Benchmarking report for MOOD 2020 - Pixel-Level

created by challengeR v0.3.3
Wiesenfarth, Reinke, Landman, Cardoso, Maier-Hein & Kopp-Schneider (2019)

15 Oktober, 2020

This document presents a systematic report on a benchmark study. Input data comprises raw metric values for all algorithms and test cases. Generated plots are:

- Visualization of assessment data: Dot- and boxplots, podium plots and ranking heatmaps
- Visualization of ranking robustness: Line plots
- Visualization of ranking stability: Significance maps
- Visualization of cross-task insights

Ranking of algorithms within tasks according to the following chosen ranking scheme:

aggregate using function ("mean") then rank

Ranking list for each task:

brain : Analysis based on 55 test cases which included 0 missing values.

	value_FUN	rank
Algorithm_1	0.4491865	1
Algorithm_2	0.4161859	2
Algorithm_4	0.2732214	3
Algorithm_3	0.2110740	4
Algorithm_6	0.2039788	5
Algorithm_5	0.2014513	6
Algorithm_7	0.1595155	7
Algorithm_8	0.0209868	8

colon : Analysis based on 36 test cases which included 0 missing values.

	value_FUN	rank
Algorithm_1	0.3937558	1
Algorithm_2	0.2877517	2
Algorithm_5	0.2385403	3
Algorithm_3	0.2211918	4
Algorithm_4	0.2171832	5
Algorithm_7	0.0718899	6
Algorithm_6	0.0137616	7
Algorithm_8	0.0137616	7

Consensus ranking according to chosen method euclidean:

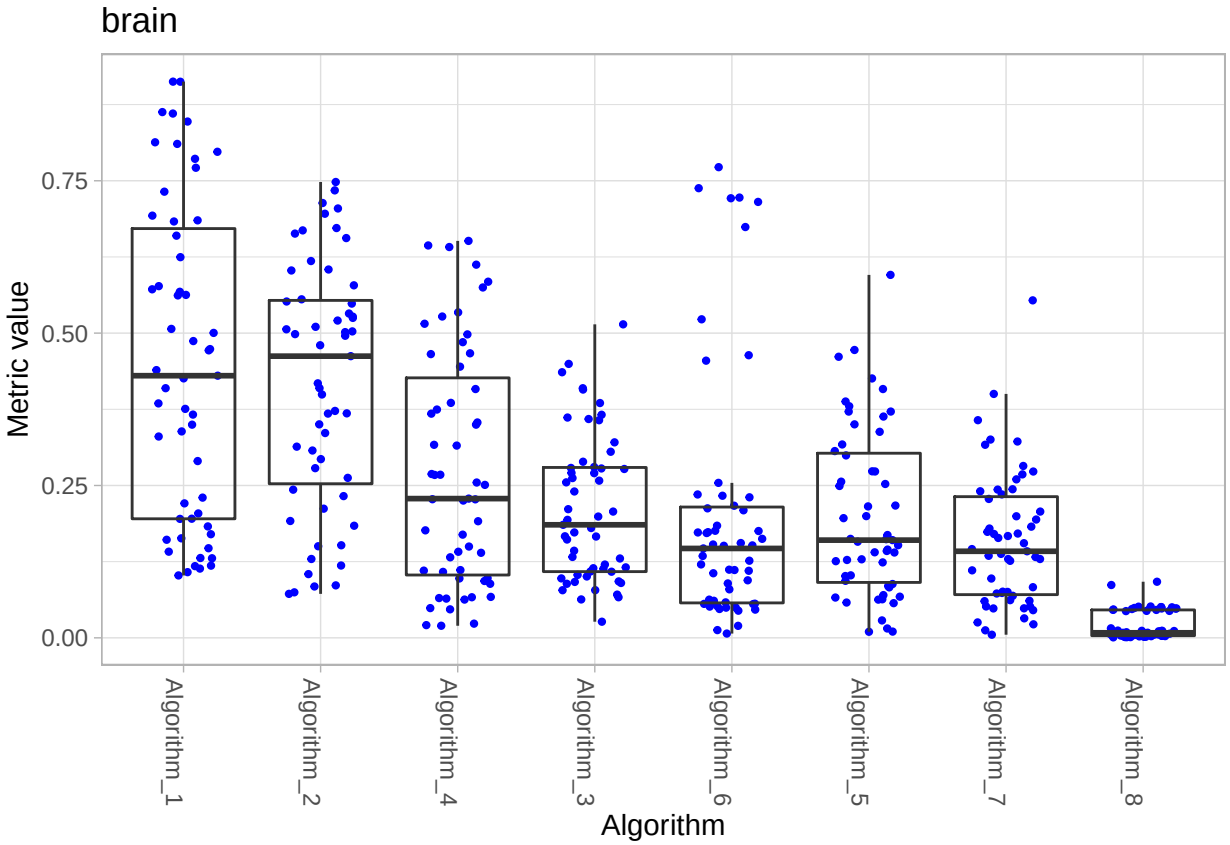
	value	rank
Algorithm_1	1.00	1
Algorithm_2	2.00	2
Algorithm_3	4.00	3
Algorithm_4	4.00	3
Algorithm_5	4.50	5
Algorithm_6	6.25	6
Algorithm_7	6.50	7
Algorithm_8	7.75	8

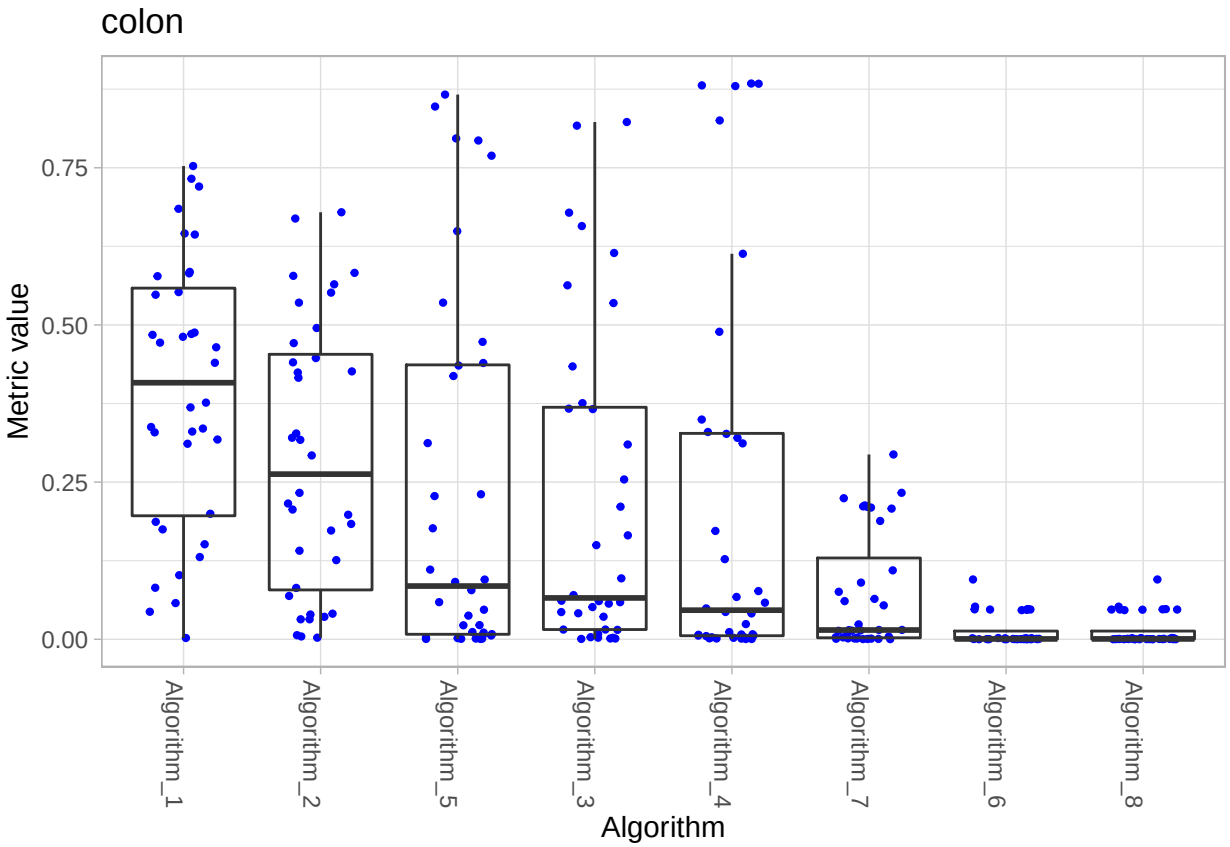
1 Visualization of raw assessment data

Algorithms are ordered according to chosen ranking scheme for each task.

1.1 Dot- and boxplots

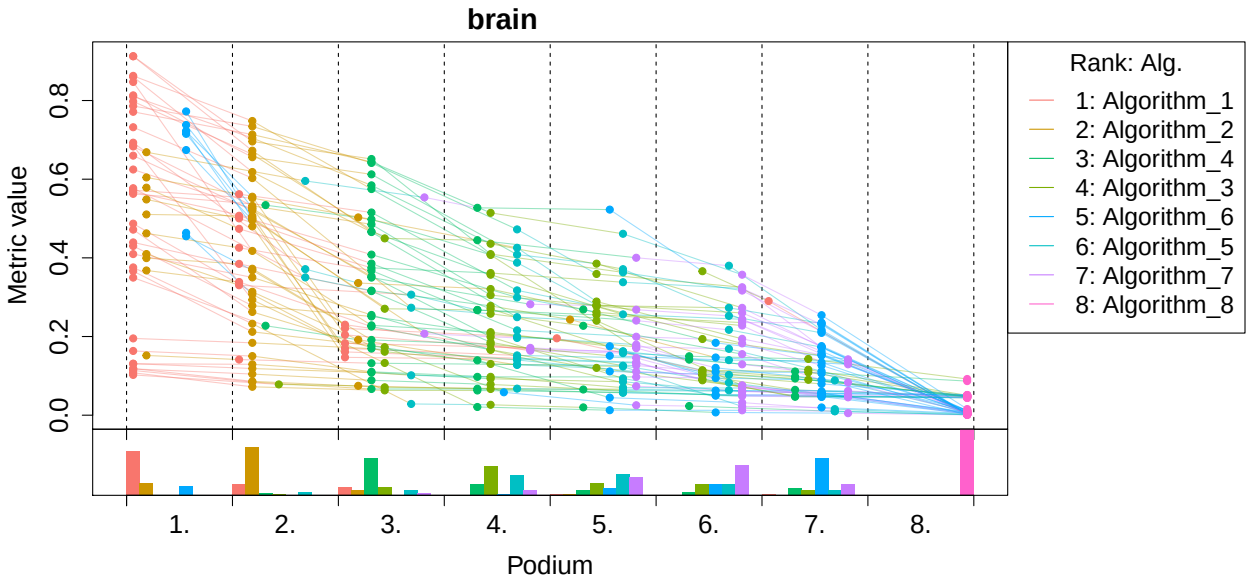
Dot- and boxplots for visualizing raw assessment data separately for each algorithm. Boxplots representing descriptive statistics over all test cases (median, quartiles and outliers) are combined with horizontally jittered dots representing individual test cases.

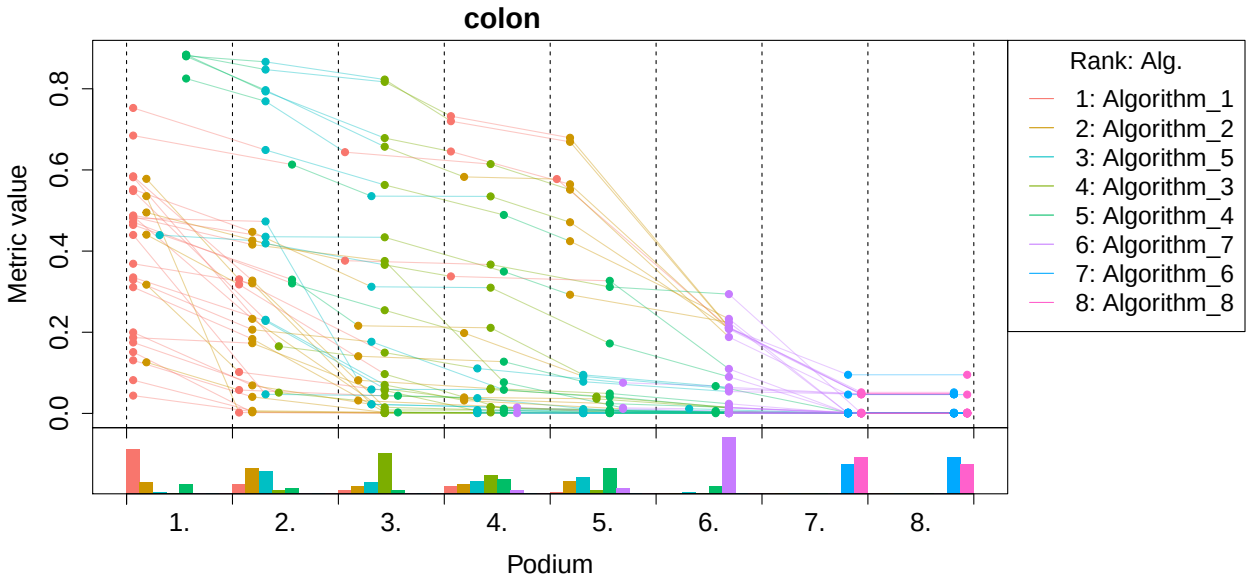




1.2 Podium plots

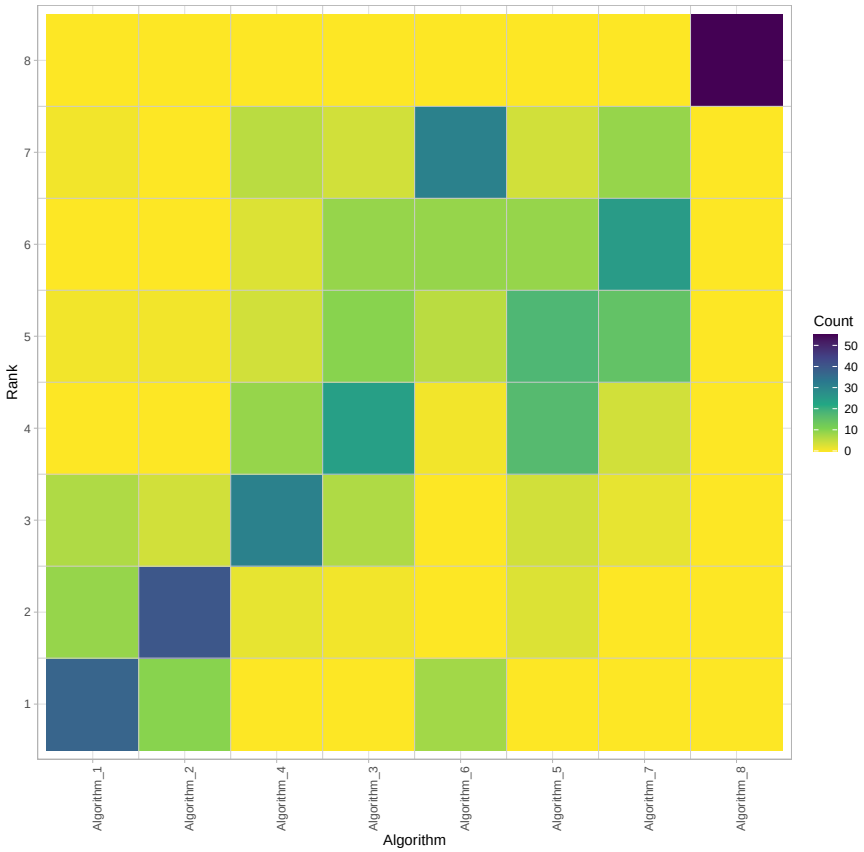
Podium plots (see also Eugster et al, 2008) for visualizing raw assessment data. Upper part (spaghetti plot): Participating algorithms are color-coded, and each colored dot in the plot represents a metric value achieved with the respective algorithm. The actual metric value is encoded by the y-axis. Each podium (here: p=8) represents one possible rank, ordered from best (1) to last (here: 8). The assignment of metric values (i.e. colored dots) to one of the podiums is based on the rank that the respective algorithm achieved on the corresponding test case. Note that the plot part above each podium place is further subdivided into p "columns", where each column represents one participating algorithm (here: $p = 8$). Dots corresponding to identical test cases are connected by a line, leading to the shown spaghetti structure. Lower part: Bar charts represent the relative frequency for each algorithm to achieve the rank encoded by the podium place.

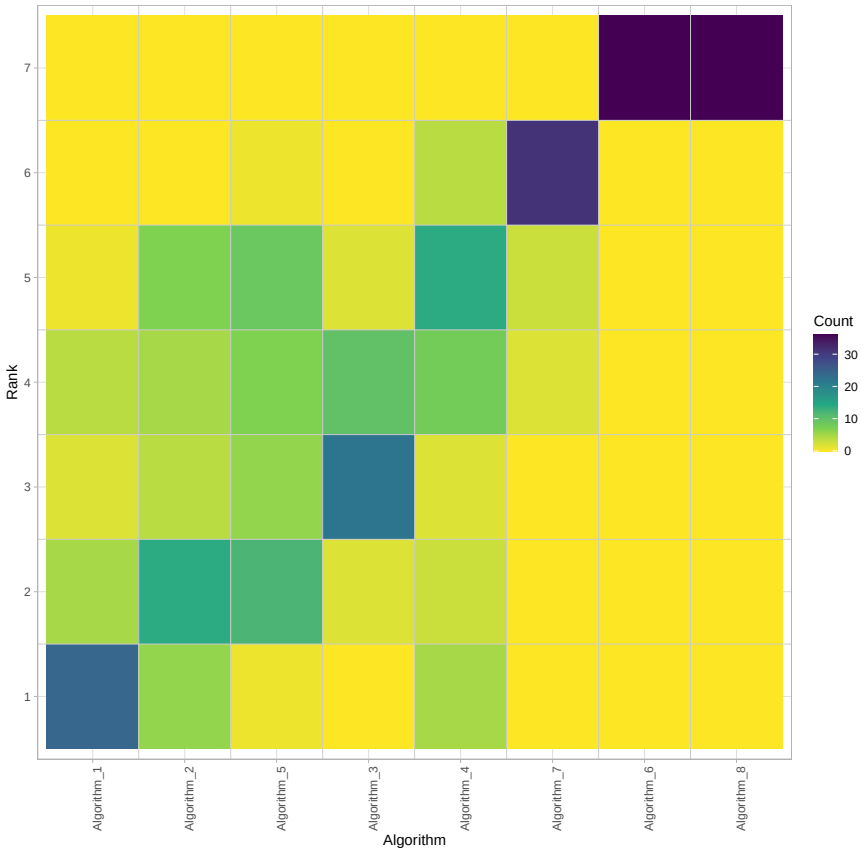




1.3 Ranking heatmaps

Ranking heatmaps for visualizing raw assessment data. Each cell (i, A_j) shows the absolute frequency of test cases in which algorithm A_j achieved rank i.

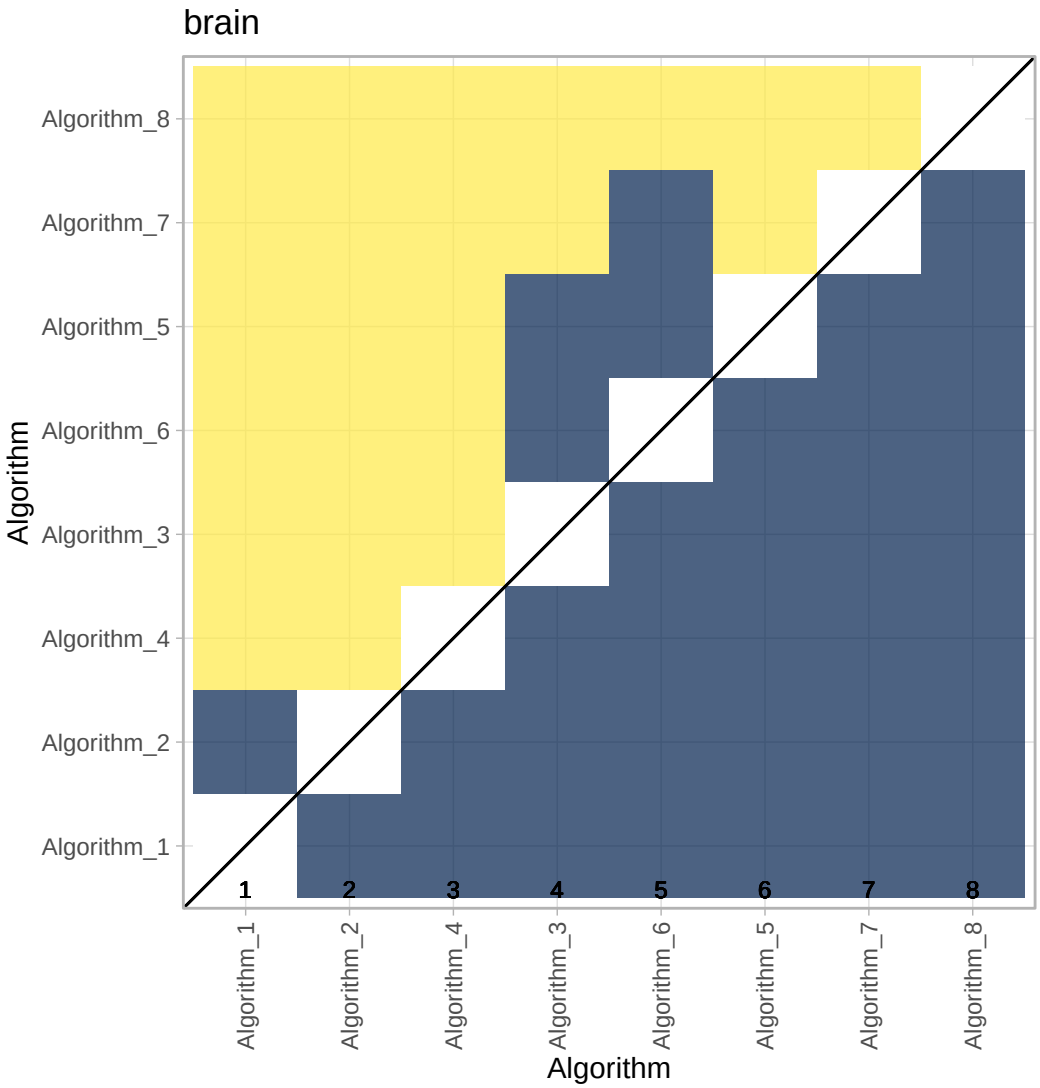

Rank
8
7
6
5
4
3
2
1
Algorithm_1
Algorithm_2
Algorithm_4
Algorithm_3
Algorithm_6
Algorithm_5
Algorithm_7
Algorithm_8
Algorithm
Count
50
40
30
20
10
0

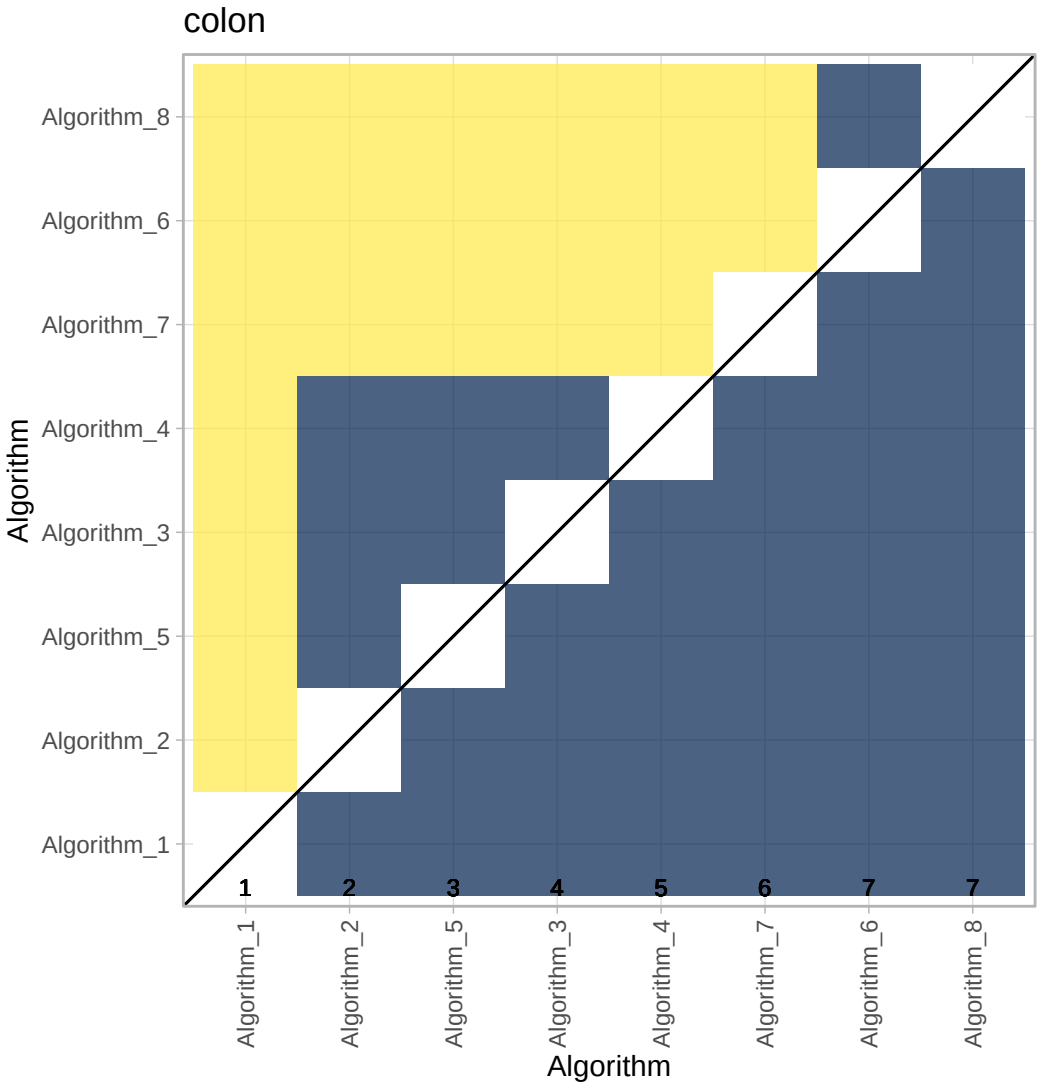


2 Visualization of ranking stability

2.1 *Significance maps* for visualizing ranking stability based on statistical significance

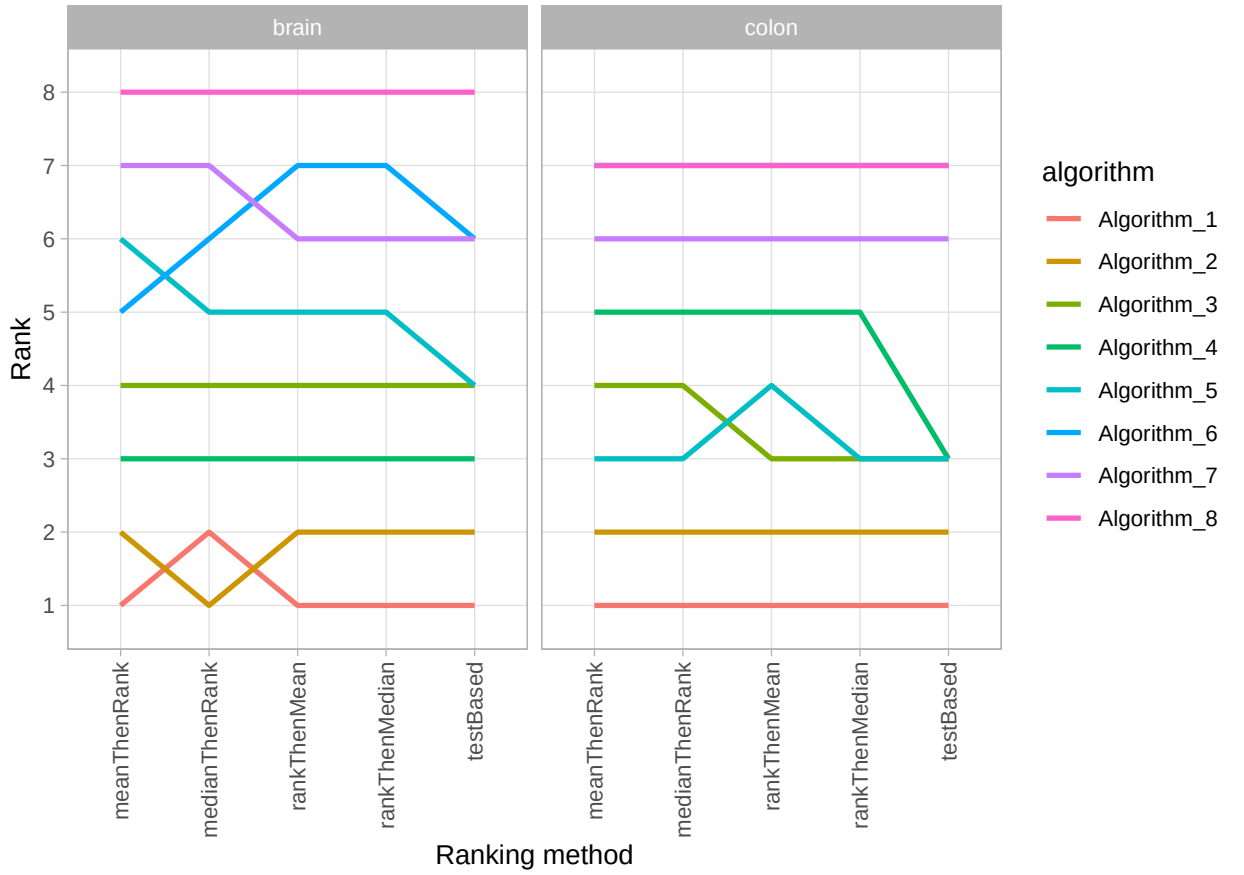
Significance maps depict incidence matrices of pairwise significant test results for the one-sided Wilcoxon signed rank test at a 5% significance level with adjustment for multiple testing according to Holm. Yellow shading indicates that metric values of the algorithm on the x-axis were significantly superior to those from the algorithm on the y-axis, blue color indicates no significant difference.


brain
Algorithm_8
Algorithm_7
Algorithm_5
Algorithm_6
Algorithm_3
Algorithm_4
Algorithm_2
Algorithm_1
Algorithm
1
2
3
4
5
6
7
8
Algorithm_1
Algorithm_2
Algorithm_4
Algorithm_3
Algorithm_6
Algorithm_5
Algorithm_7
Algorithm_8
Algorithm



2.2 Ranking robustness to ranking methods

Line plots for visualizing rankings robustness across different ranking methods. Each algorithm is represented by one colored line. For each ranking method encoded on the x-axis, the height of the line represents the corresponding rank. Horizontal lines indicate identical ranks for all methods.

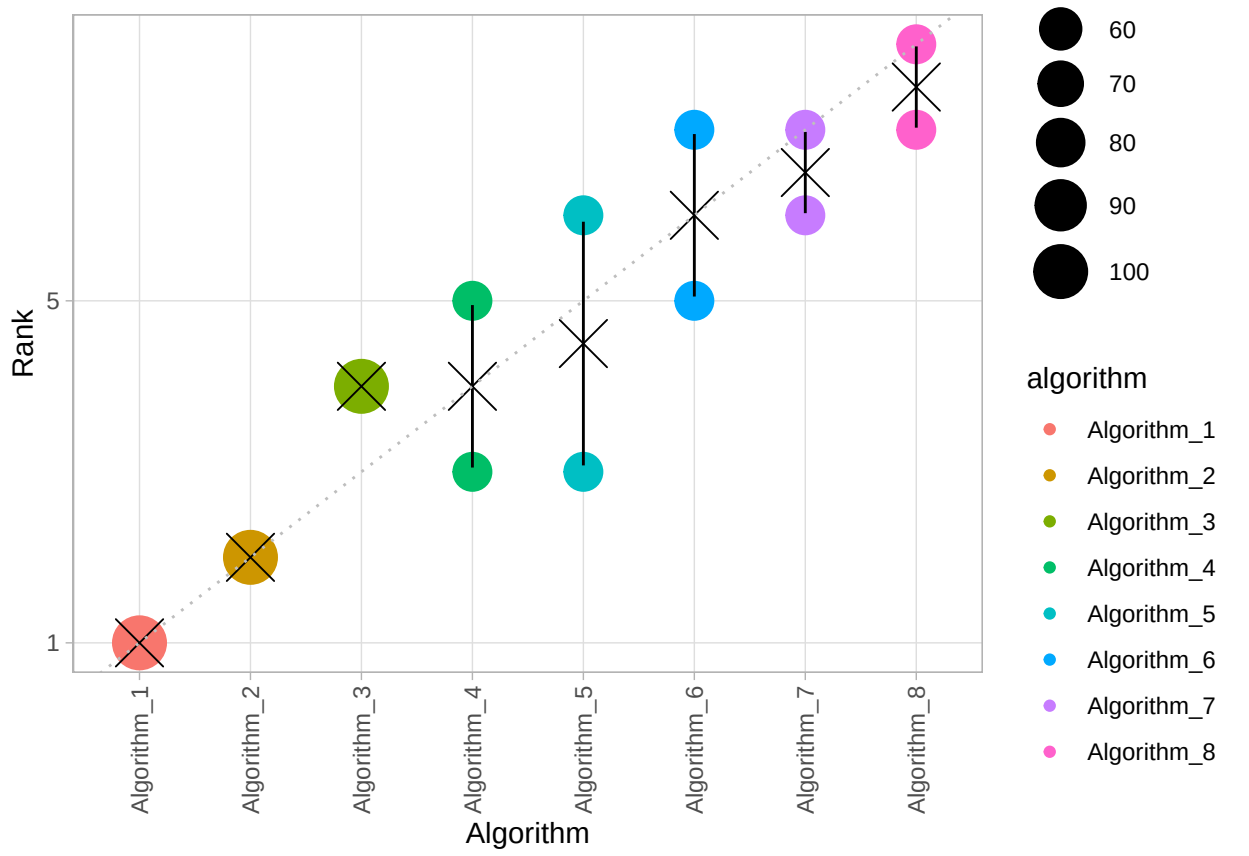


3 Visualization of cross-task insights

Algorithms are ordered according to consensus ranking.

3.1 Characterization of algorithms

3.1.1 Ranking stability: Variability of achieved rankings across tasks



3.2 Characterization of tasks

3.2.1 Cluster Analysis

Dendrogram from hierarchical cluster analysis} and *network-type graphs* for assessing the similarity of tasks based on challenge rankings.

A dendrogram is a visualization approach based on hierarchical clustering. It depicts clusters according to a chosen distance measure (here: Spearman's footrule) as well as a chosen agglomeration method (here: complete and average agglomeration).

```
##
## Cluster analysis only sensible if there are >2 tasks.
```

In network-type graphs (see Eugster et al, 2008), every task is represented by a node and nodes are connected by edges whose length is determined by a chosen distance measure. Here, distances between nodes are chosen to increase exponentially in Spearman's footrule distance with growth rate 0.05 to accentuate large distances. Hence, tasks that are similar with respect to their algorithm ranking appear closer together than those that are dissimilar. Nodes representing tasks with a unique winner are colored-coded by the winning algorithm. In case there are more than one first-ranked algorithms in a task, the corresponding node remains uncolored.

4 Reference

Wiesenfarth, M., Reinke, A., Landman, B.A., Cardoso, M.J., Maier-Hein, L. and Kopp-Schneider, A. (2019). Methods and open-source toolkit for analyzing and visualizing challenge results. *arXiv preprint arXiv:1910.05121*

M. J. A. Eugster, T. Hothorn, and F. Leisch, "Exploratory and inferential analysis of benchmark experiments," Institut fuer Statistik, Ludwig-Maximilians- Universitaet Muenchen, Germany, Technical Report 30, 2008. [Online]. Available: http://epub.ub.uni-muenchen. de/4134/.