A deep learning pipeline for the palaeographical dating of ancient Greek papyrus fragments

Graham West[3], Matthew I. Swindall[1], James H. Brusuelas[2], Francesca Maltomini[5], Marius Gerhardt[4], Marzia D'Angelo[6], John F. Wallin[1]

[1]Middle Tennessee State University, [2]University of Kentucky, [3]Meharry Medical College,
[4]Ägyptisches Museum und Papyrussammlung, Staatliche Museen zu Berlin,
[5]Università Degli Studi Firenze, [6]Università Degli Studi Di Napoli Federico II,

Correspondence: graham.west@mmc.edu

Abstract

In this paper we present a deep learning pipeline for automatically dating ancient Greek papyrus fragments based solely on fragment images. The overall pipeline consists of several stages, including handwritten text recognition (HTR) to detect and classify characters, filtering and grouping of detected characters, 24 character-level date prediction models, and a fragment-level date prediction model that utilizes the per-character predictions. A new dataset (containing approximately 7,000 fragment images and 778,000 character images) was created by scraping papyrus databases, extracting fragment images with known dates, and running them through our HTR models to obtain labeled character images. Transfer learning was then used to fine-tune separate ResNets to predict dates for individual characters which are then used, in aggregate, to train the fragment-level date prediction model. Experiments show that even though the average accuracies of character-level dating models is low, between 35%-45%, the fragment-level model can achieve up to 79% accuracy in predicting a broad, two-century date range for fragments with many characters. We then discuss the limitations of this approach and outline future work to improve temporal resolution and further testing on additional papyri. This image-based deep learning approach has great potential to assist scholars in the palaeographical analysis and dating of ancient Greek manuscripts.

1 Introduction

With the meteoric rise in deep learning technologies, many fields are rapidly adopting these tools and incorporating them into their workflow. Palaeography, the study of the handwriting in ancient and medieval manuscripts, is one such discipline that has benefited from these methods. Projects such as READ (https://eadh.org/projects/read) and DigiPal (https://eadh.org/projects/digipal), for example, have focused on applying these methods to issues of writer identification, layout analysis, and frameworks for digital palaeographical content, especially via handwritten text recognition (HTR). One important project of note is Ithaca (Assael et al., 2022) which, among other uses, can attribute a date range to an inscription. Our approach differs in that while Ithaca takes digital transcriptions as input, our pipeline relies solely on images. In this paper, we present our latest contribution to this research effort, consisting of a dataset and deep learning pipeline for dating ancient Greek papyrus fragments. This pipeline takes as input an image of an ancient Greek papyrus fragment and outputs a predicted date range. We describe the training methodologies used to create the various models constituting the pipeline as well as a number of performance metrics.

1.1 Palaeography and the Dating of Greek Papyri

The method for dating Greek papyri begins with manuscripts that can be accurately dated. This mostly pertains to documentary texts (letters, petitions, taxes, leases, etc.) that preserve their date of composition. Documentary papyri lacking a date can, of course, still be dated accurately, if they mention historical events or figures that generally locate them within a given century. Palaeographic analysis of these papyri, i.e. the study of the handwriting and the features of the characters preserved, is important for those papyri that are not dated, especially the immense number of literary and sub-literary papyri that never contain the date of their production. Those papyri must be assigned a date based on a meticulous comparison between the Greek characters they preserve and those in reliably dated papyrus manuscripts. Palaeographical handbooks containing human observations, discernible patterns, and even conjectured styles have thus been published and they con-

stitute the sources by which papyrologists assign dates to papyri (Roberts, 1955; Turner, 1987; Cavallo and Maehler, 2008).

In respect to the actual number of papyri preserved, however, these handbooks only contain a small number of manuscripts for comparison. It is not uncommon that an assigned date is later reevaluated and changed as more papyri are viewed and compared. The ability of deep learning methods to assist papyrologists in dating papyri by analyzing thousands of manuscript images holds great potential. To do so, this requires not only training models for the task at hand, but also creating a palaeography dataset to facilitate accurate dating. Previous work on the Ancient Lives Project provides a foundation for reaching these goals.

1.2 Ancient Lives & AL-ALL

Between 2011 and 2018, the Ancient Lives Project, a Zooniverse.org collaboration, enlisted the aid of citizen scientists in annotating the images of thousands of highly degraded, ancient Greek manuscripts (Williams et al., 2014). The project resulted in millions of annotations which were key to the creation of the first large-scale machine learning dataset for digital papyrology, AL-ALL (Swindall et al., 2021). This dataset consists of over 400,000 images of handwritten Greek characters on papyrus and has been successfully used to create various deep learning models. This dataset also includes images from fragments that are currently under papyrological study and have not been published. For a releasable dataset, a smaller, updated version of the published material, AL-PUBv2, has been made available at `https://www.kaggle.com/datasets/miswindall/al-pub-v2`.

1.3 HTR Models

The development of our dataset and pipeline for palaeographical dating rests on our two core HTR models, each of which perform a key HTR task: character detection and character classification.

1.3.1 Character Detection with YOLO

The character detection model is essentially an object detection model trained to locate Greek characters in images of papyri. Similar existing work refers to this process as *'character spotting'* (Majid and Smith, 2022; Mondal et al., 2022). To train this model, YOLOv5s (Ultralytics, 2023) was fine-tuned using 212 images of papyrus fragments from the Oxyhrynchus papyri (Bowman et al., 2007)

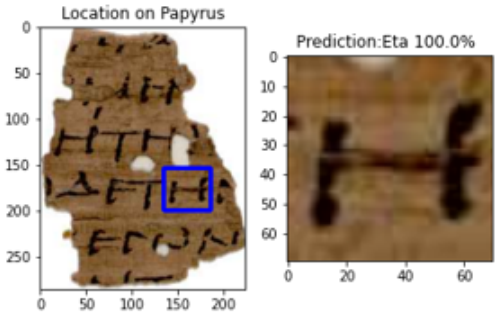

Figure 1: Example of character detection and classification using the HTR models. The YOLO model produces bounding boxes for each detected character. The bounded region is then cropped, resized, and given to the ResNet for classification.

containing 4097 character locations annotated during the Ancient Lives Project. YOLO is typically trained for multiple classes, but this model was fine-tuned to search for a single class: *Greek characters*. The model achieved precision and recall of 0.88 and 0.84, respectively, on the validation data, as well as validation box loss below 0.04. Further metrics are detailed in Figure 2.

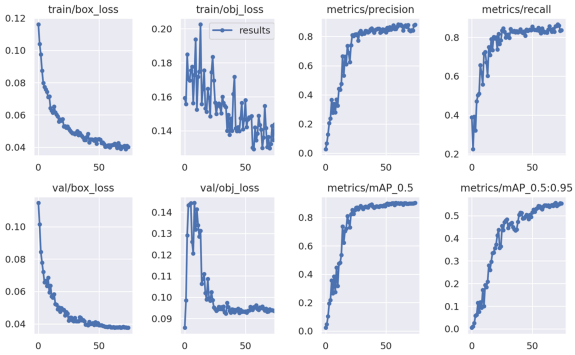

Figure 2: Training and validation metrics for the YOLO-based character detection model show that this model performs well on the task of locating Greek characters in images of damaged papryi.

1.3.2 Character Classification with ResNet

Our character classification model is a ResNet trained on the recently updated AL-ALLv2 dataset. The latest version of the dataset consists of 419,445 character images of all 24 characters in the ancient Greek alphabet, including the Lunate Sigma (C, ς) which typically replaces the more familiar Sigma (Σ, σ) in ancient papyri. This model achieved a training accuracy of 96.69% and a validation accuracy of 94.11%. Previous versions of this model

	4th-3rd BCE	2nd-1st BCE	1st-2nd CE	3rd-4th CE	5th-6th CE	Total
Fragments	299	729	3418	2002	570	7018

Table 1: The number of fragments from each century in the palaeography dataset.

	4th-3rd BCE	2nd-1st BCE	1st-2nd CE	3rd-4th CE	5th-6th CE	Total
α	6736	7582	36233	16548	4036	71135
β	302	219	1270	955	254	3000
γ	2914	2216	7765	3910	716	17521
ϵ	6508	5977	14132	7025	1795	35437
δ	903	1261	6679	2221	343	11407
ζ	354	288	2431	1158	284	4515
η	1741	3766	12873	7769	2105	28254
θ	299	729	3418	2002	570	7018
ι	4825	6951	24308	11785	4482	52351
κ	2886	2747	7343	5024	1828	19828
λ	971	2099	9766	3394	998	17228
μ	1362	2826	17883	7520	2573	32164
ν	4851	12339	28077	11578	3500	60345
ξ	62	93	653	410	106	1324
o	7011	13017	53709	24570	10146	108453
π	2293	3010	14785	7122	2043	29253
ρ	3389	4039	16570	9895	3414	37307
σ	4856	7732	28246	11490	4397	56721
τ	11823	14300	44502	23982	4910	99517
υ	2224	4727	12985	7165	1698	28799
ϕ	505	530	2652	1536	568	5791
χ	997	1944	6151	3557	1239	13888
ψ	135	108	499	265	53	1060
ω	1046	3176	20221	8912	2932	36287
Total	68993	101676	373151	179793	54990	778603

Table 2: The number of characters from each century in the palaeography dataset.

were released as a supplement to (Swindall et al., 2022), including models trained on a synthetically augmented version of AL-ALL in an effort to reduce sampling bias.

2 A Dataset for Palaeographical Dating

The development of the palaeographical dating pipeline necessitated the construction of a dataset containing images of papyrus fragments, their constituent characters, and their dates of composition. Three large papyrus databases were scraped for their fragment images and metadata (including dates of composition). The databases chosen were the Berlin Papyrus Database, Papiri della Società Italiana (PSI), and the Duke Papyrus Archive. For the first iteration of this dataset, we focused only on documentary papyri that preserve an exact date or are reliably dated within a range of a century or two. Since the format of the dates varied, the dates were processed and converted to a common format containing only the century or range of two centuries of composition. To reduce the difficulty of the dating task, we decreased the temporal resolution of the date classes from the one-century level to the two-century level: 4th-3rd BCE, 2nd-1st BCE, 1st-2nd CE, 3rd-4th CE, and 5th-6th CE (future work will consist of increasing the temporal resolution). The fragment images were then passed through our HTR models, thus obtaining cropped and classified images of each fragment's constituent characters. These character images are assigned the same date classes as the fragment on which they were written.

The character and fragment counts for each time-period in the dataset are detailed in Tables 1 and 2. As can be seen, we have examples of all 24 Greek characters from the 4th BCE to the 6th CE. There is also significant imbalance in both the characters and the dates. Concerning the characters, there are only 1,060 psis (Ψ, ψ) but 108,453 omicrons (O, o). Fortunately, transfer learning permits one to use a smaller dataset while still getting useful results since the majority of the layers have already been trained. Concerning the dates, 1st-2nd CE contains the largest number of characters and fragments (373,151 and 3,418, respectively) while 4th-3rd BCE contains the least (68,993 and 299, respectively).

It should be noted that this dataset is actually a subset of all that was scraped from the papyrus archives and run through the HTR pipeline. Many of the characters in the full dataset are of poor quality and were filtered out to create the final dataset. Filtering was done based on two factors: 1) image saturation entropy and 2) ResNet prediction entropy. The first is done in order to eliminate YOLO false positives which often consist of images with few ink pixels. Consequently, false positives of this type tend to have a low entropy in the distribution of their pixel saturation and can be reliably (though, not completely) eliminated by applying a simple threshold. The second filter removes images which

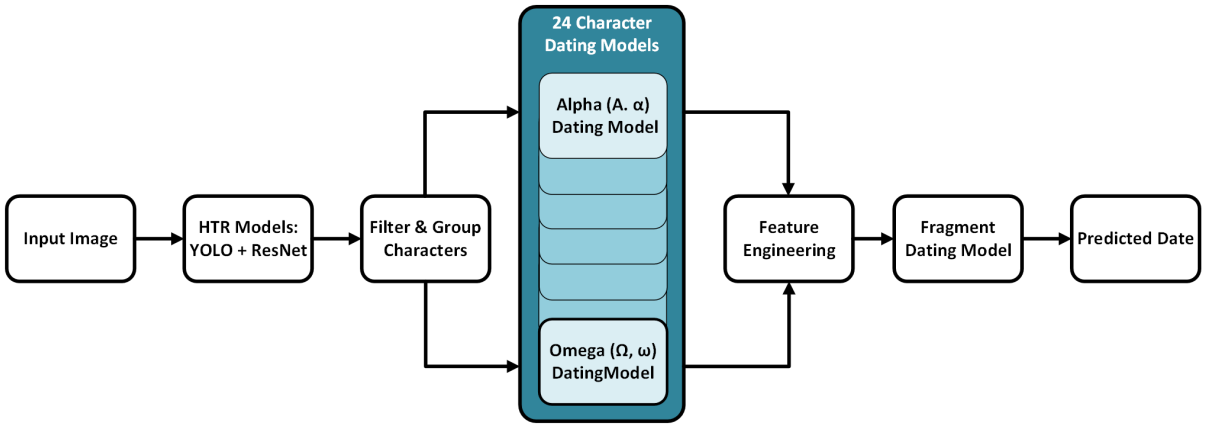

Figure 3: The palaeography pipeline performs HTR on the fragment image, obtaining images of individual characters. Poor character images (YOLO false positives, uncertain classifications, etc.) are filtered out. Remaining characters are grouped according to their character and sent to individual ResNet character dating models. These predicted character dates are then used as input to a fragment dating model.

were unreliably classified by the ResNet. Again, these can be fairly reliably eliminated by applying a threshold to the entropy of the ResNet's predicted class probabilities.

3 A Pipeline for Palaeographical Dating

Given the goal of producing a deep learning pipeline which can take an image of an ancient Greek papyrus fragment as input and output a predicted date of composition, the primary task is to determine the proper architecture for such a pipeline. Figure 3 depicts the chosen architecture, which consists of five core stages: HTR, filtering/grouping of characters, character dating, feature engineering, and fragment dating.

3.1 HTR, filtering, and grouping

The first step of the dating pipeline takes the input image and passes it through the HTR models, thus obtaining cropped and classified images of the fragment's constituent characters. The filtering steps described above are then applied to these character images to remove any unreliable samples. The characters are then grouped based on their character class (alpha, beta, etc.) before being sent to the next step of the pipeline.

3.2 Character dating models

Next, each group of characters is sent to another round of ResNet models which predict individual characters' date of composition. These models were developed via performing transfer learning on the ResNet discussed earlier. Naturally, the last two dense layers were retrained and the output layer was altered to have five output neurons (one per date class) instead of 24 (one per character).

While the transfer learning aspect was trivial, the data wrangling required to properly train these models was more complicated. Splitting the dataset into training and validation sets was done at the fragment level so that no fragment had constituent characters present in both sets of each individual ResNet model. As discussed above, there is a significant class imbalance with respect to the time periods (with 1st - 2nd CE having the overwhelming majority of samples). Thus, a great deal of balancing was performed. This was done by sampling the less frequent classes with replacement such that all classes had the same number of samples as the most frequent class. This was done separately for the training and validation sets. Additionally, data augmentation was performed using Keras's ImageDataGenerator so that there would not be identical copies of the images. The ranges for zoom, width shift, and height shift were all set to 0.1. No rotation was applied since the slant of characters is useful for determining their date of composition. This augmentation helps to increase the variability for less frequently occurring centuries which have many duplicated images due to sampling with replacement.

A custom loss function inspired by the Kolmogorov-Smirnov test was utilized since it is better suited for the ordinal nature of date labels than categorical cross entropy. Equation 1 illus-

	4th-3rd BCE	2nd-1st BCE	1st-2nd CE	3rd-4th CE	5th-6th CE
α	0	1	10	4	0
β	0	3	2	0	1

Table 3: An example of (one-hot encoded) raw features created from the output of the character dating models.

trates this loss function.

$$\text{loss} = \sum_{i=0}^{N-5} (t_i^c - p_i^c)^2 \tag{1}$$

Here, $N = 5$ is the number of classes and t_i^c, p_i^c are the true and predicted cumulative class probabilities, respectively. By comparing the cumulative probabilities, we can essentially form a metric which allows the ResNet's optimizer to take advantage of the fact that (given a true date of 5th-6th CE) a predicted date of 4th-3rd BCE is worse than 1st-2nd CE.

3.3 Feature engineering and fragment dating model

Once each character has received a predicted date, we then utilize these outputs to predict the date of the fragment as a whole. This is done via a simple dense neural network which outputs identical date classes as the ResNet in the previous step of the pipeline. Although the ordinal loss function described above worked well for the character models, it did not work well for the fragment model. Thus, categorical cross entropy was used.

For the fragment dating model's input, some clever feature engineering was done on the character dating model predictions. In what follows, all indices are assumed to start at zero. Let $C_k \in \{0, 1, 2, \cdots, 23\}$ (where k ranges over all the characters in a particular fragment) be the predicted character class. Also, let, p_{kj} (where $j = 0, \cdots, 4$ ranges over the number of date classes) be the predicted probability that character k belongs in date class j. Now, we construct the raw features X':

$$X'_{ij} = \sum_{k \ni (C_k = i)} p_{kj} \tag{2}$$

This sum adds the total probability for all α's, β's, etc. into separate columns. Table 3 shows a simplified example where, for the sake of simplicity, it is assumed that all of the probabilities are effectively one-hot encoded. These raw features are then processed with two more steps, obtaining the final features X:

$$X_{ij} = \frac{1 + X'_{ij}}{5 + \sum_j X'_{ij}} \tag{3}$$

First, Laplace's rule of succession is applied, adding a 1 to all entries of X' (we will explain the reason for this step below). Next, we normalize all of the rows (date-wise) by dividing by their sum. The normalization step ensures that all of the different fragments' feature values will be within the same range of values (between 0 and 1). The rule of succession is applied to preserve a kind of confidence that would otherwise be lost in the normalization step. Consider two fragments whose α rows are [0,0,1,0,0] and [0,0,5,0,0], respectively. Without application of the rule of succession, both columns would be normalized to [0,0,1,0,0]. Yet this is misleading since the second fragment has more α's predicted to be from 1st-2nd CE. We should want this increased confidence to be reflected in the features. Thus, by applying the rule of succession, we obtain [1,1,2,1,1] and [1,1,6,1,1] for the pre-normalization step and [0.166,0.166,0.333,0.166,0.166] and [0.1,0.1,0.6,0.1,0.1] for post-normalization. Notice how 0.6 is larger than 0.333, thus preserving the confidence due to having a larger number of characters.

Finally, as with the character-level data, there is also significant class imbalance at the fragment-level, though less severe. To combat this, we manually balance the training set of the fragment model by sampling with replacement.

4 Model Evaluation

In this section, we will discuss the performance of the models which comprise the palaeographical dating pipeline.

4.1 Character dating performance

Figure 4 shows the loss curves for each of the character dating models. Each model was trained for 200 epochs with a batch size of 256. A Keras callback was written which would store the model with the lowest validation loss in case of overfitting. The

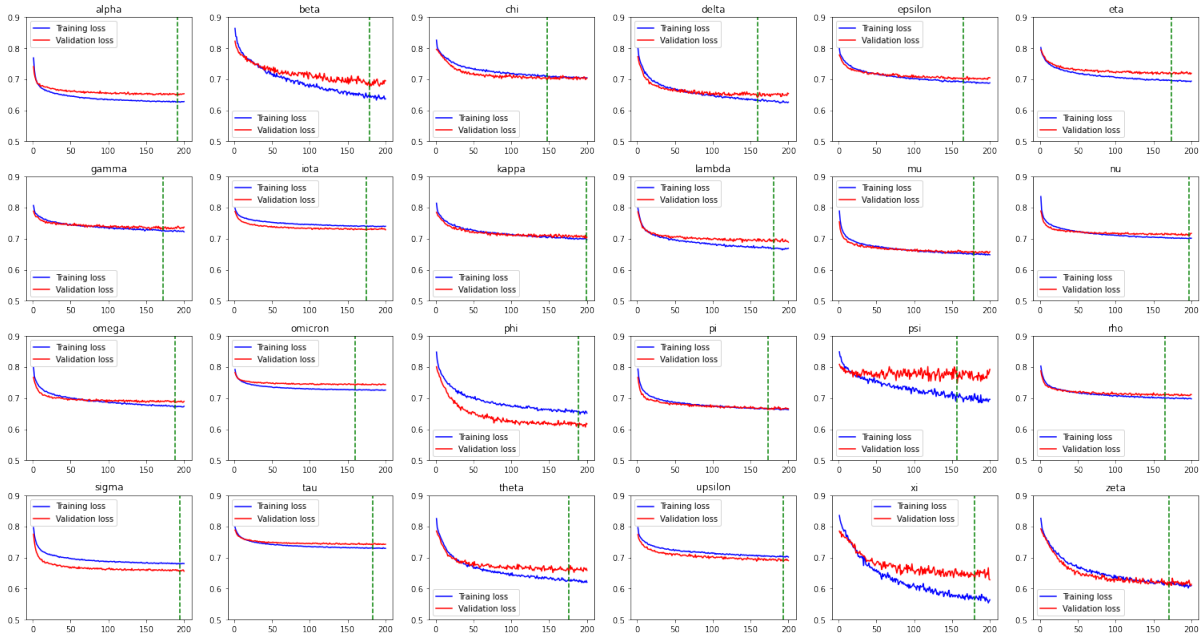

Figure 4: Loss curves for the character dating models. The vertical green line shows the epoch of lowest validation loss.

vertical green lines in the plots show when these best models were found. A range of behaviors can be seen across each of the model histories. Perhaps the most obvious is the range of optimal losses achieved. Note the difference in the values for omega and omicron. Also, some models converge in a fairly low number of epochs (such as alpha and omicron) while others have yet to converge after 200 epochs (beta and xi). This is likely due to the vast difference in sample size between the different characters. Characters with more samples are effectively trained more than those with fewer samples.

Figure 5 contains plots of the confusion matrices for each character dating model. To construct these statistics, we used Keras's ImageDataGenerator to create 1,000 augmented images per character/date combination and compare the character models' predictions on these images to the true date label (which is equivalent to that of the fragment from which the augmented character image was taken). The rows of the confusion matrices were then normalized for simplicity. All 24 models achieve overall accuracies between 35%-45%. These values are quite low, but we will see a significant increase once we see the fragment model's results.

There are several points to note about these confusion matrices. First, we can see that 3rd-4th CE is consistently the least accurately predicted date class across all characters. Second, the lower triangular portions of the matrices have a consistently higher value than the upper triangular portions. This is likely due to the fact that older handstyles can persist into the future but newer handstyles cannot retroject into the past. As such, we see a diffusion of the class probability as we move from earlier to later date classes (reading from top to bottom), causing confidence to decrease. Thirdly, we see a consistent trend in the final column which suggests that predictions of 5th-6th CE tend to have many false positives. This is likely due to the presence of a great variety of handstyles in this period, with papyri exhibiting character shapes present in older manuscripts.

4.2 Fragment dating performance

For our fragment dating model, we manually balanced the training set and performed 5-fold cross validation (five was chosen in order to keep the validation set from being too small). We present here the results obtained from a model trained on one of the folds. The accuracy of the fragment dating model depends heavily on the number of characters present in the fragment. As such, Figure 7 shows a boxplot of model accuracy (across the five folds). Note that the validation fragments have been grouped based on quartiles of the number of extant characters (which passed through the filter step). Additionally, we show in Figure 6 a set of confusion matrices for each of these groups.

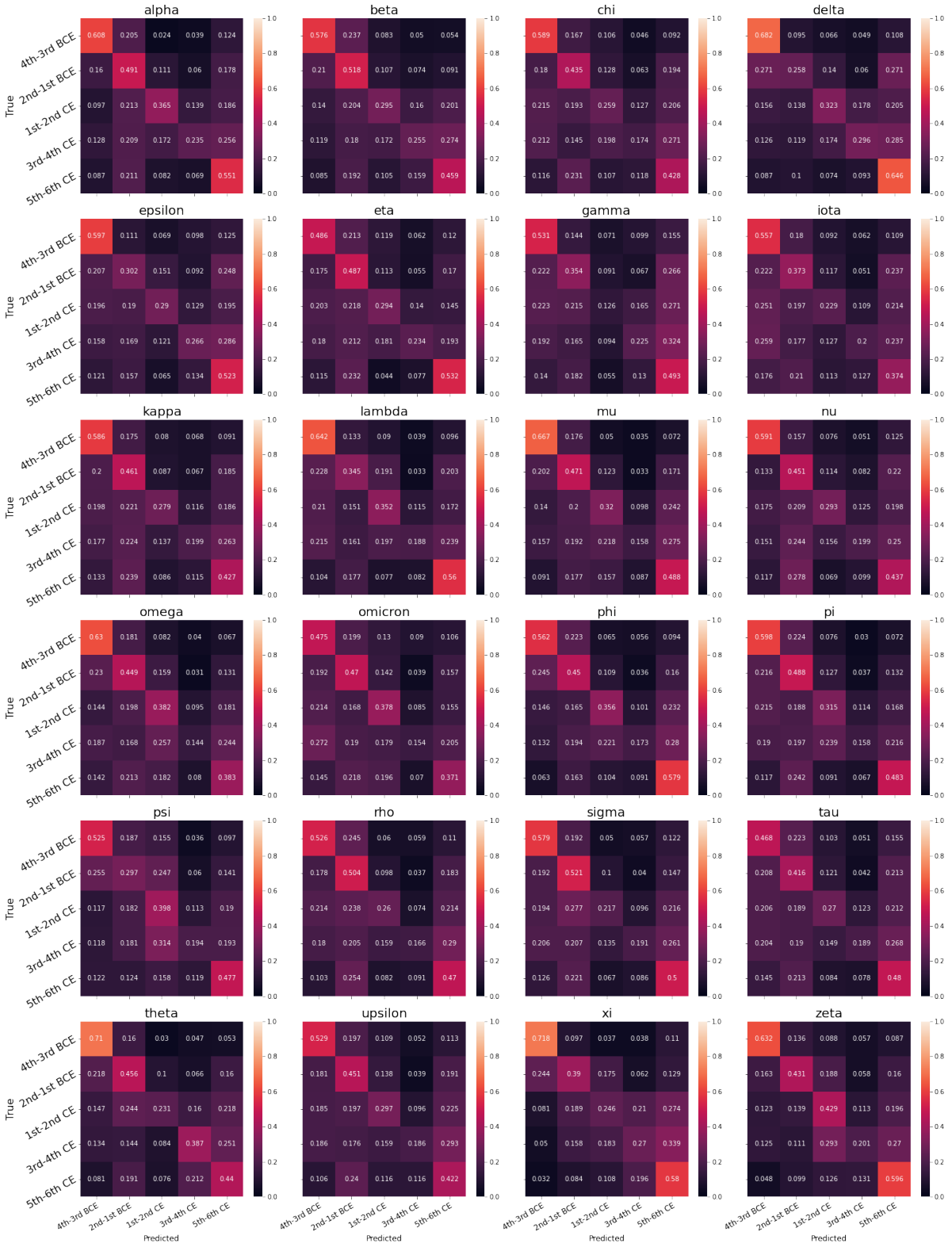

Figure 5: Confusion matrices for the character dating models.

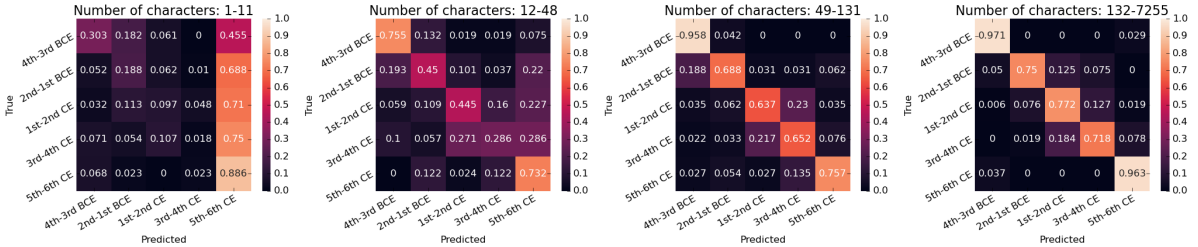

Figure 6: Confusion matrices for the fragment dating model. Each matrix contains only fragments within the specified range of number of characters.

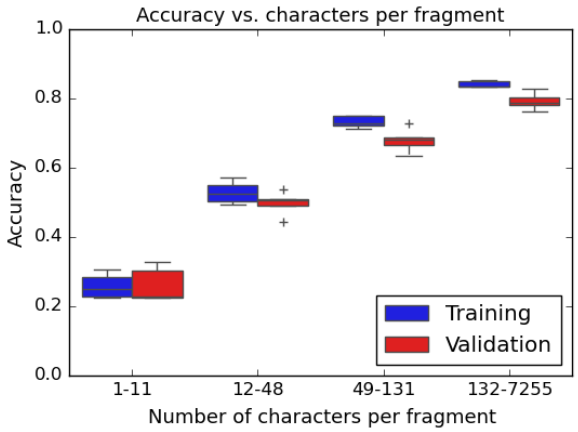

Figure 7: A box plot of the fragment dating model's accuracy of the five folds grouped based on quartiles of the number of characters in the fragments.

For the group of fragments with 1-11 characters, the average accuracy across the folds was 26% (only 6% above random chance). The confusion matrix corresponding to this group varied significantly across the folds, showing that dating fragments with a small number of characters is highly unreliable (as it is for humans). However, moving from left to right (increasing the number of characters in the fragment), we see increasing accuracy and a progressively more pronounced diagonal trend. A maximum accuracy of 79% (averaged over the folds) was achieved for fragments with between 132-7,255 characters. Thus, we can see the effect of something like the law of large numbers present in the fragment model. Although the character dating models are not very accurate, they are accurate enough that the most frequently predicted class will be correct (i.e., the probability of a correct prediction is significantly above chance). Therefore, the more characters contained within a fragment, the greater the probability of a correct date prediction.

5 Conclusions

While this research is still in its early stages, our results suggest that deep learning can perform the task of palaeographically dating ancient Greek papyri based solely on image input. This initial dataset and pipeline thus has the potential to further enhance the field of digital palaeography. Future work will pertain to increasing that temporal resolution and to leveraging the large number of individual characters in the dataset for analyzing handwriting features across time. Additionally, we plan to investigate the use of similar techniques for the location attribution of Greek papyri. More importantly, as noted above, this pipeline focuses on documentary papyri. Although these kinds of manuscripts constitute the ground truth for assigning dates to other papyri, literary and sub-literary papyri, which never preserve their date of production, have unique features of their own. How these models perform on and/or adapt to these manuscripts will also be a critical next step.

References

Y. Assael, B. Shillingford, M. Bordbar, N. de Freitas, T. Sommerschield, J. Pavlopoulos, M. Chatzipanagiotou, I. Androutsopoulos, and J. Prag. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603(7900):280–283 – 283.

Alan K. Bowman, R.A. Coles, N. Gonis, Dirk Obbink, and P. J. Parsons. 2007. *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.

Guglielmo Cavallo and Herwig Maehler. 2008. *Hellenistic Bookhands*. Walter De Gruyter.

Nishatul Majid and Elisa H. Barney Smith. 2022. Character spotting and autonomous tagging: offline handwriting recognition for bangla, korean and other alphabetic scripts. *International Journal on Document Analysis and Recognition*, 25(4):245 – 263.

R. Mondal, R. Sarkar, S. Malakar, and E.H. Barney Smith. 2022. Handwritten english word recognition using a deep learning based object detection architecture. *Multimedia Tools and Applications*, 81(1):975–1000 – 1000.

Colin H. (Colin Henderson) Roberts. 1955. *Greek literary hands, 350 B.C.-A.D. 400*. Oxford palaeographical handbooks. At the Clarendon Press.

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

Matthew I. Swindall, Timothy Player, Ben Keener, Alex C. Williams, James H. Brusuelas, Federica Nicolardi, Marzia D'Angelo, Claudio Vergara, Michael McOsker, and John F. Wallin. 2022. 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.

E. G. Turner. 1987. Greek manuscripts of the ancient world, second edition, revised and enlarged by p. j. parsons. bics supplement 46, london. *The Classical Review*.

Ultralytics. 2023. Comprehensive guide to ultralytics yolov5. https://docs.ultralytics.com/yolov5/. February 14, 2023.

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