Population Variation and Equity
Volume 1 · Number 1 · 2026

Lower Brier Error, Weak Classification: A Chronological EEG Preprocessing Audit

Jack Palmer[1] · Jeffrey Stone[2] · Paul Lewis[3]

[1] United States · [2] France · [3] United Kingdom

Original publication: November 3, 2026 | Article 645

Abstract

Preprocessing can improve one prediction metric while leaving another poor. Using one 117-second EEG recording, we fitted ridge eye-state scores on the first seventy percent of rows and evaluated the remainder. Training-only clipping lowered mean Brier error from 0.361 to 0.327; clipped paired-channel features lowered it to 0.313. Yet classification accuracies were only 33.3, 31.3, and 36.8 percent, respectively. The paired-channel Brier contrast narrowly excluded zero under a fixed HAC specification, while the raw-clipping contrast did not. This short-recording result supports metric-specific reporting and chronological checks, not clinical usefulness, new-subject generalization, or validation of a graph-sign algorithm.

Single-recording exploratory manuscript using a public EEG dataset. No diagnostic validity or new participant study is claimed; human review is pending.

Lower Brier Error, Weak Classification: A Chronological EEG Preprocessing Audit

Single-recording exploratory manuscript using a public EEG dataset. No diagnostic validity or new participant study is claimed; human review is pending.

Abstract

Preprocessing can improve one prediction metric while leaving another poor. Using one 117-second EEG recording, we fitted ridge eye-state scores on the first seventy percent of rows and evaluated the remainder. Training-only clipping lowered mean Brier error from 0.361 to 0.327; clipped paired-channel features lowered it to 0.313. Yet classification accuracies were only 33.3, 31.3, and 36.8 percent, respectively. The paired-channel Brier contrast narrowly excluded zero under a fixed HAC specification, while the raw-clipping contrast did not. This short-recording result supports metric-specific reporting and chronological checks, not clinical usefulness, new-subject generalization, or validation of a graph-sign algorithm.

1. A robust feature can still support a weak decision

Sensor preprocessing is often justified by an improvement in a numerical prediction score. That improvement does not guarantee an improvement in a thresholded decision, especially when class balance changes over time. We examine this distinction in a public eye-state EEG sequence. The question is whether training-only clipping and a compact paired-channel representation preserve useful performance in the later portion of one recording, compared with a simple raw-channel ridge baseline.

The UCI EEG Eye State release contains 14,980 chronologically ordered rows from one continuous 117-second recording (Roesler, 2013). Fourteen channels accompany a camera-derived eye-state annotation, with one indicating closed eyes and zero open eyes. This is not a multi-patient clinical cohort. Its compact scale permits a fully reproducible temporal experiment, while the single recording imposes severe limits on participant-level and clinical generalization.

The adaptive graph normalized sign approach in Peng et al., 2024 motivates interest in robust processing of related signals. We do not implement that algorithm. Our compact representation consists of seven declared left-right channel differences and a common median level. It is an interpretable feature construction, not an estimated brain-connectivity graph or a proof that the target method improves eye-state recognition. The target work supplies context for a robustness question rather than a model evaluated in the table.

2. Signal representation and evaluation scope

Graph-signal perspectives in Ortega et al., 2017 and Shuman et al., 2012 describe how relationships between channels can inform signal processing. Named electrode pairs offer a simple form of structural prior, but their subtraction does not recover functional connectivity. A channel pair can share common electrical or recording effects, and eye-state labels can correlate with artifacts as well as neural activity. We therefore call the features pair contrasts and avoid a physiological mechanism claim.

Spectral estimation in Welch, 1967, public physiological data resources in Moody and Mark, 2005, and signal-extraction work in Sameni and Clifford, 2010 illustrate other analyses of physiological recordings. We do not compute a spectral transform, extract a fetal signal, or combine recordings from those resources. Their relevance is methodological: the choice of representation and time unit should match a

clearly defined outcome. Our experiment uses the released channel values directly because it asks a preprocessing and chronological-transfer question.

Explanation tools such as Lundberg and Lee, 2017 could describe a fitted predictor, while imputation approaches including Cini et al., 2021, Cao et al., 2018, Buuren and Groothuis-Oudshoorn, 2011, and Stekhoven and Bühlmann, 2012 address different forms of missing-data modeling. None is executed here. The acquired EEG table has complete channel entries for the analyzed rows. Large amplitudes are not treated as missing values, and we do not infer that every extreme observation is a sensor fault requiring imputation.

3. Three frozen policies

The baseline policy uses all fourteen raw channel values. A second policy clips each channel at its training first and ninety-ninth percentiles before fitting the same ridge procedure. A third policy applies that clipping and then computes AF3 minus AF4, F7 minus F8, F3 minus F4, FC5 minus FC6, T7 minus T8, P7 minus P8, and O1 minus O2, together with the row's median across all channels. It therefore uses eight features instead of fourteen.

All training means, standard deviations, and clipping bounds are estimated from training rows only. Each policy receives its own training-derived standardization after its representation is formed. Ridge least squares uses a fixed penalty of ten on the standardized feature coefficients and an unpenalized intercept. The label is the binary eye-state annotation. We do not select the penalty, clipping percentiles, pair list, or decision threshold according to held-out results.

The raw regression score is clipped to the unit interval for Brier evaluation and thresholded at one half for classification. This is a declared scoring convention, not an assertion that ridge regression produces calibrated clinical probabilities. Guo et al., 2017 motivates distinguishing score calibration from discrimination, and Gneiting and Raftery, 2007 explains why a probability-sensitive loss answers a different question from classification accuracy. Both outcomes are retained so that a favorable loss contrast cannot hide a weak decision rule.

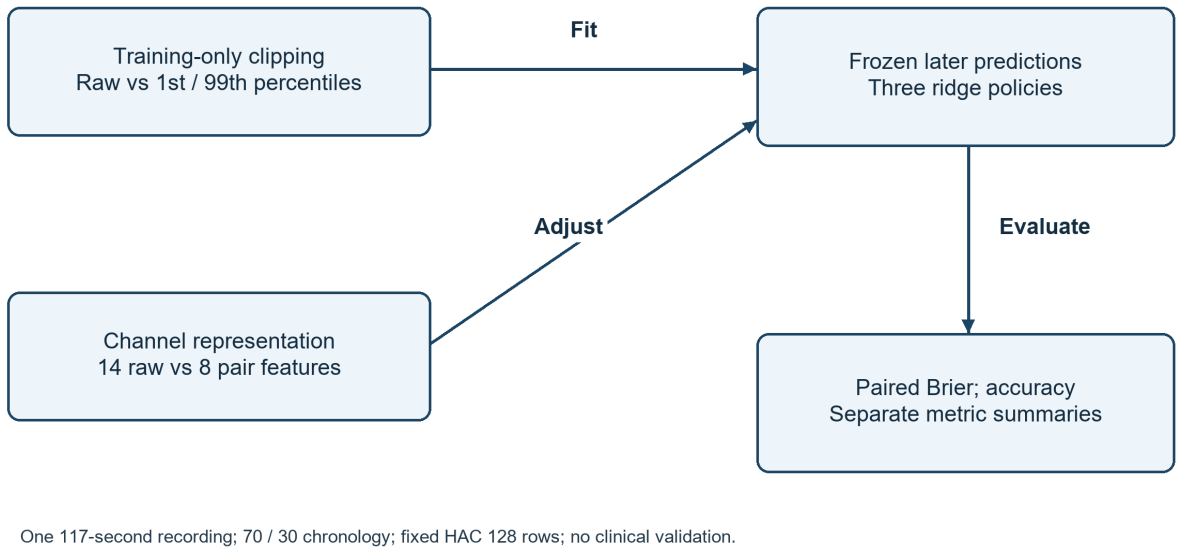


Figure 1. Training-only preprocessing produces three frozen ridge policies. Their common later-period predictions feed paired Brier contrasts and separate accuracy summaries, with no feedback from holdout labels to feature construction.

4. Chronological training and evaluation

The first 10,486 rows, exactly seventy percent of the file, form training data. The remaining 4,494 rows form the test segment. Row order is preserved throughout. Training includes 5,564 closed-eye labels, while the test segment includes 1,159. Thus closed-eye prevalence drops from about 53.1 percent to 25.8 percent. A random row split would obscure that temporal change and place neighboring observations from the same brief sequence in both partitions.

Structured-validation guidance in Roberts et al., 2017 and time-series validation discussion in Bergmeir et al., 2018 support making the chronological target explicit. Selection-bias analysis in Varma and Simon, 2006 further cautions against tuning on the test segment. Our split still does not create an independent subject or session. It tests later continuation of the same recording, where electrode placement, participant characteristics, and recording equipment are largely shared with training.

The first-and-ninety-ninth percentile clipping changes about 1.94 percent of training channel cells and 10.93 percent of test channel cells. That difference describes how often later values exceed training bounds. It does not establish why they do so. Eye state, artifacts, recording drift, or other temporal changes could contribute. The source does not specify physical channel units in the variable table, so we refer to released channel units rather than inventing microvolt scaling.

Robust-estimation ideas in Huber, 1964, Hampel, 1974, and Rousseeuw, 1984 motivate limiting extreme influence, but our clipping policy is not one of their fitted estimators. It is a fixed transformation followed by ridge regression. Clipping can suppress informative extremes as well as artifacts, and pair differences can remove common signal along with common noise. The experiment therefore evaluates a tradeoff rather than assuming that either transformation is inherently better.

5. Loss contrasts and temporal uncertainty

For every test row we compute clipped-probability squared error against the binary label. The primary contrasts subtract the raw-policy Brier error from each alternative's error on that same row. Intercept-only regressions estimate their mean differences. HAC covariance uses a fixed maximum lag of 128 retained rows, a correction, and t-based intervals. The lag is reported in rows; we do not convert it into an exact physiological time interval from a guessed sampling rate.

Newey and West, 1987 supplies the HAC covariance framework, while predictive-comparison work in Diebold and Mariano, 1995 and West, 1996 motivates pairing losses from common observations. These methods do not create independent recording sessions. A fixed lag can leave longer dependence unaccounted for, and only one brief test segment is available. The resulting intervals are conditional, exploratory summaries of this sequence under the stated temporal-covariance assumption.

Accuracy is the proportion of thresholded predictions matching the label. We additionally report sensitivity for closed eyes, specificity for open eyes, and their average as balanced accuracy. The imbalance concerns in Saito and Rehmsmeier, 2015, ROC discussion in Fawcett, 2006, and binary-prediction summary in Matthews, 1975 motivate not relying on raw accuracy alone. We do not select a threshold by an ROC curve or calculate a new best operating point after inspecting test labels.

Table 1. Alternative-minus-raw Brier differences on 4,494 common test rows. Negative values favor the alternative; uncertainty uses HAC 128 retained rows. Mean-only regressions have zero R-squared.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean contrast	-0.0340 (0.0220)	[-0.0771, 0.0090]	4494
M1: fit	R-squared = 0.0000	-	-	4494
M2	Mean contrast	-0.0489 (0.0246)	[-0.0970, -0.0007]	4494
M2: fit	R-squared = 0.0000	-	-	4494

Model labels: M1 = Clipped raw channels minus raw Brier; M2 = Clipped pair contrasts minus raw Brier.

6. Metric improvements with poor absolute behavior

Raw-channel Brier error is 0.3614. Clipped channels reduce it to 0.3274, a difference of -0.0340 with standard error 0.0220 and interval -0.0771 to 0.0090. The clipped pair-contrast policy has error 0.3125, a difference of -0.0489 with standard error 0.0246 and interval -0.0970 to -0.0007. The first contrast remains uncertain, while the second only narrowly excludes zero under the declared HAC calculation.

Thresholded accuracies are 0.3333 for raw channels, 0.3131 for clipping, and 0.3680 for pair contrasts. These are poor absolute values. Raw predictions have sensitivity 0.9802 but specificity only 0.1085, indicating a strong tendency toward the closed-eye class. Clipping produces sensitivity 0.8214 and specificity 0.1364. Pair contrasts yield sensitivity 0.8352 and specificity 0.2057. Their respective balanced accuracies are 0.5444, 0.4789, and 0.5204.

The pair representation therefore improves Brier error and ordinary accuracy relative to raw channels, but lowers balanced accuracy slightly. Raw clipping improves Brier error while lowering both ordinary and balanced accuracy. These combinations are possible because squared probability error and a fixed threshold emphasize different features of the prediction distribution. Reporting only the most favorable metric would give an incomplete account of what the policy actually does.

About 7.19 percent of raw-policy test scores require clipping into the probability range, compared with 0.13 percent for clipped channels and 0.18 percent for pair contrasts. These counts partly explain why probability-sensitive loss can improve without a strong decision boundary. They do not establish full calibration. A separate calibration curve, uncertainty analysis, and independent evaluation would be needed before interpreting any score as a trustworthy probability for operational use.

7. A baseline check on apparent progress

The test set contains many more open-eye rows than closed-eye rows. A rule that always predicts open eyes would achieve about 74.2 percent ordinary accuracy on this segment, well above all three fitted policies. We present that prevalence-based comparison descriptively, not as a model selected using test data for deployment. It makes the limited absolute performance visible and prevents a small relative Brier gain from being mistaken for a useful eye-state classifier.

Influence diagnostics in Cook, 1977 could identify particular extreme rows, but removing them after observing their prediction errors would alter the temporal test. Missing-data principles in Little and Rubin, 2002 also distinguish a missing observation from an unusual observed amplitude. We retain the complete source sequence and apply the declared training-only policies. The source's brief duration and changing class mix remain limitations that no local cleaning operation automatically resolves.

One useful extension would collect multiple sessions and participants, lock preprocessing and model choices on training sessions, and evaluate unseen sessions with a declared class-sensitive objective. Another would inspect artifacts using expert labels rather than inferring faults from amplitude alone. Either design could test whether pair contrasts preserve transferable physiological information. The present single-recording result cannot support a diagnostic claim, clinical monitoring recommendation, or general conclusion about robust graph estimators.

8. What remains reproducible

The class-balance change also shows why a frozen one-half threshold can be fragile. A model trained where closed eyes are slightly more common may assign many later rows to that class even when open eyes dominate the continuation. Adjusting the threshold using test labels would improve a retrospective number at the cost of contaminating evaluation. A prospective adaptation strategy would need a separate calibration segment or a justified unlabeled shift procedure, followed by another untouched test. No such strategy is implemented in this experiment.

Pair subtraction has an interpretable algebraic tradeoff. A component common to both members of a pair cancels, which may suppress shared disturbances but can also discard useful common activity. Adding the across-channel median retains one measure of common level, yet cannot recover every pattern removed by seven differences. The eight-feature representation is therefore a deliberate compression with limited information, not a general proof that fewer channels preserve physiological reasoning. Its behavior depends on the reader, labels, and recording conditions.

The narrow lower bound of the pair-contrast Brier interval deserves caution. A different dependence horizon or a different recording could change the interval, and the present study does not survey those choices to find the most favorable one. The fixed lag keeps the reported analysis reproducible. It does not make the temporal covariance model uniquely correct. The poor absolute classification and single-session scope remain material even when one loss contrast crosses a conventional significance boundary.

The appropriate reporting unit is therefore the full policy description: preprocessing, training split, reader, score transformation, threshold, and metric. Detaching the negative Brier difference from those choices would turn a conditional result into an unsupported claim of robust EEG classification. Preserving the row-level predictions allows later researchers to investigate alternative questions transparently, but any such extension should identify its new selection decisions and avoid presenting repeated analyses of this one sequence as independent participant evidence.

Documentation follows Gebru et al., 2021 by recording the one-recording scope, label convention, chronological boundary, pair list, training-only bounds, ridge penalty, and scoring transformations. Wilkinson et al., 2016 motivates retaining public source provenance and processed row-level scores with the numerical outputs. The study uses existing public measurements and claims no new participant recruitment, ethics approval, or medical intervention. Human scientific review is still required before any publication decision.

The result demonstrates a specific reporting hazard: lower probability-sensitive error can coexist with weak thresholded classification and no improvement in balanced accuracy. A compact representation offers a local Brier advantage under one covariance choice, but that is a limited finding within a poor-performing baseline system. The honest next step is broader validation and better task-specific modeling, rather than presenting preprocessing alone as evidence of physiological reasoning or clinical reliability.

References

1. 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.

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

3. Saito, Takaya and Marc Rehmsmeier. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.” *PLOS ONE*, vol. 10, no. 3, 2015, pp. e0118432. https://doi.org/10.1371/journal.pone.0118432.

4. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.

5. Matthews, B.W.. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA) - Protein Structure. https://doi.org/10.1016/0005-2795(75)90109-9.

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

7. Welch, Peter D. "The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging over Short, Modified Periodograms." *IEEE Transactions on Audio and Electroacoustics*, vol. 15, no. 2, 1967, pp. 70-73. https://doi.org/10.1109/TAU.1967.1161901.

8. Diebold, Francis X.; Mariano, Roberto S.. (1995). Comparing Predictive Accuracy. Journal of Business & Economic Statistics. https://doi.org/10.1080/07350015.1995.10524599.

9. 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.

10. 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.

11. Bergmeir, Christoph; Hyndman, Rob J.; Koo, Bonsoo. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis. https://doi.org/10.1016/j.csda.2017.11.003.

12. Moody, George, and Roger Mark. “MIT-BIH Arrhythmia Database.” Version 1.0.0, *PhysioNet*, 2005. https://doi.org/10.13026/C2F305.

13. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.

14. Cini, Andrea; Marisca, Ivan; Alippi, Cesare. (2021). Filling the G_ap_s: Multivariate Time Series Imputation by Graph Neural Networks. arXiv:2108.00298. https://arxiv.org/abs/2108.00298.

15. Shuman, David I; Narang, Sunil K.; Frossard, Pascal; Ortega, Antonio; Vandergheynst, Pierre. (2012). The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains. arXiv:1211.0053. https://arxiv.org/abs/1211.0053.

16. West, Kenneth D.. (1996). Asymptotic Inference about Predictive Ability. Econometrica. https://doi.org/10.2307/2171956.

17. Peng, Changran; Yan, Yi; Kuruoglu, Ercan E.. (2024). Adaptive Graph Normalized Sign Algorithm. arXiv:2405.04107. https://arxiv.org/abs/2405.04107.

18. Lundberg, Scott M., and Su-In Lee. “A Unified Approach to Interpreting Model Predictions.” *Advances in Neural Information Processing Systems*, vol. 30, 2017. https://proceedings.neurips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html.

19. Buuren, Stef van; Groothuis-Oudshoorn, Karin. (2011). mice: Multivariate Imputation by Chained Equations inR. Journal of Statistical Software. https://doi.org/10.18637/jss.v045.i03.

20. Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

21. Stekhoven, Daniel J.; Bühlmann, Peter. (2012). MissForest-non-parametric missing value imputation for mixed-type data. Bioinformatics. https://doi.org/10.1093/bioinformatics/btr597.

22. Cao, Wei; Wang, Dong; Li, Jian; Zhou, Hao; Li, Lei; Li, Yitan. (2018). BRITS: Bidirectional Recurrent Imputation for Time Series. arXiv:1805.10572. https://arxiv.org/abs/1805.10572.

23. Newey, Whitney K., and Kenneth D. West. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.” *Econometrica*, vol. 55, no. 3, 1987, pp. 703-708. https://doi.org/10.2307/1913610.

24. Rousseeuw, Peter J.. (1984). Least Median of Squares Regression. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1984.10477105.

25. Ortega, Antonio; Frossard, Pascal; Kovačević, Jelena; Moura, José M. F.; Vandergheynst, Pierre. (2017). Graph Signal Processing: Overview, Challenges and Applications. arXiv:1712.00468. https://arxiv.org/abs/1712.00468.

26. Hampel, Frank R.. (1974). The Influence Curve and its Role in Robust Estimation. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1974.10482962.

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

28. Sameni, Reza, and Gari D. Clifford. "A Review of Fetal ECG Signal Processing; Issues and Promising Directions." *The Open Pacing, Electrophysiology & Therapy Journal*, vol. 3, 2010, pp. 4-20. https://doi.org/10.2174/1876536X01003010004.

29. Huber, Peter J.. (1964). Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics. https://doi.org/10.1214/aoms/1177703732.

30. Roesler, Oliver. (2013). EEG Eye State [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C57G7J.