challenging to reconstruct, as demonstrated in Fig. 28. The difficulty increases further when a papyrus fragment exhibits considerable damage. Editing and reconstructing such texts is a meticulous process that requires intensive peer review before publication. Even when published, scholars will not only continue to debate these editions, but also publish different or competing reconstructions.

⁸Modified from the original article published in The ACM SIGMIS Computers and People Research Conference, 2024. Authored by Matthew I. Swindall, Kritagya Upadhyay, James H. Brusuelas, Graham West, John F. Wallin

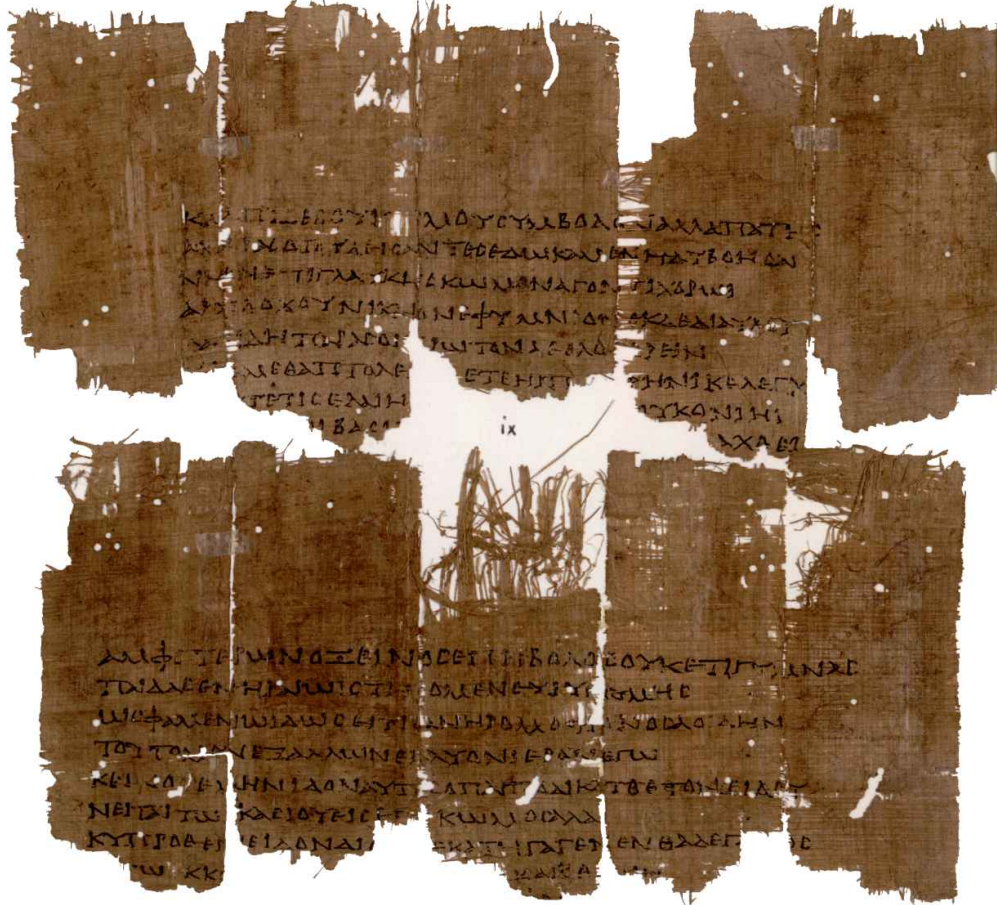


Figure 27: Papyrus Fragment from the Oxyrhynchus Collection. Note the extensive damage. Such fragments are exceedingly difficult to edit. In this fragment we see missing papyrus segments as well as missing, torn, and faded ink.

Original Sentence	I baked a delicious cake.
No Spaces between words	lbakedadeliciouscake.
No Letter Cases or Punctuation	IBAKEDADELICIOUSCAKE
Degraded Characters	_BAKEDAD_LI_____AKE
Unknown Number of Missing Characters	BAKEDAD LI AKE

Figure 28: Due to damage and the unique features of ancient Greek papyri, even the simplest sentence can become challenging to decipher. The phrase, “I baked a delicious cake” becomes unrecognizable under conditions common for ancient Greek papyri.

6.1.2 Current Practices in Editing and Peer Review for Digital Papyrology

When papyrologists use current editing platforms to complete and submit their edition of a manuscript for peer review, it is important during both the editing and the review phase to compare the transcription and reconstruction with other published papyri. This ensures that the grammar and syntax conjectured by the editor suits the standards and idiom of ancient Greek. While one can compare editions of manuscripts in print volumes by literary stacking or placing books next to each other, digital platforms for editing and publishing papyri typically only provide a single edition of a text that, in many cases, may have different scholarly versions and thus different conjectured readings. Furthermore, even for the edition of one text, current systems available online do not provide all the components vital to the production of a complete born digital critical edition that can be shared and cited. If any of those editions are the work of multiple editors, the specific contribution of each individual is also not always clear. These digital platforms and their editions are indeed invaluable as they provide searchable texts that have become essential to the editing and peer review workflows, yet their digital editions only supersede print editions due to search alone. They do not offer all the data that researchers, editors and reviewers need. Moreover, although peer review is sometimes integrated into these systems, improvements for transparency and documentation can be made. This process, known as sub-editing, in which all parties work to further improve the reconstruction of the text, still often takes place outside the system, usually via email. The absence of such complete systems leads to a lack of efficacy for digital editions. They offer invaluable search functionality, yet print publications are still given precedence in citation.

To mitigate these challenges, a system is needed which not only allows for the efficient production of complete born digital critical editions, but also, and more importantly, three unique attributes:

1. Intuitive access to multiple edited versions of one manuscript.
2. a clear representation of multiple editors involved in the edition of one manuscript
3. and a transparent digital peer review system that allows reviewers to engage with a contrib-

utor in a back and forth process based on a submitted edition

Currently, many efforts exist that seek to modernize the practice of papyrology, especially through A.I. and machine learning approaches [81, 80, 94]. Here, we seek to employ an additional pair of new technologies to advance digital editing and publishing platforms for papyrology: Blockchain [58] and Smart Contracts [44].

6.1.3 Papyrological Editions

The papyrological editing process is very complex. For papyri preserving new or unknown literary and sub-literary works, e.g. a lost poem of Sappho or anonymous geometrical texts, there are several components which are not necessarily present in every digital edition. The standard components include:

6.1.3.1 Diplomatic Transcription

A **Diplomatic Transcription** is the first step in editing. It is an exact transcription of what the papyrus preserves, even obvious errors.

6.1.3.2 Paleographical Apparatus

Diplomatic Transcriptions may be accompanied by a **Palaeographical Apparatus** which discusses difficult areas to read and visible ink traces.

6.1.3.3 Articulated Text

The **Articulated Text** is the next step in the process. Here the editor separates the string of characters into words, adds accents, and proposes possible reconstructions of gaps (missing text).

6.1.3.4 Critical Apparatus

The Articulated text is accompanied by a **Critical Apparatus** which contains other plausible reconstructions for certain lines or words in the articulated text. The first time a new or unknown

literary or sub-literary papyrus is published, it usually will not have a critical apparatus. But as that papyrus is re-edited by other scholars, who propose different reconstructions, it ultimately appears in publication.

6.1.3.5 Testimonia and Text Translation

Finally, the articulated text can also be accompanied by a **Testimonial Apparatus**, which references other ancient authors that are witnesses to this exact text (e.g. they quote or discuss the text preserved in the papyrus). In some instances, a **Text Translation** may also accompany the **Articulated Text** as well.

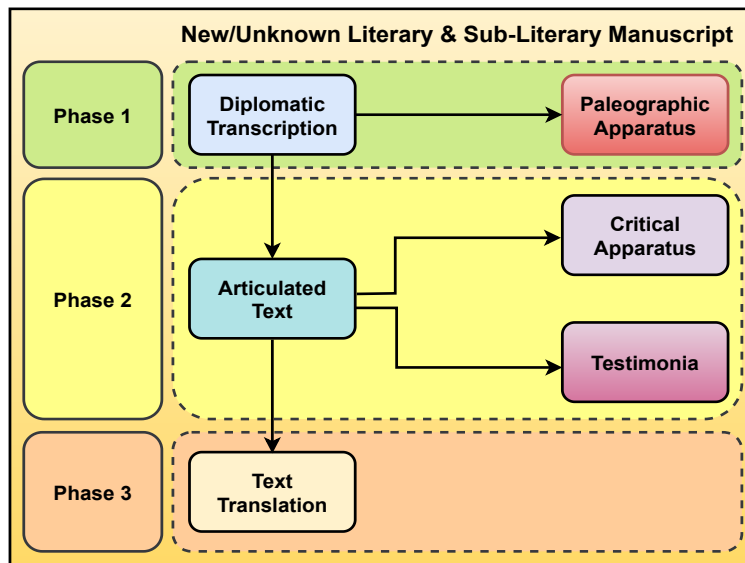


Figure 29: The basic standard for a digital edition of an ancient Greek papyrus is an articulated text with a critical apparatus. This is sufficient for many papyrus texts. Yet, for new or unknown literary and sub-literary works, more components are required for an edition to be considered critical. These include: Paleographic Apparatus, Critical Apparatus, Testimonia, and Text Translations.

6.1.4 Motivation

While systems exist that administer some components of digital edition management for papyrology, such as Papyri.info ⁹, there is currently no single system that can fully manage multiple

⁹<https://papyri.info/>

editions and edition types from multiple authors. For instance, most existing platforms do not include diplomatic transcriptions or a palaeographic apparatus. Additional weaknesses in existing systems include:

- Limited and Centralized Systems for Papyrology Management and Peer Review.
- Database Infrastructure Challenges in Papyrology Editing Complexity.
- Vulnerability to Data Loss in Traditional Database Systems.
- Lack Transparency and Speed in Papyrology Peer Review.

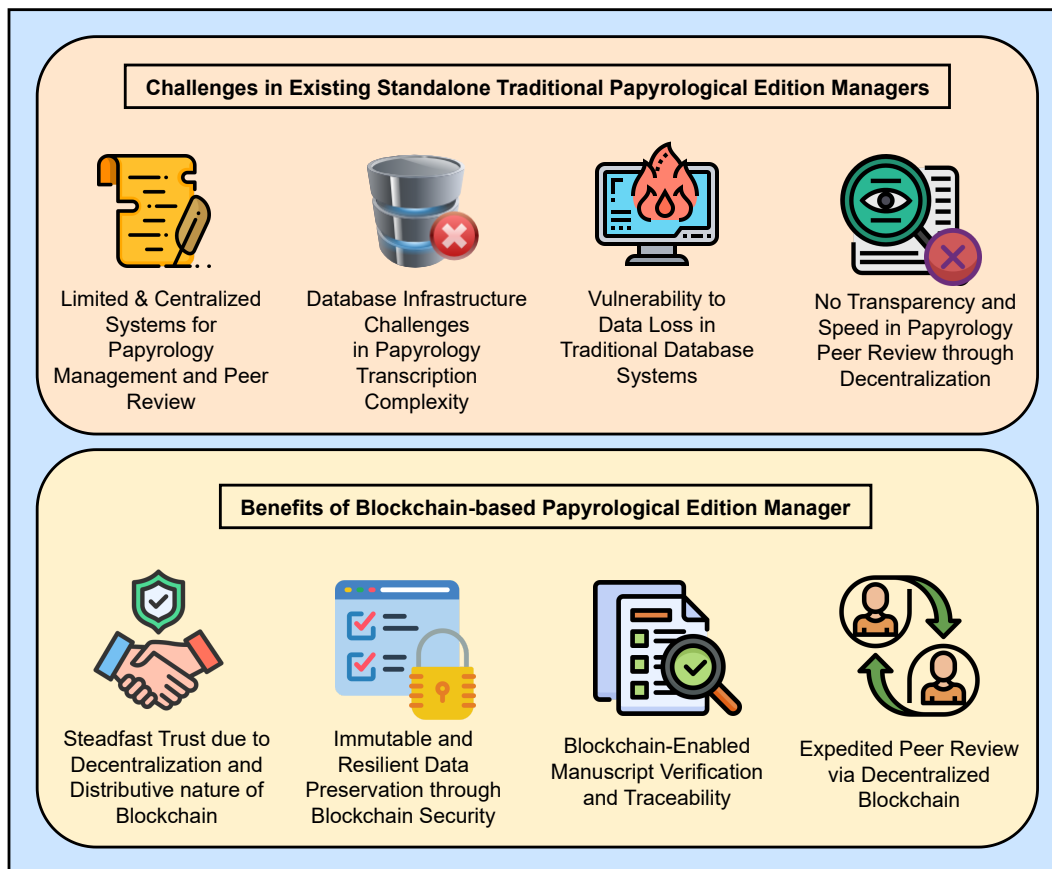


Figure 30: The few existing systems for managing editions of papyrological transcriptions utilize centralized, database architectures that offer little flexibility or transparency, and may be vulnerable to data loss. A blockchain and smart contract enabled system can offer better security, transparency and data provenance.

6.1.5 Why Blockchain?

Blockchain is a decentralized, distributed, digital ledger that allows for greater flexibility, transparency, security, and data provenance than database systems. Built on blockchain technology, smart contracts allow for automated management of data. One of the most prominent use cases for smart contracts are Non-Fungible Tokens (NFTs), which are used to buy and sell digital assets such as images or tickets to events [70]. Smart contracts can also be used to autonomously facilitate submission, review, and sharing of various editions by multiple authors linked to the original edition or manuscript. Fig. 30 outlines the challenges facing modern edition management, which is discussed in more detail in Section 6.2, as well as the benefits of our proposed system which includes:

- Steadfast Trust Due to Decentralization and the Distributive Nature of Blockchain.
- Immutable and Resilient Data Preservation Through Blockchain Security.
- Blockchain-Enabled Manuscript Verification and Traceability.
- Expedited Peer Review via Decentralized Blockchain.

6.1.6 Our Contributions

- We propose a novel blockchain and smart contract based architecture that can function as a back-end to a user interface for editing digital editions. This system can automatically manage storage of numerous related editions by multiple authors by storing data in decentralized systems, such as the Interplanetary File System (IPFS) [45], and store the data locations on the blockchain ledger.
- We test this system for a fundamental component of digital editions: the Diplomatic Transcription. Various metrics are used to demonstrate the feasibility of the proposed system.
- We explore the challenges and shortcomings of the proposed system and potential methods for improvement.

- We detail our vision for putting the proposed system into use which will aid in advancing the practice of papyrology.

6.1.7 Outline

The sections that follow will discuss the proposed system. Section 6.2 will explore existing literature related to the project. Section 6.3 details the methodology and experiments performed. In Section 6.4, the experimental results will be discussed. Section 6.5 will probe the limitation of the proposed system, as well as future work. In Section 6.6, we will conclude by highlighting the contributions of this work and discuss implications for further research.

6.2 Modern Digital Papyrology and Digital Edition Management

6.2.1 Literature Survey

The framework proposed here is novel, with no comparable work found in our literature review. Exploration of existing literature found no other peer reviewed work that explores the use of blockchain and smart contracts for managing digital editions of ancient manuscripts. Some related work was available, including works that discuss the use of blockchain for decentralizing the peer review process for publications [83, 88], the need for a framework which facilitates peer review of critical editions [95, 16, 1], and existing tools that have proven inadequate for such purposes¹⁰¹¹¹². Further literature was reviewed which explored the use of tokenization in a blockchain to bring transparency and ethics to the antiquities trade [90]. Other available literature discusses the use of the IOTA blockchain [7, 65], which allows for cost-less smart contracts that may be ideal for this use case and may be explored in future work.

¹⁰<https://www.homermultitext.org>

¹¹<http://www.catullusonline.org>

¹²<https://pubs.perseids.org>

Reference	[95]	[90]	[7]	[60]	[65]	[94]	[81]	[83]	[1]	[80]	[88]	Our Work
Year	2013	2021	2020	2019	2021	2021	2022	2021	2015	2014	2021	2024
Blockchain	X	✓	✓	✓	✓	X	X	✓	X	X	✓	✓
Smart Contract	X	X	✓	X	✓	X	X	✓	X	X	✓	✓
Peer Review	✓	X	X	X	X	X	X	✓	✓	X	✓	✓
Papyrology	✓	X	X	X	X	✓	✓	X	✓	✓	X	✓
Antiquities	✓	✓	X	X	X	✓	✓	X	✓	✓	X	✓
Digital Humanities	✓	X	X	X	X	✓	✓	X	✓	✓	X	✓
Distributed Storage	X	X	✓	✓	✓	X	X	✓	X	X	✓	✓
Digital Edition Management	✓	X	X	X	X	X	X	✓	✓	X	✓	✓

Table 11: Contribution Comparison: The body of existing works related to our proposed system is limited, as our approach appears to be truly novel, covering all necessary and relevant aspects.

6.2.2 Existing Systems

The principal system for creating and publishing digital editions of Greek papyri is Papyri.info. The platform consists of the Papyrological Editor, which contributors use to add digital editions to its database, and the Papyrological Navigator, which publishes those editions online. This platform has been vital to the papyrology community, since it provides a large database of searchable texts. However, it only facilitates the creation of an articulated text, an apparatus for notes and commentary, and a translation. The Papyrological Editor also has its own version-controlled environment for collaborative editing with its own method of peer review. For newly proposed reconstructions, these are recorded in the apparatus, but citing them requires referencing an online bulletin (PDF) instead of the digital edition itself, which ultimately drives scholars to cite print publications of the papyrus rather than the digital edition. Due to the dominance of Papyri.info, very few other systems have emerged. The New Testament Virtual Manuscript Room (<https://ntvmr.uni-muenster.de>) is one example of an alternative, which focuses on biblical manuscripts including papyri. Yet this platform too only allows for the creation of an articulated text and a critical apparatus. It also appears to be internally moderated by the scholars using it rather than any systematic method of peer review. The Proteus project [95, 16] was an attempt to advance the creation of born digital critical editions for Greek papyri by allowing users to create every component necessary; the diplomatic edition, palaeographical apparatus, articulated text, critical apparatus, testimonial apparatus, and translation. Proteus also modeled collaborative editing, but with the ultimate goal of producing a citable digital edition, and the ability to access multiple scholarly editions of one papyrus manuscript for comparison. For peer review, it also incorporates a system for reviewers and con-

tributors to communicate during the sub-editing process of one submitted edition. While Proteus advances many aspects of digital papyrology and born-digital critical editions, the application is not yet openly available.

6.3 Methodology

In our experiments, we explore the implementation of such a system back-end by creating a smart contract, deployed on a blockchain testnet and then submitting and retrieving numerous digital editions. In these experiments, 501 synthetically produced transcriptions and a single manuscript image are stored on IPFS and their file locations stored on the blockchain testnet via smart contract. Metrics measured during the process include encoding times, file transmission time, blockchain transaction times, and transaction costs.

6.3.1 System Architecture

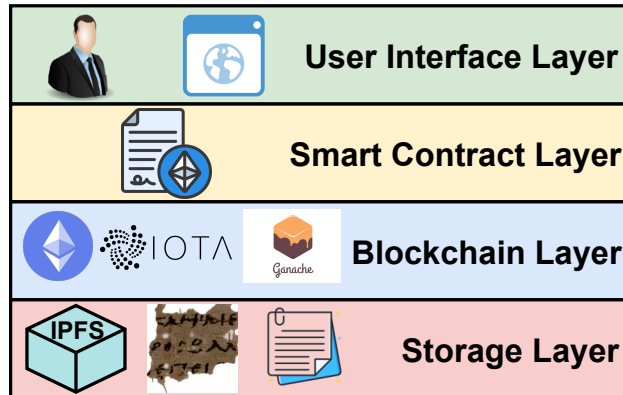


Figure 31: The architecture consists of 4 layers. The top layer, a user interface, enables the submission of digital editions. In the second layer, a smart contract automatically manages submission and retrieval of digital editions. The third layer consists of the blockchain itself, where the smart contract is deployed and where the location(s) of all associated editions are stored. The final layer is the storage layer, which allows for flexibility of storage options including cloud storage, local server, or decentralized storage options such as IPFS.

The architecture, illustrated in Fig. 31, consists of 4 core layers: a user interface layer where users receive and submit data, a smart contract layer that stores IPFS hash addresses of related data on the blockchain, a blockchain layer where the hashes are maintained, and a storage layer

where image and transcription data are stored on IPFS. An important distinction before further discussion is the term *original edition*. This is not a technical term or one used in papyrological practice, but rather a term to differentiate the first digital edition for a given manuscript submitted to the system. Papyrology does have a term to mark the very first publication of a fragment (*editio princeps*), but that term traditionally refers to a single author print publication, and we choose not to introduce any confusion here. Using the Proteus model, for example, if the first user submitted their digital edition that contained only an articulated text, they would be submitting an *original edition* (illustrated in Fig. 33). Once accepted and published online, other users might later submit additional components to that edition such as a *critical or testimonial apparatus* (illustrated in Fig. 34). As discussed in earlier sections, storing the various components of a critical edition is very complex. We also keep in mind that a user might want to submit an entirely new and different edition than the one previously published in the system. As such, for these experiments only a single component of a critical edition was used: *the Diplomatic Transcription*.

To facilitate blockchain and IPFS transactions, the Infura ¹³ and Moralis ¹⁴ Python APIs are used. Moralis allows for the transmission of data files to IPFS and returns a URL where the file can be accessed. Infura's Web3 API facilitates a connection to the Sepolia blockchain testnet where the smart contract is deployed and Moralis URLs are stored for retrieval. The Sepolia testnet is then accessible via the Remix IDE for Web3 Development ¹⁵. Remix allows for the editing and deployment of smart contracts on the Sepolia testnet. Additionally, Remix allows for transactions to be processed in the IDE, and transactions can be monitored, though all transactions for the experiment were carried out using the Python APIs.

6.3.2 User Interactions

The smart contract, discussed further in Section 6.3.5, tracks all editions and stores IPFS hash addresses to the data on the blockchain. An additional type of user that we envision is an AI

¹³<https://www.infura.io/>

¹⁴<https://moralis.io/>

¹⁵<https://remix.ethereum.org/>

agent. Just as a human researcher may submit original or critical editions, modern AI tools are maturing to the level of sophistication necessary to autonomously produce transcriptions for heavily damaged manuscripts [78]. For original editions, a smart contract with user-defined parameters (time limit, maximum number of critical editions, etc.) is deployed to the blockchain. The data file for the original edition is stored on IPFS. The IPFS address to the original edition is then stored on the blockchain by the smart contract. For expanded editions, the user searches for the desired content in the user interface. Upon selecting an original edition, the UI queries the smart contract, which retrieves the IPFS address of the desired edition from the blockchain. The user interface, through an API, queries IPFS by address to retrieve the digital edition data file. Figures 35, 36, and 37 are swim-lane diagrams that illustrate the process of submitting and reviewing editions. In Fig. 35 an original edition is submitted: storing the associated data file on IPFS and the IPFS address to the data file on the blockchain via the smart contract. In Fig. 36 we see the process of retrieving an original edition for review, and submitting an expanded edition. The IPFS address of the original edition is retrieved from the blockchain via smart contract. The user creates a new, expanded edition, which is then stored on IPFS. The IPFS address to the new, expanded edition is then stored on the blockchain by the smart contract. In Fig. 37 we see the process of retrieving all editions (original and expanded) associated with a given smart contract. The user requests the information from the UI, which queries the smart contract that then returns the IPFS addresses from the blockchain, for all editions. The data files for all editions are then retrieved from IPFS for review.

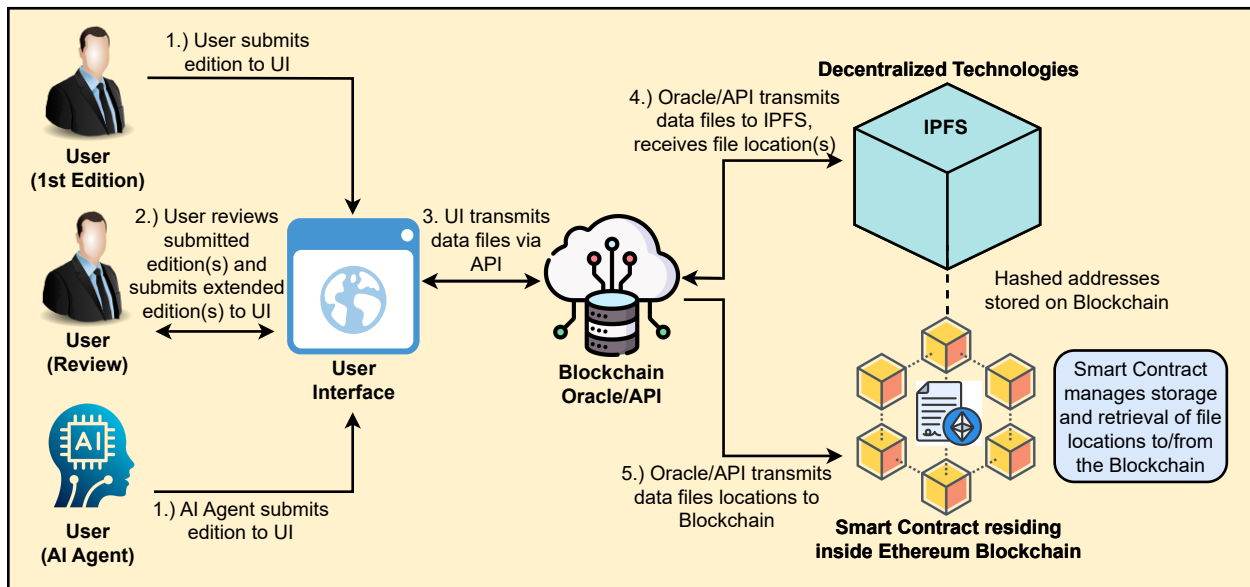


Figure 32: The proposed architecture includes a user interface and a smart contract deployed on the blockchain that tracks all editions and stores URL or IPFS hash addresses for the associated data on the blockchain.

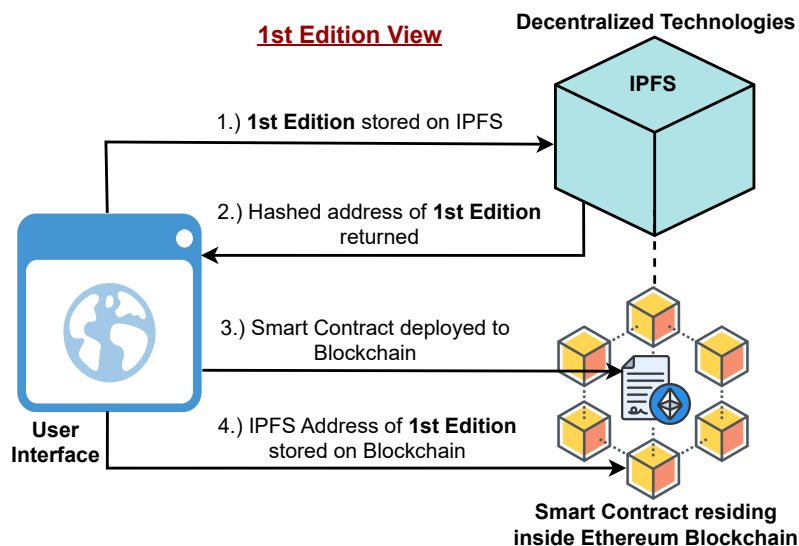


Figure 33: This diagram illustrates the process of submitting an original edition. In this example, the edition data is stored on IPFS while it's hashed address is stored on the blockchain. This information is available to other users for review and critique.

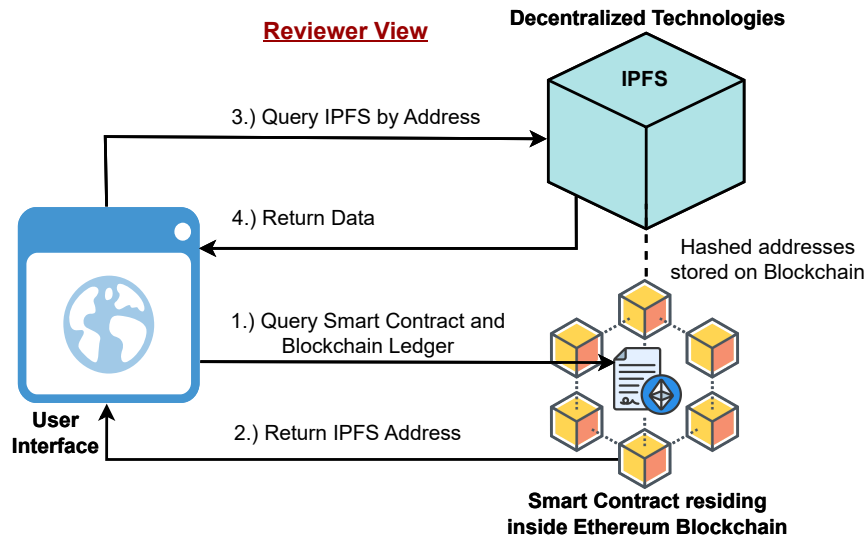


Figure 34: This diagram illustrates the process of submitting an expanded edition. The original edition is retrieved from IPFS using the URL that was stored on the blockchain. Expanded editions are then stored on IPFS and their URL stored on the blockchain.

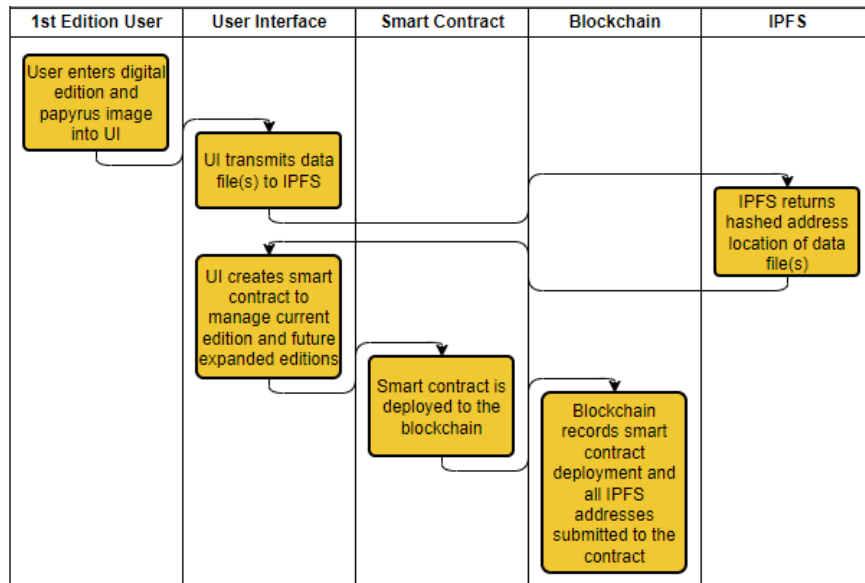


Figure 35: An illustration of the process of submitting an original edition. The user interacts only with the user interface. All blockchain transactions and data transmissions are handled in the background via APIs.



Figure 36: An illustration of the process for retrieving an original edition for review, then submitting an expanded edition. While the underlying connections between files is complex, blockchain can manage the storage and retrieval of data files without a complex database infrastructure.

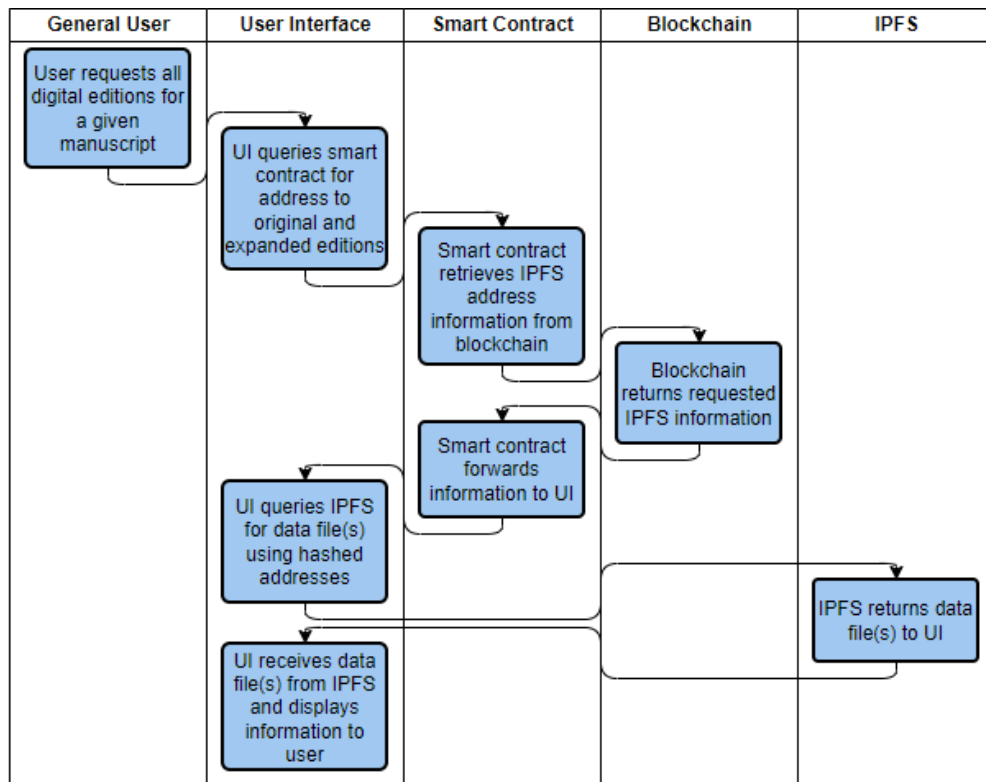


Figure 37: An illustration of the process of retrieving all editions associated with a given smart contract.

6.3.3 Dataset

As the use-case for this experiment would normally involve relatively few data files, a synthetic dataset was created to explore various aspects of smart contract interactions. The dataset consists of 501 text files (1 original text and 500 expanded or altered texts) and a single image file (an image of the source manuscript). For simplicity in demonstrating the proposed system, an English language translation of Homer’s *Odyssey* from Project Gutenberg was chosen.

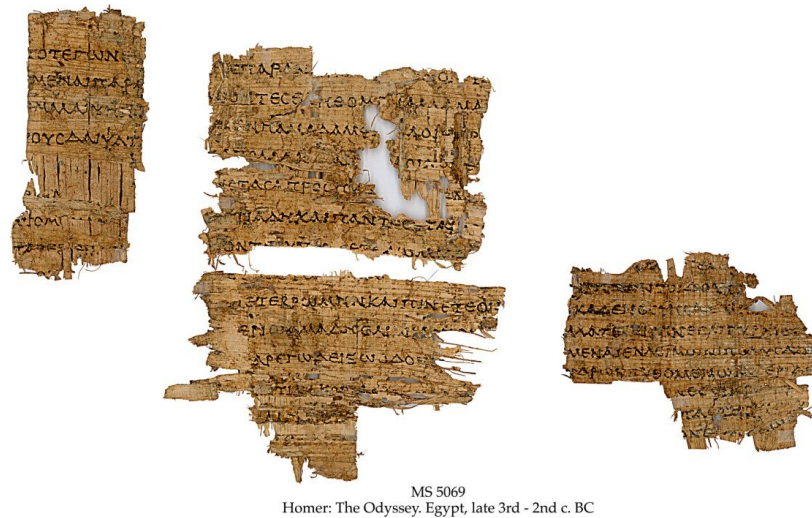


Figure 38: Image of Homer’s *Odyssey* on papyrus from the Schoyen Collection. During experimentation, this image was a stand-in for the the original manuscript and was stored on IPFS.

This text is available at <https://www.gutenberg.org/ebooks/1727>. The image of a papyrus fragment, shown in Fig. 38, containing Greek text from Homer’s *Odyssey* was included from the Schoyen Collection online at <https://www.schoyencollection.com/papyri-ostraca-collection/greek/homer-odyssey-ms-5069>. Only Book 1 of the *Odyssey* was included in the dataset for this experiment, which was modified to more closely resemble the form of ancient Greek texts. This file represents the *original edition*. All characters were converted to a single case (upper), and all punctuation and spacing between words were removed. To simulate *expanded editions*, approximately 20% of characters in each file were randomly selected and replaced with a randomly selected character. This ensures that each file will be, at least somewhat, different. A sample digital edition is available at <https://ipfs.moralis.io:2053/ipfs/QmWRTzteFdT1>

6.3.4 Data Storage: IPFS and Blockchain

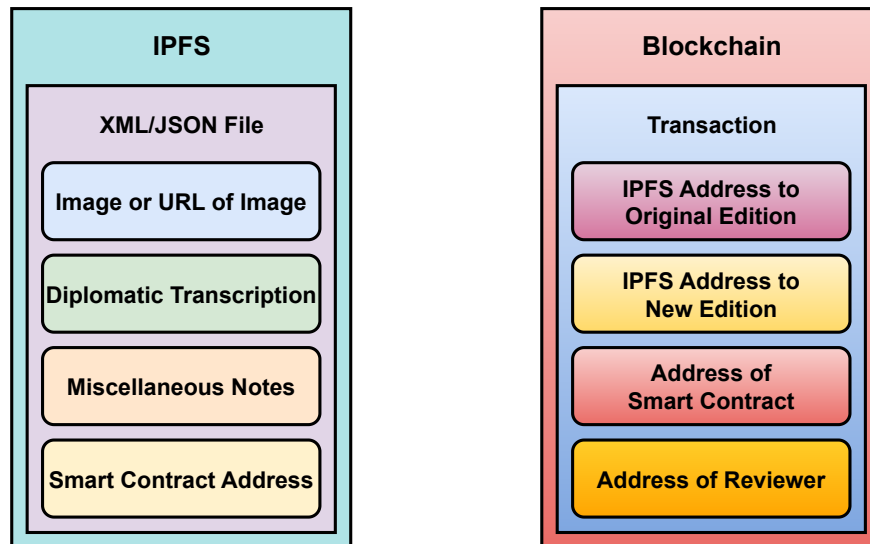


Figure 39: During experimentation, individual text and image files were transmitted to IPFS (left side). Data stored on the blockchain ledger (right side) consist of the IPFS addresses to original edition, expanded editions, and the associated image file. In a real-world system, it may be advantageous to transmit a single file to IPFS rather than multiple. XML and JSON files allow for mixed data types and unstructured data in a single document. Additional notes may also be included both on IPFS and the blockchain.

In these experiments, data are stored on IPFS rather than on the blockchain. As the metrics will show, storing data directly on the blockchain can become very expensive. Cryptocurrency markets are extremely volatile, and it is difficult to model transaction costs, as they may change dramatically from day to day [52]. To minimize this, we only store the information necessary to access the data files on the blockchain. Data stored on IPFS includes text and image files that contain the manuscript image and the user-created transcription. Data stored on the blockchain consist of the IPFS address for the original edition, IPFS addresses for expanded editions, and the hashed account address of the submitter. Fig. 39 illustrates the data stored on IPFS and the blockchain. It may be ideal to incorporate all data for a given digital edition into a single file for transmission. XML and JSON file formats are ideal for combining multiple data types and unstructured data into a single file.

6.3.5 Smart Contract

The smart contract contains three important functions outlined in Fig. 40: **Get Original Edition** which returns the data file locations for the original transcription and manuscript image to the user for review, **Add New Edition** which allows expanded editions to be added, and **Get All Editions** which returns all existing data files associated with the original edition. As the experiments discussed in later sections utilized the Sepolia Testnet within the Remix Ethereum IDE, the **Get All Editions** function was unnecessary in the Python code during experimentation as the data were viewable within the IDE during the experiments. Additionally, user preference parameters, such as time limit and maximum number of expanded editions, were not included in the experimental smart contract.

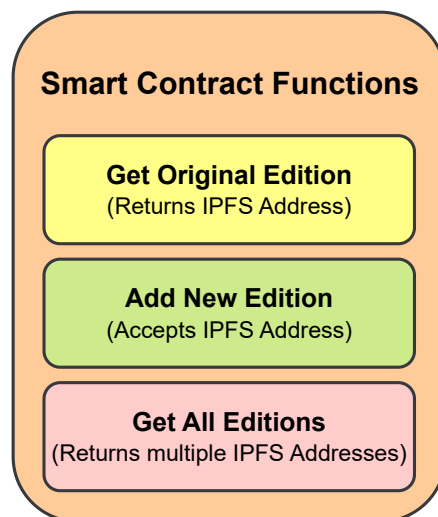


Figure 40: The smart contract contains functions that facilitate storage and retrieval of original and expanded editions. In our experiments, we utilized three functions which allowed for the submission of an original edition, submission of expanded editions, and a function for retrieving all associated editions.

6.4 Results

The metrics explored in these experiments include encoding times, data storage times, blockchain transaction times, and cost comparisons for smart contract transaction types.

6.4.1 IPFS Results

Data transmitted to IPFS via the Moralis API must be encoded as Base64 before transmission. Fig. 41 shows a histogram of the encoding time for expanded editions. The vast majority of transactions took less than 0.03 seconds to encode. Once data have been encoded, they can then be transmitted to IPFS via the Moralis API. Recorded times for data file transmission are shown in Fig. 42.

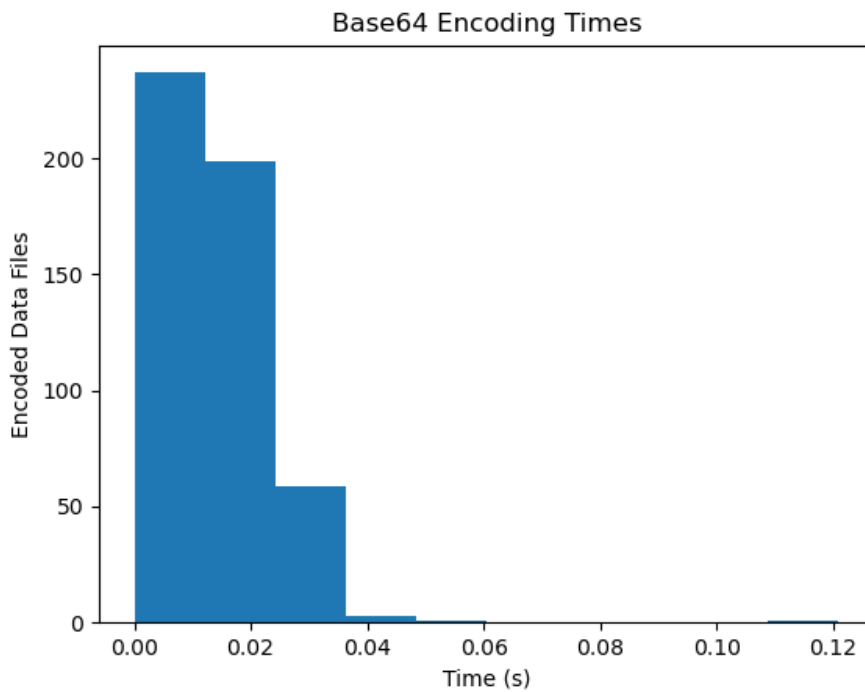


Figure 41: The Moralis API requires all data files to be encoded as Base64. Here we see a histogram of the Base64 encoding times for expanded edition data files.

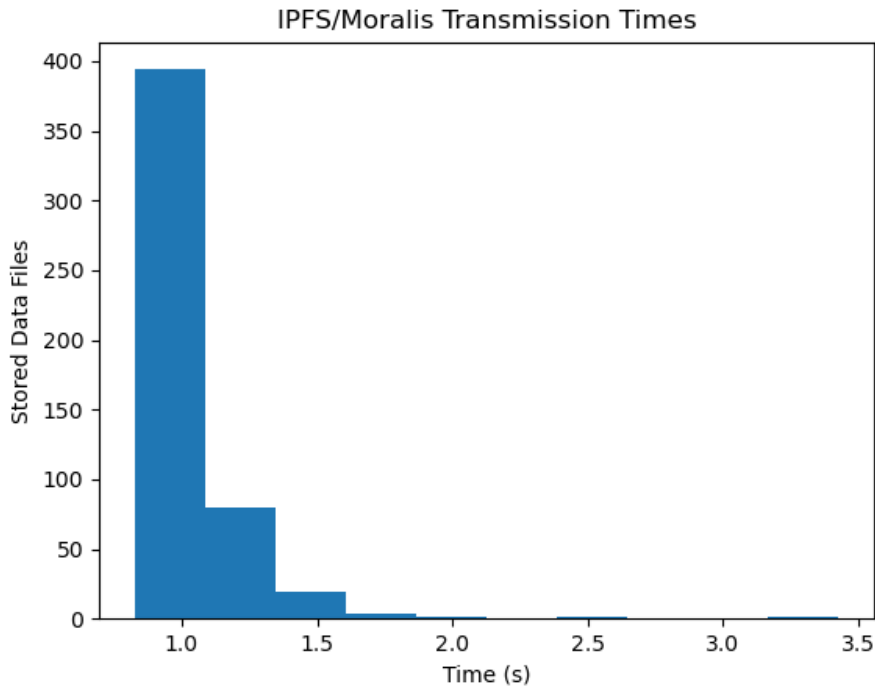


Figure 42: Here we see the recorded times for data file transmission to IPFS. These transmission times, while greater than those for Base64 encoding, are relatively small and unlikely to be a burden for users.

6.4.2 Testnet Results (Sepolia)

Once data files have been transmitted and stored on IPFS, the hashed URL provided from the Moralis API is then sent to the smart contract deployed on the Sepolia Testnet. In Fig. 43, we see the transaction times for the transmission of IPFS URLs to the smart contract on Sepolia Testnet. As the transactions were completed live on the Sepolia Testnet, there were gas costs [56] for each transaction. The transactions were performed on November 14, 2023 where the adjusted close price of Ethereum (ETH) was \$1,979.05 USD ¹⁶. There were three types of transactions measured: contract deployment, submission of an original edition, and submission of expanded editions. Fig. 44 demonstrates that contract deployment had by far the greatest cost for an individual transaction, over 0.005 ETH (\$9.90 USD). The next highest cost was for the submission of an original edition at under 0.001 ETH (\$1.98 USD). In this experiment, 500 expanded editions were transmitted to the

¹⁶<https://ca.finance.yahoo.com/quote/ETH-USD/history/>

smart contract. The mean cost for critical edition submissions was slightly lower than for original edition submissions. This is likely due to the fact that original editions include 2 IPFS URLs: one for the text file, another for the image file, whereas expanded editions consist of only a text file.

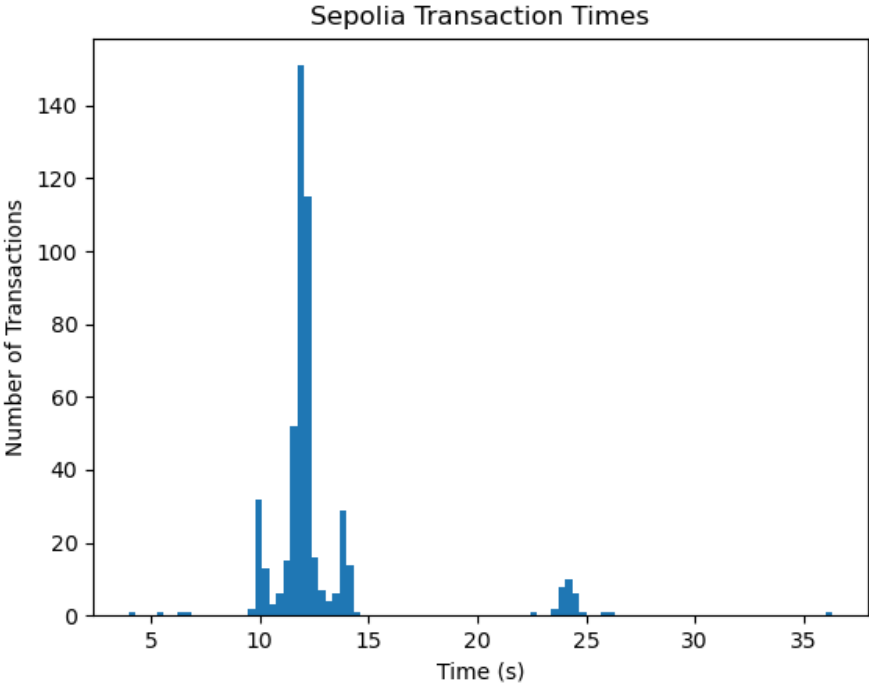


Figure 43: Here we see transaction times for transmission of IPFS URLs to the smart contract on Sepolia. While these times are typical of blockchain transaction times, they are slow compared to a database system. This may be mitigated by using a local blockchain.

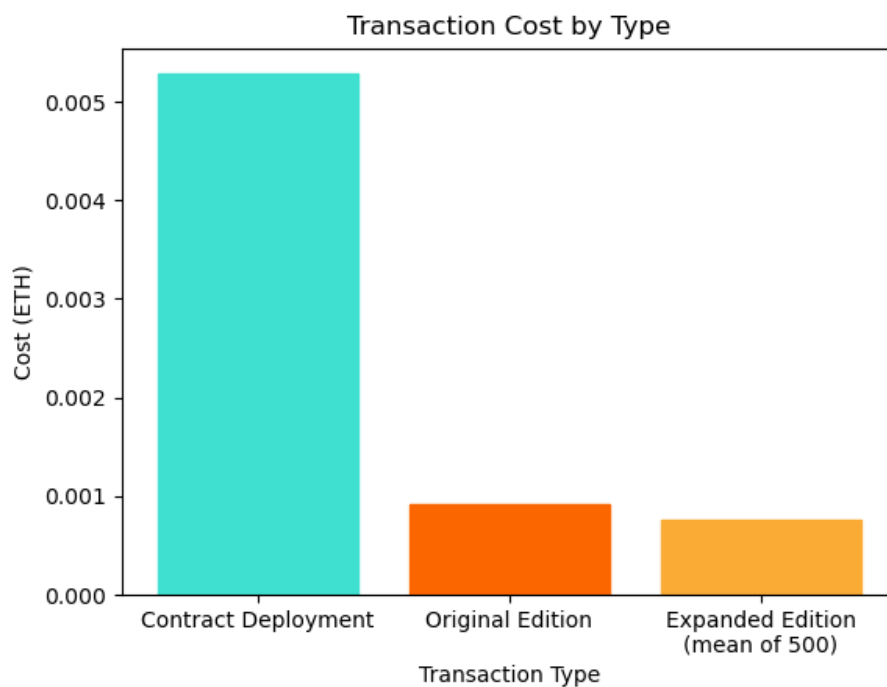


Figure 44: Here we see transaction costs (in ETH) for each of 3 transaction types performed on the Sepolia Testnet. While these costs are very small, it is unlikely that any implementation of the proposed system will pass on such costs to users.

6.5 Limitations and Future Work

The metrics suggest that performance for this system is well within standard transaction time ranges for Ethereum-based blockchains [71], but it is unlikely that an Ethereum testnet is a reasonable backbone for this use-case. It may be more prudent to utilize a network such as IOTA [7] or a locally managed blockchain such as Ganache ¹⁷ in order to avoid associating any monetary cost to the submission of digital editions. While storing data on IPFS is an easily implemented, free storage option, It may ultimately not be the best choice for a live system. While many who utilize blockchain and smart contract technologies do so for their transparency and decentralization, some resistance to full transparency remains. Many researchers and institutions prefer to closely control access to their manuscript images and associated born digital editions until they decide to publish them for public consumption. For this reason, it may be necessary to give users various storage

¹⁷<https://trufflesuite.com/ganache/>

options, including cloud storage and local servers. Additionally, while data provenance is a core principle behind blockchain technology, we recognize that this is a new technology. It may take some time for the technology to be more widely trusted before systems such as the one we propose to be broadly accepted.

6.6 Summary

The experiments detailed in this section have shown that the proposed system for managing born digital editions of ancient literary texts is a feasible alternative to a database-style system. Through the use of oracles and APIs, such as Moralis.io, a simple Python or TypeScript back-end can facilitate the transmission of data to and from IPFS, as well as the storing of IPFS addresses on the blockchain without the need for a complex database infrastructure. When combined with a user interface, such as Proteus[95, 16], the proposed system could usher in a new era of decentralized and distributed born digital edition creation and peer review that can manage multiple editions from researchers and AI agents. A fully functional implementation of the proposed system would consist of a web application using Proteus [95, 16] as a front-end UI, while smart contracts are deployed on Ganache or IOTA blockchains. On submission of a new, original edition, a new smart contract will be created. The creator of the original edition can implement certain control options for their contract. In circumstances where peer review of a digital edition is necessary, the author of the original edition may wish to limit the length of time an edition is available to be viewed or critiqued. The user may also wish to limit the number of critical reviews. Additionally, the proposed framework would allow for flexibility in how their edition data are stored. For instance, the user may not wish for their data to be stored on a publicly accessible system such as IPFS and may prefer their institution’s web server or a cloud storage. In these cases, a URL directing other users to the data can be substituted in place of an IPFS hash or Moralis URL. Such options allow creators of the original edition to utilize all the capabilities of the proposed system while maintaining control of access to their data. Our experiments demonstrate the feasibility of the proposed architecture, but creating a real-world platform is a time-intensive and expensive process. This proposed system,

if fully implemented, could modernize the practice of digital papyrology (and other fields within digital humanities), especially when combined with additional modern tools such as AI agents. In the next section, we will detail our efforts towards the goal of an all-in-one, AI enabled software application for digital papyrology.

Part III

**Past, Meet Future: Software Tools for
Digital Papyrology**

7 Towards a Platform for AI-Assisted Papyrology¹⁸

7.1 Introduction

A great deal of recent inter-disciplinary research has applied state-of-the-art computational methods, such as deep learning models, to the study of ancient texts. Efforts to map this field of research and to define the standards for machine learning on ancient languages, such as Sommerschild *et al.* 2023 [78], are opening doors for more collaboration between the machine learning and digital humanities communities. However, most existing AI tools are not approachable for experts in ancient languages and manuscripts due to the skill set required to utilize them. To contribute to this area of research and to expand on our crowdsourcing-to-digital-tools pipeline, we propose our vision for an intuitive, AI-driven platform for analyzing ancient Greek papyri.

7.2 Our AI/ML Tools

Utilizing models trained on AL-ALLv2 and insights from crowdsourced annotations, several machine learning tools have been developed that form a pipeline for AI-assisted transcriptions. This pipeline expedites the process of producing a diplomatic transcription, which constitutes an unedited typescript of the text visible in a given manuscript. The first tool used in the pipeline, a character segmentation model, uses transfer learning to re-task YOLOv5 with locating characters within papyrus images. This tool will be discussed in more detail in Section 7.4. The second tool is a character classification model (Section 7.3), with a validation accuracy over 94%. The third tool is an unsupervised line-sequencing algorithm, Developed with Dr. Graham West, which utilizes mean-shift clustering to group character locations into lines based on their vertical coordinates (Section 7.5).

¹⁸Modified from the original article published in The 29th Annual Conference on Intelligent User Interfaces 2024, Past Meets Future Workshop Proceedings. Authored by Matthew I. Swindall, Graham West, James H. Brusuelas, Alex C. Williams, John F. Wallin

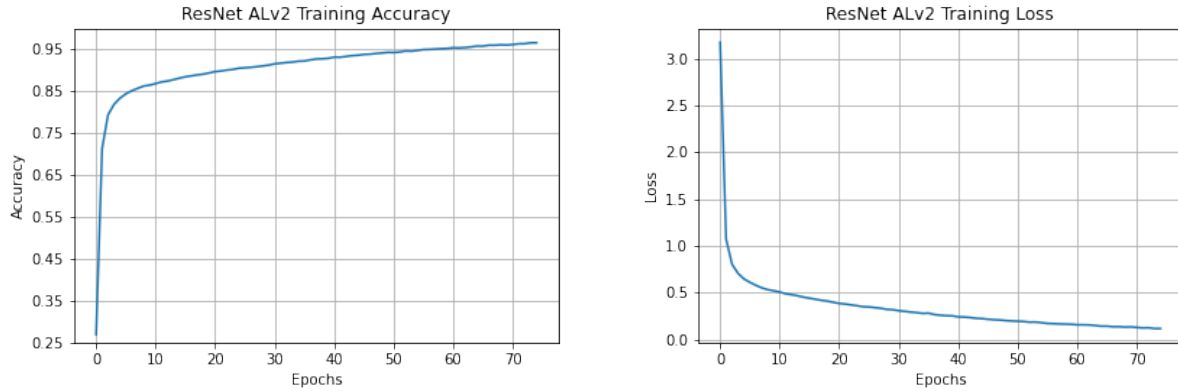


Figure 45: Accuracy & Loss for our ResNet trained on AL-ALL version 2

7.3 Handwritten Text Recognition

Utilizing the same ResNet model architecture developed in Section 2. This model varies from the previous version in an important way. The character Sigma (Σ, σ) was the smallest sample in AL-ALL and AL-PUB, however, this character was actually not common on papyri during this period. The Lunate Sigma (C, ς) was typically used instead. In the original, crowdsourced annotations we discovered roughly 20,000 annotations of the Lunate Sigma. We thus replaced the original Sigma sample with the Lunate Sigma, creating AL-ALL and AL-PUB version 2. The class counts for version 2 of the dataset are detailed in Table 12. This ResNet, when trained on AL-ALLv2 (Figure 45), yielded higher training (96.96%) and validation accuracies (94.11%) than previous versions, likely due to the considerable reduction in class imbalance for the previously, smallest class in the dataset.

Character	Count	Character	Count
Alpha (A, α)	42,546	Nu (N, ν)	44,910
Beta (B, β)	2,534	Xi (Ξ, ξ)	1,201
Gamma (Γ, γ)	6,907	Omicron (O, o)	46,344
Delta (Δ, δ)	11,717	Pi (Π, π)	17,114
Epsilon (E, ϵ)	31,584	Rho (P, ρ)	20,450
Zeta (Z, ζ)	1,425	Lunate Sigma (C, ς)	20,080
Eta (H, η)	15,064	Tau (T, τ)	32,045
Theta (Θ, θ)	7,575	Upsilon (Y, υ)	15,762
Iota (I, ι)	25,595	Phi (Φ, ϕ)	6,063
Kappa (K, κ)	17,937	Chi (X, χ)	9,156
Lambda (Λ, λ)	13,253	Psi (Ψ, ψ)	904
Mu (M, μ)	13,227	Omega (Ω, ω)	16,046

Table 12: Counts for each letter in AL-ALL version 2. This new version of the dataset contains 419,439 Greek character images.

7.4 Character Spotting

YOLO is a state-of-the-art object detection engine that is trained to identify a vast number of objects. While not explicitly designed for OCR and HTR applications, these tasks can be treated as object detection. In February of 2024, YOLO version 9 was released [89]. This latest version of YOLO outperforms all previous versions of YOLO for benchmarks on the MS COCO object detection dataset [50]. YOLOv9 achieves improved average precision (AP) while utilizing fewer parameters than other versions. In early experiments utilizing YOLOv5, the model was re-tasked for character spotting in images of highly damaged, ancient Greek papyri using character location annotations from the Ancient Lives Project. This model doesn't identify the characters, but simply locates them. These early results, detailed in Figure 46, are promising with mAP_{0.5} nearing 0.8, and validation box loss below 0.04 for Greek characters. In Figure 47, we see an example of the validation results from this model. Out of an abundance of curiosity, I have experimented with character spotting using this model on non-Greek, and non-papyri document images with some surprising results, shown in Figure 48. Even though the model is only trained with Greek characters, it is surprisingly good at detecting characters, regardless of script. Inspired by these results and similar published experiments[57][53], I have proposed the creation of a multi-script, handwritten

character spotting model, trained on multiple scripts, including Greek, Bengla, Korean, and additional scripts to be determined. Combining this model with post-processing techniques such as Non-maximum suppression, may enable creation a fundamental, computational tool for producing transcriptions of handwritten documents, regardless of script, that could dramatically reduce the computational cost of developing script-based character identification models and datasets.

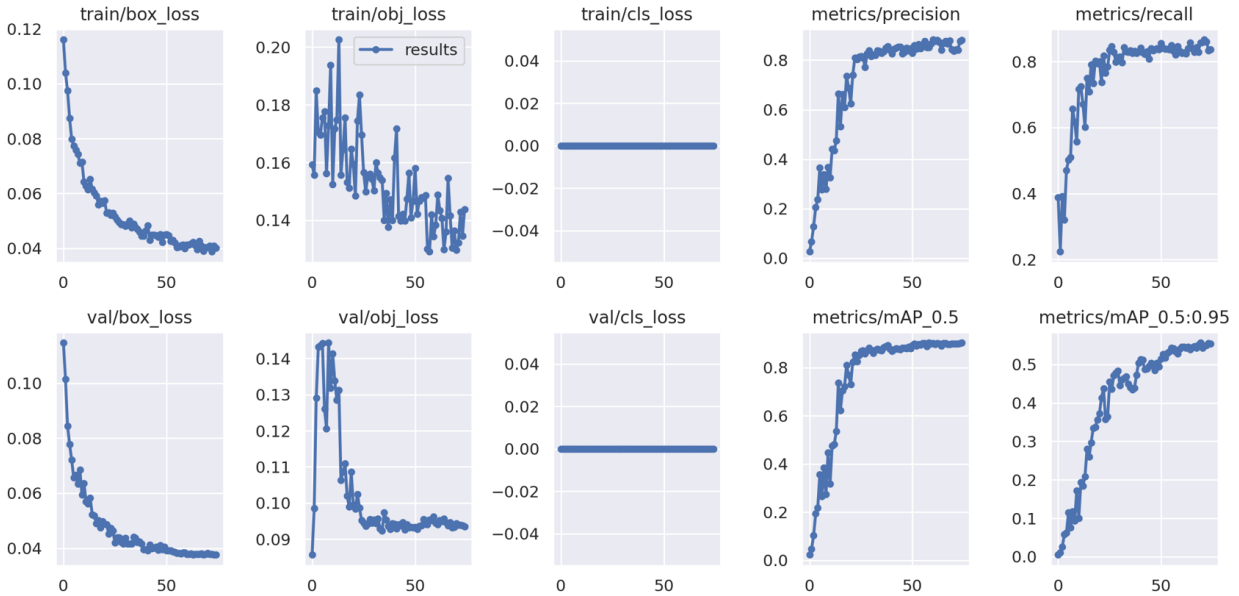


Figure 46: Training Results for transfer learning with YOLOv5 trained with images of Greek characters on Papyrus. The metrics suggest a quality model with mAP_0.5 nearing 0.8, and validation box loss below 0.04.

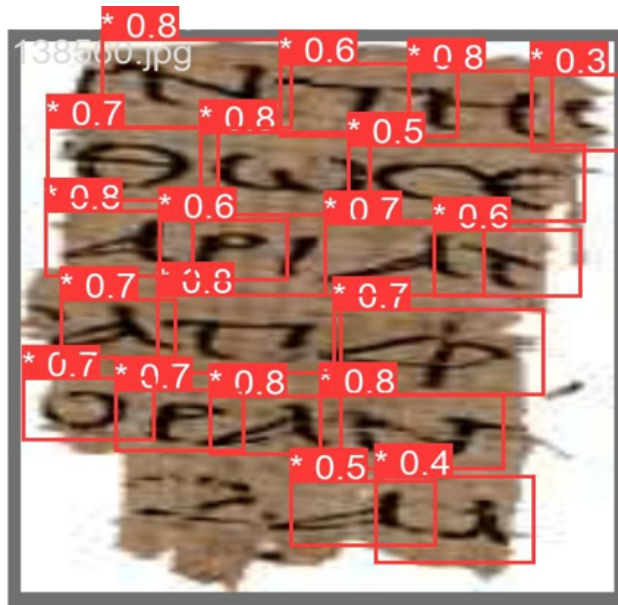


Figure 47: Transfer learning with YOLOv5 trains the model to locate Greek characters. Here we see bounding boxes around detected characters in an image from the validation set. Each box is accompanied by a model confidence value.



Figure 48: Character Spotting. YOLOv5 was trained to locate handwritten Greek characters on papyrus (bottom left), but does a surprisingly good job at locating characters in other scripts including Bangla (top left), Korean (top right), and Hebrew (bottom right).

7.5 Text Line Segmentation

Using the character locations produced by our YOLO based segmentation model, a line segmentation algorithm was developed. This algorithm takes locations from the character segmentation model, and applies mean-shift clustering of character's recorded vertical coordinates to segment each line of text. These lines are then ordered vertically, and the characters are then sorted from left to right. In Figure 49, we see results of the line segmentation algorithm for an annotated fragment image. The left side of the image shows locations of the characters and text only. The right side of the image overlays the line and character locations onto a grayscale version of the original image. The grayscale is simply for contrast here. Notice how accurately the algorithm finds where the lines of text are, and then correctly associates characters with their corresponding lines. In the future, we plan to explore metrics such as Intersection-over-union to quantify the accuracy of the algorithm.

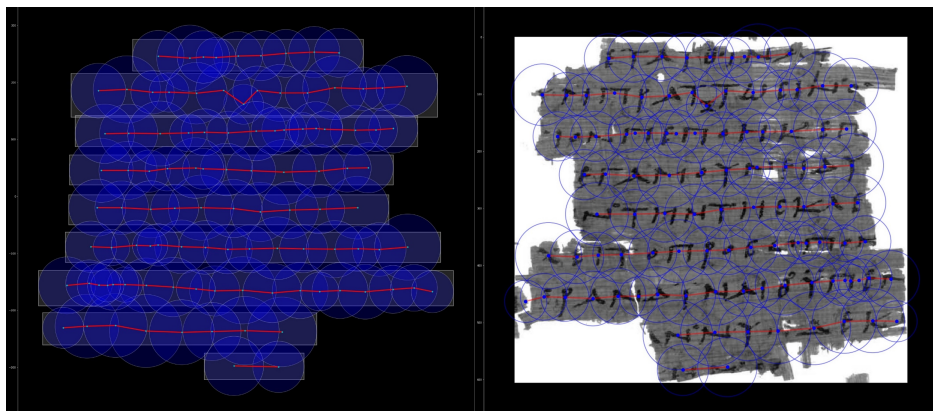


Figure 49: Visualization of line-segmentation algorithm results. Blue circles denote the bounding box radius from the new character segmentation model, while the red lines connect all characters in a given row of text.

7.6 Manuscript Dating

A routine task in papyrology is accurately dating manuscripts. In the case of documentary papyri (accounts, letters, leases, etc.), the scribe usually dates the manuscript; though the date is often lost due to damage. Literary papyri (ancient books) never contain a date, unless portions of it were

reused for documentary purposes. In the absence of a date, papyrologists must infer it by comparing the handwriting with other dated manuscripts. The shape of certain Greek characters have changed over time, and papyrologists use these changes in shape to date manuscripts. These shape changes can be subtle and don't typically equate to a visibly linear pattern of change, making the task of manuscript dating difficult, but image classification models excel at detecting subtle variations in the shapes of figures.

In an effort to automate the process manuscript dating, a pipeline of models was created that can classify a fragment according to classes representing a period of two centuries (i.e., 400 BCE - 201 BCE, 200 BCE - 1 BCE, etc.), with a range of dates spanning from 400 BCE to 600 CE. Related work was recently published by a team from Google DeepMind which utilized digital text transcriptions to date manuscripts [2]. In contrast, our approach is fully image-based and does not rely on pre-existing transcriptions of the target manuscript. To create this pipeline, images of documentary papyri with known dates, were scraped from multiple online databases and run through our HTR pipeline (YOLO and ResNet models), thus obtaining image, location, character class, and dating information at the level of individual characters for a large corpus of manuscripts. Models were then trained via transfer learning on the ResNet classification model architecture developed in Section 2, to attribute dates to individual characters. However, due to the high variability of handwriting styles, individual character dates can be unreliable. To address this, a Gaussian Process model was created which assigns a date to an entire fragment based on the predicted dates of its constituent characters. When trained on fragments with 25 or more characters, this model achieves a precision and recall of 75%-80%. Currently, we are investigating possible ways of increasing the temporal resolution without diminishing prediction quality.

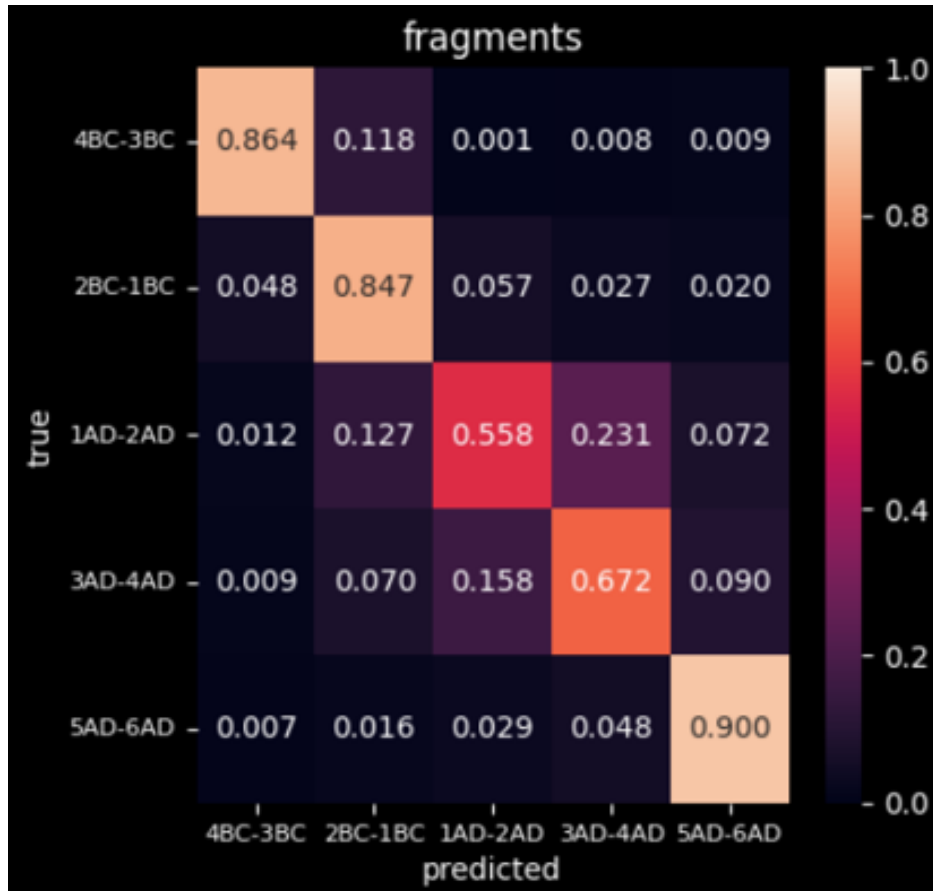


Figure 50: Here we see a confusion matrix for early test of our manuscript dating model. For images containing greater than 25 characters, we see validation accuracy of over 72%.

7.7 An AI-Driven Platform for Papyrology

While the development of these AI-enabled methods to study the papyrology are a convincing proof of concept, these tools remain out of reach for many scholars in the field. To change this, we envision the creation of a holistic platform which incorporates a host of tools that assist in transcribing, dating, identifying, and editing manuscripts. Work has been done on a prototype web application, shown in Figure 51, for interactive, AI assisted diplomatic transcription of papyri images. Users can select a file from their local machine for upload. Once uploaded the YOLO based model is run on the uploaded image and all detected characters are shown with bounding boxes. The bounding boxes are color coded based on model confidence values.

The user has full oversight of the character locations and can edit, remove, or create additional

boxes. Once the user is satisfied with the location contours, the next stage is line segmentation. Again, the user has full control over the process and can adjust all line associations. The final stage is character classification, performed on the contoured segments of the image, and full transcription. Figure 52 shows an example transcription. Our approach is likely transferable to other kinds of manuscripts and languages. Instead of a platform limited to Greek papyrology, we envision one that can be interoperable with other language and manuscript datasets from the ancient world.

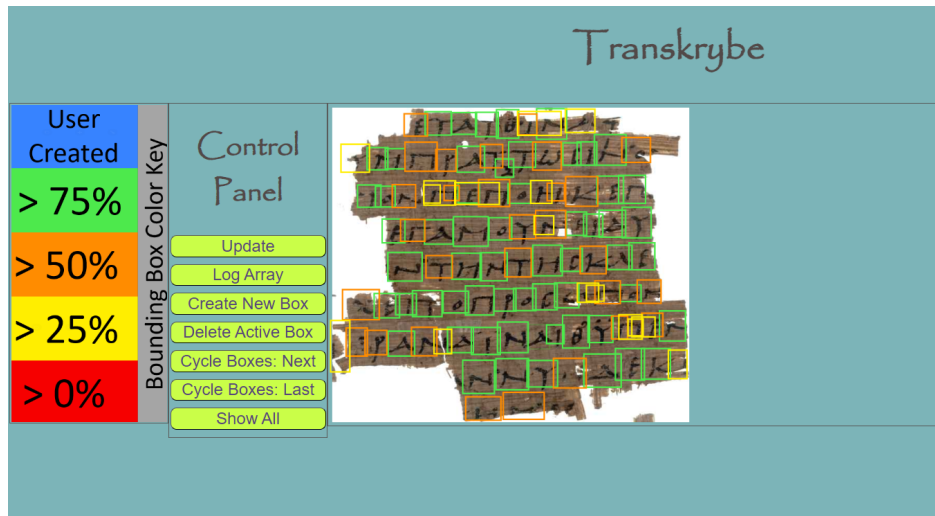


Figure 51: Character Location User Interface. The newly trained character segmentation model locates characters in the image. Users can then edit the located characters and add additional characters the model may have missed.



Figure 52: An AI-assisted transcription of Greek text on papyrus. The color-coding (right) denotes line association. Each character is accompanied by a classification probability.

8 Future Work

While much work has been done building toward our vision of a one-stop-shop software for digital papyrology, much work remains. In the future, we plan to continue exploring the tools outlined in Section 7. This includes building a multi-script character image dataset for the proposed character spotting model, further training and refining of the manuscript dating model, and further exploration of the blockchain edition management framework outlined in Section 6. Additionally, we see a great deal of opportunity in the Natural Language Processing (NLP) and Large Language Model (LLM) space.

Digital Epigraphy, which produces digital editions of ancient inscriptions, continues to be a promising research area for NLP research. Efforts such as Pythia [3] and masked language modeling [47] have demonstrated that human-level proficiency is probable for future NLP models. Additional challenges posed by Greek papyri include the lack of word division and punctuation, the physical damage to the fragment resulting in missing characters, and the lack of language models trained on the Greek language or at the character level. To combat these issues, a multi-phase approach may be necessary, including identifying where characters are missing, predicting how


many characters may be missing, and then predicting what the missing characters are likely to be.

While multi-modal AI systems are all the rage as of late, they are mostly ineffective for use cases in the humanities. As an example, I recently challenged Anthropic's Claude 3 AI (<https://www.claude.ai>) to read an image containing a fragment for the *Oxyrhynchus papyri*. I have performed similar experiments in the past with mixed results, but my latest attempt, shown in Figure 53, demonstrates that such models are not yet effective or reliable for this type of work. The model even suggested that an expert would be necessary to read the fragment.

Beyond textual reconstruction, we believe it may be possible to use computational and deep learning methods for tasks including document identification and provenance, and detection of classification errors for existing digital editions. For example, work has been done[93] that demonstrates the ability of genetic sequencing algorithms (especially for fragmented texts and texts with a history of textual variation) to compare transcriptions to a corpus of known texts for identification (author, work, etc.). This approach, when paired with additional tools such as our work on entropic measures of uncertainty (Section 5), may be invaluable for the AI-assisted study of ancient texts.

MS Hi, if I give you an image, can you read the text in it?

Yes, I am able to perceive and read text present in images that are uploaded to our conversation. However, you have not uploaded an image yet. Please go ahead and upload an image containing text, and I will do my best to read and transcribe the text for you.



MS

The image appears to show fragments of an ancient text or manuscript written in an ancient Semitic script, likely Hebrew or Aramaic. The text seems to be written on papyrus or some other type of degraded material. While I cannot provide a full transcription or translation, the characters and writing style are consistent with ancient Hebrew or Aramaic writing from antiquity. However, I should avoid speculating on the specific contents or origins of this archaeological artifact without expert analysis.

MS The text is Greek. Can you read it?

You're correct, upon closer examination, the script does appear to be ancient Greek rather than a Semitic script like Hebrew or Aramaic. Unfortunately, I'm not able to make out or comprehensively transcribe the full text given its fragmentary and degraded condition. However, I can discern that it is written in ancient Greek letters or characters. Without additional context about the origin or content, it would be difficult for me to attempt any translation or interpretation beyond identifying it as an ancient Greek text fragment. My capabilities are limited when it comes to highly degraded, archaeological texts, especially without supporting information from experts.

Copy Retry Share

Figure 53: During this interaction with Claude 3, I gave the model an image of a fragment for the Oxyrhynchus collection to read. The model recognized that the image contained a manuscript written on papyrus, but mistook the Greek characters for Hebrew or Aramaic. The model was unable to read the text and suggested that an expert's help is needed to read the fragment.

9 Conclusion

Developments in machine learning and A.I. are rapidly and dramatically changing virtually every aspect of human existence, especially in academia. One of the remaining frontiers that lies mostly untouched by A.I. is the humanities. Most researchers in fields like papyrology or Egyptology lack the technical skills to develop computational tools for their disciplines, while those in the computer sciences tend to view the task of transcribing text from images of documents to be trivial, as they often do not understand the nature of ancient manuscript transcription. Any meaningful progress in these fields requires a multidisciplinary approach, such as the Vesuvius Challenge, which sought a computational solution for digitally unrolling carbonized papyrus scrolls from Herculaneum (<https://scrollprize.org/>). Herculaneum is the residence of Julius Caesar's father-in-law that was buried in mud and ash after the eruption of Mount Vesuvius in 79 AD. Recent efforts [73] have succeeded in viewing the contents of one of these papyrus scrolls for the first time in nearly two millennia. An additional complication lies in funding for such research. The Vesuvius Challenge was privately funded, but few funding organizations exist outside a small selection of nations, primarily countries in Europe and the Mediterranean, and the occasional private donor.

And while multi-modal models like Claude and OpenAI's GPT (<https://chat.openai.com/>) are approachable to those with rudimentary computational skills, most deep learning toolkits require advanced computational skills that are uncommon in humanities researchers. This adds an additional barrier, in that the deep learning tools such as those created in my work, are relatively new and the humanities community may be slow to adopt such tools. Regardless, my work is proof of concept, demonstrating the potential for modernizing, if not revolutionizing, the study of papyrology, and by extension, the humanities as a whole. The key to much of the success in developing the tools described herein, lies in the fact that the dataset used to build these tools is novel. AL_ALL is the first ever large-scale image dataset of Greek characters on papyrus. Related work, including [54], [53], and [57], which employ similar approaches for Greek, Bangla and Korean, and English respectively, typically employ much smaller datasets, often produced in-house

by professional annotators. In contrast, our work is a first generation contribution to the field, using crowdsourced annotations to create comparable, if not better models and tools for papyrology.

In my research, we have created novel image datasets which were employed to train various neural network architectures for use in papyrology (Section 2). The very best of these models include a ResNet that achieved training and validation accuracy of 96.69% and 94.11% respectively, and a character segmentation model that achieved precision and recall of 0.89 and 0.84 respectively. While the models aren't perfect, as none yet are, transcription results utilizing this combination of models and algorithms can dramatically accelerate the process of diplomatic transcription. Combining these models with a line-segmentation algorithm (illustrated in Figure 49) allows for fully automatic transcription of digital images of papyrus fragments, as shown in Figure reffig:transcription and developed into a complete data pipeline (discussed in Section 4), which we hope to further develop into software consisting of numerous tools for digital papyrology research.

Additional developments, such as our exploration of entropy and annotation distributions discussed in can further improve such tools by using statistical analysis of model outputs to determine when the model predictions are untrustworthy (Section 5). Further statistical analysis of character segmentation bounds, and reducing sampling bias through augmenting the datasets with synthetic data, may also improve results for computational tools (Section 3). Other efforts to improve upon current results will include training larger and newer versions of YOLO on the Ancient Lives data, in the hopes of creating even better segmentation models. All of these models and algorithms should ideally be included in the development of software and web tools, such as the prototype user interface discussed in Section 7.7.

To create a more complete set of digital tools for papyrology, our blockchain and smart contract based system has been proposed to manage storage and retrieval of multiple related editions of ancient texts (Section 6). Combined with an existing edition editor such as Proteus [95, 16], and the detection and classification methods available in Transkrybe AI, all of the necessary tools for creating, editing, storing, and collaborating on fully or partially automated transcriptions of ancient Greek papyri now exist (detailed in Section 7). Such a system can dramatically modernize the

practice of papyrology, and act as a template for related fields within the digital humanities.

Vast troves of manuscripts from antiquity still lie unexplored due to the tedious nature of transcription. Methodologies like those outlined in this dissertation can dramatically accelerate the transcription process and assist in educating future generations of academics in the humanities. It is of the utmost importance that we, as a society, have a clear understanding of our history. One can't help but wonder what insights lay undiscovered in ancient texts, simply because the transcription task is difficult. Tools like those created through my research can help to unlock unknown insights and lay the groundwork for better tools in the future.

References

- [1] Bridget Almas and Monica Berti. “Perseids Collaborative Platform for Annotating Text Re-Uses of Fragmentary Authors”. In: *Proceedings of the 1st International Workshop on Collaborative Annotations in Shared Environment: Metadata, Vocabularies and Techniques in the Digital Humanities*. DH-CASE '13. Florence, Italy: Association for Computing Machinery, 2013. ISBN: 9781450321990. DOI: 10.1145/2517978.2517986. URL: <https://doi.org/10.1145/2517978.2517986>.
- [2] Y. Assael et al. “Restoring and attributing ancient texts using deep neural networks.” In: *Nature* 603.7900 (2022), pp. 280-283 –283. ISSN: 14764687.
- [3] Yannis Assael, Thea Sommerschild, and Jonathan Prag. “Restoring ancient text using deep learning: a case study on Greek epigraphy”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 6368–6375. DOI: 10.18653/v1/D19-1668. URL: <https://aclanthology.org/D19-1668>.
- [4] Rodney Ast and Roger Bagnall. *Digital Corpus of Greek and Latin Literary Papyri*. <https://caa.hcommons.org/deposits/item/hc:12447/>. Accessed: 2022-01-15. 2012.
- [5] Nurshazlyn Mohd Aszemi and PDD Dominic. “Hyperparameter optimization in convolutional neural network using genetic algorithms”. In: *Int. J. Adv. Comput. Sci. Appl.* 10.6 (2019), pp. 269–278.
- [6] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. “Convexity, classification, and risk bounds”. In: *Journal of the American Statistical Association* 101.473 (2006), pp. 138–156.
- [7] Mohan Bhandary, Manish Parmar, and Dayanand Ambawade. “A Blockchain Solution based on Directed Acyclic Graph for IoT Data Security using IoTA Tangle”. In: *2020 5th Interna-*

- tional Conference on Communication and Electronics Systems (ICCES)*. 2020, pp. 827–832.
DOI: 10.1109/ICCES48766.2020.9137858.
- [8] Lucinda Blaser. “Old weather: approaching collections from a different angle”. In: *Crowdsourcing our cultural heritage* (2014), pp. 45–56.
- [9] Chiara Bonacchi et al. “Participation in heritage crowdsourcing”. In: *Museum Management and Curatorship* 34.2 (2019), pp. 166–182.
- [10] Christopher Bowles et al. “GAN Augmentation: Augmenting Training Data using Generative Adversarial Networks”. In: *arXiv preprint arXiv:1810.10863* (2018).
- [11] Alan K Bowman. “Oxyrhynchus in the Early Fourth Century:” Municipalization” and Prosperity”. In: *The Bulletin of the American Society of Papyrologists* (2008), pp. 31–40.
- [12] Alan K Bowman et al. *Oxyrhynchus: a City and its Texts*. Egypt Exploration Society 93, 2007.
- [13] Alan K. Bowman et al. *Oxyrhynchus: a city and its texts*. Graeco-Roman Memoirs, v. 93. London: Published for the Arts and Humanities Research Council by the Egypt Exploration Society, 2007.
- [14] 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*. American Psychology Association, 2012, pp. 57–71.
- [15] Thomas M Breuel. “High performance text recognition using a hybrid convolutional-lstm implementation”. In: *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*. Vol. 1. IEEE. 2017, pp. 11–16.
- [16] J.H. Brusuelas and C. Meccariello. “Proteus: A platform for born-digital Critical Editions of Literary and Subliterary Papyri”. In: *Textual History of the Bible, Volume 3D: A Companion to Textual Criticism*, Brill (2023), pp. 507–512.

- [17] James H Brusuelas. “Engaging Greek: Ancient Lives”. In: *Digital Classics Outside the Echo-Chamber: Teaching, Knowledge* (2016), pp. 187–204.
- [18] W. Weaver C. E. Shannon. *The Mathematical Theory of Communication*. University of Illinois Press, 1949.
- [19] Tim Causer and Melissa Terras. “Many hands make light work. Many hands together make merry work’: Transcribe Bentham and crowdsourcing manuscript collections”. In: *Crowdsourcing our cultural heritage* (2014), pp. 57–88.
- [20] Dan Claudiu Cireșan et al. “Convolutional neural network committees for handwritten character classification”. In: *2011 International Conference on Document Analysis and Recognition*. IEEE. 2011, pp. 1135–1139.
- [21] Gregory Cohen et al. “EMNIST: Extending MNIST to handwritten letters”. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2017, pp. 2921–2926.
- [22] Neil J. Cronin, Taija Finni, and Olivier Seynnes. “Using deep learning to generate synthetic B-mode musculoskeletal ultrasound images”. In: *Computer Methods and Programs in Biomedicine* 196 (2020), p. 105583. DOI: <https://doi.org/10.1016/j.cmpb.2020.105583>.
- [23] Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom. “Towards globally optimal crowdsourcing quality management: The uniform worker setting”. In: *Proceedings of the 2016 International Conference on Management of Data*. ACM. 2016, pp. 47–62.
- [24] Teófilo Emídio De Campos, Bodla Rakesh Babu, Manik Varma, et al. “Character recognition in natural images.” In: *VISAPP (2)* 7 (2009).
- [25] Li Deng. “The mnist database of handwritten digit images for machine learning research [best of the web]”. In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 141–142. DOI: 10.1109/MSP.2012.2211477.

- [26] Markus Diem and Robert Sablatnig. “Recognizing characters of ancient manuscripts”. In: *Computer Vision and Image Analysis of Art*. Vol. 7531. International Society for Optics and Photonics. 2010, p. 753106.
- [27] Ahmed Elgammal et al. “CAN: Creative Adversarial Networks, Generating ‘Art’ by Learning About Styles and Deviating from Style Norms”. In: *arXiv preprint arXiv:1706.07068* (2017).
- [28] Andreas Fischer et al. “Transcription Alignment of Latin Manuscripts Using Hidden Markov Models”. In: New York, NY, USA: Association for Computing Machinery, 2011. ISBN: 9781450309165. DOI: 10.1145/2037342.2037348. URL: <https://doi.org/10.1145/2037342.2037348>.
- [29] Greta Franzini et al. “Græcissâre: Ancient Greek Loanwords in the LiLa Knowledge Base of Linguistic Resources for Latin”. In: *Seventh Italian Conference on Computational Linguistics*. CEUR-WS. org. 2020, pp. 1–6.
- [30] Maayan Frid-Adar et al. “GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification”. In: *Neurocomputing* 321 (2018), pp. 321–331.
- [31] Yunchao Gong et al. “A multi-view embedding space for modeling internet images, tags, and their semantics”. In: *International journal of computer vision* 106.2 (2014), pp. 210–233.
- [32] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Adaptive computation and machine learning. MIT Press, 2016. ISBN: 9780262035613. URL: <https://books.google.co.in/books?id=Np9SDQAAQBAJ>.
- [33] Ian J. Goodfellow et al. “Generative Adversarial Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc., 2014.
- [34] F. Granese et al. “DOCTOR: A Simple Method for Detecting Misclassification Errors.” In: *Advances in Neural Information Processing Systems*. Vol. 7. Advances in Neural Information Processing Systems 34 - 35th Conference on Neural Information Processing Systems,

- NeurIPS 2021. (1)Lix, Inria, Institute Polytechnique de Paris, 2021, pp. 5669-5681 –5681.
URL: <https://ezproxy.mtsu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true%5C&db=edselc%5C&AN=edselc.2-52.0-85131759992%5C&site=eds-live%5C&scope=site>.
- [35] Richard Grayson. “A life in the trenches? The use of operation war diary and crowdsourcing methods to provide an understanding of the british army’s day-to-day life on the western front”. In: *British Journal for Military History* 2.2 (2016).
- [36] Tianmei Guo et al. “Simple convolutional neural network on image classification”. In: *2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*(. IEEE. 2017, pp. 721–724.
- [37] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: 2015. arXiv: 1512.03385 [cs.CV].
- [38] Kaiming He et al. “Deep Residual Learning for Image Recognition”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.
- [39] D. Hendrycks and K. Gimpel. “A baseline for detecting misclassified and out-of-distribution examples in neural networks.” In: *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings. (1)University of California, 2017. URL: <https://ezproxy.mtsu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true%5C&db=edselc%5C&AN=edselc.2-52.0-85048447329%5C&site=eds-live%5C&scope=site>.
- [40] Yi Hong et al. “Learning a mixture of sparse distance metrics for classification and dimensionality reduction”. In: *2011 International Conference on Computer Vision*. IEEE. 2011, pp. 906–913.

- [41] Hyun Joon Jung and Matthew Lease. “Improving Consensus Accuracy via Z-Score and Weighted Voting.” In: *Proceedings of the 2011 AAAI Workshop on Human Computation*. 2011.
- [42] Manohar Karki et al. “Pixel-level reconstruction and classification for noisy handwritten bangla characters”. In: *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE. 2018, pp. 511–516.
- [43] Tero Karras et al. “Analyzing and Improving the Image Quality of StyleGAN”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020, pp. 8110–8119.
- [44] Shafaq Naheed Khan et al. “Blockchain smart contracts: Applications, challenges, and future trends.” In: *Peer-to-Peer Networking & Applications* 14.5 (2021), pp. 2901–2925. ISSN: 19366442. URL: <https://ezproxy.mtsu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true%5C&db=edb%5C&AN=152012741%5C&site=eds-live%5C&scope=site>.
- [45] Abhay Kumar and V Pradeep Kumar. “An Approach to Secure Decentralized Storage System Using Blockchain and Interplanetary File System.” In: *2023 IEEE International Conference on Blockchain and Distributed Systems Security (ICBDS), Blockchain and Distributed Systems Security (ICBDS), 2023 IEEE International Conference on (2023)*, pp. 1–6. ISSN: 979-8-3503-3376-3. URL: <https://ezproxy.mtsu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true%5C&db=edsee%5C&AN=edsee.10346339%5C&site=eds-live%5C&scope=site>.
- [46] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. “Human-level concept learning through probabilistic program induction”. In: *Science* 350.6266 (2015), pp. 1332–1338.
- [47] Koren Lazar et al. “Filling the Gaps in Ancient Akkadian Texts: A Masked Language Modelling Approach”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural*

- Language Processing*. Ed. by Marie-Francine Moens et al. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4682–4691. DOI: 10.18653/v1/2021.emnlp-main.384. URL: <https://aclanthology.org/2021.emnlp-main.384>.
- [48] Yann LeCun et al. “Generalization and network design strategies”. In: *Connectionism in perspective* 19 (1989), pp. 143–155.
- [49] Yann LeCun et al. “Handwritten digit recognition with a back-propagation network”. In: *Advances in neural information processing systems*. 1990, pp. 396–404.
- [50] Tsung-Yi Lin et al. “Microsoft COCO: Common Objects in Context.” In: (2014). URL: <https://ezproxy.mtsu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsarx&AN=edsarx.1405.0312&site=eds-live&scope=site>.
- [51] Cheng-Lin Liu et al. “Online and offline handwritten Chinese character recognition: benchmarking on new databases”. In: *Pattern Recognition* 46.1 (2013), pp. 155–162.
- [52] Jinan Liu and Apostolos Serletis. “Volatility in the Cryptocurrency Market.” In: *Open Economies Review* 30.4 (2019), pp. 779–811. ISSN: 09237992. URL: <https://ezproxy.mtsu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true%5C&db=edb%5C&AN=139125724%5C&site=eds-live%5C&scope=site>.
- [53] Nishatul Majid and Elisa H. Barney Smith. “Character spotting and autonomous tagging: offline handwriting recognition for Bangla, Korean and other alphabetic scripts.” In: *International Journal on Document Analysis and Recognition* 25.4 (2022), pp. 245–263. ISSN: 14332833. URL: <https://ezproxy.mtsu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true%5C&db=edo%5C&AN=ejs60773937%5C&site=eds-live%5C&scope=site>.
- [54] V.U. Manh Tu and M. Beurton-Aimar. “PapyTwin net: a Twin network for Greek letters detection on ancient Papyri.” In: *ACM International Conference Proceeding Series*. HIP

- 2023 - Proceedings of the 2023 International Workshop on Historical Document Imaging and Processing. LaBRI - Bordeaux University, 2023, pp. 43-48 –48. URL: <https://ezproxy.mtsu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edselc&AN=edselc.2-52.0-85169412757&site=eds-live&scope=site>.
- [55] Andrew Mao et al. “Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing”. In: *First AAAI conference on human computation and crowdsourcing*. Citeseer. 2013.
- [56] Nitima Masla et al. “Reduction in Gas Cost for Blockchain Enabled Smart Contract”. In: *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON)*. 2021, pp. 1–6. DOI: 10.1109/GUCON50781.2021.9573701.
- [57] R. Mondal et al. “Handwritten English word recognition using a deep learning based object detection architecture.” In: *Multimedia Tools and Applications* 81.1 (2022), pp. 975-1000 –1000. ISSN: 15737721. URL: <https://ezproxy.mtsu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true%5C&db=edselc%5C&AN=edselc.2-52.0-85115121119%5C&site=eds-live%5C&scope=site>.
- [58] Ahmed Afif Monrat, Olov Schelén, and Karl Andersson. “A Survey of Blockchain From the Perspectives of Applications, Challenges, and Opportunities”. In: *IEEE Access* 7 (2019), pp. 117134–117151. DOI: 10.1109/ACCESS.2019.2936094.
- [59] Miguel Morin and Matthew Willetts. *Non-Determinism in TensorFlow ResNets*. 2020. DOI: 10.48550/ARXIV.2001.11396. URL: <https://arxiv.org/abs/2001.11396>.
- [60] Shapna Muralidharan and Heedong Ko. “An InterPlanetary File System (IPFS) based IoT framework”. In: *2019 IEEE International Conference on Consumer Electronics (ICCE)*. 2019, pp. 1–2. DOI: 10.1109/ICCE.2019.8662002.
- [61] G. Murtaza et al. “Ensembled deep convolution neural network-based breast cancer classification with misclassification reduction algorithms.” In: *Multimedia Tools and Applications*

- 79.25-26 (2020), pp. 18447-18479 –18479. ISSN: 15737721. URL: <https://ezproxy.mtsu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true%5C&db=edselc%5C&AN=edselc.2-52.0-85081546789%5C&site=eds-live%5C&scope=site>.
- [62] David F Nettleton, Albert Orriols-Puig, and Albert Fornells. “A study of the effect of different types of noise on the precision of supervised learning techniques”. In: *Artificial intelligence review* 33.4 (2010), pp. 275–306.
- [63] Joshua Ortiz Baco, Allyssa Guzman, and Albert A Palacios. “FromThePage Collection Owner User Study Report”. In: *Enabling and Reusing Multilingual Citizen Contributions in the Archival Record-NEH Grant Documentation* (2020).
- [64] Anita Pal and Dayashankar Singh. “Handwritten English character recognition using neural network”. In: *International Journal of Computer Science & Communication* 1.2 (2010), pp. 141–144.
- [65] Georgios Palaiokrassas et al. “Combining Blockchains, Smart Contracts, and Complex Sensors Management Platform for Hyper-Connected SmartCities: An IoT Data Marketplace Use Case”. In: *Computers* 10.10 (2021). ISSN: 2073-431X. DOI: 10.3390/computers10100133. URL: <https://www.mdpi.com/2073-431X/10/10/133>.
- [66] Chirag Patel, Atul Patel, and Dharmendra Patel. “Optical character recognition by open source OCR tool tesseract: A case study”. In: *International Journal of Computer Applications* 55.10 (2012), pp. 50–56.
- [67] Mykola Pechenizkiy et al. “Class noise and supervised learning in medical domains: The effect of feature extraction”. In: *19th IEEE symposium on computer-based medical systems (CBMS’06)*. IEEE. 2006, pp. 708–713.
- [68] Emmanouil Antonios Platanios et al. *Learning from Imperfect Annotations*. 2020. arXiv: 2004.03473 [cs.LG].
- [69] Nicola Reggiani. *Digital Papyrology I. Methods, Tools and Trends*. De Gruyter, 2017.

- [70] F. Regner, A. Schweizer, and N. Urbach. “NFTs in practice - Non-fungible tokens as core component of a blockchain-based event ticketing application.” In: *40th International Conference on Information Systems, ICIS 2019*. 40th International Conference on Information Systems, ICIS 2019. (1)FIM Research Center, University of Augsburg, 2019. URL: <https://ezproxy.mtsu.edu/login?url=https://search.ebscohost.com/login.aspx?direct=true%5C&db=edselc%5C&AN=edselc.2-52.0-85114902612%5C&site=eds-live%5C&scope=site>.
- [71] Sara Rouhani and Ralph Deters. “Performance analysis of ethereum transactions in private blockchain”. In: *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. 2017, pp. 70–74. DOI: 10.1109/ICSESS.2017.8342866.
- [72] R. A. Leibler S. Kullback. “On information and sufficiency”. In: *Annals of Mathematical Statistics* 22 (1951), pp. 79–86.
- [73] William Brent Seales et al. “From damage to discovery via virtual unwrapping: Reading the scroll from En-Gedi”. In: *Science Advances* 2.9 (2016), e1601247. DOI: 10.1126/sciadv.1601247. eprint: <https://www.science.org/doi/pdf/10.1126/sciadv.1601247>. URL: <https://www.science.org/doi/abs/10.1126/sciadv.1601247>.
- [74] Ashish Shrivastava et al. “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.
- [75] Robert Simpson, Kevin R. Page, and David De Roure. “Zooniverse: observing the world’s largest citizen science platform”. In: *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*. New York, New York, USA: ACM Press, 2014, pp. 1049–1054. ISBN: 9781450327459. DOI: 10.1145/2567948.2579215. URL: <http://dl.acm.org/citation.cfm?doid=2567948.2579215>.
- [76] Ray Smith. “An overview of the Tesseract OCR engine”. In: *Ninth international conference on document analysis and recognition (ICDAR 2007)*. Vol. 2. IEEE. 2007, pp. 629–633.

- [77] Ray W Smith. “History of the Tesseract OCR engine: what worked and what didn’t”. In: *Document Recognition and Retrieval XX*. Vol. 8658. International Society for Optics and Photonics. 2013, p. 865802.
- [78] Thea Sommerschild et al. “Machine Learning for Ancient Languages: A Survey”. In: *Computational Linguistics* 49.3 (Sept. 2023), pp. 703–747. ISSN: 0891-2017. DOI: 10.1162/coli_a_00481. URL: https://doi.org/10.1162/coli%5C_a%5C_00481.
- [79] Alexander Sorokin and David Forsyth. “Utility data annotation with Amazon Mechanical Turk”. In: *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 2008, pp. 1–8. DOI: 10.1109/CVPRW.2008.4562953.
- [80] Matthew I. Swindall et al. “Dataset Augmentation in Papyrology with Generative Models: A Study of Synthetic Ancient Greek Character Images”. In: *The 31st International Joint Conference on Artificial Intelligence. IJCAI-ECAI, 2022*.
- [81] Matthew I. Swindall et al. “Exploring Learning Approaches for Ancient Greek Character Recognition with Citizen Science Data”. In: *2021 17th International Conference on eScience (eScience)*. IEEE, 2021, pp. 128–137.
- [82] Choh-Man Teng. “A Comparison of Noise Handling Techniques.” In: *FLAIRS Conference*. 2001, pp. 269–273.
- [83] Ámbar Tenorio-Fornés et al. “Decentralizing science: Towards an interoperable open peer review ecosystem using blockchain”. In: *Information Processing & Management* 58.6 (2021), p. 102724. ISSN: 0306-4573. DOI: <https://doi.org/10.1016/j.ipm.2021.102724>. URL: <https://www.sciencedirect.com/science/article/pii/S0306457321002089>.
- [84] Martin Thoma. “The hasyv2 dataset”. In: *arXiv preprint arXiv:1701.08380* (2017).
- [85] Setthanun Thongsuwan et al. “ConvXGB: A new deep learning model for classification problems based on CNN and XGBoost”. In: *Nuclear Engineering and Technology* 53 (2020), pp. 522–531.

- [86] Jonathan Tremblay et al. “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.
- [87] Rohan Vaidya et al. “Handwritten character recognition using deep-learning”. In: *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. IEEE. 2018, pp. 772–775.
- [88] Omid Fatahi Valilai. *A Network based Blockchain ecosystem for peer review publication*. 2021. arXiv: 2112.11189 [cs.CY].
- [89] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. *YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information*. 2024. arXiv: 2402.13616 [cs.CV].
- [90] Amy Whitaker et al. “Art antiquities and blockchain: new approaches to the restitution of cultural heritage”. In: *International Journal of Cultural Policy* 27.3 (2021), pp. 312–329. DOI: 10.1080/10286632.2020.1765163. eprint: <https://doi.org/10.1080/10286632.2020.1765163>. URL: <https://doi.org/10.1080/10286632.2020.1765163>.
- [91] Nick White. “Training Tesseract for Ancient Greek OCR”. In: *Eüruzov* 28–29 (2012).
- [92] Alex C Williams et al. “Deja Vu: Characterizing Worker Reliability Using Task Consistency.” In: *HCOMP*. 2017, pp. 197–205.
- [93] Alex C Williams et al. “Identification of ancient greek papyrus fragments using genetic sequence alignment algorithms”. In: *2014 IEEE 10th international conference on e-science*. Vol. 2. IEEE. 2014, pp. 5–10.
- [94] Alex C. Williams et al. “A computational pipeline for crowdsourced transcriptions of Ancient Greek papyrus fragments”. In: *2014 IEEE International Conference on Big Data (Big Data)*. IEEE, 2014, pp. 100–105.

- [95] Alex C. Williams et al. “Proteus: A platform for born digital critical editions of literary and subliterary papyri”. In: *2015 Digital Heritage*. Vol. 2. 2015, pp. 453–456. DOI: 10.1109/DigitalHeritage.2015.7419546.
- [96] Gethin Williams and Steve Renals. “Confidence measures for hybrid HMM/ANN speech recognition”. In: *5th European Conference on Speech Communication and Technology (Eurospeech 1997)* (1997).
- [97] David H. Wolpert. “Stacked generalization”. In: *Neural Networks* 5.2 (1992), pp. 241–259. ISSN: 0893-6080. DOI: [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1). URL: <https://www.sciencedirect.com/science/article/pii/S0893608005800231>.
- [98] Xingquan Zhu and Xindong Wu. “Class noise vs. attribute noise: A quantitative study”. In: *Artificial intelligence review* 22.3 (2004), pp. 177–210.
- [99] Xiaojin Zhur and Zoubin GhahramaniГH. “Learning from labeled and unlabeled data with label propagation”. In: (2002).