Journal of open archaeology data

Handwritten Babylonian Numerals (HBN) Dataset

DATA PAPERS

LOAY ALZUBAIDI
JABER JEMAI
ASHRAF ALKARAKI
RAJESH THOMAS
*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

The Handwritten Babylonian Numerals (HBN) Dataset is an open, balanced image collection of 14,000 original handwritten samples contributed by 100 participants. It represents 14 numeral classes: 1–9, 10, 20, 30, 40, and 50. Each participant reproduced every class ten times on a structured form. Forms were digitized at 300 dpi, segmented using predetermined coordinates, stripped of printed frames, converted to grayscale, and standardized to 128 × 128 pixels. The multi-writer design preserves variation in symbol shape, stroke formation, spacing, orientation, and handwriting style. HBN supports reproducible research in ancient numeral recognition, image classification, structural analysis, digital humanities, and computational archaeology.

CORRESPONDING AUTHOR:
Loay Alzubaidi
Faculty of Computing and Information Systems, Higher Colleges of Technology, United Arab Emirates
lalzubaidi@hct.ac.ae

KEYWORDS:
Babylonian numerals; cuneiform; handwriting recognition; image dataset; ancient scripts; computational archaeology; digital humanities

TO CITE THIS ARTICLE:
Alzubaidi L, Jemai J, Alkaraki A, Thomas R 2026 Handwritten Babylonian Numerals (HBN) Dataset. *Journal of Open Archaeology Data*, 14: 29, pp. 1–11. DOI: https://doi.org/10.5334/joad.236

DATASET

Title: Handwritten Babylonian Numerals (HBN) Dataset.

DOI or unique identifier: 10.5281/zenodo.22843086 (version 1.0.0).

Dataset creator: Loay Alzubaidi. Dataset contributors (data curation): Jaber Jemai; Ashraf Alkaraki; Rajesh Thomas.

Publisher/repository: Zenodo (archival dataset record); GitHub (related repository).

Location: https://zenodo.org/records/22843086; DOI: https://doi.org/10.5281/zenodo.22843086; related GitHub repository: https://github.com/drloay40/HBN-Dataset.

Dataset License: Creative Commons Attribution 4.0 International (CC BY 4.0). The published Zenodo record includes README.md and LICENSE.txt alongside HBN-Dataset.zip.

SUMMARY

BACKGROUND AND SCOPE

Babylonian numerical notation belongs to the sexagesimal (base 60) mathematical tradition of ancient Mesopotamia and occurs in mathematical, astronomical, administrative, and commercial records [1–3]. Its written forms are constructed from wedge-shaped signs whose number and spatial arrangement encode numerical values. Computational analysis of cuneiform material has demonstrated the potential of image-based methods for ancient-script research [4], while benchmark handwriting datasets have shown the value of standardized, reusable image collections for recognition research [5]. The HBN acquisition design and handwritten domain scope are further documented in [7].

HBN was created to address the limited availability of structured image resources dedicated specifically to handwritten Babylonian numerals. It contains modern handwritten reproductions of selected Babylonian numeral forms rather than photographs or scans of archaeological clay tablets. This distinction is important: HBN is intended as a controlled computational resource for studying visual representation and recognition of ancient numeral forms, not as a substitute for archaeological tablet imagery. The controlled handwriting design complements, rather than replaces, archaeological image collections: it isolates the visual structure of numeral forms from confounding factors such as tablet texture, erosion, lighting, and three dimensional wedge impressions. This makes the collection useful for experiments that focus specifically on symbol geometry, stroke arrangement, and writer variability before models are tested on more difficult archaeological domains. The need for clearly documented image collections is also reflected in surveys of historical document image datasets, which emphasize dataset characteristics, accessibility, and suitability for comparative document analysis [10].

Distinct contribution. Unlike generic handwriting benchmarks, HBN targets 14 selected Babylonian numeral forms and uses a balanced 100-writer, ten-repetition-per-class acquisition design. Unlike archaeological cuneiform image collections, its samples are contemporary handwritten reproductions without clay texture or excavation-related variation. This controlled domain supports focused study of numeral shape, stroke arrangement, and writer variation, but cannot establish performance on authentic archaeological tablets.

The collection contains 14,000 original images from 100 participants. Fourteen classes represent the values 1–9, 10, 20, 30, 40, and 50. Each participant contributed ten repetitions per class, producing exactly 1,000 images per class. As shown in Figure 1, the dataset covers the reference numeral forms for all 14 HBN classes. The public HBN dataset repository is identified in the Dataset Description section and provides direct access to the released resource.

SPATIAL COVERAGE

The handwriting data were collected at the Higher Colleges of Technology, United Arab Emirates. The dataset does not encode archaeological findspots or geographic provenance because its records are newly collected handwritten reproductions rather than excavated artefacts.

TEMPORAL COVERAGE

The represented notation derives from the ancient Babylonian sexagesimal tradition. The handwritten Babylonian numeral samples were collected at the Higher Colleges of Technology, United Arab Emirates, between September 2023 and December 2024. The released images represent contemporary handwritten reproductions of selected Babylonian numeral forms rather than authentic archaeological inscriptions. The collection period refers to the acquisition of modern handwriting samples and does not represent the historical period of the Babylonian numeral system.

DATA DESCRIPTION

The following specifications summarize the scope, acquisition design, access conditions, and related publications for the HBN Dataset (Table 1).

DATASET FILES AND ORGANIZATION

Object Name

Handwritten Babylonian Numerals (HBN) Dataset/HBN-Dataset.

Data Type

Primary image data: original handwritten numeral samples collected on structured paper forms and processed into grayscale image files.

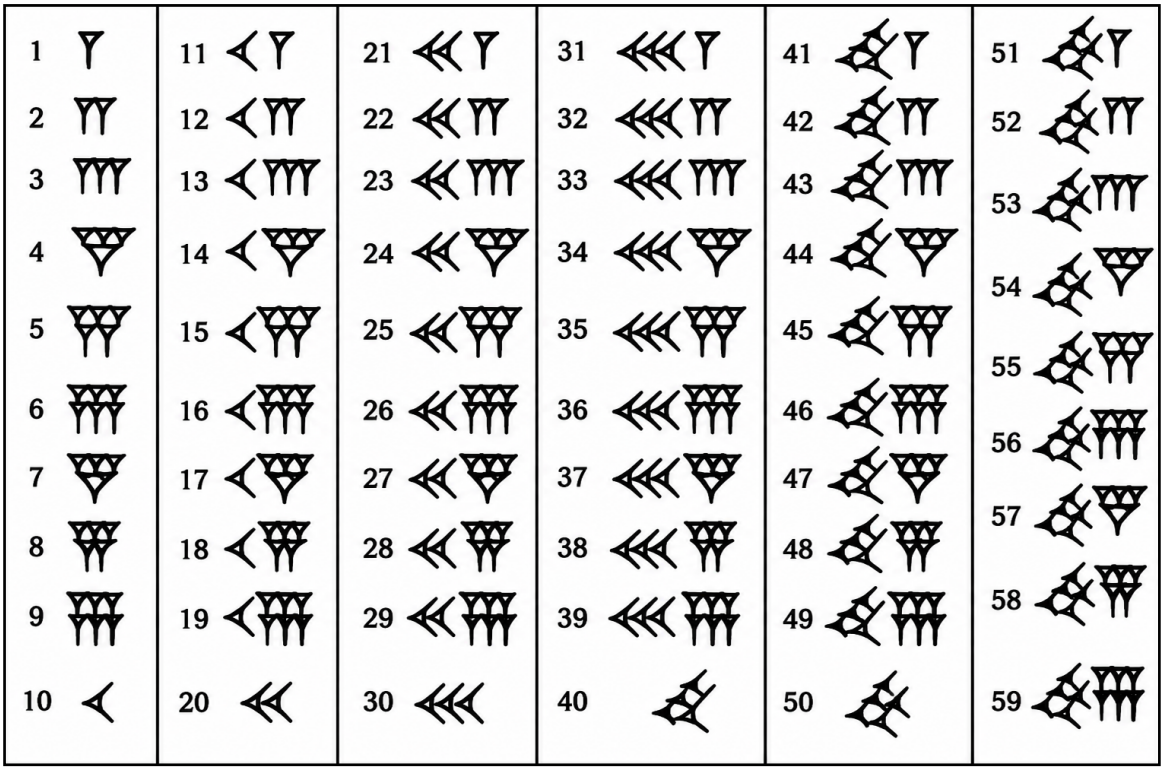

Figure 1 Babylonian numeral reference forms for the 14 HBN classes.

SUBJECT	COMPUTER SCIENCE
Specific subject area	Image processing, pattern recognition, historical script recognition, historical document analysis, and handwritten Babylonian numeral recognition.
Type of data	Grayscale images.
Data collection	Each of 100 participants completed a structured form containing 14 base numeral classes (1–9, 10, 20, 30, 40, and 50). Each class was written 10 times, giving 140 samples per participant and 14,000 original handwritten samples in total [6–8].
Data source location	Higher Colleges of Technology, United Arab Emirates.
Data accessibility	Published Zenodo record (version 1.0.0): https://zenodo.org/records/22843086; DOI: 10.5281/zenodo.22843086. Files: HBN-Dataset.zip, README.md, LICENSE.txt. License: CC BY 4.0. Related GitHub repository: https://github.com/drloay40/HBN-Dataset.
Related research articles	Alzubaidi et al., Algorithms (2026), DOI: 10.3390/a19040322 [8]. Alzubaidi et al., Applied Computational Intelligence and Soft Computing (2026), DOI: 10.1155/acis/4724541 [7]. Alzubaidi et al., IEEE ICOCO (2024), DOI: 10.1109/ICOCO62848.2024.10928249 [6].

Table 1 Specifications of the HBN Dataset.

Dataset Structure

Images are organized into class-specific directories corresponding to the 14 Babylonian numeral classes. Each image filename follows a structured naming convention that encodes the anonymized participant identifier, numeral class, and repetition number (Table 2). For example, the filename **01.jpg_1_1.png** represents participant 01, numeral class 1, and repetition 1. The embedded ".jpg" string is retained from the original naming procedure, while ".png" indicates the final processed image format.

This naming convention facilitates sample traceability, dataset integrity verification, and reproducible participant disjoint evaluation. In participant-disjoint evaluation, all samples belonging to a particular participant are assigned exclusively to one dataset partition (training, validation, or testing), thereby preventing writer level data leakage and supporting a more reliable assessment of generalization to previously unseen handwriting styles.

File Formats and Image Specifications

Processed dataset images are grayscale PNG files at 128 × 128 pixels. Example released filenames include 01.jpg_1_1.png; the embedded ".jpg" is part of the filename, while ".png" is the file extension.

Dataset Creators

Zenodo lists Loay Alzubaidi as dataset creator (data collector) and Jaber Jemai, Ashraf Alkaraki, and Rajesh Thomas as data-curation contributors. All four

PROPERTY	VALUE
Numeral system	Babylonian (base 60)
Number of classes	14
Class values	1–9, 10, 20, 30, 40, 50
Participants	100
Repetitions per class per participant	10
Samples per participant	140
Original samples per class	1,000
Total original images	14,000
Image type	Grayscale
Image size	128 × 128 pixels
Scanning resolution	300 dpi
Augmentation included	No
License	CC BY 4.0

Table 2 Summary of the HBN Dataset.

are authors of this Data Descriptor; author-specific manuscript contributions are given below.

License

Creative Commons Attribution 4.0 International (CC BY 4.0).

Repository Location and Data Access

The archived, publicly accessible HBN Dataset (version 1.0.0) is published on Zenodo at https://zenodo.org/records/22843086 (DOI: 10.5281/zenodo.22843086). Its files include HBN-Dataset.zip, README.md, and LICENSE.txt, released under CC BY 4.0. The related GitHub repository is https://github.com/drloay40/HBN-Dataset. The broader cuneiform field also benefits from established digital infrastructures such as the Cuneiform Digital Library Initiative [13] HBN-Dataset (GitHub).

Example Data

As illustrated in Figure 2, representative segmented HBN samples for numeral values 1, 2, 5, and 10 retain the handwritten symbols after removal of the printed collection-form frames.

Dataset Constraints and Limitations

HBN was collected on structured paper forms and represents selected handwritten numeral classes rather than authentic clay tablet imagery or complete cuneiform texts. It therefore does not reproduce tablet texture, erosion, three dimensional wedge impressions, illumination effects, or other archaeological acquisition conditions. The collection covers 14 selected classes rather than complete sexagesimal numeral sequences. Models trained on HBN should not be assumed to transfer directly to archaeological tablet imagery without additional domain-specific data and evaluation. These constraints define the appropriate interpretation of results: strong performance on HBN demonstrates recognition of controlled handwritten numeral forms, not automatic equivalence to performance on excavated cuneiform tablets (Table 3).

METHODS

SAMPLING STRATEGY

One hundred students from the Higher Colleges of Technology, United Arab Emirates, participated in the handwriting data collection conducted between September 2023 and December 2024. Participants were introduced to the 14 selected Babylonian numeral symbols (1–9, 10, 20, 30, 40, and 50) during the data-collection process. Each participant completed a structured handwriting form containing predefined writing regions for all 14 numeral classes, with ten repetitions allocated to each class. Consequently, each participant contributed 140 handwritten samples, resulting in a balanced dataset of 14,000 original images, with 1,000 samples per numeral class (Table 4).

The structured acquisition procedure was designed to capture natural variations in handwriting characteristics, including symbol shape, stroke formation, thickness, spacing, and orientation. Although participants were introduced to the Babylonian numeral symbols during data collection, their prior familiarity with the numeral system was not formally assessed. Participant demographic characteristics were not reported; therefore, no claims regarding demographic diversity are made.

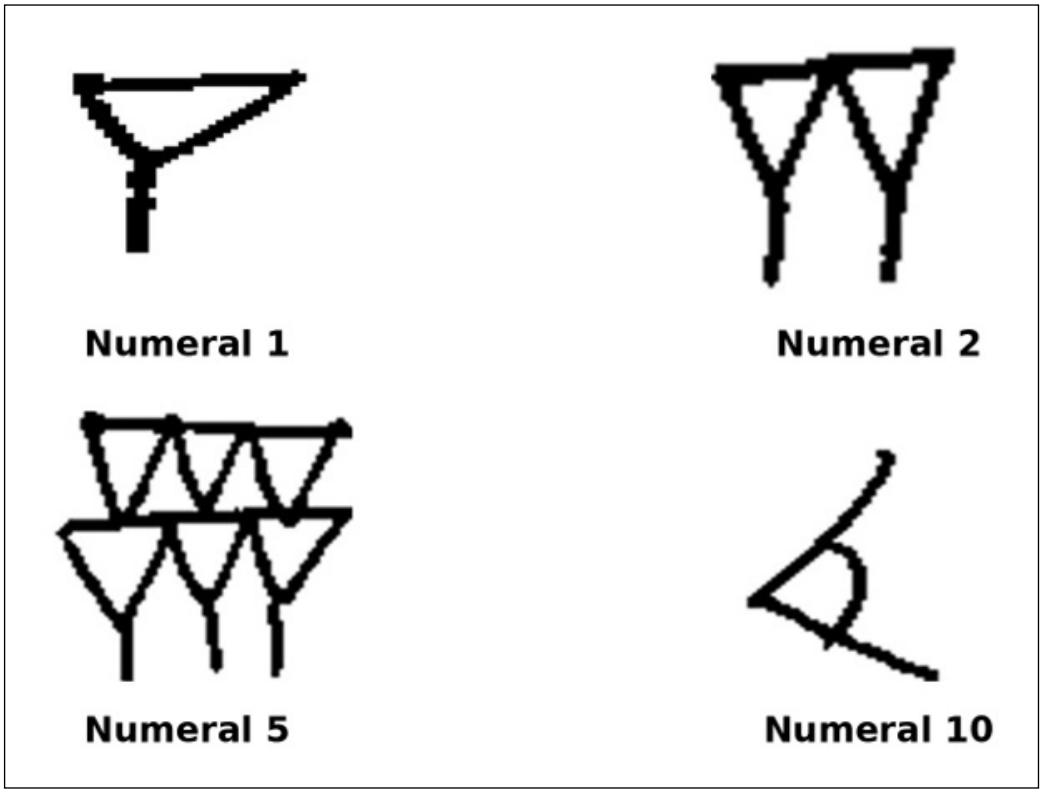

Figure 2 Segmented HBN examples for values 1, 2, 5, and 10 after printed frame removal.

DATASET CHARACTERISTIC	IMPLICATION	RECOMMENDED PRACTICE
Modern handwritten reproductions	Domain differs from archaeological tablets	Report HBN results as handwriting-domain results
Structured collection forms	Acquisition is more controlled than field imagery	Test robustness separately on uncontrolled images
14 selected classes	Does not represent every sexagesimal expression	State the evaluated label space explicitly
100 writers	Enables writer variation studies	Prefer participant disjoint evaluation splits
Balanced acquisition	Each class has equal original representation	Use per class metrics in addition to overall accuracy
No archaeological provenance per image	Images are newly created research data	Do not infer findspot, period, or tablet provenance

Table 3 HBN limitations and recommended reuse strategies.

DESIGN ELEMENT	VALUE	ROLE IN THE DATASET
Participants	100	Provides multi-writer handwriting variation
Numeral classes	14	Values 1–9, 10, 20, 30, 40, and 50
Repetitions per class per participant	10	Captures within writer variation
Samples per participant	140	14 classes × 10 repetitions
Samples per class	1,000	100 participants × 10 repetitions
Total original samples	14,000	100 participants × 140 samples
Scanning resolution	300 dpi	Preserves stroke and symbol detail during digitization

Table 4 HBN acquisition design and expected sample counts.

DATASET GENERATION PROCEDURE

The dataset generation procedure comprised the following sequence: The structured grid is central to reproducibility because the position of each writing cell determines its class label. Coordinate based extraction therefore reduces manual annotation while preserving the participant-class-repetition relationship encoded by the collection design (Table 5).

Structured form completion: each participant wrote all 14 numeral classes ten times.

Digitization: completed forms were scanned at 300 dpi.

Coordinate based localization: predetermined row and column boundaries identified the 140 writing regions on each form.

Individual extraction: each predefined region was cropped to obtain one handwritten numeral.

Position based labelling: the class label was assigned from the known position of the writing region.

IMAGE PREPROCESSING

Following segmentation, the printed square frames were removed from the extracted images to eliminate unwanted form boundaries while retaining the handwritten numeral symbols. The images were subsequently converted to grayscale, and surrounding background regions were reduced while preserving the complete symbol structure. Each image was then standardized to a resolution of 128 × 128 pixels to ensure consistent dimensions for subsequent image-processing and machine-learning applications.

The preprocessing procedure was designed to preserve the essential geometric characteristics of Babylonian numeral symbols, including stroke arrangement, relative positioning, shape variation, and individual handwriting characteristics. The initial scanning resolution of 300 dpi supported the capture of handwritten stroke details before image standardization. However, the effect of resizing on fine stroke thickness was not quantitatively evaluated.

A sample of the processed images was visually inspected to assess whether the handwritten numeral symbols remained visible and whether printed collection-form boundaries had been excluded. This inspection provided a qualitative assessment of the preprocessing output but did not constitute a comprehensive image-by-image validation of the entire dataset.

Identification and storage: Participant, numeral value, and repetition identifiers were associated with each image, and the images were organized by numeral class.

Dataset verification: Expected sample totals were checked against the acquisition design. Additional release-level integrity checks and their verification status are discussed in the Quality Control section.

Figure 3 illustrates the structured handwriting collection form, including the 14 numeral classes and ten repetitions allocated to each class. Figure 4 summarizes the HBN Dataset generation procedure, beginning with participant data collection and 300 dpi scanning, followed by segmentation, preprocessing, class labeling, and dataset organization, and ending with the final collection of 14,000 original handwritten numeral images.

No data augmentation was used to generate the 14,000 images included in HBN Dataset version 1.0.0. Any augmentation performed in subsequent machine-learning experiments constitutes an optional downstream operation that produces derivative samples separate from the original released dataset.

QUALITY CONTROL

The structured handwriting collection forms provided a consistent framework for data acquisition, segmentation, and position-based labeling. Each predefined writing region corresponded to a specific Babylonian numeral class, reducing the need for manual class annotation and supporting labeling consistency. During preprocessing, a sample of the extracted images was visually inspected to assess whether the handwritten numeral symbols were retained and the printed collection-form boundaries were excluded.

The processed images were specified as grayscale PNG files with standardized dimensions of 128 × 128 pixels.

STAGE	OPERATION	PURPOSE/OUTPUT
Digitization	Scan completed forms at 300 dpi	Create high resolution digital source forms
Grid localization	Use predefined row/column positions	Locate the 140 expected writing cells
Region extraction	Crop one numeral from each cell	Produce individual numeral images
Position-based labeling	Assign class from known cell position	Avoid manual class annotation
Frame removal	Exclude printed collection box boundaries	Retain handwriting rather than template graphics
Grayscale conversion	Standardize image representation	Reduce channel complexity
Spatial standardization	Resize to 128 × 128 pixels	Provide consistent model input dimensions
File organization	Store by class with structured identifiers	Supporting reproducible loading and partitioning

Table 5 Processing stages and their reproducibility purpose.

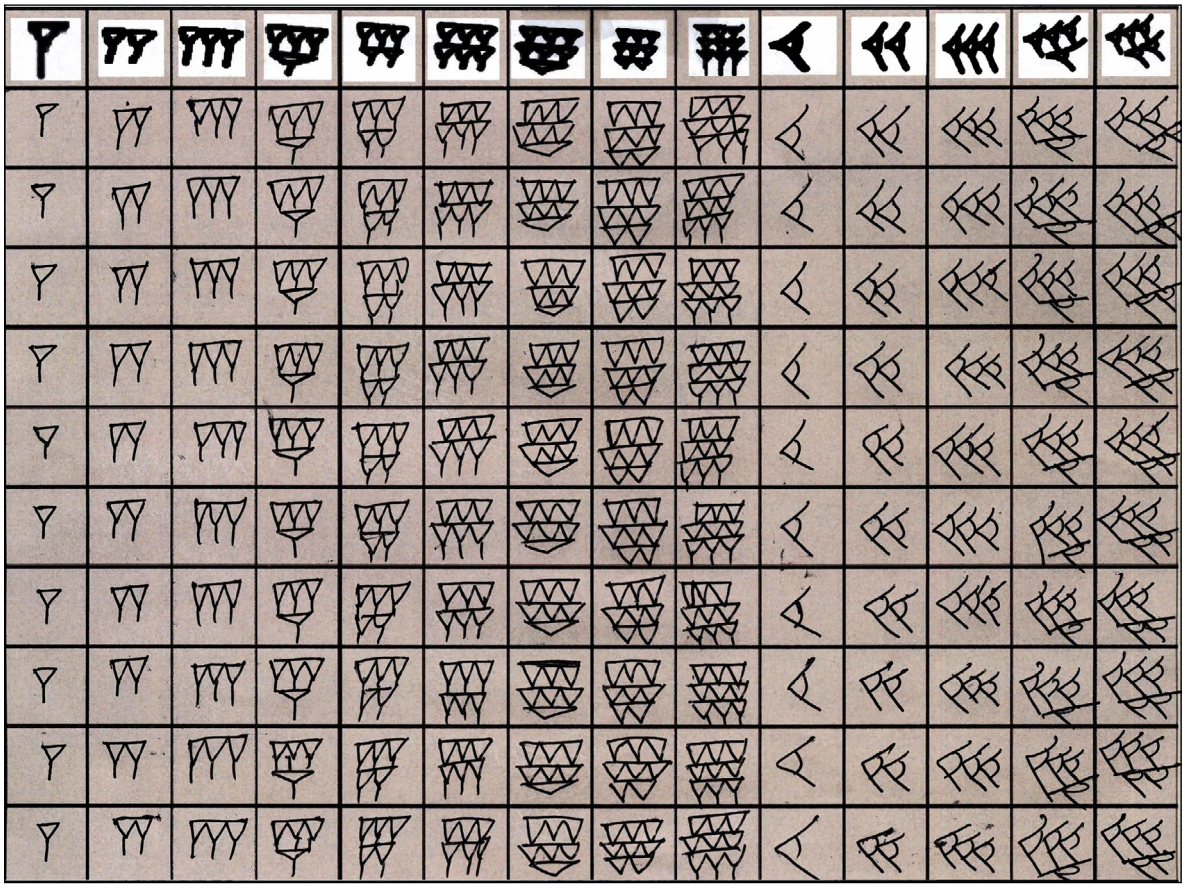

Figure 3 Example structured HBN data collection form with 14 classes and ten repetitions per class.

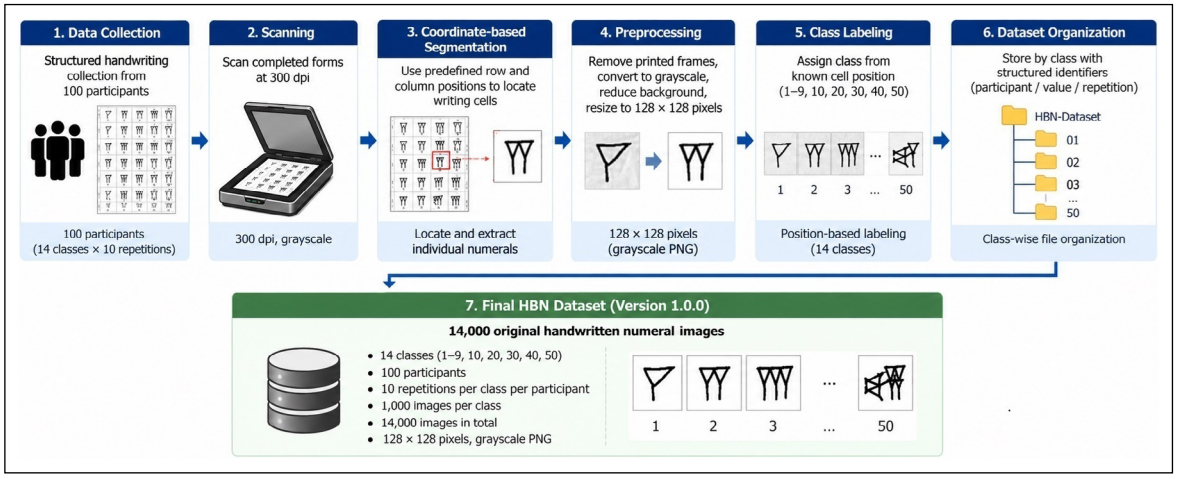

Figure 4 Overview of the HBN Dataset generation pipeline. The procedure includes structured handwriting collection from 100 participants, scanning at 300 dpi, coordinate-based segmentation, image preprocessing, position-based labeling, dataset organization, and generation of the final collection of 14,000 original handwritten numeral images across 14 classes. Data augmentation is not included in HBN Dataset version 1.0.0. Any downstream augmentation produces derivative samples separate from the original dataset.

However, the supplied documentation does not establish that every file in the published HBN Dataset version 1.0.0 archive has undergone independent integrity verification. Accordingly, the expected dataset counts and image specifications are distinguished from the results of completed release-level checks.

A comprehensive archive-level audit would involve verifying the total number of readable images, participant and class identifiers, image dimensions, filename consistency, missing records, and unintended duplicates. The verification status of these checks is summarized in Table 6. Checks for which completed results are unavailable are explicitly identified as not independently verified.

The HBN Dataset has additionally been used in three recognition studies [6–8], demonstrating its practical utility for handwritten Babylonian numeral recognition. These publications are cited as examples of dataset reuse

INTEGRITY CHECK	EXPECTED CONDITION	VERIFICATION STATUS
Participant count	100 anonymized participants	Documented acquisition design; archive verification pending
Class count	14 numeral classes	Documented acquisition design; archive verification pending
Per-participant count	140 images	Expected; archive verification pending
Per-class count	1,000 images	Expected; archive verification pending
Total image count	14,000 original images	Documented dataset total; independent audit pending
Image dimensions	128 × 128 pixels	Specified preprocessing output; full archive verification pending
Image readability	All images readable	Not independently verified
Filename identifiers	Valid participant, class, and repetition identifiers	Verification pending
Missing records	No missing expected records	Not independently verified
Label consistency	Correct correspondence between image and numeral class	Position-based labeling documented; full verification pending
Unintended duplicates	No unintended duplicate files	Not independently verified

Table 6 Dataset integrity checks derived from the HBN acquisition design.

rather than as new experimental results of the present study. In particular, StructFusionNet reported 97.39% accuracy, 97.39% precision, 97.44% recall, and 97.40% F1-score for the 14-class classification task [7]. The multi-symbol recognition study [8] generated composite images from HBN glyphs; these synthetic composites are separate from the original 14,000-image dataset and are not included in the released collection.

USER NOTES

The Handwritten Babylonian Numerals (HBN) Dataset provides a focused benchmark resource for research in ancient script recognition, computational archaeology, image processing, historical document analysis, and machine learning. Its principal value lies in the following characteristics:

- The collection contains 14,000 original labeled grayscale images contributed by 100 participants. The multi-writer design captures natural variation in symbol shape, stroke formation, spacing, orientation, and handwriting style [7].
- The acquisition design is balanced: every participant wrote each of the 14 selected numeral classes ten times, yielding 140 samples per participant and exactly 1,000 original images per class. This structure supports controlled comparison across classes and writers [6–8].
- A structured acquisition and coordinate-based extraction procedure reduces manual labeling effort. The known position of each numeral on the collection form supports reproducible segmentation and position based class assignment.
- The dataset supports numeral classification, writer-independent recognition, structural feature analysis, representation learning, historical document processing, and computational studies of ancient numeral forms. Broader work on machine learning for ancient languages and historical document datasets demonstrates the relevance of reusable, well documented image resources for these tasks [9, 10].
- The HBN Dataset has already supported several research directions, including CNN-SVM recognition [6], structural feature fusion with StructFusionNet [7], and multi-symbol recognition with positional decoding [8]. These published uses provide evidence of practical reuse beyond the data paper itself.
- The standardized 128 × 128 grayscale representation and class-based organization make the images suitable for common computer vision workflows in frameworks such as MATLAB, TensorFlow, and PyTorch. Researchers should nevertheless report their own preprocessing and partitioning procedures explicitly.
- The public availability of HBN supports reproducibility, comparative evaluation, teaching, and future extension. It can also serve as a controlled handwritten source domain for studies that later evaluate transfer to authentic cuneiform or archaeological imagery [4, 11, 12].
- The dataset should not be interpreted as a substitute for archaeological tablet imagery. Because the samples are modern handwritten reproductions collected on paper forms, they do not contain clay texture, erosion, three-dimensional wedge impressions, or archaeological provenance. This clear domain boundary is important when reporting reuse and model generalization.

REUSE POTENTIAL

HBN provides a balanced, openly reusable multi-writer resource for computational work on Babylonian numeral forms. Its 100 writer structure enables research on writer independent recognition, representation learning, transfer learning, structural feature extraction, and robustness to handwriting variation. Because every class contains exactly 1,000 original samples, researchers can design balanced experiments without class reweighting caused by the acquisition process. Its demonstrated use spans hybrid CNN-SVM recognition, structural-feature fusion, and multi symbol decoding (Table 8) [6–8]. More generally, machine learning research on ancient languages shows how computational resources can support recognition, restoration, and analysis across historical writing systems (Table 7) [9].

The dataset is also relevant to digital humanities and computational archaeology. It can support methodological studies that examine how computer-vision systems represent repeated unit and tens wedge structures, as well as teaching activities involving ancient numeral systems, image preprocessing, classification, and reproducible dataset design. Its structured filenames allow researchers to construct participant independent training, validation, and test partitions, which is important for avoiding writer leakage.

HBN has already supported CNN-SVM recognition [6], structural feature fusion [7], and downstream multi-symbol recognition and positional decoding [8]. The present data paper complements those research articles by documenting the original 14,000-image resource, its acquisition, organization, processing, constraints, and conditions for reuse. It does not duplicate the model contributions or synthetic composite datasets reported in those studies.

Future users may combine HBN with domain-specific cuneiform or archaeological image resources to investigate transfer learning and domain adaptation. Such work should preserve the distinction between modern handwritten reproductions and archaeological tablet imagery. Such transfer oriented evaluation follows naturally from the distinction between handwritten HBN samples and archaeological tablet imagery emphasized in prior HBN work [7]. DeepScribe provides a complementary example in which deep learning is applied to the localization and classification of signs in authentic Elamite cuneiform material [11]. CuReD further demonstrates the use of deep learning based optical character recognition for cuneiform text editions and legacy materials [12].

CONCLUDING USER NOTES

The HBN Dataset provides 14,000 original handwritten Babylonian numeral images from 100 participants across 14 selected numeral classes. Its balanced acquisition design, structured extraction procedure, standardized image representation, and demonstrated reuse in published recognition studies support its value as a benchmark for handwriting recognition, structural feature analysis, and computational research on ancient numeral forms. Because HBN contains modern handwritten reproductions rather than archaeological tablet imagery, future work should evaluate transfer to authentic cuneiform material separately and report that domain distinction explicitly.

RESEARCH TASK	HBN FEATURE THAT SUPPORTS IT	SUGGESTED EVALUATION
Numeral classification	14 balanced labeled classes	Accuracy, macro precision/recall/F1
Writer-independent recognition	100 anonymized writers	Participant disjoint train/test split
Structural feature analysis	Wedge/stroke geometry in 128 × 128 images	Ablation of structural vs. learned features
Representation learning	Repeated samples across writers/classes	Embedding separation and downstream classification
Transfer/domain adaptation	Controlled handwritten source domain	Evaluate separately on archaeological target domain images
Multi-symbol studies	Base glyph inventory	Clearly distinguish generated composites from original HBN records

Table 7 Example research tasks enabled by HBN.

STUDY	RESEARCH USE OF HBN	DEMONSTRATED ROLE
Alzubaidi et al. [6]	CNN-SVM hybrid recognition	Single symbol Babylonian numeral classification
Alzubaidi et al. [7]	StructFusionNet structural feature fusion	Evaluation of structural and convolutional representations
Alzubaidi et al. [8]	Multi symbol recognition and positional decoding	Downstream construction and decoding of more complex numeral expressions

Table 8 Published research directions that have used the HBN dataset.

DATA ACCESSIBILITY STATEMENT

HBN Dataset v1.0.0 is available at https://doi.org/10.5281/zenodo.22843086 under CC BY 4.0. Related repository: https://github.com/drloay40/HBN-Dataset.

ETHICS AND CONSENT

Participation in the data collection process was voluntary. Participants were informed about the purpose of the collection and that their handwritten Babylonian numeral samples could be used for academic research and released publicly as part of an anonymized research dataset. No directly identifiable or sensitive personal information was collected, stored, or released with the dataset. Participant identifiers used for dataset organization were anonymized and do not reveal individual identities. Ethical approval was not required for this data collection under the applicable institutional requirements of the Higher Colleges of Technology, United Arab Emirates.

ACKNOWLEDGEMENTS

The authors thank the volunteers who contributed handwriting samples and acknowledge the colleagues and research assistants who supported dataset preparation, digitization, segmentation, and quality verification.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Loay Alzubaidi: Conceptualization; Methodology; Data curation; Investigation; Supervision; Writing, original draft; Writing, review & editing. Jaber Jemai: Validation; Formal analysis; Writing, review & editing. Ashraf Alkaraki: Validation; Data curation; Writing, review & editing. Rajesh Thomas: Validation; Visualization; Writing, review & editing.

AUTHOR AFFILIATIONS

Loay Alzubaidi orcid.org/0009-0003-5790-5732
Faculty of Computing and Information Systems, Higher Colleges of Technology, United Arab Emirates

Jaber Jemai orcid.org/0000-0003-2363-1625
Faculty of Computing and Information Systems, Higher Colleges of Technology, United Arab Emirates

Ashraf Alkaraki orcid.org/0009-0009-5866-2318
Faculty of Computing and Information Systems, Higher Colleges of Technology, United Arab Emirates

Rajesh Thomas orcid.org/0000-0002-6513-7994
Faculty of Computing and Information Systems, Higher Colleges of Technology, United Arab Emirates

REFERENCES

1. **Robson E.** *Mathematics in Ancient Iraq: A Social History*. Princeton, NJ: Princeton University Press; 2008.
2. **Friberg J.** *A Remarkable Collection of Babylonian Mathematical Texts: Manuscripts in the Schøyen Collection: Cuneiform Texts I*. New York: Springer; 2007. DOI: https://doi.org/10.1007/978-0-387-48977-3
3. **Muroi Y.** The Long Journey to Number. *Math Horizons*. 2009;16(1):22–25. DOI: https://doi.org/10.1080/10724117.2009.11974827
4. **Dencker T, Klinkisch P, Maul SM, Ommer B.** Deep learning of cuneiform sign detection with weak supervision using transliteration alignment. *PLOS ONE*. 2020;15(12):e0243039. DOI: https://doi.org/10.1371/journal.pone.0243039
5. **LeCun Y, Bottou L, Bengio Y, Haffner P.** Gradient based learning applied to document recognition. *Proc IEEE*. 1998;86(11):2278–2324. DOI: https://doi.org/10.1109/5.726791
6. **Alzubaidi L, Al-Qudah I, Bouazza KE.** Deep Learning Approaches to Babylonian Numeral Recognition: A CNN-SVM Hybrid Model. *2024 IEEE International Conference on Computing (ICOCO)*, Kuala Lumpur, Malaysia, 12–14 Dec. 2024. pp. 140–145. DOI: https://doi.org/10.1109/ICOCO62848.2024.10928249
7. **Alzubaidi L, Alnahdi A, Ahmed A, Waheed N, El Hassan A.** *StructFusionNet: Structural Feature Fusion in CNNs for Handwritten Babylonian Numeral Recognition*. Applied Computational Intelligence and Soft Computing. 2026; Article ID 4724541.acis/4724541. DOI: https://doi.org/10.1155/acis/4724541
8. **Alzubaidi L, Bouazza KE, Al-Qudah I.** A Hybrid Deep Learning Framework for Multi Symbol Recognition and Positional Decoding of Handwritten Babylonian Numerals. *Algorithms*. 2026;19(4):322. DOI: https://doi.org/10.3390/a19040322
9. **Sommerschield T, Assael Y, Pavlopoulos J, Stefanak V, Senior A, Dyer C, Bodel J, Prag J, Androutsopoulos I, de Freitas N.** Machine Learning for Ancient Languages: A Survey. *Computational Linguistics*. 2023;49(3):703–747. DOI: https://doi.org/10.1162/coli_a_00481
10. **Nikolaidou K, Seuret M, Mokayed H, Liwicki M.** A survey of historical document image datasets. *International Journal on Document Analysis and Recognition*. 2022;25(4):305–338. DOI: https://doi.org/10.1007/s10032-022-00405-8
11. **Williams EC, Su G, Schloen SR, Prosser MC, Paulus S, Krishnan S.** DeepScribe: Localization and Classification of Elamite Cuneiform Signs via Deep Learning. *ACM Journal*

on Computing and Cultural Heritage. 2025;18(2):Article 31, 1–32. DOI: https://doi.org/10.1145/3716850

12. **Homburg T, Chiarcos C.** CuReD: Deep Learning Optical Character Recognition for Cuneiform Text Editions and Legacy Materials. *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*. 2024:130–140. DOI: https://doi.org/10.18653/v1/2024.ml4al-1.14
13. **Rattenborg R, Pagé-Perron É.** *The Cuneiform Digital Library Initiative: A Primer*. Version 1.0. Zenodo; 2025. DOI: https://doi.org/10.5281/zenodo.15309738

TO CITE THIS ARTICLE:
Alzubaidi L, Jemai J, Alkaraki A, Thomas R 2026 Handwritten Babylonian Numerals (HBN) Dataset. *Journal of Open Archaeology Data*, 14: 29, pp. 1–11. DOI: https://doi.org/10.5334/joad.236

Submitted: 15 September 2026 **Accepted:** 09 November 2026 **Published:** 27 November 2026

COPYRIGHT:
© 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See https://creativecommons.org/licenses/by/4.0/.

Journal of Open Archaeology Data is a peer-reviewed open access journal published by Ubiquity Press.