Operational Evidence and Outcomes
Volume 1 · Number 1 · 2026

A Better Reconstruction Can Still Lose a Correct Digit: Linking Hidden-Block Error to Classification Outcomes

Henry Richardson[1] · Brandon Webb[2] · James Harrison[3]

[1] Spain · [2] Portugal · [3] Italy

Original publication: November 8, 2026 | Article 610

Abstract

Lower reconstruction error helped digit classification on average, but it did not protect every correct decision. We restored a central sixteen-feature block in 1,797 optical-digit records using a fixed latent model or training-mean filling. Latent reconstruction reduced mean hidden-block MSE by 0.02424 and raised prototype-classification accuracy by 9.63 percentage points. Nevertheless, 42 records had lower reconstruction error while changing from a correct to an incorrect class. The two outcomes generally improved together at the aggregate level without becoming interchangeable criteria for individual restored items.

Paired restoration and downstream-decision study on optical-digit features. Source digit labels define the operational task; historical inscription meaning and human readability are unmeasured.

A Better Reconstruction Can Still Lose a Correct Digit: Linking Hidden-Block Error to Classification Outcomes

> Paired restoration and downstream-decision study on optical-digit features. Source digit labels define the operational task; historical inscription meaning and human readability are unmeasured.

Abstract

Lower reconstruction error helped digit classification on average, but it did not protect every correct decision. We restored a central sixteen-feature block in 1,797 optical-digit records using a fixed latent model or training-mean filling. Latent reconstruction reduced mean hidden-block MSE by 0.02424 and raised prototype-classification accuracy by 9.63 percentage points. Nevertheless, 42 records had lower reconstruction error while changing from a correct to an incorrect class. The two outcomes generally improved together at the aggregate level without becoming interchangeable criteria for individual restored items.

A numerical improvement and a task decision

Restoration is often evaluated by how closely reconstructed values match a hidden reference. A downstream workflow may instead need the restored item to retain a recognizable class. Those goals are related, but an average numerical improvement does not logically guarantee that every classifier decision is preserved. A small change near a decision boundary can matter more than a larger change in an irrelevant feature direction.

We ask whether improving masked-feature reconstruction also preserves correct digit decisions, and how often lower target error accompanies a lost classification. The experiment holds the observed input coordinates and downstream classifier fixed while changing only the hidden-region reconstruction rule. This creates a paired view of numerical loss and operational consequences for the same source records.

Meng et al., 2025 studies zero-shot inpainting of oracle-bone inscription rubbings using a visual Mamba block. Our experiment does not execute that architecture or establish readability of historical characters. It addresses an adjacent evaluation issue: whether a reconstruction loss can serve as a complete proxy for the interpretive task that motivates restoration.

Decision-oriented evaluation in Gneiting and Raftery, 2007 links a criterion to the outcome it is intended to measure. Elkan, 2001 further emphasizes that different errors can have different decision costs. Here the operational task is deliberately narrow: choosing one of ten source digit labels. We assign no human or archival cost to a particular misclassification and do not infer that all errors would matter equally in a real application.

Why the two outcomes need not agree

Mean squared error weights deviations in the hidden coordinates according to their numerical size. A nearest-prototype classifier weights the resulting complete feature vector through its distance to class means. Those two geometries are not the same. Moving a reconstruction closer to its reference can still cross a classifier boundary in a direction that favors another class.

The image-representation traditions in Turk and Pentland, 1991 and Belhumeur et al., 1997 illustrate how discrimination and reconstruction can motivate different projections. We do not apply their face-recognition models to digits. Their conceptual relevance is that representing an image accurately and separating its class are distinguishable objectives, even when they use related numerical features.

Pearson, 1901 provides a closest-fit perspective, while Tipping and Bishop, 1999 formalizes a probabilistic principal-component representation. Hoerl and Kennard, 1970 motivates regularization of the latent fitting equation. These sources locate the simple reconstruction model used here; they do not imply that minimizing its squared error should optimize a separate classifier's discrete decisions.

Agreement analysis in Bland and Altman, 1986 and Lin, 1989 offers another useful distinction. Two outcomes can be strongly related without being interchangeable for individual items. We therefore retain paired image-level losses and class transitions instead of presenting one aggregate correlation as proof that reconstruction quality is an adequate semantic proxy.

A fixed source split and an artificial missing block

The public data are Alpaydin and Kaynak, 1998, optical digits represented by an eight-by-eight grid of sixty-four block-count features. Values are divided by sixteen. The official 3,823 training records determine the reconstruction basis, training means, and class prototypes. All 1,797 official test records remain in the evaluation, including records on which both restoration rules fail the downstream task.

The hidden region is the central four-by-four block, sixteen features in total. The surrounding forty-eight coordinates remain exactly observed. This is an imposed numerical mask rather than an actual damaged document. The reconstruction target is a normalized block count, not a pixel-level photograph of an inscription rubbing or a direct representation of the full character stroke process.

The mean-fill baseline assigns each hidden feature its mean over the training records. The latent route estimates coefficients from the same forty-eight visible values and predicts the hidden block from a frozen training-derived basis. Neither route receives the test record's true digit label or hidden target values when producing its completed vector.

Known test labels are used only to evaluate the downstream decision. Leakage considerations in Kapoor and Narayanan, 2023 motivate keeping that boundary explicit, particularly because source labels would make class-specific restoration easy to construct. No privileged class-conditioned fill is included in this experiment. The reconstruction basis is learned without digit identities.

Reconstruction and classification as separate fitted components

The latent basis consists of the first sixteen principal directions of the centered complete training features. Its coefficient estimate minimizes visible squared residuals plus a ridge penalty of 0.01. We clip predicted hidden values to the known zero-to-one feature range. The mean-fill baseline already lies within that range, so both routes share the same bounded output convention.

Each completed vector preserves all visible coordinates and changes only the central target block. The downstream classifier compares this full sixty-four-feature vector with ten class prototypes. Each prototype is the complete training mean for one source digit label. The nearest prototype supplies the predicted class, with no test-data tuning or adaptation of the classifier after restoration.

The classifier is intentionally simple and imperfect. We also apply it to the complete unmasked source test vectors as a descriptive reference. That result is not an oracle upper bound on every possible completion or recognition method. It indicates how this particular fixed decision rule behaves when its original input values are available.

Open-set recognition discussed in Scheirer et al., 2013 concerns situations beyond a fixed known label menu. Our classifier has no unknown-character option and must choose among the ten source digits. Its

accuracy therefore cannot be read as a verified rejection policy or a confidence measure for arbitrary historical symbols.

Restoration literature does not supply the downstream result

Pathak et al., 2016 uses learned context for completion, while Liu et al., 2018 and Yu et al., 2018b address masked visual inputs through different network operations. These works motivate treating visible context as informative, but their published image results do not determine how a fixed digit prototype will respond to our feature predictions.

Suvorov et al., 2021 and Lugmayr et al., 2022 pursue image-completion mechanisms with their own priors. Nazeri et al., 2019 emphasizes structural completion, and Ronneberger et al., 2015 supplies a general dense-prediction architecture. We do not compare their outputs or count them as executed baselines. Their role is to place the question of target error versus downstream interpretation within a broader restoration setting.

Denoising representation learning in Vincent et al., 2008 likewise involves learning how corrupted observations relate to useful structure. The current experiment has one fixed mask and a frozen linear latent representation. A favorable mean classification gain is not evidence that this lightweight model matches a modern restoration network or handles naturally damaged artifacts.

Estimating paired changes and retaining exceptions

For each test record, we compute the mean squared hidden-feature error under both restoration rules. We separately record whether each completed vector receives the correct source digit label. The two regression outcomes are latent-minus-mean-fill MSE and latent-minus-mean-fill correct-class indicator. Their units remain separate throughout the analysis.

Each outcome is summarized by intercept-only OLS across 1,797 paired records with HC3 standard errors and t1796 intervals. The robust covariance rationale comes from White, 1980. No controls or fixed effects enter these mean models. R-squared is zero because they estimate differences rather than a predictive relationship between reconstruction and classification.

Unknown writer identifiers prevent writer-clustered uncertainty. The fitting split, prototype classifier, mask, and basis are held fixed, so their uncertainty is not propagated. Carroll et al., 2006 motivates caution when a measured feature loss is used as a proxy for a broader construct. Stevens, 1946 further requires distinguishing normalized squared error from an accuracy fraction.

We also tabulate class recoveries and losses, and identify records whose hidden MSE improves but whose classification changes from correct to wrong. This final category directly answers the individual-level research question. It is not selected to refit either method, and its existence is not converted into an estimate of human semantic damage.

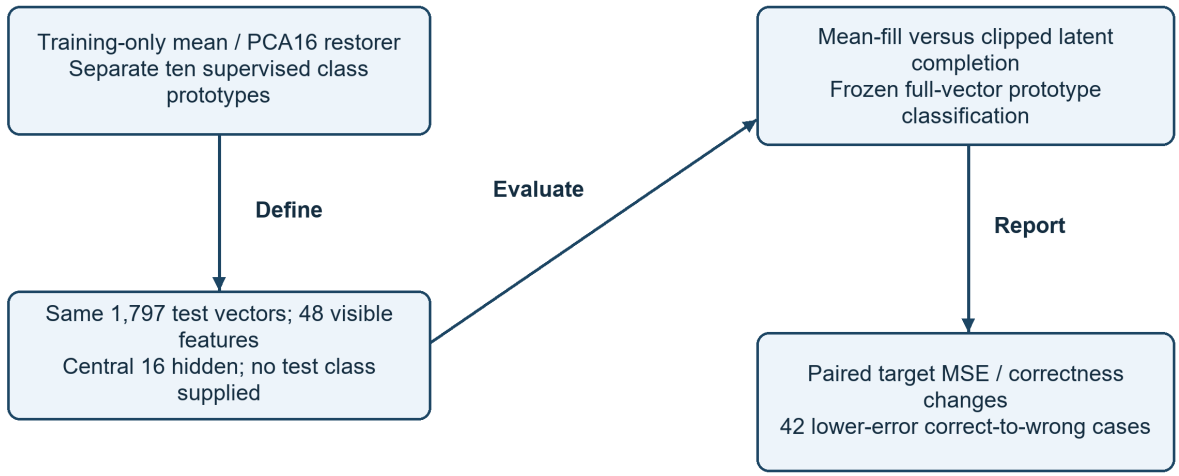

Figure 1. *The same restored vector supports two different outcomes. Source labels evaluate classification after reconstruction and never enter hidden-block prediction.*

Improvements in both averages, with discordant individual changes

Mean-fill hidden MSE was 0.144916 and latent-model MSE was 0.120672. The mean difference was -0.024243, with standard error 0.001754 and working interval -0.027683 to -0.020804. The latent reconstruction therefore improved the selected numerical target substantially within the fixed source experiment.

Mean-fill classification was correct for 1,133 records, or 0.630495. Latent reconstruction raised that count to 1,306, or 0.726767. Their mean accuracy difference was 0.096272, with interval 0.073200 to 0.119343. The complete-source classifier identified 1,606 records correctly, or 0.893712, as a separate descriptive input reference.

Table 1. *Latent reconstruction minus mean filling on hidden-block MSE and downstream accuracy. N equals 1,797 paired test records. Intercept-only OLS, HC3 covariance, t1796 intervals; no controls or fixed effects. The two rows measure different outcomes and are not interchangeable effect scales.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	-0.024243 (0.001754)	[-0.027683, -0.020804]	1797
M1: fit	R-squared = 0.0000	-	-	1797
M2	Mean outcome/contrast	0.096272 (0.011763)	[0.073200, 0.119343]	1797
M2: fit	R-squared = 0.0000	-	-	1797

Model labels: M1 = Latent restoration minus mean-fill hidden MSE; M2 = Latent restoration minus mean-fill classification accuracy.

Latent reconstruction recovered 318 previously incorrect classifications and lost 145 previously correct ones, producing the net gain of 173 decisions. It lowered hidden MSE in 1,184 records. Within that improved-error group, 42 records nevertheless changed from a correct class to an incorrect class. Aggregate gains therefore coexisted with directly observed exceptions to the claim that lower reconstruction loss protects every decision.

The 42 cases are not an accusation that the reconstructed vectors are visually worse or historically misleading. The experiment contains no human reading study. They establish disagreement between two operational criteria under a particular fixed classifier. Other recognition rules could place boundaries differently and produce a different exception count.

What a downstream audit adds

A restoration report containing only the negative MSE difference would correctly describe the numerical target and omit the class-transition consequences. A report containing only the positive accuracy difference would hide the 145 lost decisions and the distinction between recovered and damaged cases. Retaining both outcomes makes the practical result more specific without declaring either criterion universally sufficient.

The oracle problem in Barr et al., 2015 offers a broader warning that an executable or numerical check need not exhaust intended correctness. Our source labels provide an explicit task oracle for ten digits, but they do not turn a restored block-count vector into a complete semantic interpretation. That boundary is particularly important when drawing analogies to historical writing with uncertain or context-dependent meaning.

Threshold evaluation discussed in Fawcett, 2006 would require a score and operating rule suited to a specified decision. Our forced nearest-prototype labels do not provide a validated accept-or-review threshold. No human reviewer intervenes in the lost cases, and no measured review cost is attached to them. Such a workflow would require a separate experiment rather than being inferred from class accuracy alone.

Influence analysis in Cook, 1977 motivates preserving record-level transitions instead of deleting awkward examples. Wasserstein and Lazar, 2016 cautions that small p-values for average differences do not settle the interpretation of individual outcomes. The strong aggregate evidence for lower MSE and higher accuracy is compatible with the practical fact that some correct decisions were lost.

Why a class boundary can overturn the numerical gain

The prototype classifier compares complete-vector distances, including the forty-eight unchanged observed features. Those visible values can already place a record near the boundary between two class means. Changes to the hidden block then affect which prototype is closest. The hidden reconstruction error averages all sixteen target-coordinate discrepancies, while the class boundary responds to their direction relative to the competing prototypes. A smaller average squared discrepancy need not move the vector toward the correct side of that boundary.

This explanation does not require labeling the classifier or the reconstruction as erroneous whenever they disagree. A classifier can be imperfect, and a reconstruction can optimize a different objective successfully. The controlled experiment evaluates their combination as a workflow. The 42 discordant records show why a task result should be measured directly rather than guaranteed by an improvement in a surrogate loss.

The complete-source accuracy is useful for the same reason. It describes the fixed classifier on the original feature vector, with 191 errors still present. It is not a ceiling that every imputed vector must obey. A completion could accidentally move a difficult example into the correct prototype region even when the original source was misclassified. Such an event would improve this operational decision without proving a more faithful reconstruction of the hidden target.

The mean comparisons also weight source records equally. They do not show that every digit category benefits equally or that the gain would persist under a different class mixture. Class prototypes are training means, but their downstream success depends on the query distribution and geometry. No class-specific tuning or selective removal of difficult digits was performed to create the reported aggregate gain.

Nor is the 42-case category a causal subgroup defined before reconstruction. Membership requires observing both a loss improvement and a class transition. It identifies a concrete incompatibility between two evaluation criteria, with no claim about the rate at which a future independent restoration will damage meaning. Its value is to keep individual consequences visible alongside favorable averages and to motivate an outcome appropriate to the actual interpretive task.

Conclusion

Yan, 2026 compares greedy and GAN-based compressed sensing for MRI reconstruction, an example of a reconstruction comparison whose interpretation depends on the chosen fidelity objective. Our digit-feature experiment additionally scores a frozen downstream classifier, without acquiring MRI, assessing clinical interpretation, or running those reconstruction methods. Lower hidden-block error and correct class recovery are distinct measured outcomes here. Documentation in Gebru et al., 2021 keeps the restorer, classifier and transition counts linked, including cases where numerical improvement accompanies loss of a correct label. This supports evaluating the intended downstream task alongside numerical fidelity within the present controlled setting.

References

1. 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.
2. Belhumeur, P.N.; Hespanha, J.P.; Kriegman, D.J.. (1997). Eigenfaces vs. Fisherfaces: recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/34.598228.
3. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.
4. Tipping, Michael E.; Bishop, Christopher M.. (1999). Probabilistic Principal Component Analysis. Journal of the Royal Statistical Society Series B: Statistical Methodology. https://doi.org/10.1111/1467-9868.00196.
5. 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.
6. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
7. Alpaydin, E.; Kaynak, C. (1998). Optical Recognition of Handwritten Digits [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C50P49.
8. 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.
9. 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.
10. 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.
11. Fawcett, Tom. (2006). An introduction to ROC analysis. Pattern Recognition Letters. https://doi.org/10.1016/j.patrec.2005.10.010.
12. 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.

13. 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.

14. Turk, M.A.; Pentland, A.P.. (1991). Face recognition using eigenfaces. Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.1991.139758.

15. Yan, Biyun. (2026). A Comparative Study: Greedy versus GAN-Based Compressed Sensing for Fast MRI Reconstruction. International Journal of Advance in Applied Science Research, 5(3), 36-42. https://h-tsp.com/index.php/ijaasr/article/view/264.

16. Hoerl, Arthur E.; Kennard, Robert W.. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. https://doi.org/10.1080/00401706.1970.10488634.

17. Pearson, Karl. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science. https://doi.org/10.1080/14786440109462720.

18. 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.

19. 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.

20. 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.

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

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

23. Scheirer, W. J.; de Rezende Rocha, A.; Sapkota, A.; Boult, T. E.. (2013). Toward Open Set Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/tpami.2012.256.

24. 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.

25. 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.

26. 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.

27. 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.

28. 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.

29. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

30. 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.