Towards a Dataset for Paleographic Details in Historical Torah Scrolls

Laura Frank[a], Germaine Götzelmann[b] and Danah Tonne[c]
Karlsruhe Institute of Technology, Germany
{*laura.frank, germaine.goetzelmann, danah.tonne*}*@kit.edu*

Keywords: Dataset, Hebrew Letter Decoration, Labeling, Image Classification.

Abstract: Historical textual witnesses used in religious practice have been a research interest for a long time but still remain mysterious. In particular, medieval Torah scrolls show irregularities in the scripture, whose intentions have not yet been revealed. In this paper, we assess the analysis of letter decorations from the perspective of computer vision and investigate the possibilities of extending qualitative research in Jewish Studies by quantitative analysis methods of computer science. For this purpose, we introduce a methodological approach to obtain a reproducible and extensible dataset of Hebrew letters and present a set of labels usable for various machine learning tasks. The evaluation of the dataset in terms of decoration recognition shows promising prediction accuracy rates of up to 90% with standard transfer learning methods and architectures.

1 INTRODUCTION

Torah scrolls are an essential component of Jewish religion and ceremonies. Writing a Torah scroll, e.g. the Pentateuch text, is considered extremely holy and remains a manual endeavor to this day. Although the Torah scroll has a rich past and has been subject to research for centuries, the object and intention of writing still hold mysteries to investigate (Perani, 2022).

One example are decorations that are added to the Hebrew letters. These can either occur as so-called *tagin* ('little crowns'), most often situated on top of the letter, or as other so-called *otiyyot meshunnot* ('special letters'), changing the shape or size of the letter or attaching curls, bows, or flags in various places. As tagin already show a broad and complex range of variations, see Fig. 1 and Fig. 2, we focus on this phenomenon as a pivotal representative for otiyyot meshunnot variations.

In this paper we assess the analysis of letter decorations from the perspective of computer vision, evaluating the potential to enhance traditional, qualitative research of this phenomenon in the domain of Jewish studies with quantitative methods in computer science. However, introducing a digital approach yields several obstacles, both rooted in characteristics of the historical material and the challenge to design an appropriate computer vision solution.

[a] https://orcid.org/0000-0001-6286-2771
[b] https://orcid.org/0000-0003-3974-3728
[c] https://orcid.org/0000-0001-6296-7282

(a) Cod.heb. 488. Nun and Shin: 3 tagin.

(b) Ms. Heb. 4°7156. Nun and Shin: 3 tagin, Chet: 1 tag. Resh and Aleph: thin strokes on top (serifs or tagin).

Figure 1: Part of Lev 23:18 in two Torah scrolls.

***Tagin* as a Subject of Historical Research.** From a modern perspective, the appearance of tagin on Torah scroll letters presents itself as a stable phenomenon with a clear ruleset and uniform distribution of decorations: always three crowns on the letters Nun/Nun sofit, Shin, Ayin, Teth, Tsade/Tsade sofit, Gimel, and Zayin (so called *sha'atznez getz* letters), always one crown on the letters Qof, Dalet, Yod, He, Bet, and Chet (so called *bedeq chaya* letters), no crowns on the remaining letters. However, these rules have developed over a long period of time, resulting in an immense variety of differing rule sets being applied to historical material of Torah scrolls. Rule books for scribes like the *Sefer Tagin* outlined not only the amount of decoration but also the specific places in the Torah text where a certain version of tagin decoration would be allowed to occur (defining more than 1900 specific letter ornaments per scroll (Michaels, 2020)), making it even more difficult to follow the rules precisely. Especially medieval scribes show a tendency to decorate letters in irregular ways for em-

Frank, L., Götzelmann, G. and Tonne, D.
Towards a Dataset for Paleographic Details in Historical Torah Scrolls.
DOI: 10.5220/0013173800003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 926-933
ISBN: 978-989-758-728-3; ISSN: 2184-4321

phasis or to add a hidden layer of meaning to the script. The amount of variation is so great and the impact on the purity of the scroll is so critical that the common deviations from the rule have influenced the rule making itself. An example of varying decorations in two manuscripts can be seen in Fig. 1.

***Tagin* as a Computer Vision Problem.** In terms of a classification problem, both the regular and irregular tagin may indicate a certain bias in the decoration practice. While the first one is more clearly visible in the data, tipping the quantitative scale towards a dominating amount of 'rule-abiding' decorations, the second bias may not only be more subtle, but remains completely unknown. Due to the complex tradition and meaning of ritualistic Torah writing, we can expect that the real-world distribution of tagin on letters in the historical material is neither fully deterministic nor random (Michaels, 2020). Researching tagin as a codicological phenomenon is accompanied by additional variety in the material due to origin of the scroll and its script style (such as Ashkenazi, Sephardic, Oriental ...), the individual scribal hand, the scrolls material and ink, and surrounding circumstances both in historical transmission and acquisition of the digital material (damages of scroll or ink, heterogeneous digital reproduction and image processing).

Research on the historical formation process of the tagin currently lacks suitable datasets as foundation for computer-assisted analysis of the little letter crownlets, making it a challenging real-world data task in need of novel concepts for collecting the necessary historical document data, appropriate data labels, and fitting classification applications.

2 STATE OF THE ART

Exploiting information from historical manuscripts has become of great interest in recent years due to the progressive digitization of cultural assets. Computer-assisted methods for recognizing text or finding patterns in handwriting styles to identify scribes (Faigenbaum-Golovin et al., 2022) are already successfully applied to historical Hebrew manuscripts.

Character Recognition and Document Analysis. Several digital approaches have been performed in recent years. In terms of automatic classification of script types and modes, experiments have been carried out (Droby et al., 2022), resulting in a classification of paleographic classes. Regarding the recognition of text, a recent case study uses the platform Transkribus to analyze medieval Hebrew scripts (Prebor, 2024). Furthermore, the eScriptorium project (Kiessling et al., 2019) offers an open source platform. In close cooperation, the BiblIA model (Stökl Ben Ezra et al., 2021b) is developed as a page segmentation and recognition model. Another digital early-stage research to classify Hebrew characters on cuneiform tablets with the help of the yolov8 computer vision model has recently been conducted in (Saeed et al., 2024) (preprint). Despite their similarity to our research objective, none of the mentioned approaches targets the recognition of tagin or otiyyot meshunnot.

Datasets. The Pinkas dataset (Kurar Barakat et al., 2019) was published as a first benchmark for Hebrew historical documents and consists of 30 images of a single medieval Hebrew manuscript. The segmentation and respective transcription cover page, line, and word level, but do not break down to single characters. The more extensive BiblIA dataset (Stökl Ben Ezra et al., 2021a) includes 202 document pages of almost 100 manuscripts. It comprises Ashkenazi, Sephardic, and Italian scripts dated between 11^{th} and 16^{th} centuries. Script styles and time period fit our needs, but segmentation and transcription are limited to line level. Neither the Pinkas nor the BiblIA dataset includes Torah scrolls, therefore the occurrence of tagin or other decorations is unlikely. In rare cases, letters in Hebrew bibles may be decorated, but even if the BiblIA dataset contains such decorations in bible pages they are not annotated due to the focus on the line transcription of the content. The HHD dataset (Rabaev et al., 2020) offers character segmentation and transcription, but only includes modern Hebrew handwriting and no historical documents. Another approach is presented in the VML-HP dataset (Droby et al., 2021), which provides paleographic information, e.g. styles and modes of script, on page level of medieval Hebrew manuscripts.

The existing Hebrew datasets target different scopes and objectives, such as text line recognition and transcription, writer identification, and script classification. But none of them focuses on letter decorations. All Hebrew datasets for medieval manuscripts lack character-level segmentation as well as annotations of letter decorations. Consequently, the state of the art datasets are not sufficient to recognize and classify decorated letters. The dataset presented in this paper aims to fill this gap.

3 DATASET

With our dataset we aim for a comprehensive transcribed collection of images of Hebrew letters stemming from Medieval handwritten manuscripts, ini-

tially focusing on Ashkenazi square script. While our primary interest targets special decorated letters in the dataset, we want to introduce a more general dataset, which could in principle be extended, adjusted, and (re-)used for different tasks in various research directions. Therefore, we introduce a larger dataset of segmented letters enriched with metadata and the respective transcription as well as labeled subsets with regard to tagin. The resulting dataset consists of manuscript-specific csv files.

For the creation of the dataset, a semi-automatic, reproducible workflow was chosen allowing for additions and corrections. The image data of the chosen scrolls (cf. 3.1) was obtained and uploaded to eScriptorium (Kiessling et al., 2019). The software's standard segmentation model was fine-tuned by training on the first image of ten different Torah scrolls. After the segmentation step (text regions and line masks) the chosen scrolls were transcribed using the state-of-the-art model (Stökl Ben Ezra et al., 2021b) for historical Hebrew Handwritten Text Recognition (HTR), resulting in transcribed lines as well as estimated character positions provided by the Kraken model. The transcription results were automatically aligned with passim (Smith et al., 2014) to a full text of the Pentateuch[1] to allow automatic assessment of the transcription quality. From this workflow a set of cropped letters was derived (cf. 3.2). Afterwards, the letter annotations were manually labeled to classify tagin decorations (cf. 3.3).

3.1 Selection of Scrolls

Obtaining a comprehensive overview of medieval Torah scrolls is a complicated task – scrolls and fragments are still to be digitized or made publicly available, metadata is sparse and/or error-prone and especially dating is an ongoing challenge for scholars and librarians.[2] Thus, this topic will remain a field of active research for the future, and additional scrolls are expected to appear. For the dataset in this paper we have therefore defined the following requirements for a scroll to be included:

- The image data and metadata are accessible via IIIF API (Snydman et al., 2015).
- The resolution and image quality are suitable for the analysis of small paleographic letter details.
- The holding institution provides a license allowing reuse of the material for research purposes.
- The digitized object contains a certain amount of readable, consecutive text with the potential to align it to the Pentateuch text (excluding extremely fragmented and heavily damaged scrolls).

The limitation on Torah scrolls available via IIIF API was chosen not only for easy import and use of the data in all workflow steps and tools but also to improve the potential to share data and results from all workflow steps without the necessity to include the image data into datasets. Since IIIF allows for dynamic cropping of images via URL parameters, it is sufficient to share cropping coordinates together with the image URL of the according library collection. To rate the image resolution, we compare the width of an exemplary cropped letter Aleph. Manuscripts with a letter width below 30px generally showed too little detail for correct data labeling and further use. The selection of applicable licenses especially excludes digitized scrolls made available online but with special reuse restrictions such as prohibition "to copy the digital copy of the manuscript", unclear distribution permissions, and watermarked images. While we plan to extend the dataset to include all common script types, we decided to initially focus on a single script type to facilitate clearer conclusions from the workflow results. Due to the focus of the associated research project, the scroll data was limited to the script type of Ashkenazi origin in regards to the main scribal hand. All scrolls of Ashkenazi type available to us show letter decorations to some extent.

The resulting data selection can be found in Tab. 1 along with their selection criteria (for the topic of scholarly annotations see 4.3). The final selection contains Torah scrolls with dating estimations between the 13th century (Ms. or. fol. 1218) and the 19th century (Ms. Heb. 4°1459), providing a wide range of decoration customs.

3.2 Generation of Letter Data

The digitized material of the scrolls remains a challenging task for out-of-the-box HTR. Since we did not focus on high-quality full transcriptions, our dataset is based on transcription quality that can currently be expected without putting resources towards quality improvement by additional training and manual corrections. Major obstacles for fully automatic transcription are calligraphic oddities like very widely written letters (sometimes as wide as a full word); partial columns due to digitizing the scroll by unwinding it bit by bit; material aspects of the parchment scrolls (various background colors, damages, stitches, ...). Faulty transcription may result in either assigning the wrong letter or a suboptimal character position to

[1] Text source: tanach.us

[2] As an example case see the date remark for Torah scroll *MS. or. Fol. 133* in the *hebrewpal* database (URL https://www.hebrewpalaeography.com/data/api/items/24/).

Table 1: Selection criteria for dataset creation. Letter width measured on one exemplary Aleph. Manuscripts marked in bold selected for final dataset. For manuscript sources see Appendix A.

Library ID	char width	rejected	shelved for testing
2° Ms. theol. 1	40 px		scholar's annotation
2° Ms. theol. 303	68 px		fragmentary
Christ Church MS 201a	9 px	image quality	
Cod. hebr. 225	83 px		fragmentary
Cod. hebr. 226	92 px		fragmentary
Cod. hebr. 240	115 px		fragmentary
Cod. Parm. 3598	-	reuse restriction	
Cod.hebr. 488	48 px		
Hs. or. 14091	96 px		
BL Add. 11828	30 px		
BL Or. 1085	38 px		fragmentary
Ms. Heb. 24°9084	14px	image quality	
Ms. Heb. 4°1408	53 px		
Ms. Heb. 4°1459	38 px		
Ms. Heb. 4°6066	24 px	image quality	
Ms. Heb. 4°7156	46 px		
Ms. Heb. 4°7247	58 px		
Ms. Heb. 4°8457	-	reuse restriction	
Ms. Heb. 4°9859	41 px	faulty IIIF data	
Ms. Hebr. 34°8421	56 px		fragmentary
Ms. Oct. 19	28 px	image quality	
Ms. or. fol. 1216	76 px		scholar's annotation
Ms. or. fol. 1217	68 px		scholar's annotation
Ms. or. fol. 1218	75 px		
Ms. or. fol. 133	156 px		scholar's annotation
Ms. or. fol. 134	102 px		
Ms. Rhineland 1217	-	reuse restriction	

(a) (b) (c) (d) (e)

Figure 2: The letter Shin in different variations, images from (a) Ms. Heb. 4°7247, (b) Ms. Heb. 4°7156, (c) Ms. Heb. 4°1459, (d) BL Add. 11828, and (e) Cod.hebr. 488.

a glyph, both potentially influencing our letter data. To reduce such problems despite mediocre transcription quality, we used the Pentateuch reference text to filter out rigorously all text lines that were not a perfect match with the reference text (cf. Tab. 2). On average, ∼17% of all transcribed lines and ∼23% of all alignments were perfect matches, resulting in ∼455500 character regions (without white-spaces).

Despite the comparatively low amount of extracted letters they proof to represent the original Torah text well. On average, perfectly aligned letters account for 14.9% of the corresponding letter appearance in the Pentateuch with a standard deviation of 0.01. Due to the somewhat arbitrary distribution of perfect alignments, we can assume to have captured the content of the scrolls in a representative manner.

3.3 Class Definition and Labeling

Challenges. Identifying the various paleographic details in Hebrew handwriting is a skill taught in deep scholarly curricula specifically tailored to careful and experienced observation of scribal handwriting. Trying to train computer vision machine learning on such a subject of course must lack in granularity and requires a stricter and more well-defined class definition than scholarly annotation. Especially two concepts in the scribal composition appear strikingly familiar: serifs ("Small stroke added to the basic letter component (e.g. upper horizontal bar) once the component had been traced") and tagin ("Additional strokes added to the letters of some Torah scrolls as a part of an ancient scribal-mystical tradition.")[3]. Both are usually applied after the initial drawing of the letter, both appear i.a. on top of the head line, both are (nearly) attached to the letter, and both can appear either as nearly invisible hairline strokes or as thick patches of ink. It is a topic of ongoing research to determine whether a clear distinction between serif and tagin can always be made and, if so, what context information might be needed for a precise decision. Therefore, an extensive collection of letter material is the first and maybe most crucial step to reach consensus on categorization and classification.

Table 2: Number of transcribed lines, successful alignments of the reference texts and perfect matches counted per Torah scroll in the dataset.

	lines trans	lines align	perfect align
Ms. or. fol. 1218	5312	4522	1586
Hs. or. 14091	11881	7163	650
Ms. Heb. 4°1459	11925	9541	2037
Ms. Heb. 4°7247	11951	4062	303
BL Add. 11828	13181	8439	677
Ms. Heb. 4°7156	14564	11734	5018
Ms. Heb. 4°1408	14659	8674	995
Ms. or. fol. 134	14970	12470	5664
Cod.hebr. 488	15564	10805	2651
sum	114007	77410	19581

Material Sampling. Each Torah scroll is unique in its letter decoration, and even from visual assessment of exemplary cases it is hard to determine the specific amount and characteristics of the contained tagin. To quantify the contents of our dataset, we chose to sample the letter data and to discuss and to ultimately count letter decorations in general (including serifs, flags, bows, and other unspecified additions to the base letter). The sampling process, which involved a detailed discussion on all categorizations, was limited to the 15 sha'atznez getz letters and bedeq chaya letters. Sha'atznet getz are contemporarily written without serif on their heads and with 3 tagin, bedeq chaya are written with optional serif and with 1 tagin in a top left position of the letter. We can derive from

[3] *Glossary of palaeographical concepts*, URL: https://www.hebrewpalaeography.com/help/1/.

the material sampling that the sha'atznet getz in our historical material are mostly following the expected decoration practice of being written tagin (not taking into account the 'correct' number of tagin) with $\frac{9}{10}$ letters showing decorations (Tab. 3a). The bedeq chaya on the other hand show a much more complex and unclear picture, where closer to $\frac{2}{5}$ of the sampled letters clearly has some additional stroke added to the base letter, but only a very low number can with absolute certainty be labeled as tagin (Tab. 3b). Sampling results show the varying imbalance between undecorated letters and their decorated counterparts as well as the varying difficulty of class assignments.

Labeling Decision. As described in 3.2 the letter data for labeling are retrieved by aligning the eScriptorium transcription with the Pentateuch reference text. Consequently, each letter image is assigned the respective letter label, creating 27 categories. In addition to the transcription label, we decide on a binary classification design and introduce the labels 'tagin' and 'none' for decorated and undecorated letters, respectively. Manual labeling is performed according to the dual control principle. The formerly mentioned classification challenges, letter imbalance, and limitation in dataset size prevent a more fine-grained classification of specific tagin counts and their position. A controlled vocabulary of letter decorations is a current work in progress to enable further research.

4 EVALUATION OF THE DATASET

With regard to our specific interest in letter decorations, we conduct a preliminary evaluation on a subset of labeled letter images (cf. 4.1) as a proof of concept, and to assess the potential of our collected data. Therefore, we design a transfer learning model suitable for the task (cf. 4.2) and train it on a variety of data subsets for comparison. Evaluation of the results can be derived from applying it on task specific test data as well as real scholarly annotations (cf. 4.3).

4.1 Training Data

Dataset I: Full Labeling of a Single Letter. The sha'atnez getz letters, as shown in section 3.3, present themselves in a mostly clear distinction between tagin and no-tagin versions. As a good representation for this subclass of letters, we have chosen the letter Shin as a sample case. For the labeling, 24600 occurrences were manually assessed in completeness and labeled in two classes. After discarding unsuitable images

Table 3: Letter samples (sample size = 100) and their fraction of letters with decoration (dec) and with tagin.

(a) sha'atnez getz letters.

letter	dec (tagin)
Shin	0.92 (0.92)
Ayin	0.93 (0.93)
Teth	0.96 (0.96)
Nun	0.93 (0.93)
Nun sofit	0.91 (0.89)
Zayin	0.89 (0.89)
Gimel	0.89 (0.89)
Tsade	0.86 (0.85)
Tsade sofit	0.94 (0.94)

(b) bedeq chaya letters

letter	dec (tagin)
Bet	0.31 (0.01)
Dalet	0.41 (0.02)
Qof	0.36 (0.04)
Chet	0.84 (0.38)
Yod	0.40 (0.03)
He	0.33 (0.09)

(i.e. badly cropped chars and undecidable classification cases), the final dataset consists of 23187 images (20949/2238 class split). It is noteworthy that the minority class (without tagin) is additionally very unbalanced (87% of all labels) towards one document, Ms. or. fol. 1218. The scroll only provides 17 images for the class of letters with tagin.

Dataset II: Representative Labeling of Sha'atznez Getz and Bedeq Chaya. The second chosen subset includes the 15 sha'atznez getz and bedeq chaya letters (cf. Tab. 3a and Tab. 3b), based on the discussions in 3.3. The classes are assigned manually. For each letter we aimed for 100 images per class, 'tagin' and 'none', respectively. If the amount of images of the minority class for a specific letter is less than 100, the counterpart class is reduced to the same amount. For each image of the minority class, a letter as similar or close as possible is chosen as a counterpart. In most cases, this allows for good comparison between both variants by the same scribal hand. The small class size for each letter eases the problem of imbalance in decoration practices while not making it disappear completely. In practice it was not possible to balance the label dataset both by class and by document most of the time—in these cases class balance and aiming for the target amount for labeled data per letter took precedence. While some letter classes could be filled with a high number of easily distinguishable representations, others proved too difficult to be labeled without input of a domain expert. In total, 1396 images (688/708 class split) are considered for the training data. As shown in the visualization of the dataset (Fig. 3), different Torah scrolls take precedence in different letter categories, overall diminishing the naturally occurring unbalanced class distributions.

On all data input the following preprocessing steps were performed: conversion to greyscale, resizing (128x128px), normalization in [0,1] range. For usage as training data a 80/20 training-validation split is

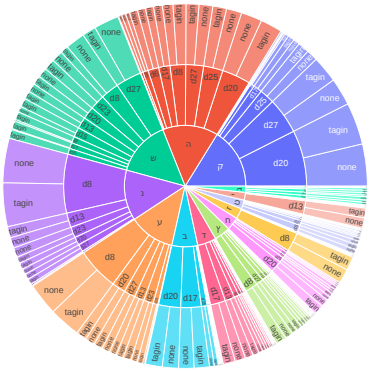

Figure 3: Labeled data for multiple letters. Dataset used for training scenarios SC3 and SC4. Inner ring: labeled letter. Middle ring: contributing documents (cf. Appendix A). Outer ring: label class. All rings sorted by element count.

used in all cases. The data is fed in batch size of 32.

4.2 Model Architecture and Training

For our currently binary classification task, we chose a deep learning architecture. With regard to the small size of the training datasets as well as to limited number of Torah scrolls, transfer learning on an existing model is preferred rather than training a model from scratch. Therefore, our model is fine-tuned from a pre-trained Xception model (Chollet, 2017), initially trained on the ImageNet dataset. The Xception model has been shown as solid choice for character recognition with small datasets accounting for high performance rate independent of dataset size (Benaissa et al., 2022). The three fully-connected layers on top of the base model were dropped to allow for flexibility in input size. On top of the base model we apply a model of the following architecture: Conv2D layer (32 3x3 filters, ReLU activation), MaxPooling2D layer (2x2), Dropout layer (0.25 dropout rate), GlobalAveragePooling2D layer, Dense layer (64 units, ReLU activation), Dropout layer (0.5 dropout rate), Dense layer (1 unit, sigmoid activation function). We utilize dropout layers to help prevent overfitting on the small datasets. The model is compiled with adam optimization and binary crossentropy loss function with label smoothing (0.1) to improve robustness of the model and to encourage model generalization. Label smoothing specifically fits the described difficulty of choosing clean and distinct labels for the problem of letter decorations.

The training process comprises two phases. In phase 1 the Xception base model is frozen while fitting the model to the new training data (20 epochs). In phase 2 all layers of the base model are set to trainable to allow adaptation to differences in input format of our data compared to the ImageNet data (200 epochs). In this phase a low training rate (1×10^{-5}) is chosen to prevent significant alteration of the pre-trained weights. Main goal of our model architecture is applicability to divers subsets of our training data with differences in class sizes and balancing, aiming for assessment of dataset characteristics and value.

SC1: Dataset I (balanced by Documents and by Classes). We use the Shin label set with a balancing factor of 1, fitting the majority class (Shin with tagin) with same amount of files as the minority class (undecorated Shin) by a random seeded sample. We get a resulting input of 2238/2238 files (tagin/no tagin).

SC2: Dataset I (Balanced by Classes). We use the Shin label set with a balancing factor of 2, but additionally balancing by document source; fitting majority class per Torah scroll in the dataset with amount of files from minority class in the same scroll (2 to 1, up to maximum number of files in one scroll). We get a resulting input of 547/329[4] (tagin/no tagin) files.

SC3: Dataset II (Shin Only). We use the mixed letter label set, utilizing the labels for letter Shin. We get a resulting input of 104/102 files (tagin/no tagin).

SC4: Dataset II (Full Dataset). We use the complete mixed letter label set. We get a resulting input of 708/688 files (tagin/no tagin).

4.3 Testing Data

The model is applied to unseen letter images of a test dataset. Only images of manuscripts not included in the scroll selection (cf. 3.1) are considered, creating a challenging task for model prediction and making it easy to assess overfitting on scribal features in the dataset. The test datasets consist of scholarly annotations as well as images retrieved similar to the workflow described in 3.2.

Scholarly Annotation Data [ANNO]. Manual annotation of scholars [5] leads to high-quality data with very fine-grained classification. However, due to limited human resources, only letters with irregular tagin are annotated. These annotations cannot provide training data for a generalized model, but serve as a valuable test case for our training scenarios. The annotation data is shaped by the following characteristics: 2158 files (tagin only), 14 dif-

[4]To reduce the class imbalance, the minority class was extended by 30 manually cropped letters from a single sheet fragment (Hs. or. 13467, Berlin State Library).

[5]The authors thank Juan E. Mora, Dr. Emese Kozma, Aram Abu-Saleh und Konstantin Paul for their detailed manual annotations.

Table 4: Prediction accuracies for the training scenarios.

	SC1	SC2	SC3	SC4
Shin ANNO	0.80	0.93	0.77	0.97
Shin TTEST	0.72	0.77	0.68	0.81
All ANNO	0.70	0.72	0.53	0.92

Table 5: Confusion Matrices for TTest Results.

SC1		Predicted	
		none	tagin
True	none	0.74	0.26
	tagin	0.31	0.69

SC2		Predicted	
		none	tagin
none		0.95	0.05
tagin		0.42	0.58

SC3		Predicted	
		none	tagin
True	none	0.69	0.31
	tagin	0.33	0.67

SC4		Predicted	
		none	tagin
none		0.87	0.13
tagin		0.25	0.75

ferent letters; source material from Torah scrolls as well as *Sefer Tagin* (ancient scribal manual); Ashkenazi and Sephardic script types; irregular versions of tagin only; fine-granular classes (>180); independent annotation process/training (different cropping style). The main benefits of this dataset are its variety and size as well as its quality, while the drawback is the complete lack of undecorated letters in the dataset.

Task-specific Test Data [TTEST]. To fill the gap provided by the scholarly annotations and for equal assessment of all training scenarios we additionally provide a test dataset on the single letter Shin, enabling comparison of performance on the one single letter included in all training datasets. The test data are fairly split between the two classes (36 tagin/39 no tagin images). The test data contains only texts of the Ashkenazi script type as identical characteristic to the training data. However, the sources of the letter crops go beyond material from Torah scrolls, adding other Hebrew text genres such as materials from Bible, Talmud or Mezuza. Therefore, the test data is challenging and assesses applicability of our proof-of-concept training scenarios beyond research on Torah scrolls.[6]

4.4 Performance

All training scenarios were tested for prediction accuracy on the single letter in the scholarly annotations (Shin ANNO, 186 files), the task specific test dataset (Shin TTEST) and the full set of scholarly annotations (All ANNO) (cf. Tab. 4). The confusion matrices for all training scenarios tested on TTest are shown in Tab. 5. The scenarios with training on the single letter Shin only and a reasonable size of training data (SC1, SC2) show a surprising potential for generalization towards tagin on other letters. But they are clearly outperformed by SC4 (trained on a mix of letters) in all tests, even on Shin specific test data. The training history of SC3 did not compare well to the other trainings with troublesome stabilization over multiple runs, unclear convergence of the training loss and very heterogeneous accuracies on the different test sets (so the single run results above have to be assessed carefully). Since SC3 is the smallest dataset, this result is not surprising. The generalization potential from the small set of Shin data to other letters is low. Although the accuracy of SC2 (largest dataset, only balanced by classes, not by contribution of Torah scrolls) seems promising, the confusion matrix shows a high amount of false positives for undecorated letters on the TTEST result. Due to overwhelming and very unbalanced influence of a single Torah scroll (cf. 4.1), the lack of generalization is not completely surprising. With regard to the corresponding long training time and the high amount of resources, such a large dataset of a single letter does not seem advantageous if it does not provide representative labeling.

All scenarios show a tendency to lean towards no tagin prediction, which results in a higher false-negative rate on tagin of the TTEST. This general behaviour might be caused by the source material, the shape of the classification task, or biases in the labeling process. Further investigation of the phenomenon is needed. Overall, a manual check of false predictions on the ANNO test set shows a slight lack of robustness in regards to unexpected image cropping, which could be addressed by introduction of data augmentation in the preprocessing step.

5 CONCLUSION

Letter decorations in historical Torah scrolls expose numerous research possibilities in the field of computer vision. In this paper, we provide the concept for a novel, reproducible dataset of segmented letters labeled for tasks in historical Hebrew palaeography. Our work provides a more general and comprehensive dataset of segmented letters labeled with the corresponding letter transcription and enriched with metadata. We introduce a labeled subset providing a wide variety of historical decoration practices and enabling first-time experimentation with classification of typical letter decorations. Although our dataset appears small compared to other computer vision datasets, it can be considered large from a domain perspective. Due to it's medium-sized real data corpus as source material and its high-quality of segmentation and an-

[6]Test data sources: Ms. or. fol. 1216 (State Library Berlin), Cod.hebr. 501 (BSB Munich), Cod.hebr. 2 (BSB Munich), Cod.hebr. 212 (BSB Munich), Cod.hebr. 153(2,1 (BSB Munich), Ms. Hebr. 34°8421 (NLI).

notation, our data set is of great value for the research field of Digital Jewish Studies.

We have shown that applying decoration labels to such a real-world dataset is not only resource-intensive but also a process of very varying difficulty and consensus. Our machine learning scenarios indicate a need for distinctive, high-quality labeling data despite the very unbalanced decoration data. Further concepts of semiautomatic labeling might be necessary to facilitate high-quality, large-scale data input and to enable scholarly input and plausibility checks.

The evaluation shows already promising results in terms of decoration recognition. We find it encouraging that smaller, yet more clear-cut labeling sets outperform larger datasets with less careful balancing and selection. Counterintuitively, creating data for multiple letter classes simultaneously helps with classification training of a single letter and makes training with these data more robust. Overall, this encourages working in the direction of high-quality labels including detailed scholarly input and finer classification of tagin variations to move beyond a mere glimpse on the scribal intentions towards a more comprehensive insight into the tradition of historical Torah scrolls.

ACKNOWLEDGEMENTS

This work was funded by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01UL2202B and supported by the Helmholtz Association Initiative and Networking Fund on the HAICORE@KIT partition.

REFERENCES

Benaissa, A., Bahri, A., El Allaoui, A., and Bourass, Y. (2022). Character recognition using pre-trained models and performance variants based on datasets size: A survey. *ITM Web Conf.*, 43:01008.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807.

Droby, A., Kurar Barakat, B., Shapira, D., Rabaev, I., and El-Sana, J. (2021). Vml-hp: Hebrew paleography dataset. In *Document Analysis and Recognition – ICDAR 2021*, pages 205–220.

Droby, A., Rabaev, I., Shapira, D., Kurar Barakat, B., and El-Sana, J. (2022). Digital hebrew paleography: Script types and modes. *Journal of Imaging*, 8(5).

Faigenbaum-Golovin, S., Shaus, A., and Sober, B. (2022). Computational handwriting analysis of ancient hebrew inscriptions—a survey. *IEEE BITS the Information Theory Magazine*, 2(1):90–101.

Kiessling, B., Tissot, R., Stokes, P., and Stökl Ben Ezra, D. (2019). escriptorium: An open source platform for historical document analysis. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 19–19.

Kurar Barakat, B., El-Sana, J., and Rabaev, I. (2019). The pinkas dataset. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 732–737.

Michaels, M. (2020). *Sefer Tagin Fragments from the Cairo Genizah: A Critical Edition, Commentary and Reconstruction. Cambridge Genizah Studies Series, Volume 12*. Brill, Leiden, The Netherlands.

Perani, M. (2022). *Chapter 11 The Tagin: Their Origin, Use, and Oscillating Evolution between Embellishment and Mystical Signifier. New Light from the Ancient Bologna Sefer Torah*, pages 297 – 348. Brill, Leiden, Niederlande.

Prebor, G. (2024). From digitization and images to text and content: Transkribus as a case study. *Manuscript Studies*, 9(1):72–89.

Rabaev, I., Kurar Barakat, B., Churkin, A., and El-Sana, J. (2020). The hhd dataset. In *2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 228–233.

Saeed, E. A., Jasim, A. D., and Malik, M. A. A. (2024). Hebrew letters detection and cuneiform tablets classification by using the yolov8 computer vision model. *eprint arXiv*.

Smith, D. A., Cordel, R., Dillon, E. M., Stramp, N., and Wilkerson, J. (2014). Detecting and modeling local text reuse. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 183–192.

Snydman, S., Sanderson, R., and Cramer, T. (2015). The international image interoperability framework (iiif): A community & technology approach for web-based images. *Archiving Conference*, 12(1):16–21.

Stökl Ben Ezra, D., Brown-DeVost, B., Jablonski, P., Kiessling, B., Lolli, E., and Lapin, H. (2021a). Biblia – an open annotated dataset.

Stökl Ben Ezra, D., Brown-DeVost, B., Jablonski, P., Lapin, H., Kiessling, B., and Lolli, E. (2021b). Biblia - a general model for medieval hebrew manuscripts and an open annotated dataset. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, HIP '21, page 61–66, New York, NY, USA. Association for Computing Machinery.

Appendix A: Image Sources

Internal ID	Library ID	Library	IIIF Manifest
	2° Ms. theol. 1	UB Kassel	https://orka.bibliothek.uni-kassel.de/viewer/api/v1/records/1337850581405/manifest/
	2° Ms. theol. 303	UB Kassel	https://orka.bibliothek.uni-kassel.de/viewer/api/v1/records/1314262537823/manifest/
	Christ Church MS 201a	Bodleian	https://iiif.bodleian.ox.ac.uk/iiif/manifest/ab202c0a-5a65-4da8-a56c-2fcdc8d8c784.json
	Cod. hebr. 225	Austrian National Library	https://iiif.nli.org.il/IIIFv21/DOCID/PNX_MANUSCRIPTS990037225270205171-1/manifest
	Cod. hebr. 226	Austrian National Library	https://iiif.nli.org.il/IIIFv21/DOCID/PNX_MANUSCRIPTS990038531430205171-1/manifest
	Cod. hebr. 240	Austrian National Library	https://iiif.nli.org.il/IIIFv21/DOCID/PNX_MANUSCRIPTS990026657902051 71-1/manifest
	Cod. Parm. 3598	Biblioteca Palatina Parma	https://iiif.nli.org.il/IIIFv21/DOCID/PNX_MANUSCRIPTS990017459802051 71-1/manifest
d8	**Cod.hebr. 488**	BSB Munich	https://api.digitale-sammlungen.de/iiif/presentation/v2/bsb00151486/manifest
d13	**Hs. or. 14091**	Berlin State Library	https://content.staatsbibliothek-berlin.de/dc/1809307333/manifest
d6	**BL Add. 11828**	London British Library	https://iiif.nli.org.il/IIIFv21/DOCID/PNX_MANUSCRIPTS990001223160205171-1/manifest
	BL Or. 1085	London British Library	https://iiif.nli.org.il/IIIFv21/DOCID/PNX_MANUSCRIPTS990001223180205171-1/manifest
	Ms. Heb. 24°9084	NLI Jerusalem	https://iiif.nli.org.il/IIIFv21/DOCID/PNX_MANUSCRIPTS990044275740205171-1/manifest
d15	**Ms. Heb. 4°1408**	NLI Jerusalem	https://iiif.nli.org.il/IIIFv21/DOCID/PNX_MANUSCRIPTS990000448750205171-1/manifest
d17	**Ms. Heb. 4°1459**	NLI Jerusalem	https://iiif.nli.org.il/IIIFv21/DOCID/PNX_MANUSCRIPTS990000449050205171-1/manifest
	Ms. Heb. 4°6066	NLI Jerusalem	https://iiif.nli.org.il/IIIFv21/DOCID/PNX_MANUSCRIPTS990025691070205171-1/manifest
d20	**Ms. Heb. 4°7156**	NLI Jerusalem	https://iiif.nli.org.il/IIIFv21/DOCID/PNX_MANUSCRIPTS990000417490205171-1/manifest
d23	**Ms. Heb. 4°7247**	NLI Jerusalem	https://iiif.nli.org.il/IIIFv21/DOCID/PNX_MANUSCRIPTS990025626850205171-1/manifest
	Ms. Heb. 4°8457	Klein Charitable Foundation	
	Ms. Heb. 4°9859	NLI Jerusalem	https://iiif.nli.org.il/IIIFv21/DOCID/PNX_MANUSCRIPTS997012714752051 71-1/manifest
	Ms. Hebr. 34°8421	NLI Jerusalem	https://iiif.nli.org.il/IIIFv21/DOCID/PNX_MANUSCRIPTS990035376570205171-1/manifest
	Ms. Oct. 19	UB Frankfurt Main	https://iiif.nli.org.il/IIIFv21/DOCID/PNX_MANUSCRIPTS990001378080205171-1/manifest
	Ms. or. fol. 1216	Berlin State Library	https://content.staatsbibliothek-berlin.de/dc/666097267/manifest
	Ms. or. fol. 1217	Berlin State Library	https://content.staatsbibliothek-berlin.de/dc/666097291/manifest
d27	**Ms. or. fol. 1218**	Berlin State Library	https://content.staatsbibliothek-berlin.de/dc/666097333X/manifest
	Ms. or. fol. 133	Berlin State Library	https://content.staatsbibliothek-berlin.de/dc/859630935/manifest
d25	**Ms. or. fol. 134**	Berlin State Library	https://content.staatsbibliothek-berlin.de/dc/859632946/manifest
	Ms. Rhineland 1217	Private collection	