

DConfusion

An R Package for Cross-Study Evaluation of Binary Classifiers

Aadya Pawar (pawar17@purdue.edu) | Surya Gundavarapu (suryagsrnlr32@gmail.com) | Matthew A. Lanham (mlanham1@butler.edu)



OVERVIEW

Evaluating and comparing binary classifiers across studies remains a persistent challenge, as authors frequently report inconsistent subsets of performance metrics, making direct cross-study comparison difficult or impossible. DConfusion is a fully documented, open-source R package developed from scratch, implementing the Bowes et al. (2014) framework with extensions including bootstrap inference, cost-sensitive analysis, and consistency checking not available in any existing R tool.

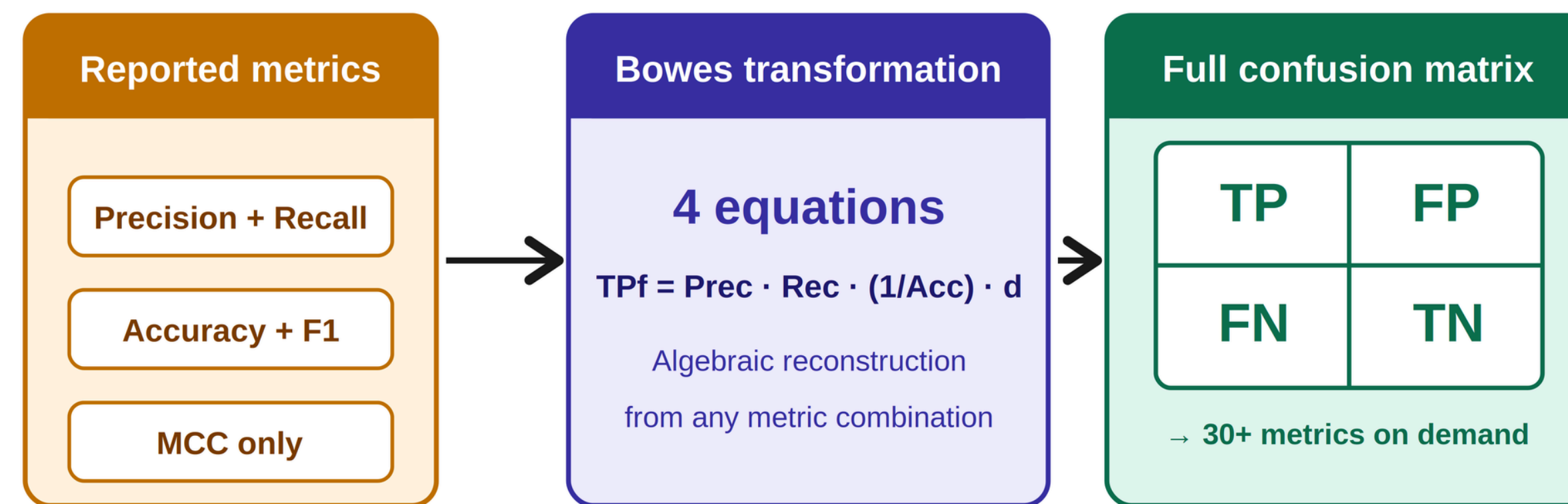
Confusion Matrix

n = 500 | Accuracy = 92.6% | F1 = 0.84

		Predicted Label	
		Positive	Negative
Actual Label	Negative	False Positive (FP) 22	True Negative (TN) 378
	Positive	True Positive (TP) 85	False Negative (FN) 15

Bowes, Hall & Petrić (2014) proved that a full confusion matrix can often be reconstructed from reported metric combinations — enabling legitimate cross-study aggregation. DConfusion makes this method accessible to R users.

Bowes et al. (2014) Framework

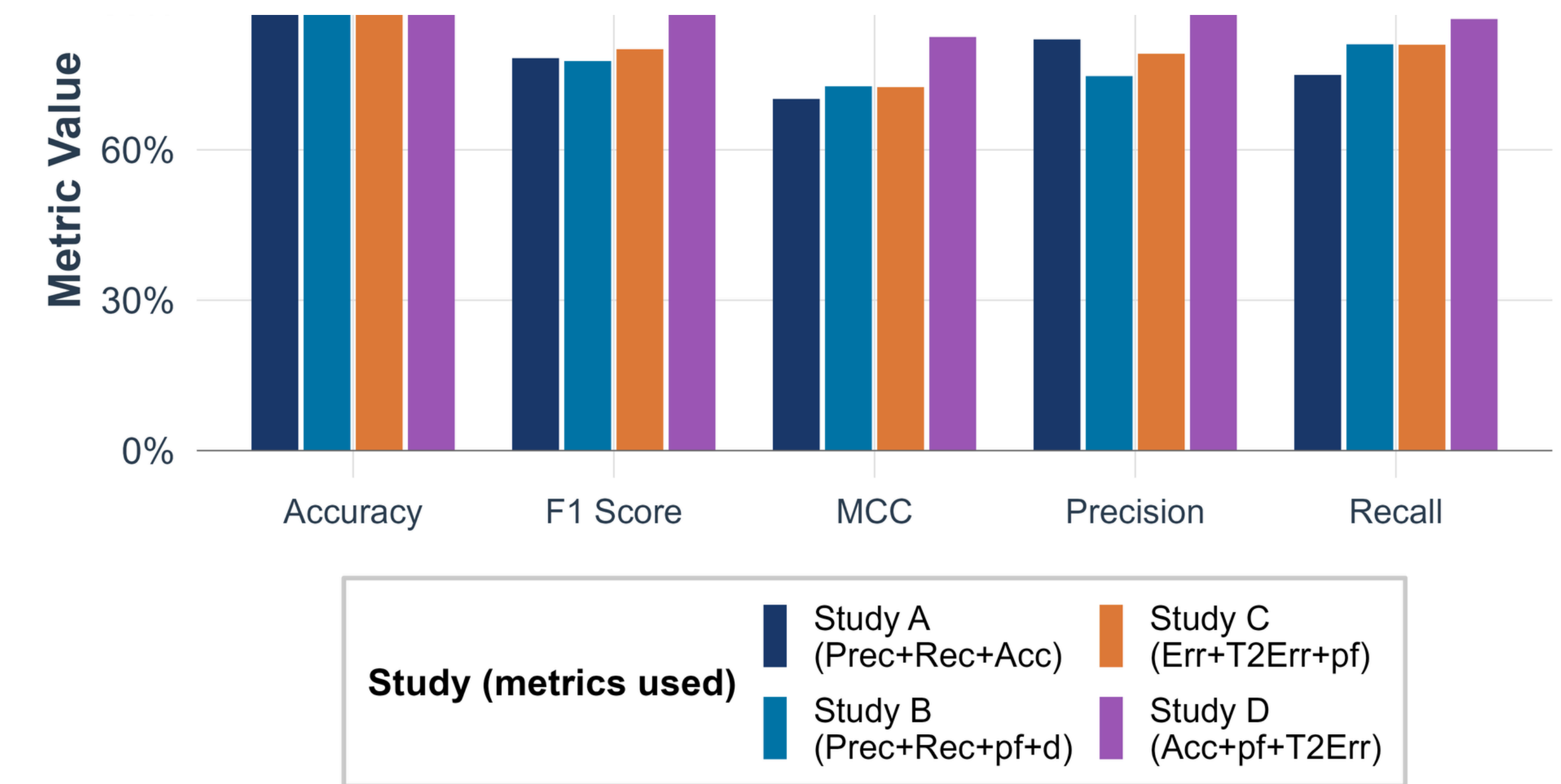


Matrix Reconstruction

- Input:** Raw counts (TP/FP/FN/TN), prediction vectors, or any subset of reported metrics
- Solve:** Apply Bowes et al. algebraic transformation to infer missing cell counts
- Validate:** Consistency check — verify reported scores are mathematically coherent
- Compute:** Derive full metric suite from reconstructed matrix
- Infer:** Bootstrap CIs, McNemar's test, cost-sensitive analysis

From this single object, any of 30+ performance metrics can be derived on demand. The package warns automatically on small sample sizes ($N < 100$) and severe class imbalance ($IR > 10$).

Cross-Study Comparison via DConfusion Transformation



The Cross-Study Comparison Problem

Empirical software engineering and ML research suffer from a widespread reproducibility gap: authors report inconsistent subsets of classification metrics, making direct comparison across studies difficult or impossible.

Study A reports Precision + Recall
Study B reports Accuracy + F1
Study C reports MCC only

Without a full confusion matrix, how do you compare them?

What DConfusion Provides

- Matrix construction** from raw counts, vectors, or metric sets via unified API
- Metric suite:** Accuracy, Precision, Recall, F1, MCC, Balanced Accuracy, and 25+ more
- Consistency checking:** audits whether a paper's reported scores are internally coherent
- Statistical inference:** bootstrap confidence intervals and McNemar's test for significance
- Cost-sensitive analysis:** identifies optimal metric under asymmetric FP / FN costs
- Automated warnings** for small N and class imbalance before any metric is reported
- Visualization:** Confusion matrix plots, ROC & PR curves

Novel contributions vs. existing R tools

- Bootstrap CIs
- Consistency Audit
- Cost Sensitivity
- Imbalance Warnings
- McNemar Test
- Matrix Reconstruction

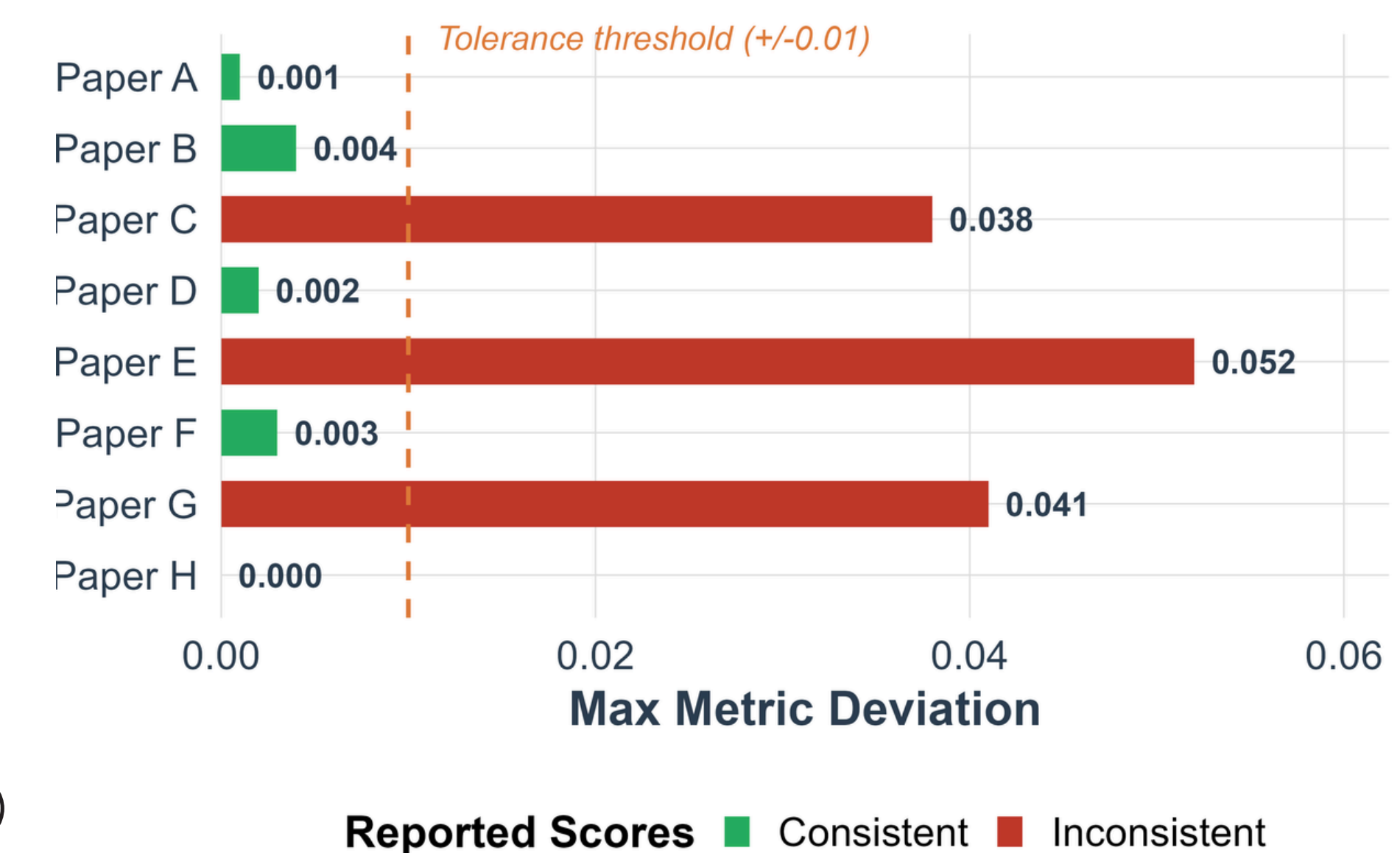
Use Cases

- Systematic literature reviewers
- Empirical researchers
- Peer reviewers / auditors
- Meta-analysts
- Practitioners

Evidence for need

- Shepperd et al. (2019)
- Mahmood et al. (2018)
- Fazekas (2024)
- Chicco et al. (2021–23)
- Lovell et al. (2023)

Consistency Validation of Published Results



CONCLUSION

- First production R implementation of Bowes et al. (2014)
- Enables reproducible cross-study classifier benchmarking
- Lowens barrier to transparent reporting in fault prediction research