APPENDIX-B: DATA AUGMENTATION

In need of more handwritten data and in connection with the original research theme of binarization (in Chapter 3), this thesis also explores tangential research of text-data augmentation. Due to the limited number of manuscripts that can be used on a particular experiment, different data augmentation techniques are studied and implemented within the scope of this thesis. Data augmentation techniques generate additional data by applying various transformations to the existing data. Most of the chapters of this thesis already took advantage of text-data augmentation. As further research, this thesis also studies the effects of character-Level data augmentation on style-based dating of historical manuscripts, including the Dead Sea Scrolls (DSS), early Aramaic manuscripts, and the French Royal Chancery collection (for DSS, the radiocarbon data is not used for the study of this chapter). By utilizing augmentation techniques, generating more realistic data to improve existing models is possible.

This chapter is added as an appendix of the thesis to enable future studies to incorporate the information from text-data augmentation of the DSS and French Royal Chancery collection (and thus other historical manuscripts) into various palaeographic research.

PROVENANCE
This chapter is an adaptation of two previously published articles:

AUTHOR CONTRIBUTIONS
Dhali conceived the idea for the experimental setup, monitored and moderated the analysis and experiments with other co-authors, and contributed to the final manuscript with the co-authors.
**Abstract**

One of the main questions paleographers aim to answer while studying historical manuscripts is when they were produced. Automatized methods provide tools for a more accurate and objective date estimation. Many of these methods are based on the hypothesis that handwriting styles change over periods. However, the sparse availability of digitized historical manuscripts poses a challenge in obtaining robust systems. The presented research extends previous research that explored the effects of data augmentation by elastic morphing on the dating of historical manuscripts. Linear Support Vector Machines were trained on k-fold cross-validation on textural and grapheme-based features extracted from the Medieval Palaeographical scale, early Aramaic manuscripts, the Dead Sea Scrolls, and volumes of the French Royal Chancery collection. Results indicate training models with augmented data can improve the performance of historical manuscript dating by 1% - 3% in cumulative scores but also diminish it. Data augmentation using elastic morphing can both enhance and decrease the date prediction of historical manuscripts and should be carefully considered. Moreover, further enhancements are possible by considering models tuned to the features and documents’ scripts.

### 9.1 Introduction

Historical manuscripts, i.e., handwritten accounts, letters, and similar documents, contain essential information to understand historical events. Paleographers study such documents, seeking to understand their social and cultural contexts. They specifically seek to identify the script(s), author(s), location, and production date of historical manuscripts, often by studying handwriting styles, writing materials, and contents. To do so, specific domain knowledge is needed. Additionally, these methods are time costly and lead to subjective estimations. Moreover, the repetitive physical handling of the documents contributes to further degradation.

Digitizing historical manuscripts by, e.g., scanning contributes to their preservation. It has allowed for the application of machine learning to help paleographers answer the aforementioned questions more objectively. Historical manuscript dating, in particular, can benefit from these automatized methods, as it can be necessary to resort to physical processes that can be destructive and have limited reliability.

Automatized methods are commonly based on the hypothesis that handwriting styles change over a period [111], allowing for the documents’ date estimation. These methods thus aim to estimate dates by identifying characteristics in handwriting specific to time periods.

Due to the limited availability of historical manuscripts, research has mainly focused on statistical feature-extraction techniques. These statistical methods extract the handwriting style by capturing attributes such as curvature or slant or representing the general character shapes in the documents [38]. However, for reliable results,
manuscripts need a sufficient amount of handwriting to extract the handwriting styles.

Both traditional and automatized methods must deal with data sparsity and the degradation of ancient materials; new data can only be obtained by digitizing or discovering more manuscripts. Data augmentation can be a potential solution to this issue.

Data augmentation is commonly used in machine learning to generate additional realistic training data from existing data to obtain more robust models. However, this must be done carefully to prevent the loss of information on handwriting styles. This occurs when using standard techniques, such as rotating or mirroring the images. On the other hand, data augmentation on the character level could generate realistic samples that simulate an author’s own variability in their handwriting.

In previous research [153], we investigated the effects of character-level data augmentation on the style-based dating of historical manuscripts. The current chapter extends this by introducing another medieval data set containing images from the French Royal Chancery (FRC) registers and applying the previously presented techniques.

Specifically, manuscript images taken from the Medieval Palaeographical Scale (MPS) collections, the FRC (HIMANIS) collection, the Bodleian Libraries of the University of Oxford, the Khalili collections, and the Dead Sea Scrolls were augmented with an elastic rubber-sheet algorithm [35]. The first collection, MPS, has medieval charters produced between 1300 and 1550 CE in four cities: Arnhem, Leiden, Leuven, and Groningen. Additionally, 840 images from 28 books of the FRC collection were used that date between 1302 and 1330 CE. Several early Aramaic, Aramaic, and Hebrew manuscripts were taken from the last three collections. Several statistical feature-extraction methods on the textural and character level were used to train linear Support Vector Machines (SVM) with only non-augmented images and with both non-augmented and augmented images.

9.2 RELATED WORKS

Scripts have their characteristics, and some features represent these better than others. Consequently, feature selection is one of the challenges in style-based dating. Digital collections of historical manuscripts are available in various languages and scripts, such as the Medieval Palaeographical Scale [186], and the Svenskt Diplomatariums huvudkartotek (SDHK) data sets², which are in Dutch and Swedish, respectively. Additionally, the early Aramaic and Dead Sea Scrolls collections [10] contain ancient

1 French Royal Chancery, Paris, Archives nationales de France, JJ35-JJ211, shared with us in the HIMANIS project (https://eadh.org/projects/himanis) [276]
2 https://sok.riksarkivet.se/SDHK
texts in Hebrew, Aramaic, Greek, and Arabic, dating from the fifth century BCE (Before the Common Era) until the Crusader Period (12th–13th centuries CE).

Statistical feature-extraction techniques are commonly divided into textural-based and grapheme-based features that capture information on the handwriting across an entire image on the textural and character level, respectively. One example of a statistical feature is the ‘Hinge’ feature, which captures a handwriting sample’s slant and curvature. It describes the joint probability distribution of the two hinged edge fragments and forms the basis of the family of ‘Hinge’ features [38]. The feature has been extended to, i.e., co-occurrence features QuadHinge and CoHinge, which emphasize the handwriting’s curvature and shape information, respectively [113].

Next to Hinge features, other textural features have been used for historical manuscript datings, such as curvature-free and chain code features [114, 263]. Additionally, features can be combined. The authors of [100], for instance, investigated combinations of Gabor filters [86], Local Binary Patterns [123], and Gray-level Co-occurrence Matrices [102], showing an improved performance compared to the individual features on the MPS data set.

Grapheme-based features represent character shapes. They are obtained through the extraction of graphemes from a set of documents and used to train a clustering method. The cluster representations form a codebook, from which a probability distribution of grapheme usage is computed for each document to represent the handwriting styles.

One grapheme-based feature is Connected Component Contours (CO3) [255], which describes the shape of a fully connected contour fragment. This was extended to Fraglets [38], which parts the connected contours based on their minima to deal with cursive handwriting. Other extensions are k stroke and k contour fragments, which partition CO3 in k stroke and contour fragments respectively [110]. In [120], Junclets was proposed, which represents the junctions in writing contours.

The MPS data set is often used in research on historical manuscript dating. He et al. specifically proposed several methods. In [111], dates were predicted through local and global Support Vector Regression on extracted Fraglets and Hinge features. Later they extended their work and proposed i.a. k contour fragments and Junclets. Moreover, the temporal pattern codebook was proposed in [109] for maintaining temporal information lost with the Self-Organising Map (SOM) commonly used for training codebooks. Finally, He et al. compared various statistical feature-extraction methods for historical manuscript dating in [115].

The MPS data set is often used because it is relatively clean. However, it is not representative of other collections of historical manuscripts. Due to aging, they are usually more degraded, as is the case with the Dead Sea Scrolls. An initial framework for the style-based dating of these manuscripts using both statistical and grapheme-based features was proposed in [66]. The Dead Sea Scrolls posed a
challenge due to them being fragmented and containing eroded ink traces. Moreover, there are few labeled manuscripts available.

Next to statistical hand-crafted features, deep learning approaches have been proposed. These have applied transfer learning, i.e., fine-tuning pre-trained neural networks using new data on a different task than they were initially intended for. Transfer learning has the potential for historical manuscript dating as it requires less data than standard deep learning methods. In [304], the Google ImageNet-network was fine-tuned on the SDHK collection. However, this was done using 11000 images, which is large for a data set of historical manuscripts. Additionally, a group of pre-trained neural networks was fine-tuned on the 3267 images from the MPS data set in [99]. Here, statistical methods were outperformed by the best-performing model.

While deep learning approaches show promising results, statistical methods are still relevant. To train a neural network, the manuscripts’ images need to be partitioned into patches, possibly leading to loss of information. To solve this problem, [99] ensured that each patch contained “3 to 4 lines of text with 1.5 to 2 words per line” to extract the handwriting style. While this was a solution for the MPS data set, it may not be for smaller and more degraded collections, such as the Dead Sea Scrolls. In contrast, statistical feature extraction does not require image resizing and considers the handwriting style over the entire image.

More recently, studies have fused statistical features with those extracted using neural networks. Hierarchical fusion was proposed by [3]. Specifically, they combined Gabor filters, Histogram of Orient Gradients, and Hinge with features extracted from a pre-trained ResNet through joint sparse representation. This method outperformed the individual statistical and ResNet features on a collection of Arabic manuscripts.

9.3 Methods

This section will present the dating model along with data description, image processing, and feature extraction techniques.

Data

The initial research used two datasets: MPS and EAA [153]. In connection to those, the current article extends to the FRC dataset. These three datasets are briefly presented in the following subsections.

9.3.0.1 Medieval Paleographical Scale (MPS) collection

The current research uses the MPS data set [108, 110, 111, 186]. Non-text content, such as seals, supporting backgrounds, color calibrators, etc., have been removed.
This data set consequently provides relatively clean images. Yet, it also includes more degraded images and images where parts of a seal or ribbon remain present. The data set is publicly available via Zenodo\(^3\).

The MPS data set contains 3267 images of charters collected from four cities signifying four corners of the medieval Dutch language area. Charters were commonly used to document legal or financial transactions or actions. Additionally, the manuscripts’ production dates have been recorded. For these charters, usually parchment and sometimes paper was used. Figure 9.1 shows an example image.

The charters date from 1300 CE to 1550 CE. Due to the evolution of handwriting being slow and gradual, documents from 11 quarter century key years with a margin of ± five years were included in the data set. Hence, the data set consists of images of charters from the medieval Dutch language area in the periods 1300 ± 5, 1325 ± 5, 1350 ± 5, up to 1550 ± 5. Table 9.1 contains the number of charters in each key year.

<table>
<thead>
<tr>
<th>key year</th>
<th>1300</th>
<th>1325</th>
<th>1350</th>
<th>1375</th>
<th>1400</th>
<th>1425</th>
<th>1450</th>
<th>1475</th>
<th>1500</th>
<th>1525</th>
<th>1550</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of charters</td>
<td>106</td>
<td>164</td>
<td>199</td>
<td>386</td>
<td>311</td>
<td>323</td>
<td>501</td>
<td>423</td>
<td>372</td>
<td>241</td>
<td>241</td>
</tr>
</tbody>
</table>

Figure 9.1: A document image from the Medieval Paleographical Scale (MPS) collection. Image taken from [153].

9.3.0.2 Early Aramaic and additional (EAA) manuscripts

In addition to the MPS data set, 30 images from the early Aramaic, Aramaic, and Hebrew manuscripts were used. For ease of refereeing to this second dataset, EAA is used in the rest of the article, even though EAA contains Aramaic and Hebrew in addition to early Aramaic scripts. A list of the EAA images used in this study can be found in the supplementary materials at the end of this chapter (see Table 9.7). These images are publicly available through the Bodleian Libraries, University of Oxford\(^4\).

---

\(^3\) https://zenodo.org/record/1194357\#YrL-UoEBy3l

\(^4\) https://digital.bodleian.ox.ac.uk/
the Khalili collections\textsuperscript{5}, and the Leon Levy Digital. For the selected manuscripts from the EAA dataset, the dates were directly inferred from dates or events recorded in the manuscripts (i.e., internally dated). Their dates span from 456 BCE to 133 CE. In addition, the data set contains several degraded manuscripts with missing ink traces or only two or three lines of text. An example image is shown in Figure 9.2.

Figure 9.2: An Early Aramaic (EA) manuscript from the Bodleian Libraries, the University of Oxford (Pell. Aram. I). Image taken from [153].

9.3.0.3 French Royal Chancery (FRC) registers

As a third data set, and extending our previous work in [153], the FRC data set was used [253]. This was a sub-set of 840 images of the French Royal Chancery collection, also known as Registres du Trésor des Chartes. A total of 28 volumes from this collection were used, or specifically, their first 30 pages. This was to avoid the increasing use of drawings, crossed-out texts, marginal notes, etc., that is often observed in later pages of manuscripts. The list of volumes used in the current research can be found in the supplementary materials at the end of this chapter (see Table 9.8).

Most volumes contained the years they were written in (the ground truth) on the first page. However, these pages also contained many artifacts and little handwriting. Hence, these pages were removed after the ground truths were determined. Additionally, empty pages were removed. This led to the removal of 48 images, resulting in a data set of 792 images. An example image from the FRC collection is shown in Figure 9.3.

\textsuperscript{5} https://www.khalilicollections.org/all-collections/aramaic-documents/
Preprocessing

9.3.0.4 Label refinement

**EAA** The set of images from the EAA collections did not contain sufficient samples for each year. Therefore, the samples were manually classified based on historical periods identified by historians⁶. The time periods and the corresponding number of samples are shown in Table 9.2.

The Persian Period contained two groups of samples spread apart for more than 30 years. Under the assumption that handwriting styles changed during this time, these samples were split into two periods: the Early and Late Persian Periods. These were based on the samples’ production years and not on defined historical periods. Images from the upper bound of the year range in Table 9.2 were included in the classes. The manuscripts from the Roman Period were excluded as they contained insufficient samples. The images were relabeled according to the median of their corresponding year ranges.

**FRC** The set of images from the FRC collection, in contrast to the EAA collections, contained sufficient samples for each year. However, many of the volumes

⁶ https://www.deadseascrolls.org.il/learn-about-the-scrolls/
Table 9.2: Division of EAA manuscripts across historical time periods. Note that these dates may not exactly be the same as defined by historians.

<table>
<thead>
<tr>
<th>Time period</th>
<th>Year range</th>
<th>Median year</th>
<th>Number of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Early Persian Period</td>
<td>540 BCE - 400 BCE</td>
<td>470 BCE</td>
<td>12</td>
</tr>
<tr>
<td>Late Persian Period</td>
<td>400 BCE - 330 BCE</td>
<td>365 BCE</td>
<td>11</td>
</tr>
<tr>
<td>Hellenistic Period</td>
<td>330 BCE - 65 BCE</td>
<td>198 BCE</td>
<td>5</td>
</tr>
<tr>
<td>Roman Period</td>
<td>65 BCE - 325 CE</td>
<td>195 CE</td>
<td>2</td>
</tr>
</tbody>
</table>

overlapped in the time ranges they were written or were too close to each other for distinguishable changes. For this reason, the images were manually classified into 4 categories, such that they contained sufficient variability to capture the handwriting style changes over a period. The images were relabeled according to the median of their corresponding year ranges. The resulting categories and their distributions are shown in Table 9.3.

Table 9.3: Division of FRC manuscripts.

<table>
<thead>
<tr>
<th>Year range</th>
<th>Median year</th>
<th>Number of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1302 - 1307</td>
<td>1305</td>
<td>171</td>
</tr>
<tr>
<td>1308 - 1314</td>
<td>1311</td>
<td>282</td>
</tr>
<tr>
<td>1316 - 1321</td>
<td>1319</td>
<td>225</td>
</tr>
<tr>
<td>1323 - 1329</td>
<td>1326</td>
<td>114</td>
</tr>
</tbody>
</table>

9.3.0.5 Data augmentation

To augment the data such that new samples simulate a realistic variability of an author’s handwriting, the ‘imagemorph’ program [35] was used. The program applies random elastic rubber-sheet transforms to the data through local non-uniform distortions, meaning that transformations occur on the components of characters. Consequently, the Imagemorph algorithm can generate a large number of unique samples. For the augmented data to be realistic, a smoothing radius of 8 and a displacement factor of 1 were used, measured in units of pixels. As images of the MPS data set required high memory, three augmented images were generated per image. Since the EAA data sets were small, 15 images were generated per image. The FRC images were each augmented 10 times for the same reason.
Figure 9.4: The binarized version of the image from Figure 9.1 with Otsu thresholding. Image taken from [153].

Figure 9.5: The binarized version of the image from Figure 9.2 using BiNet [67]. Image taken from [153].

9.3.0.6 **Binarization**

To extract only the handwriting, the ink traces in the images were extracted through binarization. This resulted in images with a white background representing the writing surface, and a black foreground representing the ink of the handwriting.

**MPS** Otsu thresholding [203] was used for binarizing the MPS images, as the MPS data set is relatively clean, and it has been successfully used in previous research with the data set. Otsu thresholding is an intensity-based thresholding technique where the separability between the resulting gray values (black and white) is maximized. Figure 9.4 shows Figure 9.1 after binarization.

**EAA** The EAA images were more difficult to binarize using threshold-based techniques. So, for the EAA images, we used BiNet from Chapter 3 [67]. Figure 9.5 shows Figure 9.2 after binarization.

**FRC** The FRC images were partly binarized using BiNet, and partly using adaptive thresholding. Here, the threshold value is a Gaussian-weighted sum of the
neighborhood, subtracted with a set constant C. The neighborhood size was set to (45x45), and a constant of \( C = 23 \) was used after experimentation. However, due to the resulting binarized documents containing many unwanted artifacts, further operations were applied for noise reduction. First, a median filter with a size of \( 5 \times 5 \) was applied. Next, dilation was applied with a kernel size of \( (7x7) \) to emphasize the locations of the handwriting. This dilated binarized image was overlayed with the initial binarized image using the bitwise AND operator. This removed the majority of the artifacts in the original binarized images. However, the ink contained many gaps in the letters, too. Hence, a closing operation and another median filter were applied, with kernel sizes \((2x2)\) and \((3x3)\), respectively. For simplicity, we will refer to this binarization process as adaptive thresholding.

A total of 398 images were binarized using adaptive thresholding, and 394 with BiNet. For 3 images \((2\) of category 1311 and \(1\) of 1319), the binarization quality was too low to obtain features for. This left 789 images of the FRC data set to be used for feature extraction. When using adaptive thresholding, the images were cropped before binarization was applied, as this enhanced the binarization quality. For the BiNet images, this was done after binarization to decrease the whitespace. From the images, 5% of the top and left sides were cut, and 10% and 20% were cut from the bottom and right sides. Moreover, images were further manually cropped when much noise was present on the outer sides. Examples of images binarized with BiNet and adaptive thresholding are shown in Figures 9.6 and 9.7, respectively.

---

Figure 9.6: Example image of the FRC collection binarized using BiNet model [67].
Feature extraction

The handwriting styles of manuscripts were described by five textural features and one grapheme-based feature. Since the MPS, FRC, and EAA data sets are written in different scripts and languages, features were chosen that are robust to this.

9.3.0.7 Textural features

Textural-based feature-extraction methods contain statistical information on handwriting in a binarized image by considering its texture. Textural-based features capture handwriting attributes like slant, curvature, and the author’s pen grip, represented in a probability distribution. More about textural features can be found in Subsection 2.3.1.2 of Chapter 2.

9.3.0.8 Grapheme-based features

Grapheme-based features are allograph-level features that partially or fully overlap with allographs in handwriting, described by a statistical distribution. The handwriting style is represented by the probability distribution of the grapheme usage across a document, computed with a common codebook. More about grapheme-based features can be found in Subsection 2.3.1.2 of Chapter 2.
9.3.0.9 Codebook

Previous research commonly used the Self-Organizing Map (SOM) [150] unsupervised clustering method to train the codebook [115]. Using SOM, however, means losing temporal information of the input patterns. The partially supervised Self-Organizing Time Map (SOTM) [247] maintains this information. In [109], SOTM showed an improved performance for a grapheme-based feature compared to SOM. Hence, the codebook was trained with SOTM.

SOTM trains sub-codebooks $D_t$ for each time period using the standard SOM [150], with handwriting patterns $\Omega(t)$ from key year $y(t)$. The key years for the MPS (in CE), EAA (in BCE), and FRC (in CE) data sets were defined as $y(t) = \{1300, 1325, 1350, \ldots, 1550\}$, $y(t) = \{470, 365, 198\}$, and $y(t) = \{1305, 1311, 1319, 1326\}$ respectively. The final codebook $D$ is composed of the sub-codebooks $D_t$: $D = \{D_1, D_2, \ldots, D_n\}$, with $n$ key years. To maintain the temporal information, the sub-codebooks are trained in ascending order. The initial sub-codebook $D_1$ is randomly initialized as no prior information exists in the data set. The succeeding sub-codebooks are initialized with $D_{t-1}$ and then trained. Algorithm 2 shows the pseudo-code obtained from [109].

The Euclidean distance measure was used to train the sub-codebooks, as it significantly decreased training times compared to the commonly used $\chi^2$-distance. Each sub-codebook was trained for 500 epochs, ensuring sufficient training occurred. The learning rate $a^*$ decayed from $a = 0.99$ following Equation (9.1). The sub-codebooks were trained on a computer cluster 7.

$$a^* = a \cdot \left(1 - \frac{\text{current epoch}}{\text{max epoch}}\right)$$  \hspace{1cm} (9.1)

A historical manuscript’s feature vector was obtained by mapping its extracted graphemes to their most similar elements in the trained codebook, computed via the Euclidean distance, and forming a histogram. Finally, the normalized histogram formed the feature vector.

Post-processing

The feature vectors of all features were small decimal numbers, varying between $10^{-2}$ and $10^{-6}$. To emphasize the differences between the feature vectors of feature type, the feature vectors were normalized between 0 and 1 based on the range of a feature’s feature vectors. A feature vector $f$ is scaled according to the following equations:

7 https://wiki.hpc.rug.nl/peregrine/start
Algorithm 2 SOTM procedure [109] [153]

\[ y \leftarrow 1 \]
randomly initialize \( D_t \)
train \( D_t \) using input patterns \( \Omega(t) \) by a standard SOM method

\[ \textbf{while } t \leq n \textbf{ do} \]
\[ t \leftarrow t + 1 \]
initialize \( D_t \) using \( D_{t-1} \)
train \( D_t \) using \( \Omega(t) \) by a standard SOM method
\[ \textbf{end while} \]
output \( D = D_1, D_2, ..., D_t, ..., D_n \)

\[
f_{\text{std}} = \frac{f - \text{min}(f)}{\text{max}(f) - \text{min}(f)}\tag{9.2}\]

\[
f_{\text{scaled}} = f_{\text{std}} \cdot (\text{max} - \text{min}) + \text{min}\tag{9.3}\]

Here, \text{max} and \text{min} are the maximum and minimum values across the whole set of feature vectors of a certain feature, and \text{max}(f) and \text{min}(f) are the maximum and minimum values of the feature vector \( f \) [207].

\[ \text{Dating} \]

9.3.0.10 Model

Historical manuscript dating can be regarded as a classification or a regression problem. As the MPS data set was divided into 11 non-overlapping classes (or key years), and the EAA and FRC data sets were partitioned into classes, it was regarded as a classification problem. Following previous research on the MPS data set [115], linear Support Vector Machines (SVM) were used for date prediction with a one-versus-all strategy.

9.3.0.11 Measures

The Mean Absolute Error (MAE) and the Cumulative Score (CS) are two commonly used metrics to evaluate model performance for historical manuscript dating. The MAE is defined as follows:

\[
MAE = \frac{\sum_{i=0}^{N-1} |y_i - \hat{y}_i|}{N}\tag{9.4}
\]
Here, \( y_i \) is a query document’s ground truth, and \( \bar{y}_i \) is its estimated year. \( N \) is the number of test documents. The CS is defined in [89] as

\[
CS = \frac{N_{e<\alpha}}{N} \cdot 100\%
\]  

(9.5)

The CS describes the percentage of test images that are predicted with an absolute error \( e \) no higher than a number of years \( \alpha \). At \( \alpha = 0 \) years, the CS is equal to the accuracy.

For all data sets, CS with \( \alpha = 0 \) years was used. Since paleographers generally consider an absolute error of 25 years acceptable, and the MPS set has key years spread apart by 25 years, CS with \( \alpha = 25 \) years was also used for the MPS data set.

9.3.0.12 Experiments

The MPS and FRC images were randomly split into a test and training set, containing 10% and 90% of the data, respectively. The EAA images were split into a test and training set of 5 and 23 images, respectively. 2 samples were included in the test set from classes 470 and 365 BCE each, and 1 image from class 198 BCE. The images were sorted based on their labels, and the first images of each class were selected for testing.

The models were tuned with stratified k-fold cross-validation for both data sets, as they were imbalanced. For the MPS and FRC data sets, \( k = 10 \). Since the training set of the EAA data set contained only four images from 198 BCE, \( k = 4 \) for this set. To obtain the models to be tuned to specific validation sets in each iteration of k-fold cross-validation, hyper-parameters were selected using the mean results of the cross-validation across six random seeds, ranging from 0 to 250 with steps of 50. The set of values considered for the hyper-parameters were \( 2^n, n = -7, -6, -5, ..., 10 \). During the process, the augmented images of those in the validation and test sets were excluded from the training.

Models were trained in two conditions. In the non-augmented condition, only non-augmented images were used, and in the augmented condition, both augmented and non-augmented images were used for training.

**Codebook** Different sub-codebook sizes can result in different model performances. Hence, various sub-codebook sizes were tested to obtain the optimal size for the Junclets feature. A codebook’s size is its total number of nodes, i.e., \( n_{\text{columns}} \cdot n_{\text{rows}} \). The full codebook \( D \) is the concatenation of the sub-codebooks \( D_i \), and thus its size will be \( \text{size}_{D} \cdot n_{\text{classes}} \). We considered the set of sub-codebook sizes \( s = \{25, 100, 225, 400, 625, 900\} \) with \( n_{\text{columns}} = n_{\text{rows}} \). These conditions were the same for all data sets. Since different codebook sizes result in different features, the sub-codebook sizes were determined based on the validation results of models trained on only non-augmented images.
Figure 9.8: MAE over sub-codebook size on non-augmented MPS data from 10-fold cross-validation. Also presented in [153].

Figure 9.9: CS with $\alpha = 25$ and $\alpha = 0$ years over sub-codebook size on non-augmented MPS data from 10-fold cross-validation. Also presented in [153].

9.4 RESULTS

Five textural and one grapheme-based feature were used to explore the effects of data augmentation on the style-based dating of historical manuscripts. Linear SVMs were trained in ‘non-augmented’ and ‘augmented’ conditions and tuned using 10-fold (MPS and FRC) and 4-fold (EAA) cross-validation. The models were tested on a hold-out set containing only non-augmented data. The test set of the MPS and FRC data sets contained 10% of the data, and that of the EAA dataset contained 17.8% (5 images) of the data.

The models were evaluated with the MAE and CS with $\alpha = 0$ years (i.e., accuracy). In addition, the MPS data set was evaluated with CS with $\alpha = 25$ years.
9.4 RESULTS

An optimal sub-codebook size needed to be selected before investigating Junclets. Results of k-fold cross-validation for sub-codebook sizes 25, 100, 225, 400, 625, and 900 were evaluated on non-augmented data. For the FRC data set, no graphemes were extracted for a total of 395 images. From these, 96, 138, 126, and 35 images were part of the 1305, 1311, 1319, and 1316 categories, respectively. Hence, the models for Junclets could only be trained on 394 images.

MPS Figures 9.8 and 9.9 show the MAE and CS for the MPS data set over sub-codebook size, respectively. The MAE shows a minimum sub-codebook size of 625. Moreover, CS with $a = 25$ and $a = 0$ years show a maximum at sub-codebook size 625. Therefore, Junclets features were obtained with sub-codebooks of size 625 on the MPS data.

EAA Figure 9.10 displays the MAE over the sub-codebook size on validation results for the EAA data. The MAE decreases until the sub-codebook size is 225,
after which it fluctuates. This is reflected in the CS with $\alpha = 0$ years (Figure 9.11). In addition, the standard deviations for the MAE and CS ($\alpha = 0$) appear the smallest here. Hence, a sub-codebook size of 225 was chosen for the EAA data.

**FRC** Figures 9.12 and 9.13 display the MAE and CS with $\alpha = 0$ years for the FRC data set over the sub-codebook size, respectively. The MAE shows a minimum sub-codebook size of 100. Likewise, the CS shows a maximum for this size. Consequently, Junclets features were obtained using a sub-codebook size of 100 for the FRC data set.

![Figure 9.12: MAE over sub-codebook size on non-augmented FRC data from 10-fold cross-validation.](image1)

![Figure 9.13: CS with $\alpha = 0$ years sub-codebook size on non-augmented FRC data from 10-fold cross-validation.](image2)
9.4 Results

Figure 9.14: MAE on MPS (unseen) test data across non-augmented and augmented conditions. Also presented in [153].

Figure 9.15: CS with $\alpha = 25$ years on MPS (unseen) test data across non-augmented and augmented conditions. Also presented in [153].

Table 9.4: k-fold cross-validation results on the MPS data set.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Non-augmented MAE</th>
<th>Augmented MAE</th>
<th>Non-augmented CS ($\alpha = 25$)</th>
<th>Augmented CS ($\alpha = 25$)</th>
<th>Non-augmented CS ($\alpha = 0$)</th>
<th>Augmented CS ($\alpha = 0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Junclets</td>
<td>10.93 ± 1.35</td>
<td>9.15 ± 1.30</td>
<td>90.35 ± 1.57</td>
<td>92.39 ± 1.54</td>
<td>73.53 ± 2.47</td>
<td>77.56 ± 2.36</td>
</tr>
<tr>
<td>TCC</td>
<td>9.47 ± 1.10</td>
<td>8.95 ± 1.20</td>
<td>91.37 ± 1.42</td>
<td>92.16 ± 1.49</td>
<td>77.00 ± 1.87</td>
<td>77.98 ± 2.15</td>
</tr>
<tr>
<td>DeltaHinge</td>
<td>20.09 ± 1.96</td>
<td>18.35 ± 1.61</td>
<td>81.59 ± 2.01</td>
<td>83.00 ± 1.85</td>
<td>61.61 ± 2.36</td>
<td>63.91 ± 2.14</td>
</tr>
<tr>
<td>QuadHinge</td>
<td>5.76 ± 0.98</td>
<td>5.74 ± 0.98</td>
<td>95.38 ± 1.19</td>
<td>95.44 ± 1.20</td>
<td>84.65 ± 1.99</td>
<td>84.53 ± 2.01</td>
</tr>
<tr>
<td>CoHinge</td>
<td>6.81 ± 0.99</td>
<td>6.48 ± 0.91</td>
<td>94.32 ± 1.26</td>
<td>94.59 ± 1.24</td>
<td>82.13 ± 1.99</td>
<td>82.64 ± 2.00</td>
</tr>
<tr>
<td>Hinge</td>
<td>11.55 ± 1.49</td>
<td>11.28 ± 1.42</td>
<td>89.42 ± 1.8</td>
<td>89.36 ± 1.81</td>
<td>73.60 ± 2.57</td>
<td>73.76 ± 2.59</td>
</tr>
</tbody>
</table>
Figure 9.16: CS with $\alpha = 0$ years on MPS (unseen) test data across non-augmented and augmented conditions. Also presented in [153].

9.4.2 Augmentation

9.4.2.1 MPS

Figure 9.14 shows the MAE per feature across augmented and non-augmented conditions. All features except TCC display a decrease in the augmented condition compared to the non-augmented condition. TCC displayed an increase.

Figure 9.15 shows the CS with $\alpha = 25$ years for both non-augmented and augmented conditions. All features display an increase in the augmented condition compared to the non-augmented condition except for TCC and Hinge, which showed a decrease. Additionally, Junclets did not change in performance across conditions.

As displayed in Figure 9.16, all features showed an increase in CS with $\alpha = 0$ years in the augmented condition compared to the non-augmented condition except DeltaHinge. DeltaHinge feature showed no change in performance across conditions on test data.

These results denote an overall increase in performance for all features, with the exception of TCC. However, the changes in performances are small. This is reflected in the validation results shown in Table 9.4, where changes between the non-augmented and augmented conditions are insignificant. This is indicated by means of the measures in augmented conditions overlapping with the ranges denoted by the standard deviations corresponding to the non-augmented conditions.

9.4.2.2 EAA Collections

Figures 9.17 and 9.18 show the MAE and CS with $\alpha = 0$ years across all features for the EAA data set. Performance increased for Junclets in the augmented condition compared to the non-augmented condition, denoted by decreases in MAE and
Table 9.5: k-fold cross-validation results on the EAA data set.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Non-augmented MAE</th>
<th>Augmented MAE</th>
<th>Non-augmented CS (a=0)</th>
<th>Augmented CS (a=0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Junclets</td>
<td>41.62 ± 14.06</td>
<td>40.72 ± 17.75</td>
<td>73.05 ± 9.04</td>
<td>70.14 ± 10.69</td>
</tr>
<tr>
<td>TCC</td>
<td>47.26 ± 14.68</td>
<td>57.92 ± 20.37</td>
<td>71.94 ± 12.24</td>
<td>65.42 ± 13.87</td>
</tr>
<tr>
<td>DeltaHinge</td>
<td>46.72 ± 11.10</td>
<td>45.54 ± 27.67</td>
<td>65.83 ± 10.42</td>
<td>75.56 ± 14.74</td>
</tr>
<tr>
<td>QuadHinge</td>
<td>38.97 ± 14.19</td>
<td>29.92 ± 17.07</td>
<td>76.67 ± 8.50</td>
<td>82.08 ± 10.22</td>
</tr>
<tr>
<td>CoHinge</td>
<td>48.18 ± 10.82</td>
<td>38.95 ± 22.30</td>
<td>64.44 ± 10.21</td>
<td>75.28 ± 14.17</td>
</tr>
<tr>
<td>Hinge</td>
<td>33.84 ± 14.82</td>
<td>26.17 ± 22.20</td>
<td>79.86 ± 9.1</td>
<td>84.44 ± 12.72</td>
</tr>
</tbody>
</table>

increases in accuracy. QuadHinge also showed an increase in performance, indicated by the decrease in MAE in the augmented condition. A decrease in performance for TCC, DeltaHinge, and Hinge features is denoted by an increase in MAE and a reduction in accuracy. CoHinge displayed no change across conditions on the test set.

The validation results do not reflect these test results (Table 9.5), as Junclets and TCC displayed a decrease in performance with a reduction in mean MAE and a decrease in mean accuracy in the augmented condition compared to the non-augmented condition. However, DeltaHinge, QuadHinge, CoHinge, and Hinge showed the opposite. Additionally, standard deviations increased significantly in the augmented condition compared to the non-augmented condition.

Figure 9.17: MAE on EAA (unseen) test data across non-augmented and augmented conditions. Also presented in [153].

9.4.3 FRC

Figures 9.19 and 9.20 show the MAE and CS with \( a = 0 \) years across all features. In contrast to the MPS and EAA data sets, these results show a decrease in model
Figure 9.18: CS with $a = 0$ years on EAA (unseen) test data across non-augmented and augmented conditions. Also presented in [153].

Figure 9.19: MAE on FRC (unseen) test data across non-augmented and augmented conditions.

Table 9.6: k-fold cross-validation results on the FRC data set.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Non-augmented MAE</th>
<th>Augmented MAE</th>
<th>Non-augmented CS ($a=0$)</th>
<th>Augmented CS ($a=0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Junclets</td>
<td>$2.05 \pm 0.73$</td>
<td>$1.85 \pm 0.68$</td>
<td>$81.4 \pm 6.01$</td>
<td>$82.8 \pm 5.9$</td>
</tr>
<tr>
<td>TCC</td>
<td>$2.66 \pm 0.57$</td>
<td>$2.79 \pm 0.54$</td>
<td>$74.05 \pm 4.56$</td>
<td>$70.35 \pm 5.27$</td>
</tr>
<tr>
<td>DeltaHinge</td>
<td>$3.7 \pm 0.67$</td>
<td>$4.49 \pm 0.71$</td>
<td>$63.64 \pm 5.51$</td>
<td>$55.39 \pm 5.65$</td>
</tr>
<tr>
<td>QuadHinge</td>
<td>$1.68 \pm 0.51$</td>
<td>$1.72 \pm 0.45$</td>
<td>$82.51 \pm 4.31$</td>
<td>$81.81 \pm 4.06$</td>
</tr>
<tr>
<td>CoHinge</td>
<td>$1.59 \pm 0.47$</td>
<td>$1.3 \pm 0.41$</td>
<td>$84.62 \pm 3.97$</td>
<td>$85.98 \pm 3.65$</td>
</tr>
<tr>
<td>Hinge</td>
<td>$2.46 \pm 0.61$</td>
<td>$2.75 \pm 0.51$</td>
<td>$75.11 \pm 5.33$</td>
<td>$71.75 \pm 4.52$</td>
</tr>
</tbody>
</table>

performance for augmented images across all features except Junclets. The Junclets
Figure 9.20: CS with $\alpha = 0$ years on FRC (unseen) test data across non-augmented and augmented conditions.

A feature showed an increase in performance, as denoted by the reduction of test MAE and increase in test CS with $\alpha = 0$.

These results are partially reflected in the validation results, shown in Table 9.6. The test results are mainly in line with the validation results, with the exception of the CoHinge feature, which displayed an overall increase in performance. However, this increase and that of Junclets appear insignificant due to the measures’ means and corresponding standard deviations overlapping.

### 9.4.3.1 Significance

Statistical tests (ANOVA, [57]) were performed to see if the results showed significant improvements for all features.

**MPS** For the MPS data, the results from Junclets and DeltaHinge features were statistically significant for both MAE and CS, with p-values much smaller than 0.005.

**EAA** The results on the EAA data did not show any significance for any of the feature extraction techniques.

**FRC** The results on the FRC data set showed a statistically significant difference ($p < 0.0005$) between the augmented and non-augmented conditions for the accuracy of the Hinge feature and the MAE of the CoHinge feature. As aforementioned, the Hinge feature displayed a decrease in model performance, while CoHinge displayed an increase in performance. The remaining features did not show statistically significant differences.
9.5 DISCUSSION

The current research explored how the style-based dating of historical manuscripts is affected by character-level data augmentation using images from the MPS, FRC, and EAA collections affect the style-based dating of historical manuscripts. Images were binarized and augmented through elastic morphing [35]. Linear SVMs were trained on five textural features and one grapheme-based feature. Experiments were conducted to determine optimal sub-codebook sizes for the grapheme-based Junclets feature, as it was obtained through mapping extracted junctions to codebooks trained with SOTM [247]. The SVMs were then trained on only non-augmented images and both non-augmented and augmented images and evaluated through the MAE and CS with $\alpha$-values of 0 and 25 years.

Key findings

MPS

Linear SVMs trained on test MPS data in the augmented condition showed an overall increased performance compared to the non-augmented condition for all features except TCC, which showed a decrease in performance. The changes in validation results were statistically significant for Junclets and DeltaHinge. Changes in validation results were insignificant for the remaining features, with the ranges of the standard deviations and mean overlapping across conditions.

The MPS images require much computer memory, and consequently, acquiring the features and models is time costly. Specifically, obtaining the Junclets features required several days. Hence, we generated only three augmented images per MPS image. Were more images generated, results might have shown a clearer or different picture of the effects of data augmentation on the style-based dating of historical manuscripts.

Another possible explanation for the aforementioned small changes in performance is that MPS images were augmented before binarization. Images were augmented using the imagemorph program, which applies a Gaussian filter over local transformations. Applying this before binarization leads to increased influence from the image’s background and, consequently, to less severe distortions. The distortions in MPS images were noticeable, yet they might have been too light to produce samples with natural within-writer variability. Whether this significantly affected the results is uncertain and should be considered in the future.

EAA Collections

Test results of models trained on EAA images showed increased performance in the augmented condition compared to the non-augmented condition for Junclets
and QuadHinge. Models for TCC, DeltaHinge, and Hinge showed a decrease, and CoHinge showed no change in performance. This is not reflected in the validation results (Table 9.5). Instead, these results show a decreased performance for Junclets and TCC in the augmented condition compared to the non-augmented condition and an increase in performance for all other features.

An explanation for these results is the increase in standard deviations across all features for models trained in the augmented conditions compared to models trained in the non-augmented condition. This indicates models were less robust to new data in the augmented condition, which may have led to diverging test results. Moreover, due to the small size of the data set, overfitting likely occurred, as suggested by the differences between test results and validation results within the conditions (e.g., QuadHinge).

The models trained on EAA data could have been less robust in the augmented condition due to the extracted features possibly following non-linear patterns. The SVMs, however, used linear kernels. While linear SVMs worked well on the MPS images, these are written in Roman script. The EAA data is written in Hebrew. Moreover, handwriting in different geographical locations might change at different rates. Data augmentation could have emphasized the non-linear patterns in Hebrew, making the linear SVMs too rigid.

FRC

Textural features showed a decrease in performance on test results, while the grapheme-based Junclets feature showed an increase in performance. These findings are in agreement with the validation results, except for CoHinge. CoHinge showed a statistically significant difference in accuracy between non-augmented and augmented conditions, indicating its increase in performance observed in the validation results was significant. On the other hand, the Hinge feature showed a decrease in performance and displayed a statistically significant difference between the conditions.

While the MPS and FRC data sets are both written in the Roman script, the results are opposing. The results on the MPS data set showed an overall increase in model performance when images were augmented, and those on the FRC images showed an overall decrease. One of the main differences between these data sets is in the pre-processing. As with the EAA images, FRC images were augmented after binarization. Additionally, the binarization of FRC images used two different methods, of which the adaptive thresholding method often resulted in thinned or broken characters. The ‘n’ and ‘m’ characters, in particular, were often disrupted.

Another possible cause for the decrease in the performance of models trained on augmented FRC data is that the augmentation method might introduce too much variety within time periods. While the elastic morphing algorithm is designed to
mimic within-writer variability, the relatively low quality of some of the binarized images might have caused too much distortion on the textural level.

9.5.1  Future research

Characteristics differ between scripts, possibly leading to differing distributions of extracted features. Similarly, different attributes of handwriting are captured by individual features. Consequently, features might follow differing temporal trends (non-linear or linear). The current research only used linear models and did not consider the differences between both features and scripts. This could have led to a decrease in model performance for augmented data. Therefore, non-linear models should be studied to optimize performance on individual features and scripts.

Often data sets do not contain equal numbers of manuscripts or writing samples from time periods. The current chapter did not consider data balancing, for which data augmentation is also a common approach. Rather than generating equal numbers of samples through data augmentation for each image, choosing the number of samples to generate with a focus on data balancing can lead to fewer misclassifications of minority classes and therefore improve overall model performances.

Some data sets purposed for historical manuscript dating can contain classes in which most samples are produced by one writer. This can pose an issue as models trained on these data risk to lean to distinguish writer-specific traits in handwriting rather than those representatives of a particular period or year. Hence, it is important to simulate variability between writers and within time periods when augmenting data. This might lead to more robust models than when simulating a realistic within-writer variability.

Previous research showed deep learning approaches outperformed statistical features on the MPS data set. Moreover, a fusion of statistical features and deep learning approaches has been shown to improve performance [3]. It would be interesting to investigate whether data augmentation might positively affect historical manuscript dating using these methods on smaller and more heavily degraded documents, such as the EAA collections. Additionally, applying such approaches to individual characters similar to grapheme-based statistical features might bypass the issues of limited data and loss of information due to image resizing.
ACKNOWLEDGMENTS

The study for this appendix (from the original article) collaborated with several research outcomes from the European Research Council (EU Horizon 2020) project: The Hands that Wrote the Bible: Digital Palaeography and Scribal Culture of the Dead Sea Scrolls (HandsandBible 640497), principal investigator: Mladen Popović. Furthermore, for the high-resolution, multi-spectral images of the Dead Sea Scrolls, we (the authors from the original article) are grateful to the Israel Antiquities Authority (IAA), courtesy of the Leon Levy Dead Sea Scrolls Digital Library; photographer: Shai Halevi. Additionally, we express our gratitude to the Bodleian Libraries, the University of Oxford, the Khalili collections, and the Staatliche Museen zu Berlin (photographer: Sandra Steib) for the early Aramaic images. We also thank Petros Samara for collecting the Medieval Paleographical Scale (MPS) dataset for the Dutch NWO project. For the inventories of the French Royal Chancery (FRC) registers, we thank the HIMANIS project (Historical MANuscript Indexing for user-controlled Search); photography credits: Paris, Archives Nationales. Finally, we thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high-performance computing cluster.
## Table 9.7: The list of EAA images used in this research.

<table>
<thead>
<tr>
<th>A6_11R</th>
<th>A6_8</th>
<th>NS_A1r</th>
</tr>
</thead>
<tbody>
<tr>
<td>A6_12R</td>
<td>B3_1</td>
<td>NS_A2r</td>
</tr>
<tr>
<td>A6_13R</td>
<td>IA01</td>
<td>NS_A4r</td>
</tr>
<tr>
<td>A6_14</td>
<td>IA03</td>
<td>NS_A5r</td>
</tr>
<tr>
<td>A6_15</td>
<td>IA04</td>
<td>NS_A6r</td>
</tr>
<tr>
<td>A6_16</td>
<td>IA06</td>
<td>NS_C1r</td>
</tr>
<tr>
<td>A6_3</td>
<td>IA17</td>
<td>NS_C4r</td>
</tr>
<tr>
<td>A6_4</td>
<td>IA21</td>
<td>WDSP1_1</td>
</tr>
<tr>
<td>A6_5</td>
<td>Mur24_1</td>
<td>WDSP2</td>
</tr>
<tr>
<td>A6_7</td>
<td>Mur24_2</td>
<td>Maresha</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ostracon</td>
</tr>
</tbody>
</table>

## Table 9.8: The list of the volumes from FRC/HIMANIS used in this research.

<table>
<thead>
<tr>
<th>FRCHANJJ_JJ035</th>
<th>FRCHANJJ_JJ043</th>
<th>FRCHANJJ_JJ050</th>
<th>FRCHANJJ_JJ059</th>
</tr>
</thead>
<tbody>
<tr>
<td>FRCHANJJ_JJ036</td>
<td>FRCHANJJ_JJ044</td>
<td>FRCHANJJ_JJ052</td>
<td>FRCHANJJ_JJ060</td>
</tr>
<tr>
<td>FRCHANJJ_JJ037</td>
<td>FRCHANJJ_JJ045</td>
<td>FRCHANJJ_JJ053</td>
<td>FRCHANJJ_JJ061</td>
</tr>
<tr>
<td>FRCHANJJ_JJ038</td>
<td>FRCHANJJ_JJ046</td>
<td>FRCHANJJ_JJ054</td>
<td>FRCHANJJ_JJ062</td>
</tr>
<tr>
<td>FRCHANJJ_JJ040</td>
<td>FRCHANJJ_JJ047</td>
<td>FRCHANJJ_JJ055</td>
<td>FRCHANJJ_JJ064</td>
</tr>
<tr>
<td>FRCHANJJ_JJ041</td>
<td>FRCHANJJ_JJ048</td>
<td>FRCHANJJ_JJ056</td>
<td>FRCHANJJ_JJ064</td>
</tr>
<tr>
<td>FRCHANJJ_JJ042</td>
<td>FRCHANJJ_JJ049</td>
<td>FRCHANJJ_JJ058</td>
<td>FRCHANJJ_JJ066</td>
</tr>
</tbody>
</table>