Benchmarking and Comparative Evaluation
Volume 1 · Number 1 · 2026

A Blank Page Scores Ninety-Two Percent: Background Imbalance in Handwritten-Document Evaluation

Logan Wilson[1] · Austin Anderson[2] · Ryan King[3]

[1] Denmark · [2] Norway · [3] Italy

Original publication: October 6, 2026 | Article 678

Abstract

An output that predicts no handwriting obtains mean pixel accuracy 0.917 on ten H-DIBCO2018 pages. It exceeds a frozen local binarization rule on two pages, despite having foreground F1 zero everywhere. Across pages, its accuracy deficit relative to that rule is small and uncertain, while its F1 deficit is large. The result demonstrates a benchmark-definition problem rather than a successful reconstruction: when background dominates, all-pixel agreement can give a superficially favorable score to an output that removes the entire foreground.

A Blank Page Scores Ninety-Two Percent: Background Imbalance in Handwritten-Document Evaluation

Abstract

An output that predicts no handwriting obtains mean pixel accuracy 0.917 on ten H-DIBCO2018 pages. It exceeds a frozen local binarization rule on two pages, despite having foreground F1 zero everywhere. Across pages, its accuracy deficit relative to that rule is small and uncertain, while its F1 deficit is large. The result demonstrates a benchmark-definition problem rather than a successful reconstruction: when background dominates, all-pixel agreement can give a superficially favorable score to an output that removes the entire foreground.

1. The benchmark needs a deliberately unhelpful reference

A handwritten-document system should preserve the information a reader hopes to recover. An evaluation metric may reward something easier: agreeing with the large amount of background in a page. A deliberately empty output makes that distinction visible. The research question is whether an all-background prediction can appear competitive under pixel accuracy on actual benchmark pages, and how that apparent standing changes when the same outputs are evaluated for foreground recovery. The experiment uses a weak reference to audit the score rather than to propose a useful document-processing method.

The practical motivation extends to restoration. Meng et al., 2025 concerns inpainting of oracle-bone inscription rubbings, where the meaning of preserved foreground is especially salient. Our analysis neither runs that model nor uses oracle-bone images. It studies binary handwriting annotations in a different public benchmark. The connection is a measurement requirement: an output's agreement with a large background region should not be casually interpreted as recovery of the content that motivates restoration. That requirement can be tested without claiming that handwriting binarization reproduces the target inpainting task.

The source pages come from Pratikakis et al., 2018, with paired grayscale images and binary foreground references. The earlier resource in Pratikakis et al., 2016 supplies the parameter-selection cohort for one local thresholding rule already frozen in this project. We reuse those retained choices openly. This is a new comparison of metric definitions on the same later pages, not a new independent document collection or an additional validation of the source-selected threshold. Reusing the pages permits an exact output comparison while limiting claims about generalization.

2. Why accuracy and foreground recovery diverge

Pixel accuracy counts agreement on both foreground and background. If a page contains little foreground, predicting background everywhere obtains a high value without detecting any mark. Foreground F1 combines foreground precision and recall, while balanced accuracy gives equal weight to foreground sensitivity and background specificity. These measures answer different questions. The threshold-evaluation perspective in Fawcett, 2006 helps explain why a high all-pixel value need not imply useful detection of a minority class, especially when the operating goal centers on that class.

The loss perspective in Gneiting and Raftery, 2007 and the cost-sensitive account in Elkan, 2001 make the institutional choice explicit. Treating each pixel equally effectively assigns importance in proportion to the source class counts. That may be defensible for some image-reproduction goals, but it is not automatically appropriate for recovering sparse writing. We have not elicited a human cost for a missed

stroke or a spurious mark. Instead, we report several defined metrics and ask whether they tell a consistent story about a deliberately empty output.

The measurement distinctions in Stevens, 1946 also prevent translating a numerical score into a stronger semantic statement. Ninety-two-percent pixel agreement is a fraction of matching binary labels, not ninety-two-percent legibility, text accuracy, or historical-information recovery. Carroll et al., 2006 adds that the reference itself is a measurement with a particular construction, rather than an error-free account of every meaningful mark. Our experiment accepts the released binary labels for the declared pixel task and does not revise ambiguous regions based on a preferred result.

Agreement methods such as those in Bland and Altman, 1986 and Lin, 1989 concern another layer of comparison: procedures may have similar aggregate values while differing at specific locations. We consequently score each actual prediction against the same reference, then average within pages. We do not infer pixel-level equivalence from the closeness of overall means. Likewise, neither a perfect background score nor a high whole-page accuracy can establish that the few foreground pixels a reader needs have been preserved in the output.

3. Source pages and three fixed outputs

There are ten official later-cohort image/reference pairs, each read at its native resolution. We convert input images to grayscale using the local image library and preserve their pixel grids. The released ground truth contains only black and white values, with black defining foreground. Every page has positive foreground. The fraction of foreground ranges from 0.032607 to 0.162175. These observed proportions make the class imbalance concrete, without requiring an invented distribution of characters or a claim about how the pages were captured.

The empty prediction sets every pixel to background. It has no parameters, training data, or image-dependent adjustment. Its accuracy equals the source background fraction. Its foreground recall is zero, its reported foreground precision follows the explicit zero-predicted-positive convention, and its F1 is zero. Balanced accuracy is one half because foreground sensitivity is zero and background specificity is one. These are consequences of the definition on the actual nonempty references, not estimates of a general population error probability.

The global comparison uses the histogram thresholding approach in Otsu, 1979, selecting the first maximizing threshold and classifying dark pixels as foreground. The local comparison uses Sauvola and Pietikäinen, 2000 with a fifty-one-pixel window, parameter 0.2, scale constant 128, and reflected borders. That pair was chosen from a nine-combination grid on the earlier ten-page cohort by maximizing average foreground F1. Its source rule is stored and unchanged. No later-page metric is used to retune the window or threshold parameter for this comparison.

Morphological processing in Haralick et al., 1987 and connected-component methods in Rosenfeld and Pfaltz, 1966 could add further output rules, but none is applied here. This restriction keeps the benchmark question simple: what do the defined metrics say about an empty image and two familiar threshold outputs? It also avoids turning the exercise into a search for a favorable postprocessing chain. The local rule does not receive the binary reference when predicting a page, and the empty prediction receives no privileged knowledge about the location of its writing.

4. Page-level model and inferential boundary

For every output/page combination, we calculate ordinary pixel accuracy, foreground F1, and balanced accuracy. We do not substitute the contest's pseudo-F or distortion measures for these declared metrics, and we do not describe the resulting table as an official entrant ranking. The comparison of interest is empty minus local: one difference in pixel accuracy and one in foreground F1 per page. Otsu's metrics provide an additional descriptive reference, preserving the distinction between a weak global baseline and the particular local rule used in the paired table.

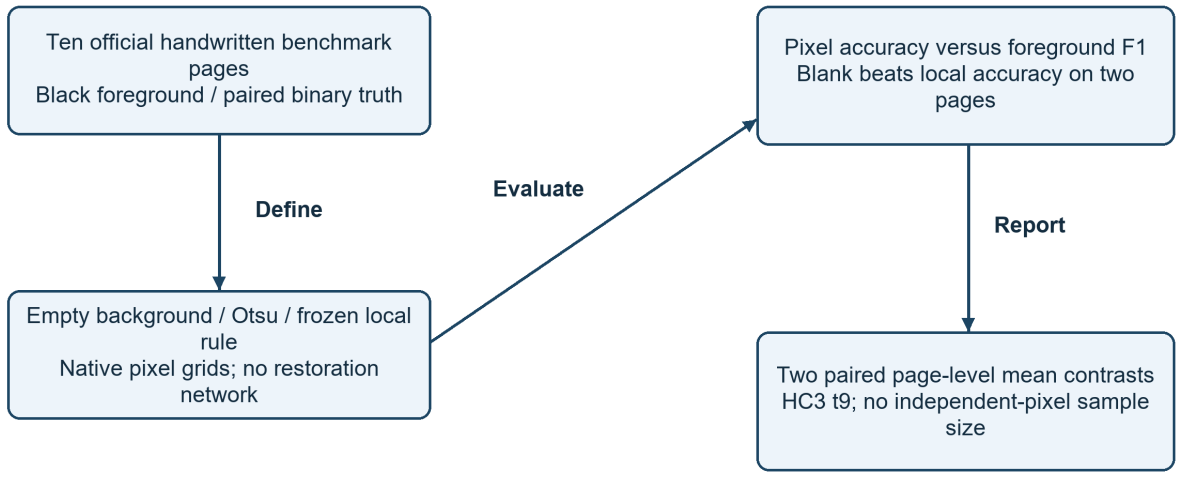

Figure 1. *Three fixed outputs are evaluated against the same binary page references. Class-insensitive pixel agreement and foreground-focused recovery produce different interpretations of the empty output.*

Each contrast is summarized by intercept-only OLS over ten page rows, using HC3 covariance and a Student reference with nine degrees of freedom. White, 1980 motivates the heteroskedasticity-aware treatment of page-to-page differences. Pixels are not treated as independent observations, and millions of pixels do not become an inferential sample size. The source does not provide a defensible scanner, writer, or archive grouping for a more complete dependence analysis. These intervals therefore condition on a very small and possibly related set of benchmark pages.

Diagnostic thinking in Cook, 1977 suggests checking unusual page influence, but no page is removed because its result is inconvenient. The reporting guidance associated with Wasserstein and Lazar, 2016 favors showing magnitudes and uncertainty rather than labeling a method useful on the basis of one significance result. In particular, a nonsignificant mean accuracy difference does not establish equivalence between blank and local outputs. Their foreground behavior is plainly different, and a small page sample can produce a wide interval around an accuracy comparison dominated by background agreement.

Table 1. *Empty-background minus frozen-local metrics, summarized by two intercept-only OLS models. N counts ten pages. Covariance is HC3, the reference distribution has nine degrees of freedom, and there are no controls. The separate constants F1 zero and balanced accuracy one half for empty outputs are descriptive identities and receive no standalone inferential model.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	-0.027381 (0.014694)	[-0.060620, 0.005858]	10
M1: fit	R-squared = 0.0000	-	-	10
M2	Mean outcome/contrast	-0.647999 (0.068752)	[-0.803527, -0.492471]	10
M2: fit	R-squared = 0.0000	-	-	10

Model labels: M1 = Empty-background minus local pixel accuracy; M2 = Empty-background minus local foreground F1.

5. High background agreement, absent foreground recovery

Mean page accuracy is 0.916526 for the empty output, 0.863111 for Otsu, and 0.943908 for the local rule. Thus the empty output exceeds the global baseline in the overall average, despite recovering no foreground. It also exceeds the local rule on two of the ten individual pages. These observations do not make emptiness an acceptable processing strategy. They demonstrate that the numerical criterion can favor an output that fails the content-recovery purpose when the source background occupies most pixels.

The empty-minus-local accuracy difference averages negative 0.027381, with standard error 0.014694 and a 95% interval from negative 0.060620 to positive 0.005858. Under the working page-level inference, the direction is uncertain. The finite-page average still favors the local output. Both facts should be retained: the estimate says which procedure wins in the observed mean, and the interval describes the limited precision of a broader mean summary under its assumptions. Neither fact establishes that the outputs have comparable practical usefulness.

Foreground F1 gives a different scale of contrast. Its mean is zero for the empty output, 0.514548 for Otsu, and 0.647999 for the local rule. The empty-minus-local difference is negative 0.647999, with an interval from negative 0.803527 to negative 0.492471. Balanced accuracy likewise separates the outputs: 0.5 for empty, 0.832114 for Otsu, and 0.858104 for local. These values quantify foreground-sensitive behavior; they still do not measure OCR accuracy, reader comprehension, or the importance of a specific missing mark.

The numerical contrast is not a paradox. Pixel accuracy includes a large correctly predicted background contribution that foreground F1 omits. Balanced accuracy changes the class weights again. These are different objectives applied to the same actual predictions and references. A high score on one cannot override the definition of another, and the observed ranking reversal does not show that any metric is arithmetically wrong. It shows that a benchmark must state whether the purpose is overall label agreement or recovery of a sparse class that carries the information of interest.

6. From binary foreground to restoration claims

Modern inpainting methods address richer tasks than the three output rules used here. Pathak et al., 2016 learns feature representations for inpainting, Suvorov et al., 2021 targets large-mask completion, and Lugmayr et al., 2022 uses a diffusion-based approach. Liu et al., 2018 and Yu et al., 2018b develop mask-aware image-completion operations, while Nazeri et al., 2019 uses edge structure in that setting. These works illustrate why a blank binary foreground baseline is an audit tool rather than an alternative inpainting model. We execute none of those networks and borrow no reported visual performance to explain our results.

Segmentation and denoising introduce other objectives. Ronneberger et al., 2015 provides a segmentation architecture, while Vincent et al., 2008 addresses recovering a signal from corrupted input. The oracle-problem perspective in Barr et al., 2015 raises the need for a suitable correctness criterion in software testing, a useful analogy for why a convenient automated score may not fully represent the desired outcome. Our released pixel reference supplies a well-defined oracle for binary agreement. It remains incomplete for semantic reading, where errors on visually small structures can carry disproportionate meaning.

The repeated-source design is also a boundary. Roberts et al., 2017 emphasizes validation under structured observations, and Kapoor and Narayanan, 2023 motivates preserving the separation between earlier parameter selection and later scoring. We honor that separation but do not claim that reusing the later pages creates fresh out-of-sample evidence. The source-selected local rule has already been studied on them for a different question. The present contribution is the explicit empty-output and metric-definition comparison, with no claim that it estimates performance on unseen archives, scripts, languages, or damage processes.

Documentation follows the aims in Gebru et al., 2021 and the reuse principles in Wilkinson et al., 2016. Page identifiers, dimensions, foreground fractions, fixed rule settings, and all metric values are retained with the source archive hashes. No binary reference is altered, no page is excluded after scoring, and no blank output is misrepresented as a reconstruction. A practical evaluation should add outcomes matching its intended use, but those additions require actual evidence. They cannot be inferred from a favorable accuracy value or from the name of a sophisticated restoration method.

Page weighting deserves attention as well. The reported averages give every page the same influence, regardless of its pixel dimensions. Pooling all pixels would answer a different question and could allow a large page with a particular foreground fraction to dominate the result. We do not substitute that alternative after observing the current comparison. The page-level table makes the primary weighting recoverable and preserves the two individual cases where the local rule has lower accuracy than an empty output.

Finally, a foreground-sensitive metric is still only one component of an evaluation. F1 ignores the locations and semantic importance of individual errors beyond their confusion counts, while balanced accuracy weights class sensitivities without representing the reading task. Their advantage in this audit is that they expose the empty output's complete foreground failure. The evidence does not justify treating either as a complete measure of historical restoration, and a workflow should avoid replacing one overly broad interpretation of a metric with another.

7. Conclusion

An all-background output achieves approximately ninety-two-percent pixel accuracy on these handwritten benchmark pages and beats the local rule on two pages, while its foreground F1 remains zero. The contrast makes the measurement priority visible: abundant background agreement can conceal complete foreground loss. Pixel accuracy can be reported as a defined quantity, but its interpretation should remain tied to the sparse-content objective that motivates document recovery and to the additional semantic evidence that the current binary benchmark does not supply.

References

1. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.
2. Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.
3. Lugmayr, Andreas; Danelljan, Martin; Romero, Andres; Yu, Fisher; Timofte, Radu; Van Gool, Luc. (2022). RePaint: Inpainting using Denoising Diffusion Probabilistic Models. arXiv:2201.09865. https://arxiv.org/abs/2201.09865.
4. Nazeri, Kamyar; Ng, Eric; Joseph, Tony; Qureshi, Faisal Z.; Ebrahimi, Mehran. (2019). EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning. arXiv:1901.00212. https://arxiv.org/abs/1901.00212.
5. Sauvola, J.; Pietikäinen, M.. (2000). Adaptive document image binarization. Pattern Recognition. https://doi.org/10.1016/s0031-3203(99)00055-2.
6. Suvorov, Roman; Logacheva, Elizaveta; Mashikhin, Anton; Remizova, Anastasia; Ashukha, Arsenii; Silvestrov, Aleksei; Kong, Naejin; Goka, Harshith; Park, Kiwoong; Lempitsky, Victor. (2021). Resolution-robust Large Mask Inpainting with Fourier Convolutions. arXiv:2109.07161. https://arxiv.org/abs/2109.07161.
7. Rosenfeld, Azriel; Pfaltz, John L.. (1966). Sequential Operations in Digital Picture Processing. Journal of the ACM. https://doi.org/10.1145/321356.321357.
8. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; 't Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.
9. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.
10. Pratikakis, Ioannis; Zagori, Konstantinos; Kaddas, Panagiotis; Gatos, Basilis. (2018). ICFHR 2018 Competition on Handwritten Document Image Binarization (H-DIBCO 2018). 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). https://doi.org/10.1109/icfhr-2018.2018.00091.
11. Liu, Guilin; Reda, Fitsum A.; Shih, Kevin J.; Wang, Ting-Chun; Tao, Andrew; Catanzaro, Bryan. (2018). Image Inpainting for Irregular Holes Using Partial Convolutions. arXiv:1804.07723. https://arxiv.org/abs/1804.07723.
12. Barr, Earl T.; Harman, Mark; McMinn, Phil; Shahbaz, Muzammil; Yoo, Shin. (2015). The Oracle Problem in Software Testing: A Survey. IEEE Transactions on Software Engineering. https://doi.org/10.1109/tse.2014.2372785.
13. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.
14. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
15. Pathak, Deepak; Krahenbuhl, Philipp; Donahue, Jeff; Darrell, Trevor; Efros, Alexei A.. (2016). Context Encoders: Feature Learning by Inpainting. arXiv:1604.07379. https://arxiv.org/abs/1604.07379.
16. Pratikakis, Ioannis; Zagoris, Konstantinos; Barlas, George; Gatos, Basilis. (2016). ICFHR2016 Handwritten Document Image Binarization Contest (H-DIBCO 2016). 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). https://doi.org/10.1109/icfhr.2016.0118.

17. Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” *Lecture Notes in Computer Science*, 2015, pp. 234-241. https://doi.org/10.1007/978-3-319-24574-4_28.

18. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

19. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.

20. Kapoor, Sayash; Narayanan, Arvind. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns. https://doi.org/10.1016/j.patter.2023.100804.

21. Elkan, Charles. “The Foundations of Cost-Sensitive Learning.” *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973-978. https://cseweb.ucsd.edu/~elkan/rescale.pdf.

22. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

23. Yu, Jiahui; Lin, Zhe; Yang, Jimei; Shen, Xiaohui; Lu, Xin; Huang, Thomas. (2018b). Free-Form Image Inpainting with Gated Convolution. arXiv:1806.03589. https://arxiv.org/abs/1806.03589.

24. Vincent, Pascal; Larochelle, Hugo; Bengio, Yoshua; Manzagol, Pierre-Antoine. (2008). Extracting and composing robust features with denoising autoencoders. Proceedings of the 25th international conference on Machine learning - ICML '08. https://doi.org/10.1145/1390156.1390294.

25. Meng, Zijie; Zeng, Yuanze; Chang, Xiang; Xu, Tianshuo; Chao, Fei; Cao, Xixin; Shang, Changjing; Shen, Qiang. (2025). Orpaint: a zero-shot inpainting model for oracle bone inscription rubbings with visual mamba block. Science China Information Sciences, 68(8), 189102. https://doi.org/10.1007/s11432-024-4493-4.

26. Haralick, Robert M.; Sternberg, Stanley R.; Zhuang, Xinhua. (1987). Image Analysis Using Mathematical Morphology. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/tpami.1987.4767941.

27. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.

28. Otsu, Nobuyuki. (1979). A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics. https://doi.org/10.1109/tsmc.1979.4310076.

29. Fawcett, Tom. (2006). An introduction to ROC analysis. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2005.10.010.

30. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.