https://doi.org/10.2352/issn.2168-3204.2025.22.1.14

Beyond Two Scribes? Column-Based Writer Identification in the 1QIsa[a] Scroll

Tabita L. Tobing, Patrick Bours; Department of Information Security and Communication Technology, NTNU – Norwegian University of Science and Technology; Gjøvik, Norway

Abstract

The 1QIsa[a] Scroll, one of the most significant manuscripts among the Dead Sea Scrolls, has long been the focus of debate over whether it was produced by a single hand or multiple scribes. In this study, we introduce a column-based writer-identification framework that combines unsupervised clustering, character-level verification, and cross-dataset evaluation, without assuming any fixed number of scribes. Benchmarking our hinge–feature–based approach against the widely recognized FIREMAKER dataset reveals its strengths and weaknesses. This exploratory analysis not only offers fresh insights into 1QIsa[a]'s scribal attribution but also underscores the need for richer or complementary features in future digital paleographic research.

Introduction

As one of the greatest archaeological discoveries of the 20th century, the 1QIsa[a] Scroll, discovered in Qumran Cave 1 in 1947, has attracted extensive scholarly interest across multiple disciplines. As the most complete and best-preserved biblical manuscript among the Dead Sea Scrolls, researchers have examined various aspects of the scroll, including its dating, textual composition, linguistic features, scribal practices, and historical significance. [1–4] Yet, one longstanding debate surrounding 1QIsa[a] concerns scribal attribution: did a single or multiple scribes copy the scroll?

Tov [5] shows a discussion of two distinct scribal hands, while long scrolls (54 columns), with a noticeable shift at column XXVII, where Scribe A (SA) left three blank lines before Scribe B (SB) continued in column XXVIII. Some scholars accept this division based on paleographical differences, while others maintain that a single scribe may have simply changed his spelling habits midway, emphasizing that handwriting variability can stem from fatigue or stylistic adaptation rather than multiple scribal possibilities.

The main challenge in the scribal attribution of 1QIsa[a] is the lack of an empirically validated method for determining the number of scribes involved in its production. Traditional paleographic analysis, which relies on expert evaluation of letterforms and scribal habits, has been invaluable in manuscript studies. However, distinguishing scribes in a long manuscript like 1QIsa[a] is challenging due to handwriting variability from writing conditions, stylistic shifts, and scribe adaptation.

To address these complexities, Popović et al. [6] apply computational analysis to detect potential scribal divisions. By utilizing a *hinge*-based clustering technique, their analysis identified a division between two sets of columns, which they interpreted as evidence of two scribes. However, despite their contribution, a methodological concern remains. While their clustering method was unsupervised, the analysis framework inherently focused on validating a two-scribe hypothesis, potentially overlooking more nuanced divisions. To address these concerns, the present study retraces the column-based writer identification method in the 1QIsa[a] Scroll, particularly in the absence of ground-truth scribal labels, and aims to:

- Explore the reliability of *hinge* features for writer identification in the 1QIsa[a] Scroll.
- Apply unsupervised hierarchical clustering to detect potential scribal transitions across columns without presupposing how many scribes contributed.
- Assess the outcome of clustering using character-level verification, comparing classification performance against the traditional two-scribe assumption.
- Use t-SNE and PCA visualizations to better understand how *hinge* features behave in simpler (FIREMAKER 2-writer versus column-based 1QIsa[a]) scenarios.

By combining unsupervised clustering, character-level verification, and cross-dataset feature evaluation, this study aims to provide a more critical and interpretable framework for *hinge*-based writer identification in the 1QIsa[a] Scroll, one that does not presuppose the number of scribes.

Materials and Methods

Dataset

For this study, we assembled a unified dataset from three sub-collections of images (see Fig. 1): FIREMAKER [7], the BiNet-based 1QIsa[a] collection [8], and the character-level 1QIsa[a] collection. For each experiment, where clustering, verification, or dimensionality-reduction visualization is required, we select one or more of these sub-collections as needed, using them individually or in combination.

FIREMAKER

FIREMAKER was originally developed as a benchmark dataset for forensic writer identification. It comprises pages of handwritten text from 250 writers, in both Dutch and English, across four task-specific pages. In this study, we utilize only Page 1, where each writer copies the provided paragraphs in their natural handwriting style. The image collection can be accessed at the following link: https://zenodo.org/records/1194612.

BiNet-Based 1QIsa[a] Images

This collection is a set of black-and-white images of the 1QIsa[a] (Great Isaiah) Scroll, created using a deep learning tool called BiNet. The images preserve only the handwritten ink, removing the parchment background, which allows for clearer visual and computational analysis. Out of the scroll's original 54 columns, this collection includes 52 columns; columns 16 and 46 are excluded due to the unavailability of digital images. The whole image collection is available at https://zenodo.org/records/4469996.

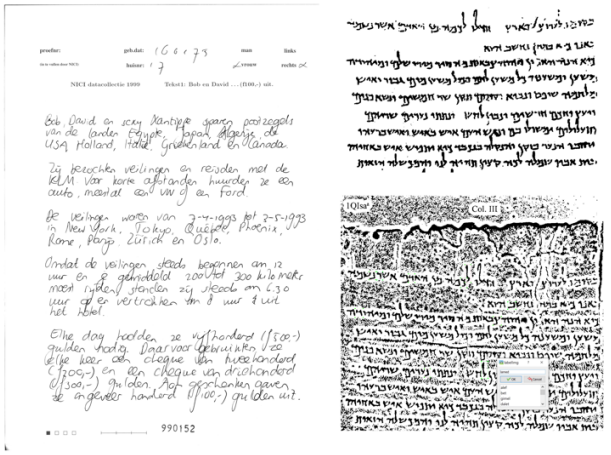

Figure 1. *Overview of the datasets used in this study.* ***Left****: A sample page from the FIREMAKER.* ***Top right****: A BiNet-based 1QIsa[a] Scroll.* ***Bottom right****: An example of character-level segmentation from Otsu-binarized 1QIsa[a] image.*

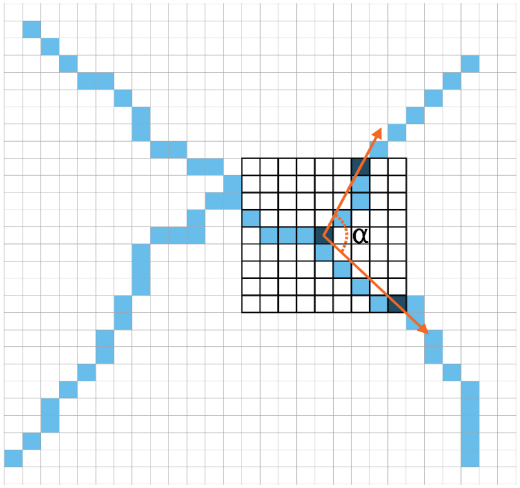

Figure 2. *Illustration of hinge feature extraction. A 9×9-pixel window slides over the 28x28-pixel binarized image to capture local stroke orientation. At each foreground (ink) pixel, the angles between pairs of contour points (α) are measured and accumulated into a histogram.*

Character-Level-Based 1QIsa[a] Images

This sub-collection contains isolated character images manually extracted from the Otsu-binarized 1QIsa[a] Scroll. The characters were segmented from the 52 BiNet-binarized column images using Otsu's thresholding method to preserve the ink shape as faithfully as possible without introducing stroke distortions. Each character image is a binary (grayscale) patch containing a single Hebrew letter, drawn from 22 distinct letter classes. We have 566 images per letter class (283 labeled SA and 283 labeled SB), except for *tet* and *samekh*, which are represented by fewer samples.

The dataset was originally compiled for a prior study on character-based writer verification in ancient Hebrew square-script manuscripts [9], where edge-directional features were applied at the single-character level. This collection enables supervised learning experiments that test whether writer-specific stroke patterns can be detected even at the scale of individual characters, offering a complementary validation layer for the column-level clustering results. Full details of the character extraction and dataset design are available in the original publication. [10]

Feature: Hinge

To capture the subtle curves and junctions in ancient Hebrew script, we use the *hinge* [11] feature, which is essentially a compact fingerprint of local stroke directions. Imagine sliding a small window—here, a 9×9-pixel square (5-pixel length)—over a 28 x 28 patch (see Fig. 2). Within each window, we look at every pixel where ink meets background and record the angle of the stroke as it "bends." By tallying how often each pair of angles occurs, we build a histogram, the *hinge* features, for that patch. Repeating this over 460 randomly chosen, non-overlapping patches per column (and later for each character) gives us a robust, 463-dimensional feature vector that encodes the scribe's characteristic stroke patterns. For a non-technical reader, these histograms can be thought of as a "texture map" of handwriting, showing how the pen changes direction. For technical readers, they represent a concise distribution of stroke orientations, filtering out pixel-level noise to produce a stable signature.

Writer Identification

Our goal is to determine how many distinct scribes, or writing "hands", were involved in copying 1QIsaa. To do this, our system performs two complementary tasks. First, it discovers groups of columns with similar *hinge* signatures via unsupervised clustering, without ever assuming a fixed number of scribes in advance. Second, it verifies those column-based groupings by testing whether individual letters can be reliably assigned back to their column clusters. In practice, that means we extract every character of 22 letter types, represent each with the same 464-bin *hinge* feature, and use a simple classification pipeline to ask: if these letters truly come from separate scribes? We then visualize the *hinge* distributions, comparing the column-based 1QIsaa results against our FIREMAKER ground-truth data, using PCA and t-SNE embeddings to highlight clustering trends.

Hierarchical Clustering

This task is conducted on the BiNet-based 1QIsa[a] collection. The manuscript originally consists of 54 columns; however, due to the absence of columns 16 and 46 in the BiNet-based images, this study analyzes a total of 52 columns.

To improve feature extraction, experiments were first conducted on the FIREMAKER image collection, where different *hinge*-based feature extraction configurations were systematically tested. These improvements guided not only the feature extraction settings but also the distance metric used to quantify handwriting differences. The finalized feature extraction settings include:

- Hinge-based feature extraction with a 5-pixel stroke length.
- Internal gradient shape representation for improved stroke characterization.
- 28×28-pixel, random, non-overlapping patches with controlled black pixel (foreground ink) density.
- 460 samples of patches per columns in 1QIsa[a], ensuring a robust representation of handwriting styles.

Following feature extraction, a distance score is computed between each pair of columns to quantify handwriting differences. The distance metric is formulated as follows:

$$D(i,j) = \sqrt{\sum_{k=1}^{d} \frac{(\mu_{i,k} - \mu_{j,k})^2}{\sigma_{i,k} \cdot \sigma_{j,k}}}, \quad (1)$$

where $D(i, j)$ represents the distance score between column i and column j , referring to the manuscript columns, excluding missing columns 16 and 46. The variable d is the total number of extracted features. The terms $\mu_{i,k}$ and $\mu_{j,k}$ denote the mean of the k -th feature in columns i and j, respectively, while $\sigma_{i,k}$ and $\sigma_{j,k}$ representing the standard deviation of the k-th feature in columns i and j. This method accounts for feature mean differences and variability, providing a normalized distance metric for clustering. This normalized metric balances average differences against variability.

We then built an average-linkage dendrogram (with optimal leaf ordering) and explored cutting it into k = 2 through 7 clusters or groups. Each clustering was scored by its silhouette coefficient (using the full 52×52 distance matrix).

Character-Level Verification

To validate, rather than assume, our column clusters, we used the isolated characters of 22 letters from Otsu-binarized column images. Each character received the same 464-bin *hinge* histogram. We then performed a dynamic, stratified cross-validation, where the number of folds was set to the smaller of five or the count of the least-represented cluster, but never fewer than two. Within each fold:

- Linear Discriminant Analysis (LDA) was fit on the training characters, choosing up to four components (but never more than n_classes−1), thus focusing on the directions that best separate the cluster labels.
- A 1-Nearest-Neighbor classifier (using cosine distance) was trained on the LDA features.

Performance was measured by one-vs-one multi-class AUC (Area Under the Curve), where for k=2 clusters we simply used the positive-class probability, and for k>2 we subset the predicted probability matrix to present classes, renormalized each row to sum to one, and averaged the pairwise AUCs. We compared two labeling schemes: the cluster-derived column labels from hierarchical clustering, and the traditional Scribe A and Scribe B split at column XXVII.

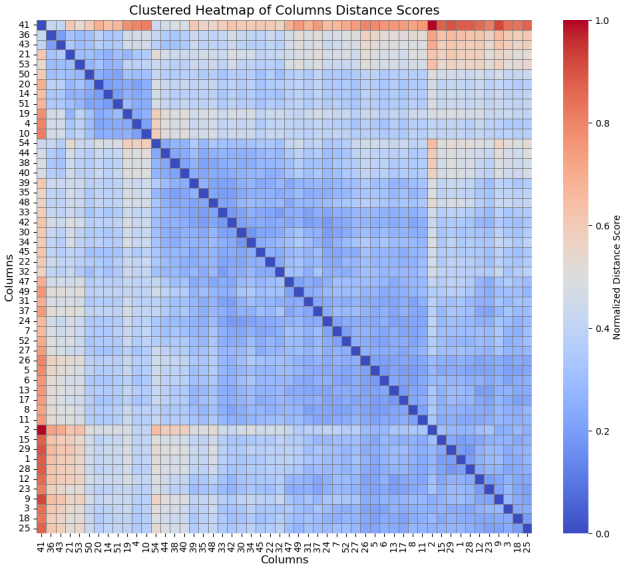

Figure 3. *Clustered heatmap of normalized pairwise distance scores between 1QIsa[a] columns, based on hinge feature histograms.*

PCA and T-SNE Visualization

To provide an intuitive view of column groupings, we applied two distribution-visualization methods to the BiNet-derived dataset and compared these embeddings with those from the FIREMAKER dataset:

- Principal Component Analysis (PCA) finds the straight-line directions of greatest overall variance in the high-dimensional *hinge* space. Plotting the first two or three principal components often reveals broad separations or gradients in writing style.
- t-Distributed Stochastic Neighbor Embedding (t-SNE) is a nonlinear method that preserves local neighborhoods: points (columns or characters) that are very similar in *hinge* space stay close, while dissimilar points move far apart.

By marking each point on the PCA and t-SNE plots with its cluster label (e.g., SA vs SB), readers can immediately judge whether the algorithmic groups align with the intuitive handwriting styles they expect, making the results accessible to both manuscript scholars and computational experts alike.

Results

Hierarchical Clustering

We began our exploration of writer attribution in the 1QIsa[a] Scroll by constructing a pairwise column-to-column distance matrix based on *hinge* features. These features were extracted using an optimized configuration derived from prior testing on the FIREMAKER dataset. Specifically, each column was represented by the mean of 460 *hinge* histograms extracted from 28×28 pixel patches, with internal gradient encoding and a fixed 5-pixel stroke length. The final distance metric (Equation 1) accounts for both mean and variability of each feature dimension, providing a normalized distance score that is sensitive to both subtle and pronounced stylistic variations.

Fig. 3 displays the clustered heatmap derived from the pairwise column-to-column distance matrix. It reveals that most columns exhibit relatively low dissimilarity (blue regions), while certain columns, including Column 41, stand out with high distance scores (red zones) relative to most others. This visual cue suggests potential scribal divergence but without a sharply defined boundary across the scroll.

To further investigate the structure revealed in the heatmap, we applied hierarchical clustering. We explored a range of cluster configurations from k = 2 to 7 to assess the granularity of potential scribal groupings. Each clustering was evaluated based on the silhouette score, which quantifies how well-separated and internally coherent the clusters are. Among all tested configurations, the 2-cluster solution shown in Fig. 4 yielded the highest silhouette score (0.666), indicating a strong overall separation between the two

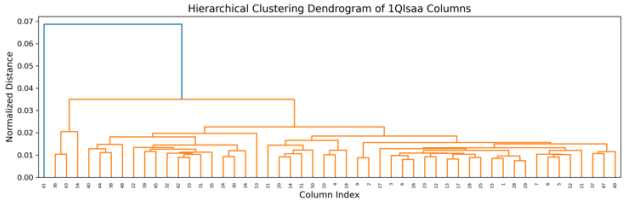

Figure 4. *Hierarchical clustering dendrogram of the 52 1QIsa[a] patch columns, cutting the tree into two clusters (highest silhouette score among k=2 through 7) yields the most coherent and balanced split of columns.*

primary writing styles or clusters. Notably, Column 41 consistently formed its own branch across multiple values of k, reinforcing its status as a potential outlier or transition point. Columns 36, 43, and 54 joined this split at higher thresholds, possibly marking a secondary transitional zone or a gradual drift in writing characteristics.

However, cluster solutions with k > 2 rapidly led to unbalanced partitions, with minor clusters often containing only one or two columns. This sparsity creates significant limitations for character-level verification, where small sample sizes from single columns make it nearly impossible to train or test reliable classifiers. Thus, while the dendrogram structure supports exploratory insights, it also highlights the fragility of clustering-based attribution in the absence of robust sample sizes or sharp stylistic boundaries.

Table 1. *Character-level writer verification results (SA/SB labeling).*

Letter	n_characters	AUC_mean	AUC_std
alef	544	0.601	0.034
ayin	546	0.632	0.039
bet	544	0.565	0.022
dalet	539	0.603	0.028
gimel	549	0.592	0.042
he	549	0.585	0.018
het	546	0.647	0.024
kaf	543	0.568	0.037
lamed	545	0.533	0.038
mem	546	0.636	0.009
nun	547	0.54	0.022
pe	545	0.549	0.041
qof	548	0.587	0.021
resh	545	0.622	0.034
samekh	451	0.569	0.062
shin	547	0.577	0.024
tav	554	0.622	0.038
tet	285	0.53	0.066
tsadi	539	0.568	0.027
vav	544	0.592	0.033
yod	544	0.565	0.016
zayin	551	0.493	0.039

Character-Level Verification

To evaluate whether the column-based clusters derived from *hinge* features represent meaningful scribal distinctions, we conducted character-level verification using the 22-letter character set previously described.

Under the traditional two-scribe assumption (split at column XXVII), we observed moderate but consistent discrimination across characters. As shown in Table 1, several letters, such as *het* (AUC = 0.647, std = 0.024), *mem* (AUC = 0.636, std = 0.009), and *ayin* (AUC = 0.632, std = 0.039), demonstrated relatively strong and stable separability. In contrast, letters like *zayin* and *tet* showed near-random classification performance. The relatively low standard deviations across most letters indicate stable classification results across folds.

In contrast, when using cluster-derived labels (from k=2 to k=7 hierarchical clustering), classification performance consistently dropped below that of the SA/SB labeling. Even the best-performing configurations reached AUC scores only around 0.516, and many configurations yielded results close to random (AUC ≈ 0.5). This suggests that clustering outputs did not map cleanly onto character-level distinctions.

A major contributing factor was the severe imbalance in cluster sizes. Many clusterings produced small groups containing only a single column, which in turn drastically limited the number of characters available for training and evaluation. For example, while the SA/SB scheme utilized characters from 51 columns, some clustering-based labels relied on just one or two columns per group. This led to fragile model generalization and unreliable cross-validation splits.

In other words, even when we zoom in to individual characters, *hinge* features still can't deliver strong discrimination: character-level classification based on *hinge* yields only modest AUCs and fails to assign letters to their true columns reliably. This reinforces the idea that *hinge*, while capturing some stroke-direction statistics, lacks the richness needed for robust writer separation. Thus, although unsupervised *hinge*-based clustering does reveal some column-level style shifts, it isn't substantiated at the character level, where the SA/SB two-scribe labels consistently outperform *hinge* groupings, suggesting that *hinge* alone is not an adequate feature for this task.

PCA and T-SNE Visualization

In this section, we discuss the *hinge* distributions revealed by the PCA and t-SNE embeddings shown in Fig. 5. In our pairwise writer experiments, we found that PCA and t-SNE behaved quite differently depending on how strongly the internal-gradient *hinge* features separated two writers. For the "most similar" pair (IDs 29501 and 42901, silhouette ≈ 0.018), both PCA and t-SNE collapsed the yellow and purple points into one blended cloud. In the PCA plot, the two writers' points overlap almost entirely with no clear gap, and in the t-SNE embedding, they intermingle completely, there were no distinct islands or boundaries. Yet when we look at their handwriting side by side, we can perceive subtle stylistic differences that *hinge* simply failed to capture. This suggests that, although *hinge* can sometimes distinguish between radically different writers, it is not sensitive enough to register more nuanced, human-perceived similarities.

By contrast, the "most different" pair (IDs 52301 and 55601, silhouette ≈ 0.627) produced two tight, well-separated clusters in both PCA and t-SNE. In PCA, their clouds lay on opposite sides of the first principal component, with minimal overlap, and in t-SNE, each writer's points formed compact islands surrounded by empty space. Here, *hinge* features did exactly what we hoped: they amplified a truly large stylistic gap into a clear separation in both embeddings.

We then treated each of the 52 columns as its own data group. First, we plotted every 28×28 patch in PCA space and saw one long, overlapping ribbon of points—only a few columns reached the edges. In t-SNE, the points broke into small islands at the periphery, but each "island" still contained patches from multiple columns. Next, to reduce noise, we replaced each column's hundreds of patches with a single averaged-feature centroid. In the PCA of these 52 centroids, the points arranged themselves along a smooth, curved arc, roughly ordered by their total gradient strength. In the t-SNE of the same centroids, they formed a loose U-shaped pattern in which adjacent columns lay nearest each other. Even after this simplification, neither embedding produced sharp, separate clusters. This makes clear that *hinge* features alone are too coarse to capture the subtle, within-page variations among columns.

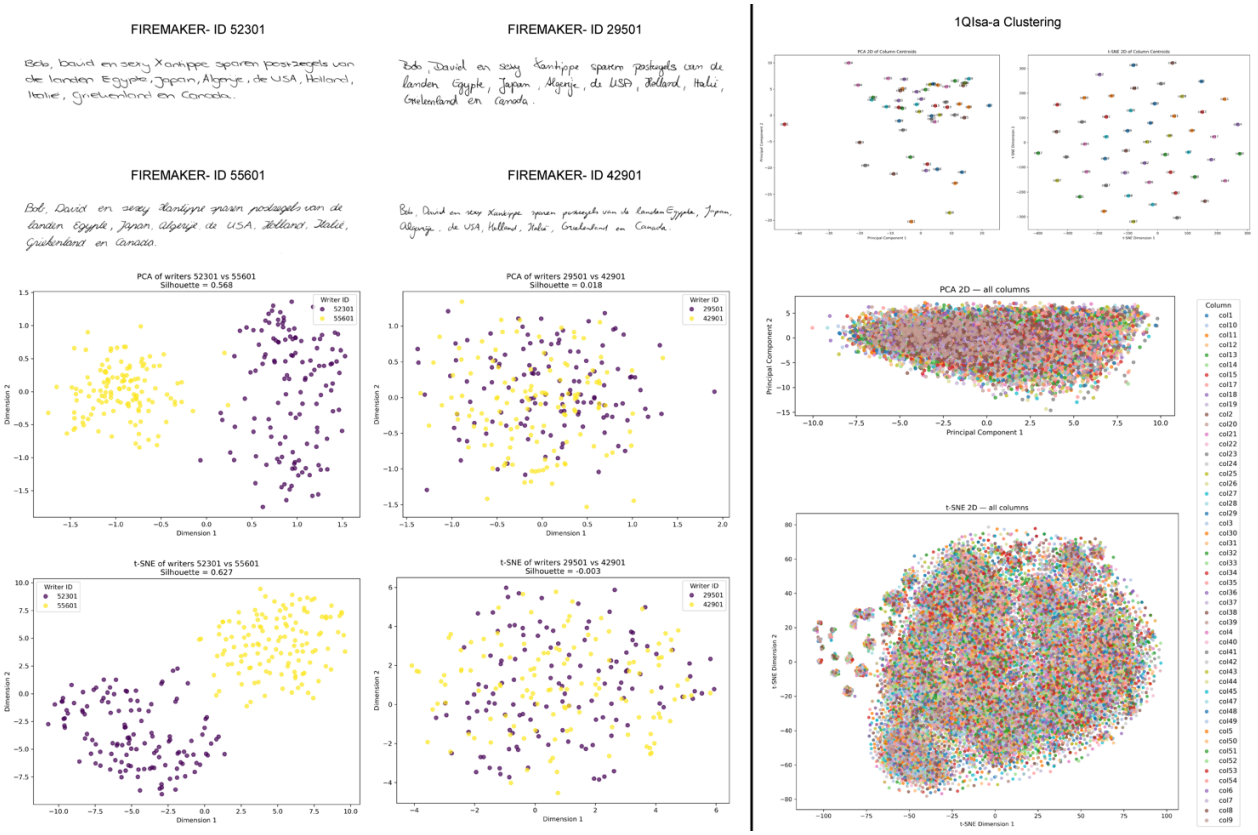

Figure 5. Three-column comparison of PCA and t-SNE embeddings alongside handwriting overviews.
Column 1: Distribution of writers 52301 vs 55601 (most different pair) in PCA (top) and t-SNE (bottom), paired with sample snippets from each writer.
Column 2: Distribution of writers 29501 vs 42901 (most similar pair) in PCA (top) and t-SNE (bottom),
Column 3: Overview of the 1QIsa[a] page: PCA (top) and t-SNE (bottom) embeddings of every 28×28 patch versus embeddings of one averaged-feature centroid per column (at the very top).

Taken together, these visualizations suggest that PCA and t-SNE are excellent at revealing cases where *hinge* features diverge strongly, highlighting the most different writers with crisp separations, but they struggle whenever the underlying feature differences are subtler. In particular, although we can numerically detect the most extreme separability, *the hinge* method misses many of the finer stylistic cues that human readers use to distinguish similar handwriting. Likewise, on the column-clustering front, even aggregated centroids only uncover broad trends rather than discrete column groupings. Going forward, combining *hinge* with additional feature types or using learned embeddings may be necessary to capture the nuanced patterns inherent in both inter-writer and intra-page variation.

Conclusions

Our investigation shows that *hinge* features, while capable of capturing pronounced stylistic differences, fall short when it comes to the subtler variations within 1QIsa[a]. Unsupervised hierarchical clustering of columns, using pairwise distances computed on *hinge*-feature summaries, yielded a coarse two-group split but produced highly imbalanced or fragile partitions beyond that. When we treated those cluster assignments as labels and attempted to verify them at the character level, the resulting classification performance was poor and often close to chance. In contrast, the traditional two-scribe (SA/SB) division produced more consistent separability, yet even those AUCs remained only in the moderate range, underscoring that neither approach provides sufficiently robust attribution on its own. PCA and t-SNE visualizations reinforce these conclusions: *hinge* embeddings cleanly separate radically different hands but collapse more similar ones and fail to resolve within-page column distinctions, even when a single aggregated centroid represents each column. Taken together, these results suggest that *hinge* features alone do not offer adequate discriminatory power for reliable scribal attribution in the 1QIsa[a] Scroll. Future work should therefore integrate complementary feature descriptors or explore alternative representations to capture the nuanced patterns of ancient Hebrew script.

References

[1] J. T. Jull, D. J. Donahue, M. Broshi, and E. Tov, “Radiocarbon Dating of Scrolls and Linen Fragments from the Judean Desert,” Radiocarbon, vol. 37, no. 1, 11–19 (1995).

[2] R. Chadwick, “Qumran Cave 1, II: The Isaiah Scrolls. Part 1: Plates and Transcriptions; Part 2: Introductions, Commentary, and Textual Variants [review] / Ulrich, Eugene and Peter W. Flint, with a contribution by Martin G. Abegg Jr..” Andrews University Seminary Studies (AUSS) 52.2 (2014). Available at: https://digitalcommons.andrews.edu/auss/vol52/iss2/31 (Accessed February 27th, 2025).

[3] E. D. Reymond, “The Scribe of 1QS, 1QSa, 1QSb, 4Q53 (4QSamc), 4Q175 and Three Features of Orthography and Phonology.” Dead Sea Discoveries, 25(2), 238-254 (2018).

[4] J. A. Manginelli, “A Suffering Messiah? Messianic Expectations in Light of the Dead Sea Scrolls: The Case of 1QISA 52:14,” Doctor of Ministry. 492, (2008). Available at: https://digitalcommons.georgefox.edu/dmin/492 (Accessed February 27th, 2025).
[5] E. Tov, Scribal Practices and Approaches Reflected in the Texts Found in the Judean Desert (Brill, Leiden, The Netherlands, 2018).
[6] M. Popovic, M. A. Dhali, and L. Schomaker, “Artificial intelligence based writer identification generates new evidence for the unknown scribes of the dead sea scrolls exemplified by the great isaiah scroll (1qisaa),” PLOS ONE 16, 1–28 (2021).
[7] L. Schomaker, L. Vuurpijl. “Firemaker image collection for benchmarking forensic writer identification using image-based pattern recognition, ” (2002).
[8] M. A. Dhali, J. W. de Wit, and L. Schomaker, “Binet: Degraded-manuscript binarization in diverse document textures and layouts using deep encoder-decoder networks,” (2019).
[9] T. Tobing, P. Škrabánek, S. Yildirim Yayilgan, S. George, and T. Elgvin, “Character-based writer verification of ancient hebrew square-script manuscripts: On edge-direction feature,” *Arch. Conf.*, vol 20, 155–158 (2023).
[10] T. L. Tobing, S. Y. Yayilgan, S. George, and T. Elgvin, “Isolated Handwritten Character Recognition of Ancient Hebrew Manuscripts,” *Arch. Conf.*, vol. 19, 35–39 (2022)
[11] M. Bulacu, L. Schomaker and L. Vuurpijl, "Writer identification using edge-based directional features," Seventh International Conference on Document Analysis and Recognition. Proceedings., Edinburgh, UK, 937-941 (2003).

Author Biography

Tabita A. M. L. Tobing received her BSc in Physics from Universitas Indonesia (2016) and an MSc in Electrical Engineering from Institut Teknologi Bandung (2021). She is currently a PhD student at NTNU under The Lying Pen of Scribes project; her research field is computational analysis for ancient Hebrew manuscripts using machine learning.

Patrick Bours earned an MSc (1990) and PhD (1994) in Discrete Mathematics from Eindhoven University of Technology. After a decade at the Netherlands National Communication Security Agency, specializing in public key cryptography and random number generation, he joined NISlab in 2005 as a PostDoc. Appointed Associate Professor in 2008 and Full Professor in 2012, his research focuses on behavioral biometrics, including gait recognition, keystroke dynamics, and continuous authentication, with applications in security and user identification.