

# Online Resource 1

The Cretogram Dataset for Isolated Character Recognition in Archaic and Classical Crete

Ilias Gotsikas · Pericles A. Mitkas · Niki Oikonomaki · Aristotle University of Thessaloniki

Dataset DOI: 10.5281/zenodo.22681485

This resource reports the class inventory, individual runs, paired comparisons and implementation settings for the frozen Cretogram experimental snapshot. Percentages are calculated from the stored out-of-fold predictions. Within each fit, training excludes the assigned validation and test glyphs. Section S6 documents a subsequent context-verified annotation correction and separate re-scoring of the unchanged predictions.

## S1 Class inventory and recognition results

**Table S1** Class coverage and mean F1 (%) in the evaluated set

| Class   | Dataset glyphs | Evaluated glyphs | Sources | HOG-SVM | CNN   | ResNet-18 | CNN control | ResNet-18 control |
|---------|----------------|------------------|---------|---------|-------|-----------|-------------|-------------------|
| Alpha   | 1574           | 1570             | 159     | 95.89   | 98.49 | 98.98     | 98.93       | 99.27             |
| Beta    | 20             | 20               | 15      | 52.94   | 84.80 | 85.48     | 86.46       | 94.87             |
| Gamma   | 141            | 140              | 62      | 76.47   | 87.71 | 90.57     | 89.90       | 91.59             |
| Delta   | 387            | 386              | 103     | 93.13   | 95.81 | 96.55     | 96.68       | 97.70             |
| Epsilon | 1148           | 1148             | 139     | 91.00   | 97.84 | 98.73     | 98.51       | 99.16             |
| Digamma | 143            | 142              | 57      | 28.81   | 74.81 | 86.97     | 82.91       | 89.64             |
| Zeta    | 15             | 15               | 11      | 40.00   | 89.10 | 84.72     | 85.36       | 90.32             |
| Heta    | 275            | 273              | 78      | 85.77   | 92.74 | 96.23     | 96.18       | 98.72             |
| Theta   | 135            | 135              | 62      | 89.82   | 97.03 | 98.50     | 97.54       | 99.26             |
| Iota    | 1351           | 1350             | 146     | 94.15   | 97.82 | 99.37     | 98.94       | 99.57             |
| Kappa   | 615            | 614              | 119     | 92.72   | 98.16 | 99.05     | 98.40       | 99.35             |
| Lamda   | 305            | 305              | 99      | 76.07   | 90.53 | 93.47     | 91.01       | 94.30             |
| Mu      | 430            | 428              | 95      | 87.50   | 96.37 | 99.45     | 97.85       | 99.14             |
| Nu      | 1021           | 1019             | 133     | 88.74   | 96.61 | 99.07     | 97.95       | 99.27             |
| Xi      | 2              | —                | 2       | —       | —     | —         | —           | —                 |
| Omikron | 1229           | 1227             | 155     | 98.08   | 98.83 | 99.29     | 99.31       | 99.66             |
| Pi      | 363            | 363              | 87      | 79.06   | 93.66 | 97.00     | 96.01       | 97.57             |
| San     | 719            | 716              | 144     | 92.65   | 98.10 | 98.77     | 98.49       | 98.93             |
| Qoppa   | 32             | 32               | 16      | 69.23   | 73.82 | 87.70     | 92.47       | 97.85             |
| Rho     | 404            | 403              | 115     | 90.17   | 95.14 | 97.07     | 96.62       | 98.30             |
| Tau     | 979            | 979              | 152     | 97.87   | 98.36 | 99.30     | 99.23       | 99.37             |
| Ypsilon | 227            | 226              | 79      | 92.06   | 97.78 | 98.24     | 97.62       | 98.60             |
| Phi     | 1              | —                | 1       | —       | —     | —         | —           | —                 |
| Psi     | 2              | —                | 2       | —       | —     | —         | —           | —                 |
| Omega   | 93             | 92               | 21      | 95.19   | 91.48 | 97.11     | 94.70       | 98.20             |

Experimental snapshot: 11,611 glyphs in 25 classes. Evaluated set: 11,583 glyphs in 22 classes. Both represent 187 inscriptions. Neural class metrics are averaged across seeds 1, 2 and 3. Xi, phi and psi are retained in the full collection and excluded from all training, validation and test partitions. The label correction in Sect. S6 does not alter the values in this table.

## S2 Results by seed

Table S2 Pooled five-fold out-of-fold metrics (%)

| Method / protocol  | Seed | Macro-F1 | Accuracy | Balanced accuracy | Inscription-macro accuracy |
|--------------------|------|----------|----------|-------------------|----------------------------|
| HOG-SVM, source    | 42   | 82.15    | 91.44    | 81.02             | 88.56                      |
| CNN, source        | 1    | 93.35    | 96.94    | 92.48             | 96.44                      |
| CNN, source        | 2    | 93.16    | 96.95    | 92.05             | 96.51                      |
| CNN, source        | 3    | 92.36    | 96.55    | 91.10             | 96.24                      |
| ResNet-18, source  | 1    | 95.89    | 98.36    | 95.35             | 98.47                      |
| ResNet-18, source  | 2    | 95.65    | 98.33    | 95.53             | 98.56                      |
| ResNet-18, source  | 3    | 95.05    | 98.20    | 94.43             | 97.81                      |
| CNN, control       | 1    | 95.33    | 97.97    | 94.37             | 96.72                      |
| CNN, control       | 2    | 95.20    | 98.02    | 94.61             | 96.92                      |
| CNN, control       | 3    | 94.61    | 97.50    | 93.93             | 96.14                      |
| ResNet-18, control | 1    | 97.16    | 98.79    | 97.08             | 98.28                      |
| ResNet-18, control | 2    | 97.39    | 98.77    | 97.43             | 98.10                      |
| ResNet-18, control | 3    | 97.35    | 98.79    | 97.28             | 98.74                      |

Source denotes inscription-disjoint evaluation. Control denotes glyph-level partitions with exactly matched class counts. Each row evaluates the same 11,583 unique glyphs once. The across-seed summaries average these metrics; they do not aggregate the predictions into an ensemble.

## S3 Paired comparisons

Table S3 Paired source-bootstrap differences, in percentage points

| Comparison                 | Metric                     | Difference | 95% interval     |
|----------------------------|----------------------------|------------|------------------|
| ResNet-18 – CNN            | Accuracy                   | +1.48      | +0.95 to +2.14   |
| ResNet-18 – CNN            | Macro-F1                   | +2.57      | +1.46 to +3.79   |
| ResNet-18 – CNN            | Balanced accuracy          | +3.23      | +1.50 to +4.91   |
| ResNet-18 – CNN            | Inscription-macro accuracy | +1.88      | +1.12 to +2.87   |
| CNN control – source       | Accuracy                   | +1.02      | +0.36 to +1.93   |
| CNN control – source       | Macro-F1                   | +2.09      | +0.27 to +4.09   |
| CNN control – source       | Balanced accuracy          | +2.43      | -0.21 to +4.64   |
| CNN control – source       | Inscription-macro accuracy | +0.20      | -0.18 to +0.56   |
| ResNet-18 – HOG-SVM        | Accuracy                   | +6.85      | +5.18 to +8.74   |
| ResNet-18 – HOG-SVM        | Macro-F1                   | +13.38     | +11.08 to +15.86 |
| ResNet-18 – HOG-SVM        | Balanced accuracy          | +14.08     | +11.69 to +16.66 |
| ResNet-18 – HOG-SVM        | Inscription-macro accuracy | +9.72      | +7.91 to +11.69  |
| ResNet-18 control – source | Accuracy                   | +0.49      | +0.08 to +1.06   |
| ResNet-18 control – source | Macro-F1                   | +1.77      | +0.25 to +3.27   |
| ResNet-18 control – source | Balanced accuracy          | +2.16      | +0.23 to +3.62   |
| ResNet-18 control – source | Inscription-macro accuracy | +0.09      | -0.40 to +0.51   |

These exploratory intervals use 5,000 accepted draws of complete inscription groups, shared across methods and seeds. Fits and folds remain fixed. Draws without every true class are rejected; the interval is conditional on class coverage. Intervals describe uncertainty within this corpus and do not include retraining or new outer partitions.

Model comparisons without a protocol qualifier use inscription-disjoint evaluation.

## S4 Individual folds

**Table S4** HOG–SVM, source results (%)

| Seed | Fold | Test n | Macro-F1 | Accuracy | Balanced accuracy | Inscription-macro accuracy | C    |
|------|------|--------|----------|----------|-------------------|----------------------------|------|
| 42   | 1    | 2312   | 75.11    | 86.98    | 74.95             | 89.33                      | 1.0  |
| 42   | 2    | 2283   | 83.82    | 92.16    | 82.67             | 86.99                      | 0.1  |
| 42   | 3    | 2331   | 83.45    | 94.17    | 82.25             | 87.25                      | 0.1  |
| 42   | 4    | 2313   | 84.21    | 92.65    | 84.14             | 88.27                      | 10.0 |
| 42   | 5    | 2344   | 82.50    | 91.25    | 82.36             | 90.58                      | 10.0 |

C was selected exclusively by validation macro-F1. SVM random state was 42. HOG was calculated once per image using fixed parameters; classifier fitting used only the designated training partition.

## Source partitions

| Fold | Train glyphs | Validation glyphs | Test glyphs | Train sources | Validation sources | Test sources |
|------|--------------|-------------------|-------------|---------------|--------------------|--------------|
| 1    | 7,840        | 1,431             | 2,312       | 128           | 23                 | 36           |
| 2    | 7,998        | 1,302             | 2,283       | 127           | 23                 | 37           |
| 3    | 7,874        | 1,378             | 2,331       | 129           | 24                 | 34           |
| 4    | 7,961        | 1,309             | 2,313       | 126           | 23                 | 38           |
| 5    | 8,001        | 1,238             | 2,344       | 123           | 22                 | 42           |

Primary partitions contain all 22 classes and no shared source identifiers, exact normalized rasters or exact horizontal reflections. Some rare classes contribute one validation glyph in a fold. Per-class partition memberships are provided in the machine-readable splits.

**S4 Individual folds (continued)****Table S5** CNN, source results (%)

| Seed | Fold | Test n | Macro-F1 | Accuracy | Balanced accuracy | Inscription-macro accuracy | Best / total epochs |
|------|------|--------|----------|----------|-------------------|----------------------------|---------------------|
| 1    | 1    | 2312   | 86.58    | 94.20    | 86.62             | 96.42                      | 20 / 30             |
| 1    | 2    | 2283   | 93.32    | 96.76    | 92.69             | 93.31                      | 14 / 24             |
| 1    | 3    | 2331   | 95.40    | 97.73    | 93.95             | 97.78                      | 54 / 60             |
| 1    | 4    | 2313   | 95.13    | 97.88    | 95.19             | 97.39                      | 26 / 36             |
| 1    | 5    | 2344   | 97.09    | 98.08    | 97.61             | 97.25                      | 19 / 29             |
| 2    | 1    | 2312   | 88.94    | 94.46    | 87.96             | 97.26                      | 22 / 32             |
| 2    | 2    | 2283   | 94.37    | 97.15    | 93.50             | 93.88                      | 39 / 49             |
| 2    | 3    | 2331   | 92.69    | 97.64    | 90.91             | 96.68                      | 20 / 30             |
| 2    | 4    | 2313   | 94.64    | 98.27    | 94.15             | 97.72                      | 22 / 32             |
| 2    | 5    | 2344   | 94.99    | 97.23    | 96.20             | 96.97                      | 23 / 33             |
| 3    | 1    | 2312   | 85.74    | 93.17    | 84.83             | 97.14                      | 47 / 57             |
| 3    | 2    | 2283   | 94.33    | 97.63    | 93.98             | 93.72                      | 39 / 49             |
| 3    | 3    | 2331   | 93.28    | 97.73    | 92.61             | 97.08                      | 30 / 40             |
| 3    | 4    | 2313   | 95.10    | 97.71    | 95.71             | 97.60                      | 29 / 39             |
| 3    | 5    | 2344   | 92.97    | 96.50    | 92.76             | 95.75                      | 18 / 28             |

Fold labels 1–5 correspond to stored indices 0–4. The main manuscript reports pooled out-of-fold metrics, not an unweighted average of these fold metrics. The best epoch was selected exclusively by validation macro-F1.

**S4 Individual folds (continued)****Table S6** ResNet-18, source results (%)

| Seed | Fold | Test n | Macro-F1 | Accuracy | Balanced accuracy | Inscription-macro accuracy | Best / total epochs |
|------|------|--------|----------|----------|-------------------|----------------------------|---------------------|
| 1    | 1    | 2312   | 92.94    | 96.37    | 93.01             | 98.26                      | 21 / 31             |
| 1    | 2    | 2283   | 96.68    | 98.55    | 96.42             | 98.41                      | 15 / 25             |
| 1    | 3    | 2331   | 95.00    | 98.97    | 93.97             | 98.71                      | 25 / 35             |
| 1    | 4    | 2313   | 97.92    | 99.09    | 98.38             | 98.94                      | 13 / 23             |
| 1    | 5    | 2344   | 96.90    | 98.81    | 97.15             | 98.10                      | 20 / 30             |
| 2    | 1    | 2312   | 91.36    | 96.45    | 92.82             | 98.59                      | 22 / 32             |
| 2    | 2    | 2283   | 97.08    | 98.51    | 96.49             | 98.32                      | 11 / 21             |
| 2    | 3    | 2331   | 96.04    | 99.06    | 95.64             | 99.15                      | 12 / 22             |
| 2    | 4    | 2313   | 96.87    | 99.05    | 97.21             | 98.44                      | 22 / 32             |
| 2    | 5    | 2344   | 97.44    | 98.59    | 98.14             | 98.36                      | 11 / 21             |
| 3    | 1    | 2312   | 90.22    | 96.45    | 89.17             | 97.85                      | 13 / 23             |
| 3    | 2    | 2283   | 94.76    | 98.07    | 93.42             | 95.34                      | 14 / 24             |
| 3    | 3    | 2331   | 96.78    | 99.10    | 97.02             | 99.27                      | 36 / 46             |
| 3    | 4    | 2313   | 95.84    | 98.62    | 96.05             | 97.71                      | 4 / 14              |
| 3    | 5    | 2344   | 97.70    | 98.72    | 98.42             | 98.84                      | 6 / 16              |

Fold labels 1–5 correspond to stored indices 0–4. The main manuscript reports pooled out-of-fold metrics, not an unweighted average of these fold metrics. The best epoch was selected exclusively by validation macro-F1.

**S4 Individual folds (continued)****Table S7** CNN, control results (%)

| Seed | Fold | Test n | Macro-F1 | Accuracy | Balanced accuracy | Inscription-macro accuracy | Best / total epochs |
|------|------|--------|----------|----------|-------------------|----------------------------|---------------------|
| 1    | 1    | 2312   | 96.66    | 98.49    | 95.50             | 98.26                      | 31 / 41             |
| 1    | 2    | 2283   | 95.12    | 98.03    | 95.26             | 97.28                      | 43 / 53             |
| 1    | 3    | 2331   | 95.28    | 97.94    | 95.06             | 97.68                      | 42 / 52             |
| 1    | 4    | 2313   | 95.19    | 97.97    | 93.41             | 96.19                      | 19 / 29             |
| 1    | 5    | 2344   | 93.38    | 97.44    | 92.62             | 97.48                      | 21 / 31             |
| 2    | 1    | 2312   | 96.19    | 98.10    | 95.24             | 97.70                      | 18 / 28             |
| 2    | 2    | 2283   | 95.93    | 98.38    | 96.47             | 97.99                      | 56 / 60             |
| 2    | 3    | 2331   | 95.77    | 98.07    | 96.63             | 98.21                      | 35 / 45             |
| 2    | 4    | 2313   | 94.48    | 97.80    | 93.57             | 96.11                      | 31 / 41             |
| 2    | 5    | 2344   | 93.50    | 97.78    | 91.70             | 97.40                      | 26 / 36             |
| 3    | 1    | 2312   | 97.45    | 98.44    | 97.07             | 97.88                      | 35 / 45             |
| 3    | 2    | 2283   | 95.30    | 98.03    | 95.97             | 97.33                      | 33 / 43             |
| 3    | 3    | 2331   | 94.55    | 97.21    | 94.66             | 96.70                      | 18 / 28             |
| 3    | 4    | 2313   | 93.75    | 97.67    | 91.64             | 96.33                      | 23 / 33             |
| 3    | 5    | 2344   | 91.30    | 96.16    | 90.48             | 94.97                      | 13 / 23             |

Fold labels 1–5 correspond to stored indices 0–4. The main manuscript reports pooled out-of-fold metrics, not an unweighted average of these fold metrics. The best epoch was selected exclusively by validation macro-F1.

**S4 Individual folds (continued)**

Table S7a ResNet-18, control results (%)

| Seed | Fold | Test n | Macro-F1 | Accuracy | Balanced accuracy | Inscription-macro accuracy | Best / total epochs |
|------|------|--------|----------|----------|-------------------|----------------------------|---------------------|
| 1    | 1    | 2312   | 98.61    | 99.09    | 98.61             | 99.01                      | 14 / 24             |
| 1    | 2    | 2283   | 97.15    | 98.90    | 98.65             | 98.78                      | 9 / 19              |
| 1    | 3    | 2331   | 97.34    | 98.76    | 97.93             | 98.91                      | 19 / 29             |
| 1    | 4    | 2313   | 95.61    | 98.27    | 94.27             | 97.81                      | 4 / 14              |
| 1    | 5    | 2344   | 96.74    | 98.93    | 95.51             | 99.09                      | 17 / 27             |
| 2    | 1    | 2312   | 98.21    | 98.96    | 98.14             | 99.11                      | 5 / 15              |
| 2    | 2    | 2283   | 97.19    | 98.86    | 98.34             | 98.50                      | 22 / 32             |
| 2    | 3    | 2331   | 97.73    | 98.84    | 97.90             | 98.69                      | 31 / 41             |
| 2    | 4    | 2313   | 96.47    | 98.49    | 95.29             | 97.75                      | 7 / 17              |
| 2    | 5    | 2344   | 97.80    | 98.68    | 97.97             | 98.68                      | 18 / 28             |
| 3    | 1    | 2312   | 98.04    | 98.92    | 98.13             | 98.43                      | 11 / 21             |
| 3    | 2    | 2283   | 96.94    | 98.69    | 98.25             | 98.36                      | 13 / 23             |
| 3    | 3    | 2331   | 98.28    | 99.06    | 98.63             | 99.39                      | 24 / 34             |
| 3    | 4    | 2313   | 96.39    | 98.57    | 95.10             | 98.31                      | 4 / 14              |
| 3    | 5    | 2344   | 97.18    | 98.72    | 96.44             | 98.60                      | 8 / 18              |

Fold labels 1–5 correspond to stored indices 0–4. The main manuscript reports pooled out-of-fold metrics, not an unweighted average of these fold metrics. The best epoch was selected exclusively by validation macro-F1.

The compact CNN and ResNet-18 use the same matched-control assignments and experimental annotations. Training, model selection and scoring follow the settings in Sect. S5.

Table S7b ResNet-18 control: pooled summary (%)

| Metric                     | Mean  | Seed SD | 95% interval   |
|----------------------------|-------|---------|----------------|
| Accuracy                   | 98.78 | 0.01    | 98.36 to 99.14 |
| Macro-F1                   | 97.30 | 0.13    | 96.31 to 98.08 |
| Balanced accuracy          | 97.26 | 0.18    | 96.11 to 98.25 |
| Inscription-macro accuracy | 98.37 | 0.33    | 97.46 to 98.98 |

Seed SD is expressed in percentage points. Intervals use 1,000 accepted whole-inscription bootstrap draws, with fits and partitions fixed.

## S5 Implementation and reproducibility

### Image preparation and temporary visual grouping

Crops were converted to grayscale and bounded by pixels with intensity below 200. The cropped intensities were inverted, scaled isotropically by  $80/\max(\text{height}, \text{width})$ , resized by bilinear interpolation and centred on a black  $92 \times 92$  canvas. Width and height were rounded with a minimum of one pixel. Source crops and final image hashes identify the inputs and outputs of these transformations.

Temporary grouping used flattened normalized pixels and a HOG-like descriptor. The latter used central-difference gradients, nine signed orientation bins over  $2\pi$ , and non-overlapping  $8 \times 8$  cells. The remaining boundary pixels outside the  $11 \times 11$  complete cell grid were omitted from that descriptor. Cell histograms were concatenated and L2-normalized. Concatenated pixel and gradient features were standardized, reduced by principal component analysis (PCA) to 160 whitened components, and L2-normalized. K-means used 30 groups, random state 42, 30 initializations, at most 600 iterations and tolerance  $10^{-4}$ . These operations organized the annotation workspace. The recognition HOG descriptor is separately specified below and has no PCA or learned scaling.

### Construction of the stored source partitions

The original source-group assignment was retained in the reported experiments and rechecked after the annotation review. Candidate outer partitions were generated by StratifiedGroupKFold with five folds, shuffling and seeds 42–57. Candidate scoring combined class-fraction, glyph-count and group-count deviations, with weights 1, 0.25 and 0.10. Candidates failing class coverage were rejected. The selected generator seed was 46. Within each outer training pool, 256 GroupShuffleSplit candidates targeted a validation fraction of 0.15 of groups, using seed 1042 plus the zero-based fold index. Validation selection required all classes in both partitions and balanced class fractions and total size. The released files are authoritative for exact membership.

The matched control was reconstructed using the frozen experimental labels. Class-wise test quotas were matched to the source-disjoint folds, followed by matched validation quotas. Its random seeds were 2042 for outer assignment and 3042 plus fold index for validation. Grouping by exact raster and horizontal-reflection hashes was retained. Training and test share 166, 163, 166, 167 and 167 source identifiers in control folds 1–5, respectively.

### Training environment and stochastic controls

The main training platform was equipped with an NVIDIA GeForce RTX 3090 (24 GB VRAM). Neural runs were executed under Python 3.10, PyTorch 2.6.0+cu124, torchvision 0.21.0+cu124, CUDA 12.4 and cuDNN 9.1.0. Numerical and image libraries were NumPy 2.2.6, SciPy 1.15.3, scikit-learn 1.6.1, scikit-image 0.25.2, Pillow 12.3.0, joblib 1.5.3 and threadpoolctl 3.6.0. Loader workers were zero; CPU threads were two. Python, NumPy and PyTorch random generators were seeded for each run. Deterministic algorithms were requested, cuDNN benchmarking and TF32 were disabled, and CUDA mixed precision was used.

## S5 Implementation and reproducibility (continued)

**Table S8** Fixed recognition settings

| Component             | Setting                                                                                                                                                              |
|-----------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| HOG input             | 92 × 92 image, symmetrically zero-padded to 96 × 96                                                                                                                  |
| HOG descriptor        | 9 orientations; 8 × 8 pixels/cell; 2 × 2 cells/block; L2-Hys; 4,356 features; no square-root transform or fitted feature scaler                                      |
| Linear SVM            | $C \in \{0.1, 1, 10\}$ ; balanced class weights; dual=False; max_iter=10,000; tolerance $10^{-4}$ ; seed 42                                                          |
| Compact CNN           | 32/64/128 channels; 3 × 3 convolutions; batch normalization and ReLU; max-pooling after first two blocks; 4 × 4 adaptive pooling; dense 128; dropout 0.3; 22 outputs |
| CNN normalization     | (grayscale / 255 − 0.5) / 0.5                                                                                                                                        |
| ResNet-18             | ImageNet-1K V1 initialization; original stem; native 92 × 92 input; grayscale repeated over three channels; 22-output classifier; all layers trained                 |
| ResNet normalization  | Channel means [0.485, 0.456, 0.406]; standard deviations [0.229, 0.224, 0.225]                                                                                       |
| Trainable parameters  | Compact CNN: 358,006; ResNet-18: 11,187,798                                                                                                                          |
| Optimization          | AdamW; unweighted cross-entropy; batch 64; weight decay $10^{-4}$ ; learning rate CNN $10^{-3}$ , ResNet-18 $10^{-4}$ ; no scheduler; gradient norm clip 5           |
| Stopping / selection  | Maximum 60 epochs; patience 10; minimum 10 epochs; best validation macro-F1; earlier checkpoint retained on ties                                                     |
| Training augmentation | Horizontal flip probability 0.5; rotation $\pm 5^\circ$ ; translation $\pm 2\%$ ; scale 0.95–1.05; bilinear resampling; black fill                                   |
| Neural seeds          | 1, 2, 3 for each model/protocol                                                                                                                                      |
| Test evaluation       | One pass using the selected checkpoint; no refit on training plus validation; no test-time augmentation                                                              |
| Uncertainty           | Whole-inscription percentile bootstrap: 1,000 accepted draws for scores, 5,000 for paired comparisons; true-class coverage retained                                  |

The ImageNet checkpoint was resnet18-f37072fd.pth (SHA-256: f37072fd47e89c5e827621c5baffa7500819f7896bbacec160b1a16c560e07ec). Adaptive pooling was implemented using the equivalent explicit-bin means to support deterministic execution. Forward/backward equivalence was tested.

### Verification and figure provenance

The analysis checked 150,579 prediction rows across 65 scientific fits, their test membership, recorded classes, source identifiers, score argmax, confusion matrices, per-class and pooled metrics, and validation-based selection. The manuscript uses these saved predictions without additional inference. Figure 2 contains the curator’s 72 selected glyphs; all displayed tiles were verified against the source pixels after integer nearest-neighbour enlargement. Figure 5 reproduces six cases selected from repeated reference-label disagreements. All glyph pixels are unchanged; all six panels retain their experimental reference labels.

## S6 Context-verified annotation correction

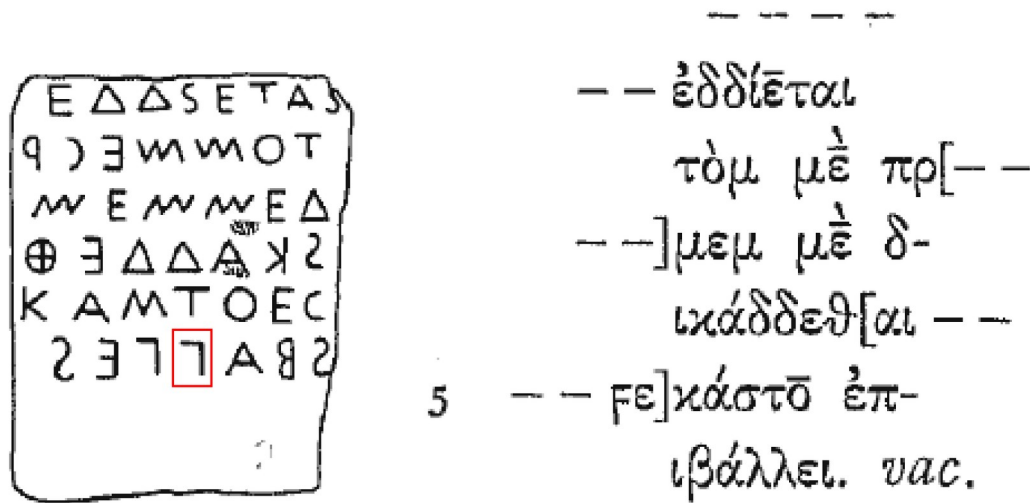
One reference-label discrepancy was resolved during the contextual examination of the figure examples. The evidence is the published transcription of IC IV.88, line 6: the glyph is the first lambda of ἐπβάλλει. It is recorded as correction E001, with its original annotation retained for reproducibility. This is a correction of a specific reading established from the inscription; no other label was changed on the basis of a classifier's output.

The source is M. Guarducci, *Inscriptiones Creticae IV* (1950), no. 88. The same transcription is accessible through PHI Greek Inscriptions, <https://epigraphy.packhum.org/text/200526>. The source drawing, crop identifiers and file hashes are recorded with E001.

|                              |                             |
|------------------------------|-----------------------------|
| Sample identifier            | ACG-V1-001676               |
| Original normalized filename | XVII-0996.png               |
| Catalogue source             | IC IV.88, line 6 (ACR-0117) |
| Original crop                | 97. IC IV.88_000949.png     |
| Experimental reference label | GAMMA                       |
| Context-verified label       | LAMDA                       |
| ResNet-18 prediction         | LAMDA in seeds 1, 2 and 3   |

Applying E001 to the catalogue changes gamma from 141 glyphs in 62 sources to 140 glyphs in 61 sources, and lambda from 305 glyphs in 99 sources to 306 glyphs in 99 sources. Within the benchmark, gamma changes from 140 to 139 glyphs and lambda from 305 to 306. The corpus and benchmark totals, source identities and image pixels are unchanged. Tables 2 and S1 retain the experimental-snapshot counts; the correction record supplies the revised reading.

The first lambda of the word is also identifiable by its neighbour: the second lambda, ACG-V1-006014, was already recorded as LAMDA. The label correction resolves a reference disagreement for ACG-V1-001676 and leaves all twelve neural predictions for that glyph unchanged. HOG-SVM predicted PI.



**Fig. S1** IC IV.88: source drawing and printed transcription, with the first lambda of ἐπβάλλει in line 6 outlined. The original strokes are unchanged. Source: Guarducci, *Inscriptiones Creticae IV*, no. 88; retained source reproduction.

### S6.1 Fixed-prediction sensitivity analysis

We re-scored the saved out-of-fold predictions after substituting LAMDA for the reference label of ACG-V1-001676 in memory. Every other label and prediction remained unchanged. Models were neither re-fitted nor used for additional inference, and validation checkpoints were not reselected. This calculation measures the direct effect of one corrected reference on scoring; it does not estimate the effect of training on revised labels.

Table S9 Scores before and after the single reference-label correction (%)

| Model / protocol    | Original macro-F1 | Re-scored macro-F1 | Original accuracy | Re-scored accuracy | $\Delta$ macro-F1 (points) |
|---------------------|-------------------|--------------------|-------------------|--------------------|----------------------------|
| HOG-SVM / source    | 82.151            | 82.158             | 91.444            | 91.444             | +0.007                     |
| CNN / source        | 92.956            | 92.979             | 96.811            | 96.820             | +0.023                     |
| ResNet-18 / source  | 95.529            | 95.551             | 98.296            | 98.305             | +0.022                     |
| CNN / control       | 95.049            | 95.072             | 97.830            | 97.839             | +0.023                     |
| ResNet-18 / control | 97.302            | 97.324             | 98.783            | 98.791             | +0.023                     |

Source denotes inscription-disjoint evaluation. Neural values are means of three seed-specific pooled out-of-fold scores. Additional decimal places show the small effect of the correction; full precision is retained in the accompanying CSV and JSON files.

For ResNet-18 under source-disjoint evaluation, the fixed-prediction correction reduces the error count from 592 to 589 prediction events across the three seeds. The number of distinct glyphs with at least one disagreement decreases from 334 to 333, and the number with disagreements under all three seeds from 93 to 92. Gamma-to-lambda disagreements decrease from 32 events on 13 distinct glyphs to 29 events on 12 distinct glyphs. These counts are separate from the original-label counts reported in Sect. 6.3.

The published bootstrap intervals refer to the original experimental annotations and have not been transferred to the re-scored values. Exact per-class matching of the primary and control partitions likewise describes the experimental snapshot. Reassigning this glyph changes the relevant gamma/lambda class quota by one in affected partitions. The fixed-prediction analysis preserves those partitions and should not be interpreted as a new matched-control training experiment.

The correction record, evidence, re-scored values and the standalone script `label_sensitivity.py` accompany this resource. The script reads the released predictions and writes a separate report, leaving the inputs unchanged. Its SHA-256 is `ad250071ee0b6a22528f3c503c471198383eb5bebc3bf9fdb4fcb3aa1317d5b`.