The Journal of Open Source Software

HERA: A Hierarchical-Compensatory, Effect-Size Driven and Non-parametric Ranking Algorithm using Data-Driven Thresholds and Bootstrap Validation

Lukas von Erdmannsdorff [1]

1 Institute of Neuroradiology, Goethe University Frankfurt

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

DRAFT

Summary

In scientific disciplines ranging from clinical research to machine learning, researchers face the challenge of objectively comparing multiple algorithms, experimental conditions, or datasets across a variety of performance metrics. This process, often framed as Multi-Criteria Decision Making (MCDM), is critical for identifying state-of-the-art methods. However, traditional ranking approaches frequently suffer from limitations: they may rely on central tendencies that ignore data variability (Benavoli et al., 2016; Demšar, 2006), depend solely on p-values which can be misleading in large samples (Wasserstein & Lazar, 2016), or require subjective weighting of conflicting metrics (Taherdoost & Madanchian, 2023).

HERA (Hierarchical-Compensatory, Effect-Size Driven Ranking Algorithm) is a MATLAB toolbox designed to automate this comparison process, bridging the gap between elementary statistical tests and complex decision-making frameworks. Unlike weighted-sum approaches that collapse multi-dimensional performance into a single scalar, HERA implements a **hierarchical-compensatory logic**. This logic integrates non-parametric significance testing (Wilcoxon signed-rank test), robust effect size estimation (Cliff's Delta, Relative Difference), and bootstrapping (e.g. Percentile and Cluster) to produce rankings that are both statistically robust and practically relevant. HERA is designed for researchers in biomedical imaging, machine learning, and applied statistics who need to compare method performance across multiple quality metrics in a statistically rigorous manner without requiring subjective parameter tuning.

Statement of Need

The scientific community increasingly recognizes the pitfalls of relying on simple summary statistics or p-values alone (Wasserstein & Lazar, 2016). In benchmarking studies, specifically, several issues persist:

1. **Ignoring Variance**: Ranking based on mean scores fails to account for the stability of performance across different subjects or folds. A method might achieve a high average score due to exceptional performance on a few easy cases while failing catastrophically on others, yet still outrank a more consistent competitor.
2. **Statistical vs. Practical Significance**: A result can be statistically significant but practically irrelevant, especially in large datasets where even trivial differences yield $p < 0.05$. Standard tests do not inherently distinguish between these cases, potentially leading to the adoption of methods that offer no tangible benefit (Sullivan & Feinn, 2012).
3. **Subjectivity in Aggregation**: Many MCDM methods require users to assign arbitrary weights to metrics (e.g., "Accuracy is 0.7, Speed is 0.3"). These weights are often chosen post-hoc or lack empirical justification, introducing researcher bias that can be manipulated to favor a specific outcome (Taherdoost & Madanchian, 2023).
4. **Distributional Assumptions**: Parametric tests (e.g., t-test) assume normality, which is often violated in real-world benchmarks where performance metrics may be skewed,

The Journal of Open Source Software

bounded, or ordinal (Romano et al., 2006).

HERA addresses these challenges by providing a standardized, data-driven framework. It ensures that a method is only ranked higher if it demonstrates a statistically significant and sufficiently large advantage, preventing "wins" based on negligible differences or noise. Unlike existing MCDM software packages such as the Python libraries pyDecision (Pereira et al., 2024) and pymcdm (Kizielewicz et al., 2023), or R's RMCDA (Najafi & Mirzaei, 2025), which often implement classical methods like TOPSIS (Hwang & Yoon, 1981), PROMETHEE (Brans & Vincke, 1985), and ELECTRE (Roy, 1968) that require user-defined weights or preference functions, HERA eliminates subjective parameterization by using data-driven thresholds derived from bootstrap resampling. Furthermore, HERA integrates statistical hypothesis testing directly into the ranking process, a feature absent in standard MCDM toolboxes. While the MATLAB ecosystem offers robust statistical functions, it currently lacks a dedicated, open-source toolbox that unifies this advanced MCDM method with bootstrap validation, forcing researchers to rely on ad-hoc scripts.

Methodological Framework

HERA operates on paired data matrices where rows represent subjects (or datasets) and columns represent the methods to be compared. The core innovation is its sequential logic, which allows for "compensation" between metrics based on strict statistical evidence.

Statistical Rigor and Effect Sizes

HERA quantifies differences using statistical significance and effect sizes to ensure practical relevance independent of sample size (Cohen, 1988; Sullivan & Feinn, 2012). A "win" always requires satisfying three conjunctive criteria, if not it is considered "neutral":

- **Significance**: $p < \alpha_{\text{Holm}}$ (Holm-Bonferroni corrected). Pairwise comparisons use the Wilcoxon signed-rank test (Wilcoxon, 1945), with p-values corrected using the step-down Holm-Bonferroni method (Holm, 1979) to control the Family-Wise Error Rate (FWER).
- **Stochastic Dominance (Cliff's Delta)**: Cliff's Delta ($d = P(X > Y) - P(Y > X)$) quantifies distribution overlap, is robust to outliers, and relates to common-language effect sizes (Cliff, 1993; Vargha & Delaney, 2000). The effect size d must exceed a bootstrapped threshold θ_d.
- **Magnitude (Relative Difference)**: The Relative Difference (RelDiff) quantifies effect magnitude on the original metric scale, normalized by the mean absolute value. This normalization is formally identical to the Symmetric Mean Absolute Percentage Error (SMAPE) used in forecasting (Makridakis, 1993) and conceptually related to the Response Ratio, which uses logarithmic ratios to compare effects across studies (Hedges et al., 1999). The metric enables scale-independent comparisons and facilitates the interpretation of percentage changes (Kampenes et al., 2007). RelDiff must exceed a threshold δ_{RelDiff}.

Dual Criteria & SEM Lower Bound HERA's complementary logic requires both dominance and magnitude, preventing "wins" based on trivial consistent differences or noisy outliers (Lakens, 2013). Thresholds are determined via Percentile Bootstrapping (lower $\alpha/2$-quantile) (Rousselet et al., 2021). To filter noise in low-variance datasets, the RelDiff threshold enforces a lower bound based on the Standard Error of the Mean (SEM), ensuring $\theta_r \geq \theta_{\text{SEM}}$. This approach is inspired by the concept of the "Smallest Worthwhile Change" (Hopkins, 2004), but adapted for HERA to quantify the uncertainty of the group mean rather than individual measurement error.

Hierarchical-Compensatory Logic

The ranking process is structured as a multi-stage tournament. It does not use a global score but refines the rank order iteratively (see Fig. 1):

DRAFT

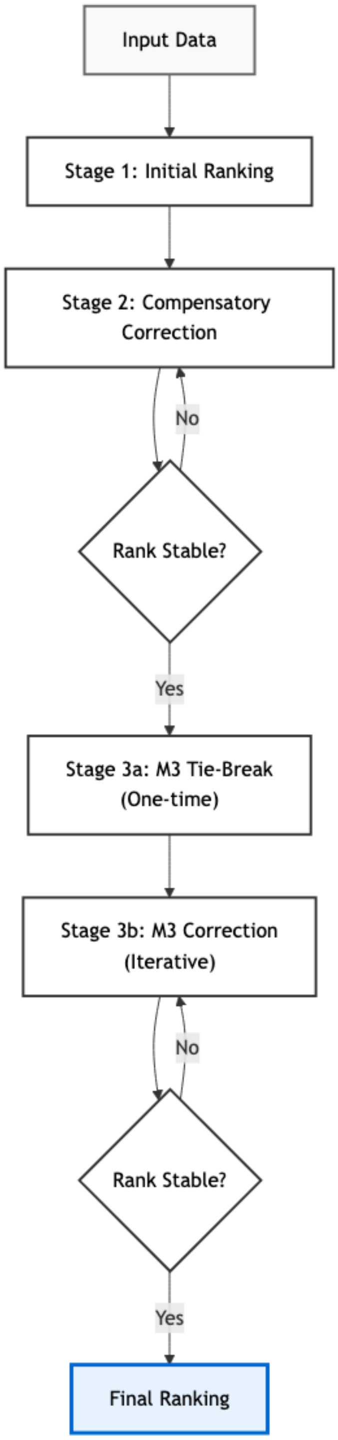

Figure 1: Hierarchical-Compensatory Ranking Logic

- **Stage 1 (Initial Sort)**: Methods are initially ranked based on the win count of the primary metric M_1. In case of a tie, Cliff's Delta is used to break the tie.
- **Stage 2 (Compensatory Correction)**: This stage addresses the trade-off between metrics. A lower-ranked method can "swap" places with a higher-ranked method if it shows a statistically significant and relevant superiority in a secondary metric M_2. This effectively

implements a lexicographic ordering with a compensatory component (Keeney & Raiffa, 1976), allowing a method that is slightly worse in the primary metric but vastly superior in a secondary metric to improve its standing.

- **Stage 3 (Tie-Breaking)**: This stage resolves "neutral" results using a tertiary metric M_3. It applies two sub-logics to ensure a total ordering:
  - **Sublogic 3a**: A one-time correction if the previous metric is "neutral" based on the HERA criteria. This handles cases where two methods are indistinguishable in the second metric while still respecting the initial ranking.
  - **Sublogic 3b**: To resolve groups of remaining undecided methods, an iterative correction loop is applied if both M_1 and M_2 are "neutral", iteratively using metric M_3 until a final stable ranking is found.

Validation and Uncertainty

HERA integrates advanced resampling methods to quantify uncertainty:

- **BCa Confidence Intervals**: Bias-Corrected and Accelerated (BCa) intervals are calculated for all effect sizes (DiCiccio & Efron, 1996).
- **Cluster Bootstrap**: To assess the stability of the final ranking, HERA performs a cluster bootstrap resampling subjects with replacement (Field & Welsh, 2007). This yields a 95% confidence interval for the rank of each method.
- **Power Analysis**: A post-hoc simulation with bootstrap estimates the probability of detecting a "win", "loss" or "neutral" in all tested metrics given the data characteristics.
- **Sensitivity Analysis**: The algorithm permutes the metric hierarchy and aggregates the resulting rankings using a Borda Count (Young, 1974) to evaluate the robustness of the decision against hierarchy changes.

Software Features

HERA offers a flexible configuration of up to three metrics (see Fig. 2). This allows users to adapt the ranking logic to different study designs and needs. It also provides a range of reporting options, data integration, and reproducibility features.

- **Automated Reporting**: Generates PDF reports, Win-Loss Matrices, Sankey Diagrams, and machine-readable JSON/CSV exports.
- **Reproducibility**: Supports fixed-seed execution and configuration file-based workflows. The full analysis state, including random seeds and parameter settings, is saved in a JSON file, allowing other researchers to exactly replicate the ranking results.
- **Convergence Analysis**: To avoid the common pitfall of using an arbitrary number of bootstrap iterations, HERA implements an adaptive algorithm. It automatically monitors the stability of the estimated confidence intervals and effect size thresholds, continuing the resampling process until the estimates converge within a specified tolerance, thus determining the optimal number of iterations B dynamically (Pattengale et al., 2010). If the characteristics of the data for bootstrapping are known, the number of bootstrap iterations can be set manually.
- **Data Integration**: HERA supports seamless data import from standard formats (CSV, Excel) and MATLAB tables, facilitating integration into existing research pipelines. Example datasets and workflows demonstrating practical applications are included in the repository.
- **Accessibility**: HERA can be easily installed by cloning the GitHub repository and running a setup script, or deployed as a standalone application that requires no MATLAB license. An interactive command-line interface guides users through the analysis without requiring programming expertise, while an API and JSON Configuration allow for automated batch processing.

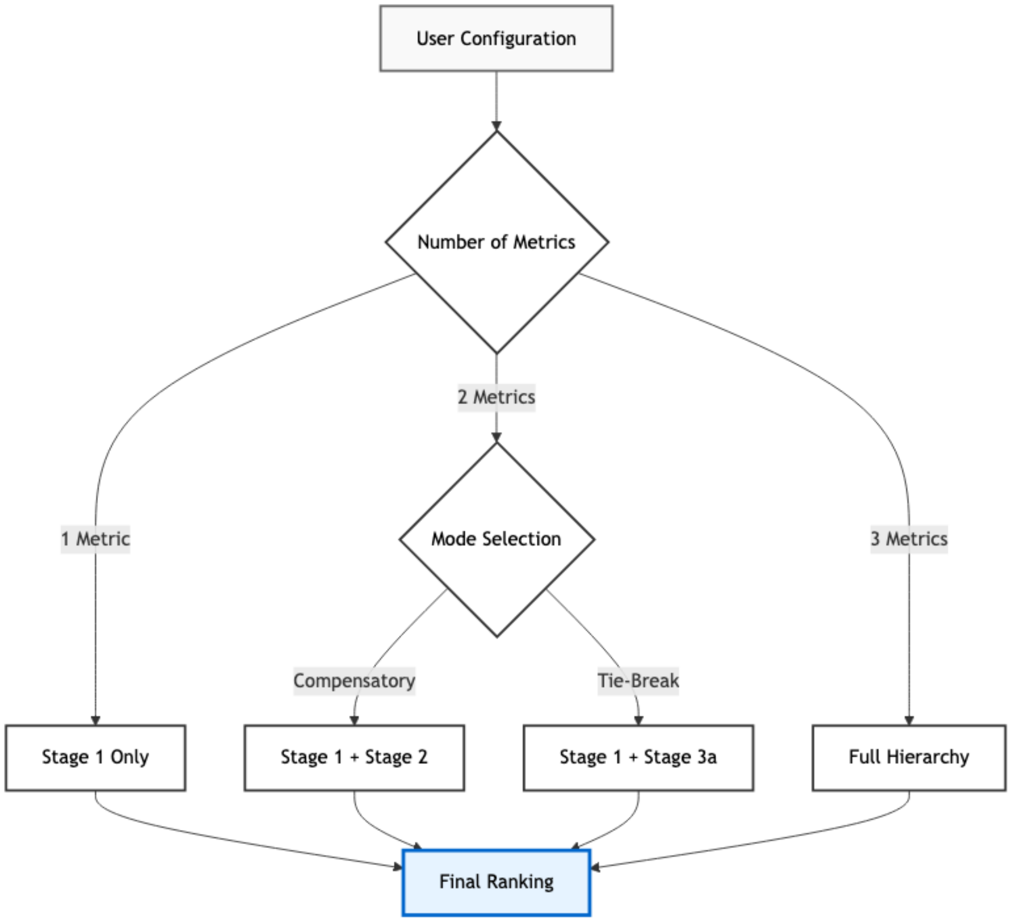

Figure 2: Flexible Configuration options for Ranking Logic

Acknowledgements

This software was developed at the Institute of Neuroradiology, Goethe University Frankfurt. I thank Prof. Dr. Dipl.-Phys. Ralf Deichmann (Cooperative Brain Imaging Center, Goethe University Frankfurt) for his support during the initial conceptualization of this project. I acknowledge Dr. med. Christophe Arendt (Institute of Neuroradiology, Goethe University Frankfurt) for his supervision and support throughout the project. I also thank Rejane Golbach PhD (Institute of Biostatistics and Mathematical Modeling, Goethe University Frankfurt) for her valuable feedback on the statistical methodology.

References

Benavoli, A., Corani, G., & Mangili, F. (2016). Should we really use post-hoc tests based on mean-ranks? *Journal of Machine Learning Research*, *17*, 1–10. https://jmlr.org/papers/v17/benavoli16a.html

Brans, J. P., & Vincke, P. (1985). A preference ranking organization method (the PROMETHEE method for multiple criteria decision-making). *Management Science*, *31*(6), 647–656. https://doi.org/10.1287/mnsc.31.6.647

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, *114*(3), 494–509. https://doi.org/10.1037/0033-2909.114.3.494

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence

Erlbaum Associates. https://doi.org/10.4324/9780203771587

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30. https://jmlr.org/papers/v7/demsar06a.html

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, *11*(3), 189–228. https://doi.org/10.1214/ss/1032280214

Field, C. A., & Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(3), 369–390. https://doi.org/10.1111/j.1467-9868.2007.00593.x

Hedges, L. V., Gurevitch, J., & Curtis, P. S. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology*, *80*(4), 1150–1156. https://doi.org/10.1890/0012-9658(1999)080%5B1150:TMAORR%5D2.0.CO;2

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. https://doi.org/10.2307/4615733

Hopkins, W. G. (2004). How to interpret changes in an athletic performance test. *Sportscience*, *8*, 1–7. https://www.sportsci.org/jour/04/wghtests.htm

Hwang, C. L., & Yoon, K. (1981). *Multiple attribute decision making: Methods and applications*. Springer. https://doi.org/10.1007/978-3-642-48318-9

Kampenes, V. B., Dybå, T., Hannay, J. E., & Sjøberg, D. I. K. (2007). A systematic review of effect size in software engineering experiments. *Information and Software Technology*, *49*(11–12), 1073–1086. https://doi.org/10.1016/j.infsof.2007.02.015

Keeney, R. L., & Raiffa, H. (1976). *Decisions with multiple objectives: Preferences and value trade-offs*. Wiley. https://doi.org/10.1017/CBO9781139174084

Kizielewicz, B., Shekhovtsov, A., & Salabun, W. (2023). Pymcdm—the universal library for solving multi-criteria decision-making problems. *SoftwareX*, *22*, 101368. https://doi.org/10.1016/j.softx.2023.101368

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Makridakis, S. (1993). Accuracy measures: Theoretical and practical concerns. *International Journal of Forecasting*, *9*(4), 527–529. https://doi.org/10.1016/0169-2070(93)90079-3

Najafi, A., & Mirzaei, S. (2025). RMCDA: The comprehensive r library for applying multi-criteria decision analysis methods. *Software Impacts*, *24*, 100762. https://doi.org/10.1016/j.simpa.2025.100762

Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R. P., Moret, B. M. E., & Stamatakis, A. (2010). How many bootstrap replicates are necessary? *Journal of Computational Biology*, *17*(3), 337–354. https://doi.org/10.1089/cmb.2009.0179

Pereira, V., Basilio, M. P., & Santos, C. H. T. (2024). Enhancing decision analysis with a large language model: pyDecision a comprehensive library of MCDA methods in python. *Journal of Modelling in Management*. https://doi.org/10.1108/JM2-04-2024-0118

Romano, J., Kromrey, J. D., Coraggio, J., & Skowronek, J. (2006). Appropriate statistics for ordinal level data: Should we really be using t-test and cohen's d for evaluating group differences on the NSSE and other surveys? *Proceedings of the Annual Meeting of the Florida Association of Institutional Research*. https://www.researchgate.net/publication/237544991

Rousselet, G. A., Pernet, C. R., & Wilcox, R. R. (2021). The percentile bootstrap: A primer with step-by-step instructions in R. *Advances in Methods and Practices in Psychological*

Science, *4*(1), 1–10. https://doi.org/10.1177/2515245920911881

Roy, B. (1968). Classement et choix en présence de points de vue multiples (la méthode ELECTRE). *Revue Française d'informatique Et de Recherche Opérationnelle*, *2*(V1), 57–75. https://doi.org/10.1051/ro/196802V100571

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Taherdoost, H., & Madanchian, M. (2023). Multi-criteria decision making (MCDM) methods and concepts. *Encyclopedia*, *3*(1), 235–250. https://doi.org/10.3390/encyclopedia3010006

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and wong. *Journal of Educational and Behavioral Statistics*, *25*(2), 101–132. https://doi.org/10.3102/10769986025002101

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968

Young, H. P. (1974). An axiomatization of borda's rule. *Journal of Economic Theory*, *9*(1), 43–52. https://doi.org/10.1016/0022-0531(74)90073-8

DRAFT