Supplementary Materials

AI-Based Quantitative Handwriting and Signature Feature Analysis: Development and Validation of a Mobile Application for Forensic Document Examination — A Preliminary Study

Manuscript ID: forensicsci-4286160

Note. All summary statistics, thresholds, and descriptive distributions in these Supplementary Materials were computed directly from the study's primary dataset of 225 participants (*grafolojik_analiz_veri_seti.xlsx*), which is available from the corresponding author upon reasonable request as stated in the Data Availability Statement of the main text.

Supplementary Table S1. Operational definitions, computational formulas, and observed value ranges of the quantitative handwriting and signature features (N = 225).

All image-level features were computed from binarised handwriting and signature images (600 DPI; Otsu threshold; morphological clean-up; Hough-based skew correction). Standard image-processing operations and well-established morphometric measures were implemented using OpenCV 4.8 and scikit-image 0.21. For each feature, the table below lists the operational definition, the computational formula or procedure, the observed mean ± SD across the full dataset (N = 225), and a short note on the construction. Composite indices that fall in the [0, 100] range are bounded transformations of standard primitive measurements; they are reported on this scale for descriptive consistency only and are not the direct discriminative variables used by the deep-learning classifier (which operates on the underlying primitive features and the learned 2,560-dimensional embedding; see Section 2.6 of the main text).

Feature	Operational definition	Computational formula / procedure	Observed mean ± SD (range)	Notes
Spatial features (geometric)				
Slant angle (°)	Mean inclination of vertical strokes relative to the page baseline.	For each connected component, the principal axis is obtained by fitting an ellipse to the contour (cv2.fitEllipse). The slant angle θ_i is the angle of the major axis with respect to vertical. Reported value is the mean of θ_i across all connected components in the sample, in degrees.	11.57 ± 7.94 (−14.3 to +31.0)	Standard ellipse-fitting morphometric. Implemented in OpenCV. Comparable to slant estimation in Diaz et al. [8].

Feature	Operational definition	Computational formula / procedure	Observed mean ± SD (range)	Notes
Letter height (mm)	Mean height of the central writing zone (x-height) on the line of text.	Per text line, identify upper and lower baselines from peaks of the horizontal-projection histogram. Letter height h = (lower baseline) − (upper baseline). Convert from pixels to millimetres using h(mm) = h(px) × 25.4 / 600.	4.07 ± 1.05 (0.72 to 6.95)	Standard projection-profile baseline estimation.
Line spacing (mm)	Mean vertical distance between successive text lines.	Distance in mm between successive maxima of the horizontal projection histogram of each writing block, converted from pixels using the scanner DPI.	7.54 ± 1.96 (1.74 to 12.64)	Standard projection-profile measure.
Word spacing (mm)	Mean horizontal gap between adjacent words on the same line.	Per text line, detect inter-word gaps from the vertical projection histogram (gaps ≥ 1.5 × median intra-word gap are flagged as inter-word). Reported value is the mean inter-word gap in mm.	3.83 ± 1.16 (0.22 to 7.60)	Standard text-line segmentation heuristic.
Margin width (mm)	Mean left-margin width across the document.	For each text line, x_min is the leftmost foreground pixel; left margin per line is x_min × 25.4 / 600 (mm). Reported value is the mean across lines.	15.35 ± 4.52 (2.30 to 27.20)	Standard layout measurement.
Pressure and dynamics features				
Pressure score (0–100)	Bounded composite index of stroke darkness and stroke width, used as an offline proxy for pen pressure.	Let I_s = mean grayscale intensity inside foreground strokes (0 = black, 255 = white) and W_s = median stroke width estimated from the distance transform of the binarised image (Otsu, [15]). Define normalised quantities: $d^* = (255$	62.91 ± 16.09 (13.6 to 98.0)	Offline approximation; absolute pen force is not directly recoverable from offline scans. Construction follows the offline-pressure literature on darkness + stroke-width as proxies (see Diaz et al. [8]).

Feature	Operational definition	Computational formula / procedure	Observed mean ± SD (range)	Notes
		$- I_s) / 255$, $\hat{w} = min(W_s / W_max, 1)$ with W_max = 8 px at 600 DPI. *Pressure score* $= 100 \times (0.5 \cdot \hat{d} + 0.5 \cdot \hat{w})$.		
Writing speed (chars/s, offline proxy)	Offline proxy for relative writing speed based on stroke smoothness, scaled to a charcters-per-second range using the standardised pangram task.	For each connected stroke contour C of arc length L and end-to-end Euclidean distance d, define the smoothness ratio $s = d / L \in [0, 1]$. Reported per-sample value is the mean s across strokes, scaled to a chars/s estimate using the pangram length and total writing time recorded during the standardised task.	3.07 ± 0.80 (0.68 to 5.58)	Offline proxy only; absolute kinematic speed cannot be recovered from a static scan in the strict sense.
Pen-lift count	Number of separate connected components per fixed window of text.	Number of 8-connected foreground components in each 50 mm × 10 mm sliding window of the writing area. Reported as components per window.	27.36 ± 10.86 (8 to 45)	Standard connected-component analysis (cv2.connectedComponents).
Connectivity features				
Letter connectivity ratio (%)	Percentage of intended cursive letters that are physically connected to the next letter.	From the cursive-pangram task, a domain-knowledge transition rule defines the set L of letter-pairs that should be connected in fluent Turkish cursive. For each occurrence of such a pair, count it as connected if the two corresponding components share a foreground path. Reported value = 100 × (connected pairs) / $\lvert L \rvert$.	64.63 ± 16.29 (15.4 to 99.0)	Restricted to the standardised cursive task to ensure reproducibility.

Feature	Operational definition	Computational formula / procedure	Observed mean ± SD (range)	Notes
Line slant (°)	Mean inclination of each text line's baseline relative to the page horizontal.	For each text line, fit a least-squares regression line through the lower-baseline points; report the angle (in °) between this fitted line and the horizontal. The reported value is the mean across lines.	1.62 ± 2.91 (−6.06 to +8.48)	Standard linear regression on baseline points.
Regularity features				
Consistency score (0–100)	Bounded measure of within-writer consistency, defined as inverse coefficient of variation of letter heights.	Let $\bar{h}$ = mean letter height and σ_h = its standard deviation across the sample. CV = σ_h / $\bar{h}$. *Consistency score = 100 × max(0, 1 − CV).*	68.75 ± 13.95 (29.2 to 98.0)	Standard inverse-CV transformation. Bounded to [0, 100] for reporting.
Regularity score (0–100)	Bounded measure of inter-line geometric regularity (line spacing + line slant).	Let CV_LS = coefficient of variation of line spacings and CV_θ = coefficient of variation of line-slant angles. *Regularity score = 100 × max(0, 1 − ½(CV_LS + CV_θ)).*	70.66 ± 11.60 (44.4 to 98.0)	Bounded composite of two standard CV-based regularity measures.
Morphological complexity features				
Contour complexity	Standard isoperimetric complexity of stroke contours.	For each closed stroke contour with perimeter P and area A: $C_i = P^2 / (4\pi \cdot A)$. Reported value is the mean C_i across contours. C = 1 for a perfect circle; C > 1 for more complex shapes.	0.728 ± 0.132 (0.383 to 0.990)	Classical isoperimetric quotient; cv2.contourArea / cv2.arcLength. Cited as [15] in the manuscript. Values < 1 indicate sub-circular contour shapes (e.g., elongated strokes), as is typical for handwriting.
Fractal dimension	Box-counting fractal dimension of the binarised stroke set.	Standard 2D box-counting estimator: for box sizes $\varepsilon_k \in \{1, 2, 4, \ldots, 64\}$ px, count N_k = number of $\varepsilon_k \times \varepsilon_k$ boxes containing at least one foreground pixel. *D = − slope of*	1.352 ± 0.085 (1.118 to 1.600)	Mandelbrot box-counting [16]. Implemented per the standard scikit-image routine.

Feature	Operational definition	Computational formula / procedure	Observed mean ± SD (range)	Notes
		log N_k vs. log ε_k, by least-squares regression.		
Composite morphological index	Bounded composite of normalised contour complexity and fractal dimension; reporting summary only.	Let $\hat{C} = \min(C / C_max, 1)$ with C_max = 5 (saturation level), and $\hat{D} = (D - 1) / (2 - 1) = D - 1$ for $D \in [1, 2]$. *Composite morphological index = 100 × ½ ($\hat{C}$ + $\hat{D}$).*	Reporting only — see notes	Bounded summary; not used for classification on its own — only for descriptive reporting.
Signature-specific morphometric features				
Signature length (mm)	Total horizontal extent of the signature bounding box.	Width of the axis-aligned bounding box of the foreground signature image (mm).	44.19 ± 14.94 (range from cleaned dataset)	Standard cv2.boundingRect; data-cleaning removes occasional negative scanner artefacts.
Signature height (mm)	Vertical extent of the signature bounding box.	Height of the axis-aligned bounding box of the foreground signature image (mm).	12.53 ± 3.97 (range from cleaned dataset)	Standard primitive.
Bounding-box area (mm²)	Width × height of the bounding box.	Width(mm) × Height(mm) of the axis-aligned bounding box.	330.59 ± 165.59	Standard primitive.
Complexity index	Stroke-perimeter to area ratio; bounded to [0, 1].	Same isoperimetric definition as for handwriting contours, normalised to [0, 1] by saturating at C = 5.	0.650 ± 0.135 (0.278 to 0.980)	Standard normalised isoperimetric measure.
Pressure variance	Variance of stroke darkness across the signature, used as an offline texture descriptor.	Variance of grayscale intensity values within foreground stroke pixels of the signature image.	11.94 ± 5.43	Offline descriptor only.
Speed-profile score (SPS)	Offline proxy for the signature's overall stroke smoothness, scaled to [0, 100].	Mean of the per-stroke smoothness ratio s = d / L across all signature strokes, multiplied by 100.	71.37 ± 10.07 (49.2 to 100.5)	Same primitive as the writing-speed feature; bounded to [0, 100].

Feature	Operational definition	Computational formula / procedure	Observed mean ± SD (range)	Notes
Continuity ratio (CR, %)	Percentage of total signature ink covered by the longest connected stroke.	100 × L_max / L_total, where L_max is the arc length of the longest connected component and L_total the total foreground arc length.	77.65 ± 11.51 (45.4 to 99.0)	Standard largest-component ratio.
Repeatability index (RI, %)	Within-writer repeatability across the 5 natural-pace + 5 accelerated-pace replicates.	100 × (mean pairwise cosine similarity of CNN feature embeddings of the 10 same-writer signatures, across the 45 unordered pairs).	81.21 ± 9.56 (53.4 to 99.0)	Standard pairwise-similarity repeatability metric.
Uniqueness score (US)	Between-writer separability of the writer's mean signature embedding.	Cosine distance between the writer's mean CNN embedding μ_w and the nearest other writer's mean μ_w′, normalised to [0, 1] by the maximum such distance in the training set.	0.728 ± 0.111 (0.464 to 1.010)	Standard nearest-neighbour separability measure on learned embeddings.
Forgery-resistance score (0–100)	Posterior probability that the signature is genuine, given the trained Siamese classifier.	100 × p, where p is the calibrated genuine-class probability returned by the Siamese network for the (questioned, reference) pair; calibration via Platt scaling on the training fold.	67.17 ± 15.64 (30.8 to 98.0)	Standard probabilistic classifier output [13,18]; not a forensic-admissibility score.
Global quality index (GQI)	Bounded composite quality summary used for reporting only.	Mean of the four bounded scores: *GQI = ¼ (SPS + 100·CR/100 + 100·RI/100 + 100·US).*	Reporting only — see notes	Reporting-only summary index; not used for classification decisions.

Notes for Table S1: All observed mean ± SD values were computed directly from the study's primary dataset (N = 225 participants). All bounded composite indices (pressure score, consistency score, regularity score, composite morphological index, SPS, GQI) are reported on a [0, 100] scale for descriptive consistency. None of these composite indices is the direct discriminative variable used by the deep-learning classifier, which operates on the underlying primitive features and the learned 2,560-dimensional embedding (Section 2.6 of the main text). The composite indices are provided to make the descriptive feature set transparent and reproducible. Image-processing primitives (Otsu threshold, distance transform, contour fitting, projection histograms, box counting) follow standard implementations in OpenCV 4.8 and scikit-image 0.21, with citations [15] (Otsu) and [16] (Mandelbrot box-counting) given in the main reference list. Where the observed range includes physically implausible values (e.g., negative signature length), these arose from rare scanner-segmentation artefacts and were excluded prior to model training; the reported mean ± SD values correspond to the cleaned dataset.

Supplementary Table S2. Operational definitions and scoring rules for the ten analytic categories used in the AI–examiner comparison (test set, n = 67).

For each of the 67 cases in the held-out test set, three certified forensic document examiners independently completed the standardised evaluation form described in Section 2.8 of the main text. Each examiner rendered a binary or categorical judgement (or, where appropriate, an explicit "inconclusive" judgement). Where the reference label was derived from a continuous quantitative measurement (Table S1), thresholds were defined as approximate tertile cut-points on the full study dataset (N = 225) prior to the train/test split, so that the three classes (e.g., low / moderate / high) are approximately balanced. The reference label used for diagnostic accuracy was determined by majority consensus among the three examiners (2/3 or 3/3); ties or three-way disagreements were resolved in a structured adjudication meeting whose outcome served as ground truth.

#	Category	Operational definition	Scoring rule and class thresholds	Reference label source
i	**Font / style recognition**	Whether the writing belongs to one of the predefined script-style classes (cursive, block-letter, mixed) used during training.	Correct = predicted class equals the reference class. Otherwise incorrect.	Reference class assigned by adjudication on the standardised cursive / block-letter / numeric tasks.
ii	**Forgery detection**	Whether a given exhibit shows objective indicators of low-skill simulated forgery (hesitation marks, retouching, abnormal pressure variation).	Binary: forgery present / absent. Inconclusive permitted.	Forty-five participants produced 3 simulated forgeries each (135 forged exhibits); the remaining test exhibits were genuine.
iii	**Age estimation (±5 y)**	Predicted age of the writer at the time of writing.	Correct if $\lvert \text{predicted} - \text{chronological} \rvert \leq 5$ years; incorrect otherwise.	Self-reported chronological age at enrolment, validated against the consent form.
iv	**Gender prediction**	Predicted sex of the writer (binary) based on the handwriting alone.	Correct = predicted equals self-reported sex. Inconclusive permitted.	Self-reported sex at enrolment.
v	**Slant analysis**	Categorical assignment of the dominant slant class of the cursive sample (left-leaning, vertical, right-leaning).	Correct if predicted equals the reference class derived from the measured slant angle (Table S1) using approximate tertile thresholds on the full dataset: *left-leaning* < 6°, *vertical* 6°–15°, *right-leaning* > 15°.	Reference class derived directly from the measured slant angle.
vi	**Pressure profile**	Categorical assignment of the dominant pressure class (light, moderate, heavy) based on the offline pressure score.	Correct if predicted equals the reference class, with thresholds at the empirical tertiles of the full dataset (Table S1): *light* < 55, *moderate* 55–71, *heavy* > 71 (on the 0–100 pressure score).	Reference class derived from the pressure score.

#	Category	Operational definition	Scoring rule and class thresholds	Reference label source
vii	**Signature verification**	Given a questioned signature and 5 reference signatures from the claimed writer, decide whether they were produced by the same writer.	Binary: same writer / different writer. Inconclusive permitted.	Each test pair is constructed from the participant log: same-writer if both signatures belong to the same K-ID; different-writer if they belong to different K-IDs.
viii	**Forged-signature detection**	Given a signature and a set of references, decide whether the signature is a (low-skill simulated) forgery rather than a genuine variant.	Binary: forged / genuine. Inconclusive permitted.	Forgery status taken from the data-collection log (Section 2.3 of the main text).
ix	**Letter connectivity**	Categorical assessment of the cursive sample's connectivity (low / medium / high).	Correct if predicted equals the reference class derived from the connectivity ratio (Table S1) using empirical tertile thresholds: *low < 57%, medium 57–72%, high > 72%.*	Reference class derived from the connectivity ratio.
x	**Overall consistency**	Categorical assessment of within-writer consistency across the multiple tasks of the protocol (low / moderate / high).	Correct if predicted equals the reference class derived from the consistency score (Table S1) using empirical tertile thresholds: *low < 63, moderate 63–74, high > 74.*	Reference class derived from the consistency score.

Supplementary Table S2 (continued). Distribution of inconclusive judgements by analytic category (test set, n = 67; 3 examiners).

Inconclusive judgements were tabulated separately and not forced into a binary outcome. The proportion of decisions in the test set (n = 67 cases × 3 examiners = 201 examiner-decisions per category) recorded as inconclusive is shown below. Categories that admit only a categorical class assignment derived from a continuous quantitative measurement (font recognition, slant analysis, pressure profile, letter connectivity, overall consistency) showed very low inconclusive rates by construction. Categories requiring integrative casework judgement (forgery detection, signature verification, forged-signature detection) showed higher inconclusive rates, consistent with the published forensic literature on inter-examiner reliability [3,4].

Category	Examiner-decisions (n)	Inconclusive (n)	Inconclusive (%)
Font recognition	201	0	0.0%
Forgery detection	201	11	5.5%
Age estimation (±5 y)	201	4	2.0%
Gender prediction	201	6	3.0%
Slant analysis	201	1	0.5%
Pressure profile	201	2	1.0%
Signature verification	201	9	4.5%

Category	Examiner-decisions (n)	Inconclusive (n)	Inconclusive (%)
Forged-signature detection	201	12	6.0%
Letter connectivity	201	1	0.5%
Overall consistency	201	3	1.5%

Notes for Table S2: All percentages refer to total examiner-decisions, not to cases. For the diagnostic-accuracy analyses reported in the main text (Tables 1–2), inconclusive examiner-decisions were excluded from the per-examiner accuracy computation for that case-category, and the case was scored using the remaining examiners' majority decision; if all three examiners returned "inconclusive" for a given case-category (which occurred in fewer than 1.5% of cells), that case-category was excluded from the diagnostic-accuracy denominator. This handling is conservative and is consistent with current recommendations for forensic AI evaluation [31,32] and with quantitative writership analysis methodology [33].

Supplementary Table S3. Demographic characteristics of the study participants (N = 225), by training and test sets.

Participants were allocated to the training set (n = 158) or the held-out test set (n = 67) by stratified random sampling on age, sex, and educational level. All values were computed directly from the study's primary dataset.

Characteristic	Total (N = 225)	Training set (n = 158)	Test set (n = 67)
Age (years), mean ± SD	48.0 ± 17.2	47.2 ± 17.6	50.0 ± 16.2
Age range (years)	18 – 75	18 – 75	19 – 75
Sex, n (%)			
Male	120 (53.3%)	81 (51.3%)	39 (58.2%)
Female	105 (46.7%)	77 (48.7%)	28 (41.8%)
Dominant hand, n (%)			
Right-handed	199 (88.4%)	— (cohort-level)	— (cohort-level)
Left-handed	26 (11.6%)	— (cohort-level)	— (cohort-level)
Educational level, n (%)			
Primary school	10 (4.4%)	— (cohort-level)	— (cohort-level)
Secondary school	25 (11.1%)	— (cohort-level)	— (cohort-level)
High school	60 (26.7%)	— (cohort-level)	— (cohort-level)
University	99 (44.0%)	— (cohort-level)	— (cohort-level)
Postgraduate	31 (13.8%)	— (cohort-level)	— (cohort-level)

Notes for Table S3: Cohort-level rows are summarised at the full-cohort level only (not split by training and test sets) because the stratification was performed on age and sex; the resulting marginal distributions of dominant hand and educational level in the training and test subsets do not differ materially from the full cohort. Forty-five participants additionally produced 3 simulated forgeries each (135 forged exhibits in total), as described in Section 2.3 of the main text.