A COMPUTATIONAL FRAMEWORK FOR EXPLAINABLE ASSISTED SIGNATURE SCREENING IN FORENSIC DOCUMENT ANALYSIS

Lissandra Kruse Fuganti

Graduate Program in Applied Computing, Universidade Estadual de Ponta Grossa (UEPG), Ponta Grossa, Paraná, Brazil - E-mail: t7426541@gmail.com

Abstract: This paper presents an explainable computational framework for assisted screening of handwritten signatures in forensic document analysis. The proposed system combines preprocessing, graphotechnical descriptor extraction, statistical normalization, compatibility scoring, structural comparison, and heuristic anti-remake risk assessment. Unlike end-to-end deep learning classifiers, the framework exposes intermediate descriptors and scores so that forensic examiners can inspect why a questioned signature is considered compatible, inconclusive, or divergent in relation to a reference set. Empirical validation was conducted on a subset of the BHSig260-Bengali offline signature dataset, comprising 20 writers, 100 reference signatures, and 400 comparison trials. The continuous compatibility score achieved an AUC of 0.9017. Using a calibrated post-hoc threshold of 85, the framework obtained 83.0% accuracy, 78.5% sensitivity for genuine signatures, and 87.5% specificity for skilled forgeries. The experiment also revealed that a fixed binary shape overlap threshold does not transfer directly across datasets and requires calibration. These results support the framework as an auditable decision-support mechanism for preliminary forensic screening, while reinforcing that its outputs are indicative and subordinate to expert human judgment.

Keywords. offline signature verification; forensic document analysis; explainable AI; decision support; graphotechnics; questioned signatures; ROC analysis; BHSig260

1. INTRODUCTION

Signature examination is one of the most frequent tasks in questioned document analysis. In both civil and criminal proceedings, questioned documents may require expert assessment of whether a handwritten signature was produced by the claimed author or constitutes a simulation. Unlike biometric authentication systems designed for operational accept-or-reject decisions, forensic examinations impose an additional requirement: the reasoning process must be transparent, traceable, and communicable to courts and opposing experts.

Computational tools for signature verification have been explored for several decades. Research has produced methods based on handcrafted descriptors, statistical classifiers, and, more recently, deep learning architectures such as convolutional neural networks and Siamese networks. While these approaches can achieve high predictive performance on benchmark datasets, their internal reasoning may be difficult to explain in forensic settings. A score produced by a neural network does not, by itself, identify which visual features of a questioned signature contributed to a conclusion of compatibility or divergence.

Explainable AI (XAI) addresses this interpretability gap through methods such as saliency maps, attention mechanisms, rule extraction, and feature importance scores. However, many explanation techniques remain external to the forensic reasoning process and may produce explanations that are post-hoc approximations rather than genuine representations of the model's decision logic. For forensic use, a stronger design choice is to build the framework around interpretable descriptors

from the outset, so that each numeric output corresponds to a visible and documentable property of the signature trace.

This paper proposes an explainable framework for assisted screening of handwritten signatures from static images. The system integrates graphotechnical descriptor extraction, statistical baseline construction, compatibility scoring, structural shape comparison, and heuristic anti-remake risk assessment. Its outputs are indicative screening results—not authorship declarations—and are intended to support expert review, organize evidence, and reduce the time required for preliminary case prioritization.

The framework was evaluated on a subset of the BHSig260-Bengali dataset. Quantitative metrics include AUC, accuracy, sensitivity, specificity, and a confusion matrix. An important calibration finding is also reported: a rigid binary shape overlap threshold did not transfer to the validation subset without dataset-specific adjustment, a result with practical implications for any forensic deployment of threshold-based classification rules.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed framework. Section 4 presents the mathematical formulation of the scoring method. Section 5 describes the experimental protocol. Section 6 reports results. Sections 7 through 9 discuss findings, present limitations, and outline future directions. Section 10 concludes the paper.

2. RELATED WORK AND POSITIONING

The literature on offline signature verification is broad and historically rich. Plamondon and Lorette (1989) established the foundational survey of automatic signature verification, describing the problem as a pattern-recognition task with global and local feature components. Impedovo and Pirlo (2008) consolidated decades of subsequent work and highlighted the importance of preprocessing quality, feature selection, intra-writer variability, and robust similarity measures. Their review remains particularly relevant because it demonstrates that verification quality depends not only on feature extraction but also on the representativeness and size of the available reference set.

Offline signature verification is inherently more challenging than online verification because temporal information—pen velocity, pressure, acceleration, and pen-up/pen-down sequences—is unavailable. The system must infer similarity from static shape, contour behavior, local texture, occupancy patterns, and spatial organization. This limitation explains why handcrafted descriptors have remained important in offline approaches even as deep learning has transformed performance ceilings in other pattern recognition domains.

Learning-based approaches have substantially advanced the field. Hafemann, Sabourin, and Oliveira (2017) demonstrated that deep convolutional representations can improve writer-independent offline verification, reducing reliance on handcrafted features. Siamese and metric-learning architectures have further improved pairwise similarity estimation. At the same time, their success deepens an old forensic concern: high-performance models may still be difficult to explain in evidential contexts.

This explainability gap has motivated work targeted specifically at forensic use. Mazzolini et al. (2021) proposed an interpretable decision-support system for dynamic signature forensics emphasizing semiautomatic support that remains auditable by human examiners. Diaz, Ferrer, and Vessio (2024) extended this direction by proposing an explainable offline verifier built around a universal background model with transparent distance assignments. Both works are directly

relevant to the present paper because they demonstrate that computational assistance in forensic signature analysis can be rigorous without becoming opaque.

Empirical studies of forensic handwriting examination are equally important for system design. Kam et al. (2001) and Sita, Found, and Rogers (2002) showed that trained forensic document examiners outperform laypersons in signature comparison tasks, supporting the view that feature comparison and interpretive judgment constitute genuine expertise domains. Hicklin et al. (2022) reported large-scale evidence on the accuracy and reliability of forensic handwriting comparisons, reinforcing that defensible conclusions require disciplined management of uncertainty and calibrated expression of confidence.

The proposed framework follows a hybrid philosophy. It borrows from biometric verification the use of explicit descriptors, score aggregation, and quantitative evaluation. At the same time, it borrows from forensic practice the requirement for traceable reasoning, feature-level reporting, and non-binary conclusions when evidence does not support stronger claims. This positions the framework as complementary to deep learning: it may serve as an interpretable screening layer, a feature-level report generator, or a calibration baseline for future hybrid systems.

Approach	Typical Strength	Main Limitation in Forensic Use	Relation to Framework
Handcrafted descriptors	Transparent and inspectable	Requires careful weighting and calibration	Core explainable feature layer
Statistical classifiers	Simple and reproducible	May underfit complex writer variability	Inform scoring strategy
Deep CNN models	High benchmark performance	Limited interpretability; data-dependent	Complementary — contrasted, not replaced
Siamese / metric-learning	Strong pairwise similarity	Indirect explanations; post-hoc	Potential future hybrid extension
Proposed framework	Auditable descriptors; explicit scoring	Requires calibration; limited training data	Designed for forensic screening and review

Table 1. Positioning of the proposed framework relative to representative signature verification approaches.

3. PROPOSED COMPUTATIONAL FRAMEWORK

The proposed framework is designed for assisted screening of questioned signatures from static images. It is not presented as a closed commercial tool or an autonomous authorship-determination system. Instead, it is a modular architecture whose outputs are intended to be auditable by a human reviewer. The central principle is that every numerical value remains connected to an interpretable feature family, so that the examiner can understand whether similarity is driven by global geometry, spatial distribution, contour behavior, apparent trace quality, or structural overlap.

The workflow is divided into seven stages, each designed to preserve interpretability while producing actionable evidence for the next stage.

3.1 IMAGE QUALITY ASSESSMENT

Input images are checked for resolution sufficiency, background interference, crop adequacy, and contour visibility. This step prevents downstream metrics from being interpreted as authorship evidence when they may reflect poor acquisition conditions. Scanned images with low DPI, heavy background noise, or clipped margins can produce misleading descriptor values even for genuine signatures.

3.2 PREPROCESSING

Preprocessing includes grayscale normalization, contrast enhancement, binarization of the ink trace, contour isolation, size normalization to a fixed canvas, and baseline alignment. The purpose is to reduce variability caused by capture scale, background differences, and document orientation. Binarization parameters are applied consistently to all reference and questioned images, which is especially important for shape-overlap measures.

3.3 DESCRIPTOR EXTRACTION

The descriptor vector combines measurements intended to reflect visible and computationally accessible properties of the signature trace. Descriptors are grouped into interpretable families:

- Geometry: aspect ratio, centroid position, component count, hole count, normalized area, contour complexity.
- Stroke distribution: fill ratio, horizontal and vertical projection profiles, baseline slope, projection entropy.
- Texture and density: mean apparent ink density, local contrast, stroke-width variation, coefficient of variation of apparent pressure.
- Stability proxies: tremor proxy, baseline RMSE, crossing density, right-angle ratio and count.

No individual descriptor is treated as definitive evidence. Each indicator provides partial evidence about form, rhythm, regularity, complexity, or trace quality. Their collective behavior relative to the reference baseline constitutes the basis for scoring.

Descriptor Group	Examples	Interpretive Role
Geometry	aspect ratio, centroid, component count, contour complexity	Broad form and spatial organization
Stroke distribution	fill ratio, projection entropy, baseline slope	Ink placement and global trace distribution
Texture and density	mean grayscale, local contrast, stroke-width CV	Apparent trace quality and uniformity
Stability proxies	tremor proxy, baseline RMSE, crossing density	Possible hesitation, irregularity, or simulation-sensitive patterns
Artificial geometry	right-angle ratio, right-angle count, fragmentation	Alerts for constructed or mechanically rigid traces

Table 2. Descriptor families and their interpretive roles in the screening workflow.

3.4 MULTISAMPLE NORMALIZATION AND BASELINE CONSTRUCTION

Because questioned signatures are compared against multiple references rather than a single template, each descriptor is contextualized against the observed reference range. For each descriptor m, the reference set provides a mean μm, standard deviation σm, minimum, and maximum. A minimum standard-deviation floor ε is applied to avoid unstable divisions when reference signatures are highly similar for a given descriptor.

3.5 DIVERGENCE-AWARE SCORING

Instead of collapsing all descriptors into a single optimistic similarity number, the framework tracks convergent and divergent descriptor families separately. A global compatibility score can be reported, but it is interpreted together with multivariate distance, structural similarity, and anti-remake risk indicators.

3.6 ANTI-REMAKE ASSESSMENT

This module applies heuristic rules designed to highlight patterns potentially consistent with tracing, slow redrawing, excessive regularity, or artificial geometry. These indicators serve as alerts for examiner attention, not as proof of forgery. The module is described in detail in Section 4.5.

3.7 STRUCTURED REPORTING

The output is a layered report containing descriptor groups, directional behavior, intermediate score summaries, structural measurements, and a final indicative label: compatible with reference variability, inconclusive, or relevant divergence. The architecture deliberately avoids the term automatic authentication. Labels are screening outputs, not expert opinions.

4. MATHEMATICAL FORMULATION

4.1 NORMALIZED DEVIATION AND PER-METRIC SCORE

For each metric m, the normalized deviation of the questioned value qm from the reference mean μm is:

$$z\_m = |q\_m - \mu\_m| / max(\sigma\_m , \varepsilon)$$

where ε is a small positive floor (ε = 0.025 in the current implementation) that prevents division by near-zero standard deviations when reference signatures are highly consistent on a given descriptor. The per-metric compatibility score is then:

$$score\_m = max(0, 100 - 22 \cdot z\_m)$$

This formulation produces a score of 100 when the questioned value exactly matches the reference mean, and decreases linearly at a rate of 22 points per standard deviation unit. A questioned value deviating by more than 4.5 standard deviations yields a score of zero. The coefficient 22 was selected so that deviations within approximately one standard deviation retain scores above 78, a value consistent with preliminary screening tolerance.

4.2 WEIGHTED GLOBAL COMPATIBILITY SCORE

The global compatibility score is computed as a weighted mean over all descriptors. Selected graphotechnical descriptors—those most directly related to form, rhythm, and spatial distribution—receive a weight wm = 1.35, while remaining descriptors receive unit weight:

$$Compatibility = \Sigma_m (score_m \cdot w_m) / \Sigma_m w_m$$

This formulation is intentionally transparent. A reviewer or examiner can inspect the contribution of each descriptor family, identify which features reduced the score, and determine whether the final value reflects broad disagreement or a few localized deviations.

4.3 MULTIVARIATE DIAGONAL DISTANCE

In addition to the weighted compatibility score, the framework computes a diagonal multivariate distance over the normalized deviations:

$$D = sqrt[(1/N) \cdot \Sigma_m z_m^2]$$

This is a conservative diagonal approximation rather than a full Mahalanobis distance, because the small reference set (five signatures) makes covariance estimation unstable. The distance score is:

$$DistanceScore = max(0, 100 - 20 \cdot D)$$

4.4 STRUCTURAL COMPARISON

Structural similarity is evaluated using binary intersection over union (IoU), comparing the binarized questioned signature mask Q with each reference mask R:

$$IoU = |Q \cap R| / |Q \cup R|$$

The framework also compares horizontal and vertical projection vectors using Pearson correlation. Projection correlation captures whether the distribution of ink along rows and columns follows a similar pattern, even when exact pixel overlap is low:

$$\rho(a, b) = cov(a, b) / (\sigma_a \cdot \sigma_b)$$

The validation results (Section 6.4) demonstrate why structural comparison must be interpreted with care: mean IoU was low for both genuine and forged samples in the BHSig260 subset, showing that shape-overlap thresholds require dataset-specific calibration.

4.5 ANTI-REMAKE RISK ASSESSMENT

The anti-remake module applies seven heuristic conditions inspired by forensic indicators of tracing, slow redrawing, and mechanical imitation:

- Low pressure variation (pressure_cv < 0.12) combined with low stroke-width variation (stroke_width_cv < 0.18) — indicator of excessive trace uniformity.
- High tremor proxy (tremor_proxy > 0.58) combined with narrow stroke width — possible hesitant construction or tracing-like behavior.
- High baseline deviation (baseline_rmse > 0.42) combined with high crossing density (crossings_density > 0.55) — structural irregularity.
- High questioned-vs-reference IoU combined with high tremor — excessive overlap with hesitation, consistent with tracing.
- Low structural IoU (< 0.22) combined with low compatibility score (< 55) — broad global divergence.
- High component count (> 0.55) and high hole count (> 0.35) — unusual fragmentation.
- High right-angle ratio (> 0.18) or right-angle count (> 0.35) — artificial geometric construction.

Each triggered condition adds to a cumulative risk score classified as low, medium, or high. This score is not proof of falsification. It is a warning layer directing examiner attention to potentially suspicious patterns, and it requires expert interpretation in context.

4.6 PSEUDOCODE

Input: questioned signature Q, reference signatures $R_1 ... R_n$

1. Preprocess Q and each R_i using the same normalization pipeline.
2. Extract normalized descriptor vectors for Q and each R_i.
3. Build reference baseline: mean, σ, min, max per descriptor.
4. For each descriptor m: compute $z_m = |q_m - \mu_m| / \max(\sigma_m, \varepsilon)$.
5. Compute per-metric score: $score_m = \max(0, 100 - 22 \cdot z_m)$.
6. Aggregate weighted global compatibility score.
7. Compute diagonal multivariate distance D and distance score.
8. Compare binary masks via IoU; compare projections via Pearson ρ.
9. Apply anti-remake heuristic rules; record triggered warnings.
10. Generate indicative label and explainable layered report.

5. MATERIALS AND METHODS

The revised methodology was designed to answer the reviewer recommendation for greater empirical validation and technical reproducibility. The present version adopts a benchmark-oriented protocol using the BHSig260-Bengali dataset and reports classification metrics derived from the continuous compatibility score.

5.1 DATASET AND SAMPLING PROTOCOL

Experiments were conducted on a subset of the BHSig260-Bengali offline signature dataset. BHSig260 contains genuine signatures and skilled forgeries organized by writer, and it is widely used in offline signature-verification research. The Bengali subset was selected because it provides a structured collection with well-defined genuine and forged samples, enabling a controlled comparison protocol.

Twenty writers were selected for validation. For each writer, five genuine signatures served as references, establishing the writer-specific baseline. The questioned set contained ten additional genuine signatures and ten skilled forgeries per writer. The resulting protocol produced 400 total comparison trials: 200 genuine questioned signatures and 200 skilled forgeries.

Parameter	Value
Dataset	BHSig260-Bengali subset
Engine version	grafotecnia-triagem-1.1.0
Number of writers	20
Reference signatures per writer	5 genuine samples
Questioned genuine per writer	10
Questioned forgeries per writer	10

Parameter	Value
Total genuine comparisons	200
Total forged comparisons	200
Total comparison trials	400
Primary discriminative variable	Compatibility score

Table 3. Validation dataset and protocol parameters.

5.2 PREPROCESSING PIPELINE

Each image was converted to grayscale and subjected to contrast normalization before binarization. The active writing region was cropped to remove excessive margins, then placed on a standardized canvas. Baseline alignment was applied to reduce gross rotation, making projection-based measures more comparable across writers. The same pipeline was applied consistently to all reference and questioned images.

The revised manuscript explicitly recognizes binarization sensitivity as a reason why fixed structural thresholds cannot be transferred across datasets. Even with consistent preprocessing, genuine signatures in the BHSig260-Bengali subset showed low mean pixel-level overlap due to natural intra-writer deformation in placement, scale, and local contour behavior.

5.3 EVALUATION METRICS

The evaluation reports accuracy, sensitivity, specificity, AUC, and a confusion matrix. Genuine signatures are treated as the positive class for sensitivity; skilled forgeries are treated as the negative class for specificity. This convention distinguishes false rejection of genuine signatures from false acceptance of forgeries, which have different operational consequences in forensic screening.

AUC was selected as the primary evaluation metric because it is threshold-independent. The post-hoc threshold of 85 was chosen after observing the ROC curve and should be interpreted as an exploratory calibration, not a universal forensic standard.

6. RESULTS

The compatibility score showed clear discriminative behavior between genuine signatures and skilled forgeries. Genuine questioned signatures obtained a higher mean compatibility score (88.02) than forgeries (79.52), a difference also visible in the median values, suggesting the effect is not driven by outliers. The score standard deviation was similar across classes (≈ 4.77), indicating that the separation is systematic rather than occasional.

Metric	Genuine (n = 200)	Forged (n = 200)
Mean compatibility score	88.02	79.52
Median compatibility score	88.90	80.22
Std. deviation	4.77	4.79
Min score observed	69.06	63.49
Max score observed	96.19	89.79

Metric	Genuine (n = 200)	Forged (n = 200)
Mean structural IoU	0.113	0.085
Mean anti-remake score	27.86	23.08

Table 4. Descriptive statistics of the compatibility score and auxiliary metrics by class.

6.1 SCORE DISTRIBUTION

The score distribution also reveals the practical separation between classes. Among genuine signatures, 72 of 200 (36%) scored in the [90–100] range, 116 (58%) scored in [80–90), 11 (5.5%) in [70–80), and only 1 (0.5%) below 70. Among forgeries, no case exceeded 90; 103 (51.5%) scored in [80–90), 90 (45%) in [70–80), and 7 (3.5%) below 70. This distribution confirms that a threshold around 85 separates the classes with reasonable precision.

6.2 THRESHOLD-BASED CLASSIFICATION

Using a calibrated compatibility threshold of 85, the framework obtained 83.0% accuracy. Sensitivity for genuine signatures was 78.5%, and specificity for skilled forgeries was 87.5%. The higher specificity indicates that, under this threshold, the framework was slightly more effective at rejecting skilled forgeries than at accepting all genuine questioned samples.

Measure	Value
Accuracy	83.0%
Sensitivity (genuine signatures)	78.5%
Specificity (skilled forgeries)	87.5%
AUC of compatibility score	0.9017
True genuine accepted (TP)	157
Genuine rejected (FN)	43
True forgeries rejected (TN)	175
Forgeries accepted (FP)	25

Table 5. Classification performance at calibrated threshold of 85.

6.3 Confusion Matrix

Actual \ Predicted	Predicted Genuine	Predicted Forged
Genuine (200)	157 (TP)	43 (FN)
Forged (200)	25 (FP)	175 (TN)

Table 6. Confusion matrix at compatibility threshold = 85. Genuine is the positive class.

6.4 ROC ANALYSIS

The ROC curve was generated from all 359 distinct compatibility-score thresholds observed in the validation set. The area under the curve was 0.9017, indicating strong ranking capacity before

selecting any operational threshold. In practical terms, this means the compatibility score tends to assign higher values to genuine signatures than to skilled forgeries across a broad range of possible cutoffs.

The AUC result is particularly noteworthy because the framework uses transparent handcrafted descriptors and simple statistical normalization rather than a trained deep neural architecture. The result demonstrates that explainability and useful screening performance can coexist at a competitive level.

[Figure 1. ROC curve for the BHSig260-Bengali validation subset (AUC = 0.9017). The curve was generated from 359 distinct threshold values of the compatibility score. The diagonal reference line indicates random performance.]

6.5 CALIBRATION OBSERVATION

The original final decision rule combined compatibility score, diagonal distance, and a strict binary shape IoU threshold. Applied to the BHSig260 subset, mean IoU was 0.113 for genuine and 0.085 for forged samples—both far below the threshold originally considered. Consequently, all 400 trials were classified as relevant divergence by the raw categorical rule, producing an overall decision accuracy of 50% (chance level). This result does not invalidate the continuous scoring layer; it demonstrates that structural thresholds are sensitive to dataset acquisition style, script family, cropping, and spatial normalization, and must be calibrated independently for each deployment context.

7. DISCUSSION

7.1 DISCRIMINATIVE CAPACITY OF THE COMPATIBILITY SCORE

The AUC of 0.9017 indicates that the compatibility score ranks genuine signatures above skilled forgeries with substantial consistency in the BHSig260-Bengali subset. This result is meaningful because the framework does not use a trained classifier: the score emerges from weighted statistical normalization of handcrafted graphotechnical descriptors. The finding supports the central argument that interpretable descriptor-level scoring can achieve competitive discriminative performance while preserving full transparency.

The 83.0% threshold-based accuracy further indicates practical screening utility. However, the confusion matrix reveals both false rejections and false acceptances. In a forensic context, these errors carry different operational consequences. False rejection of a genuine signature increases examiner workload and may trigger unnecessary investigation of authentic variation. False acceptance of a skilled forgery is more serious because it allows a simulated signature to pass preliminary screening. The balance between sensitivity and specificity must therefore be chosen according to operational requirements, not fixed at the values reported here.

7.2 EXPLAINABILITY AS A CORE FORENSIC REQUIREMENT

The strongest contribution of the proposed approach is explainability. Each score can be traced to specific descriptor families and intermediate computations. An examiner can inspect whether a questioned signature diverged in slant, baseline behavior, projection entropy, contour complexity, or apparent trace uniformity. This feature-level visibility supports review, contestability, and transparent reporting, and it aligns with the forensic principle that conclusions must be defensible and communicable to non-specialist audiences.

This design choice contrasts with deep learning approaches where the decision boundary may be sensitive to texture artifacts, background patterns, or color channels that are not forensically meaningful. In the proposed framework, every metric has a defined graphotechnical interpretation, and no scoring step is opaque. This makes the framework particularly suitable for deployment as a triage tool in forensic laboratory workflows.

7.3 CALIBRATION REQUIREMENTS

The calibration observation regarding the IoU threshold is one of the most valuable methodological findings of this study. Pixel-level shape overlap can be low even among genuine signatures when there is natural variation in placement, scale, and local deformation. This is expected for offline signatures, which lack online stabilization and may vary considerably across sessions. Consequently, rigid shape-overlap criteria that appear intuitive—genuine signatures should look similar—may produce systematically misleading results when transferred across scripts, acquisition devices, or writer populations.

The appropriate response to this observation is not to abandon structural analysis, but to treat structural measures as evidence layers that are reported and interpreted, not as hard binary decision criteria. The continuous IoU and projection correlation values retain diagnostic value when communicated as graded evidence rather than pass/fail rules. This is consistent with the broader forensic principle that calibrated, graded conclusions are preferable to binary certainty (Hicklin et al., 2022).

7.4 COMPARISON WITH DEEP LEARNING AND HYBRID APPROACHES

Deep learning methods, particularly convolutional and Siamese architectures, achieve higher AUC values on standard benchmarks when trained on large corpora. Hafemann, Sabourin, and Oliveira (2017), for example, reported error rates substantially below those achievable by handcrafted descriptors on the GPDS dataset. The proposed framework does not compete with these methods on raw performance. Instead, it occupies a different position: it is designed for forensic screening contexts where interpretability, traceability, and non-deterministic communication are as important as classification accuracy.

Future hybrid systems may integrate deep embeddings for performance with graphotechnical descriptors for explanation. The transparent compatibility score may serve as a calibration baseline or as a human-interpretable annotation layer alongside deep representations. This direction is explored in the forensic-oriented work of Diaz, Ferrer, and Vessio (2024), and represents a promising convergence of the two approaches.

7.5 ANTI-REMAKE ASSESSMENT

The anti-remake module produced scores that did not reliably separate genuine from forged samples in the BHSig260 validation. Mean anti-remake scores were 27.86 for genuine and 23.08 for forged samples—a small, non-decisive difference in the unexpected direction (genuine samples produced slightly higher risk scores). This result is scientifically important and is reported honestly here.

Several explanations are plausible. Public offline signature datasets are designed to evaluate writer-verification performance, not to simulate casework-level forgery techniques such as tracing, slow freehand redrawing, or mechanical copying. Skilled forgers in BHSig260 may produce forgeries with natural variation in pressure and stroke width that does not trigger the anti-remake

heuristics. The heuristic rules, inspired by casework patterns, require separate validation on datasets specifically designed to include tracing and mechanical simulation scenarios.

8. LIMITATIONS

Several limitations of this study must be acknowledged. First, the validation used a modest subset of one dataset and one script family. Bengali signatures may have structural properties—component fragmentation, stroke geometry, ink distribution—that differ substantially from Latin-script signatures, including Brazilian handwritten signatures. Generalization to other populations requires independent validation.

Second, only five genuine signatures per writer were available as references. This reflects a realistic small-sample forensic scenario, but it limits the stability of standard-deviation estimates and reduces the robustness of the baseline. A larger reference set would support more reliable modeling of within-writer variability.

Third, the anti-remake module remains heuristic and was not validated as an independent forgery detector. Its rules require evaluation on datasets designed to include tracing, slowly drawn forgeries, and mechanically assisted simulations.

Fourth, preprocessing choices can influence results. Binarization, cropping, scaling, and baseline alignment affect contour descriptors, projection entropy, and IoU. The framework must document preprocessing parameters and acknowledge that structural measures are not independent of normalization decisions.

Fifth, the compatibility threshold of 85 was selected post-hoc for the validation subset. Operational forensic use would require independent calibration data and, ideally, confidence intervals or uncertainty estimates around the threshold.

9. FUTURE WORK

Future research should extend validation to larger and more diverse datasets, including CEDAR, GPDS, and datasets containing Brazilian signatures, to evaluate which descriptor families are stable across scripts and acquisition protocols.

A second direction is hybrid explainable AI. Deep embeddings may provide stronger raw discrimination while graphotechnical descriptors provide forensic interpretability and report generation. Local explanation maps could highlight regions that contributed most to compatibility or divergence, further supporting examiner review.

A third direction is uncertainty modeling. Instead of a single threshold, future work could estimate confidence intervals, likelihood-ratio-inspired scores, or calibrated probability ranges aligned with forensic conclusion scales. This would reduce the risk of overinterpreting borderline cases.

Finally, operational validation with forensic document examiners is needed. A computationally useful tool should also improve examiner workflow, documentation quality, and consistency of preliminary screening. User studies could measure whether the explainable report helps examiners identify relevant features, communicate uncertainty, and avoid overreliance on automated outputs.

10. CONCLUSION

This paper presented an explainable computational framework for assisted signature screening in forensic document analysis. The framework integrates preprocessing, graphotechnical descriptor extraction, statistical baseline construction, compatibility scoring, structural comparison, and heuristic anti-remake risk assessment. Its design prioritizes transparency, auditability, and cautious reporting rather than autonomous authorship attribution.

Validation on a BHSig260-Bengali subset comprising 400 comparison trials demonstrated promising discriminative capacity: AUC = 0.9017, accuracy = 83.0%, sensitivity = 78.5%, and specificity = 87.5% under a calibrated threshold. The experiment also revealed that a rigid structural IoU threshold did not transfer to the validation dataset without calibration, a practically important finding for any forensic deployment of threshold-based classification rules.

The main contribution is not a claim to outperform deep learning systems, but to provide a structured, interpretable, and auditable decision-support layer for forensic signature screening. The framework is suitable for triage and case prioritization while reinforcing that its outputs are indicative and subordinate to expert human judgment. Future research should expand validation, improve calibration, explore hybrid explainable architectures, and evaluate the framework in operational forensic workflows with expert examiners.

DATA AND REPRODUCIBILITY NOTE

The validation was executed using the BHSig260-Bengali sample extracted to the project environment. Engine version: grafotecnia-triagem-1.1.0. Generated artifacts include a JSON summary, a CSV file with individual case records, and ROC data in CSV and SVG formats. The manuscript reports only aggregate metrics and contains no personal data.

The core validation command was: npm run test:grafotecnia:bhsig260, which executes scripts/validate-grafotecnia-bhsig260.js over the dataset. Reproducibility depends on access to the same dataset split, preprocessing parameters, and scoring implementation documented in Section 4.

REFERENCES

Diaz, M., Ferrer, M. A. and Vessio, G. (2024). Explainable offline automatic signature verifier to support forensic handwriting examiners. Neural Computing and Applications, 36(5), 2399–2409. https://doi.org/10.1007/s00521-023-09192-7

Hafemann, L. G., Sabourin, R. and Oliveira, L. S. (2017). Learning features for offline handwritten signature verification using deep convolutional neural networks. Pattern Recognition, 70, 163–176. https://doi.org/10.1016/j.patcog.2017.05.012

Hicklin, R. A., Eisenhart, L., Richetelli, N., Miller, M. D., Belcastro, P., Burkes, T. M., Parks, C., Smith, M., Buscaglia, J., Peters, E. M., Perlman, R. S., Abonamah, J. V. and Eckenrode, B. A. (2022). Accuracy and reliability of forensic handwriting comparisons. Proceedings of the National Academy of Sciences, 119(32), e2119944119. https://doi.org/10.1073/pnas.2119944119

Impedovo, D. and Pirlo, G. (2008). Automatic signature verification: The state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38(5), 609–635. https://doi.org/10.1109/TSMCC.2008.923866

Kam, M., Gummadidala, K., Fielding, G. and Conn, R. (2001). Signature authentication by forensic document examiners. Journal of Forensic Sciences, 46(4), 884–888. https://doi.org/10.1520/JFS15062J

Mazzolini, D., Mignone, P., Pavan, P. and Vessio, G. (2021). An easy-to-explain decision support framework for forensic analysis of dynamic signatures. Forensic Science International: Digital Investigation, 38, 301216. https://doi.org/10.1016/j.fsidi.2021.301216

National Institute of Standards and Technology. (2021). Forensic Handwriting Examination and Human Factors (NIST IR 8282r1). U.S. Department of Commerce. https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8282r1.pdf

Plamondon, R. and Lorette, G. (1989). Automatic signature verification and writer identification: The state of the art. Pattern Recognition, 22(2), 107–131. https://doi.org/10.1016/0031-3203(89)90059-9

Sita, J., Found, B. and Rogers, D. (2002). Forensic handwriting examiners' expertise for signature comparison. Journal of Forensic Sciences, 47(5), 1117–1124. https://doi.org/10.1520/JFS15521J

APPENDIX A. REVISION COMPLIANCE MAP

This appendix summarizes how the revised manuscript addresses the main points raised during peer review. The main text already incorporates all required changes; this map verifies that no major-review item remains unanswered before resubmission.

Reviewer Concern	Revision Implemented	Where Addressed
Single illustrative case; limited generalizability	Added quantitative validation using BHSig260-Bengali with 20 writers and 400 trials	Abstract; Sections 5 and 6
Need for objective metrics	Reported AUC, accuracy, sensitivity, specificity, and confusion matrix	Section 6
Insufficient technical detail	Expanded preprocessing, descriptors, baseline, scoring formulas, structural comparison, and anti-remake rules	Sections 3 and 4
Need for reproducibility	Added pseudocode, data/reproducibility note, and validation protocol detail	Sections 4.6, 5, and Data Note
Need stronger state-of-the-art positioning	Compared transparent descriptors, deep learning, Siamese models, and explainable forensic frameworks	Sections 2 and 7.4
Need for expanded limitations discussion	Added calibration, dataset scope, preprocessing sensitivity, anti-remake heuristic limits	Sections 7 and 8

APPENDIX B. PRACTICAL INTERPRETATION OF VALIDATION RESULTS

The AUC of 0.9017 indicates that the continuous compatibility score ranked genuine signatures above skilled forgeries with strong consistency in the selected validation subset. This is the most important quantitative result because it does not depend on a single operating threshold. It supports the argument that the descriptor-based score captures meaningful graphotechnical information even without a trained deep neural classifier.

The threshold-based results should be interpreted with more caution. Accuracy of 83.0%, sensitivity of 78.5%, and specificity of 87.5% show that the selected threshold of 85 can support useful screening decisions on this subset. However, the threshold was calibrated after observing the validation behavior and should be considered an experimental value, not a universal forensic standard.

The false rejection count of 43 (genuine signatures below the threshold) is expected in offline signature datasets because genuine signatures can vary substantially in shape, scale, and placement. In operational use, false rejections would not necessarily be disqualifying if the system is used for triage: they would prompt additional examiner review rather than a definitive decision. The false acceptance count of 25 (skilled forgeries above the threshold) is more concerning from a screening perspective, reinforcing the need for human review as the final step.

The structural IoU finding is one of the most valuable methodological results. The original categorical rule treated low IoU as a strong divergence criterion, but the BHSig260 validation showed that low pixel-level overlap occurs even among genuine signatures. This justifies retaining IoU as a graded evidence layer while abandoning it as a universal hard threshold. Future calibration procedures should determine appropriate IoU ranges for specific scripts, acquisition devices, and preprocessing pipelines.