Benchmarking report for mean_mse

created by challengeR v1.0.1

23 April, 2021

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

- Visualization of assessment data: Dot- and boxplots, podium plots and ranking heatmaps
- Visualization of ranking stability: Blob plots, violin plots and significance maps, line plots
- Visualization of cross-task insights: Blob plots, stacked frequency plots, dendrograms

Details can be found in Wiesenfarth et al. (2021).

1 Rankings

Algorithms within a task are ranked according to the following ranking scheme:

aggregate using function significance then rank

Column 'prop_significance' is equal to the number of pairwise significant test results for a given algorithm divided by the number of algorithms.

Ranking for each task:

rousset_E18_median-sub_guide : The analysis is based on 0 algorithms and 46 cases. 0 missing cases have been found in the data set.

	prop_significance	rank
1DCNN	1	1
elnet	0	2
GBM	0	2

wang_median-sub_guide : The analysis is based on 0 algorithms and 228 cases. 0 missing cases have been found in the data set.

	prop_significance	rank
elnet	0	1
GBM	0	1
1DCNN	0	1

Consensus ranking across tasks according to chosen method "euclidean":

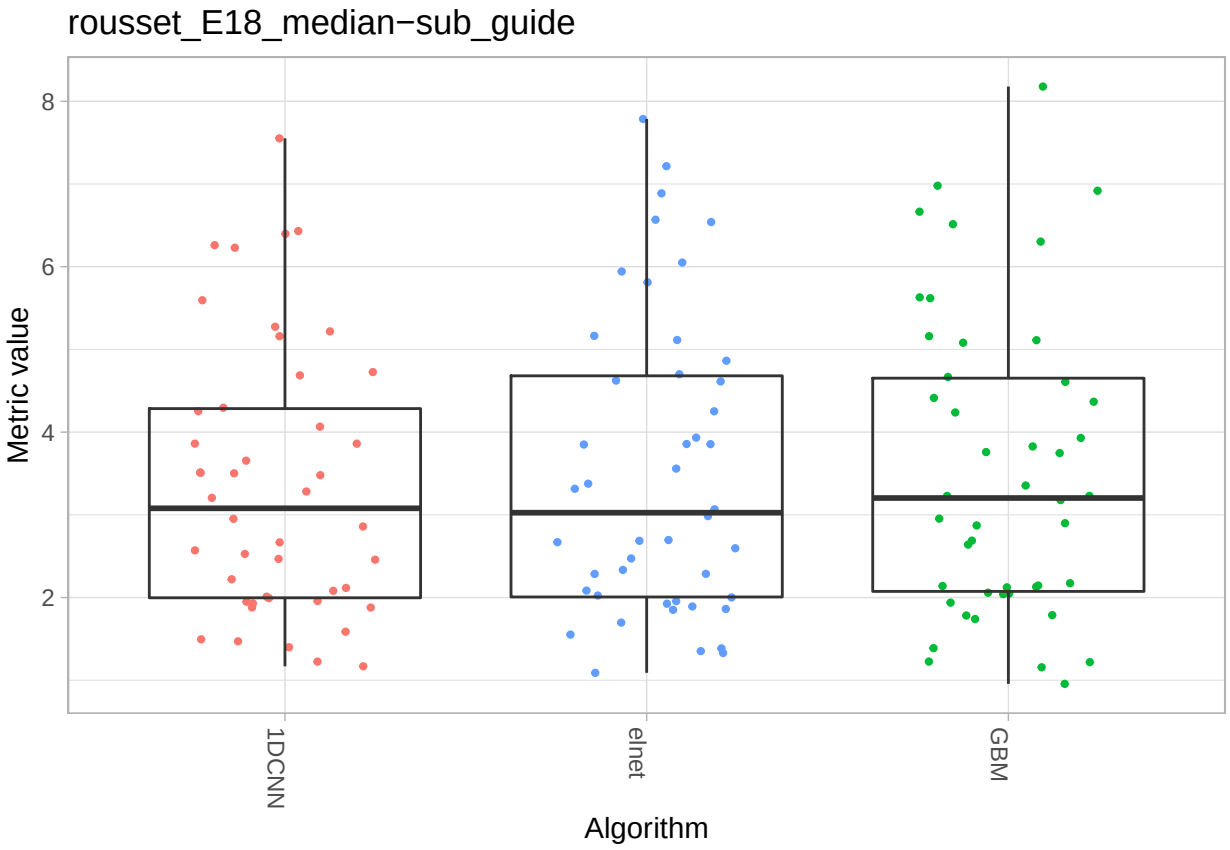
	value	rank
1DCNN	1.50	1
GBM	2.25	2
elnet	2.25	2

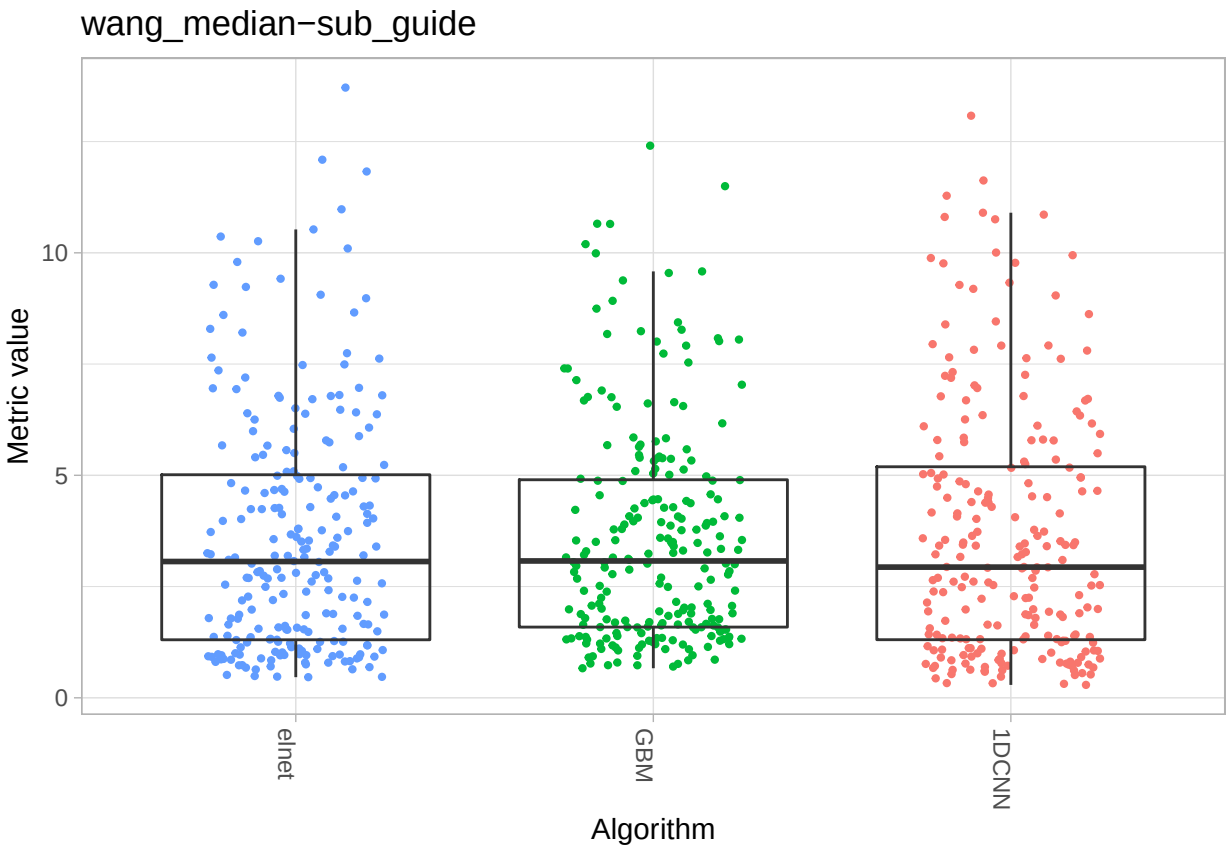
2 Visualization of raw assessment data

The algorithms are ordered according to the computed ranks for each task.

2.1 Dot- and boxplot

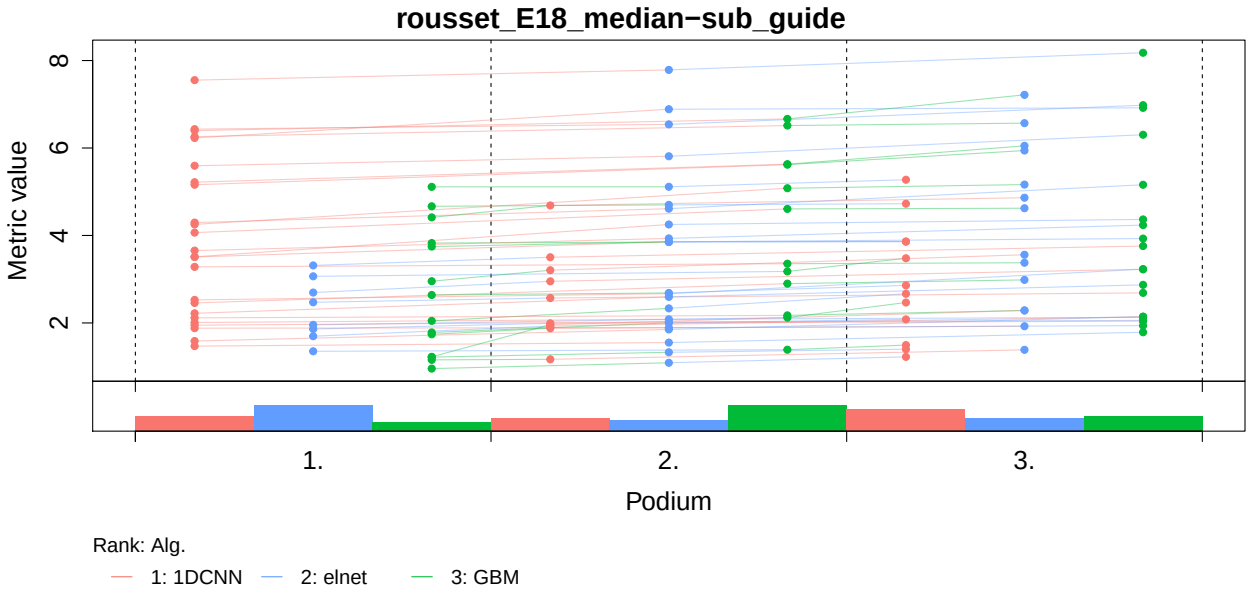
Dot- and boxplots for visualizing raw assessment data separately for each algorithm. Boxplots representing descriptive statistics over all cases (median, quartiles and outliers) are combined with horizontally jittered dots representing individual cases.

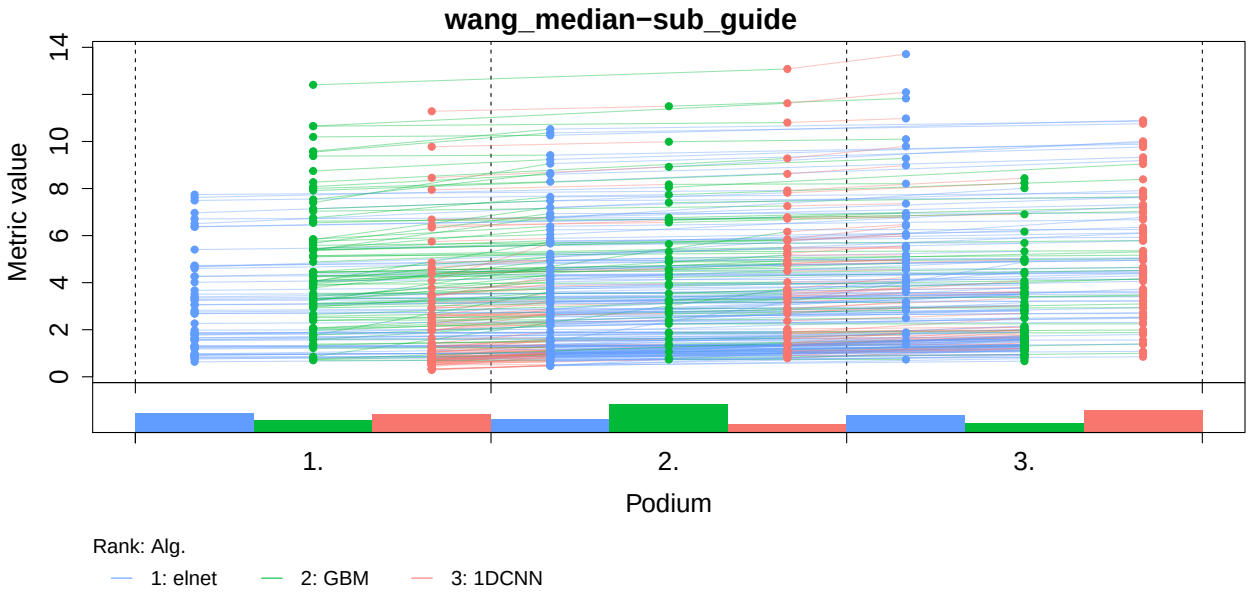



wang_median−sub_guide
Metric value
10
5
0
elnet
GBM
1DCNN
Algorithm

2.2 Podium plot

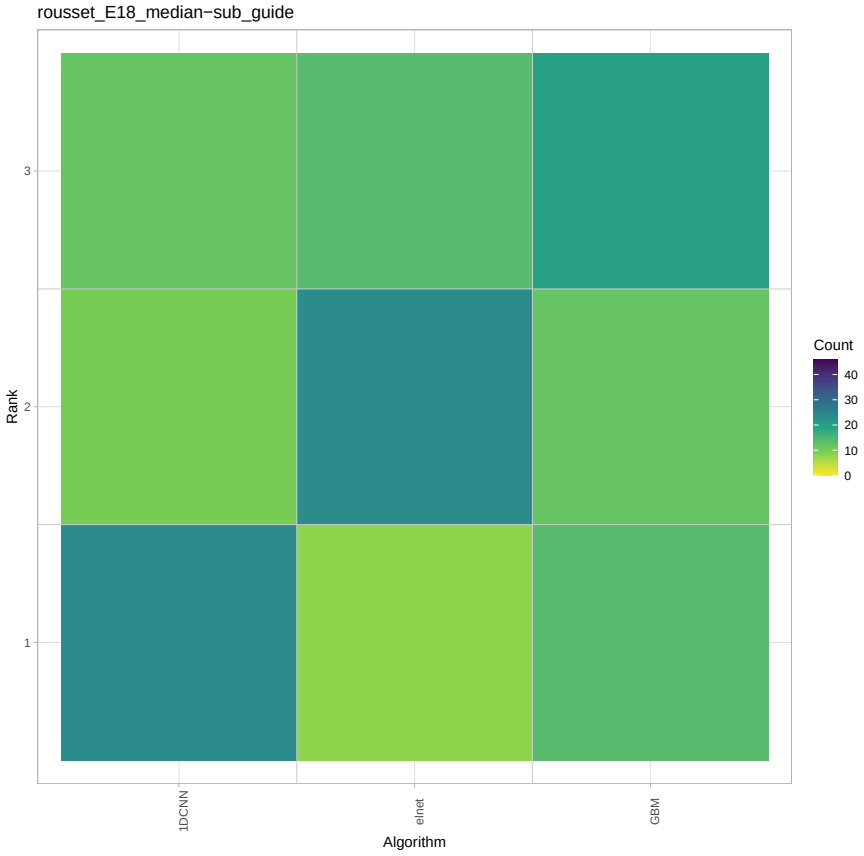
Podium plots (see also Eugster et al., 2008) for visualizing raw assessment data. Upper part (spaghetti plot): Participating algorithms are color-coded, and each colored dot in the plot represents a metric value achieved with the respective algorithm. The actual metric value is encoded by the y-axis. Each podium (here: p=3) represents one possible rank, ordered from best (1) to last (here: 3). 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 case. Note that the plot part above each podium place is further subdivided into p "columns", where each column represents one participating algorithm (here: $p = 3$). Dots corresponding to identical 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.

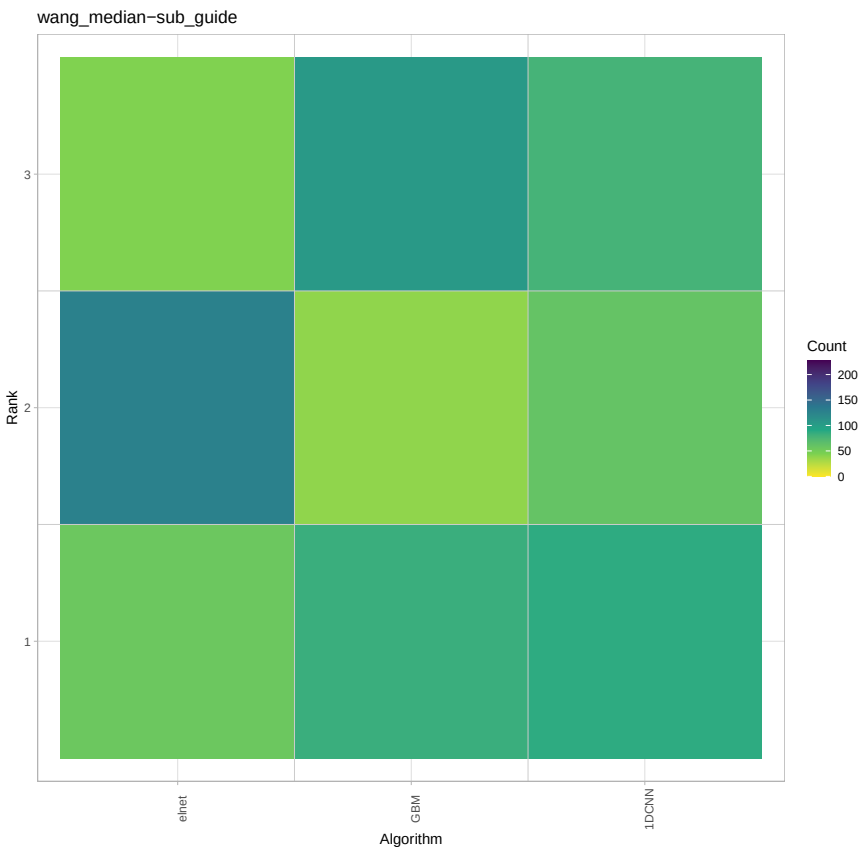



wang_median−sub_guide
Metric value
0
2
4
6
8
10
14
1.
2.
3.
Podium
Rank: Alg.
1: elnet
2: GBM
3: 1DCNN

2.3 Ranking heatmap

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

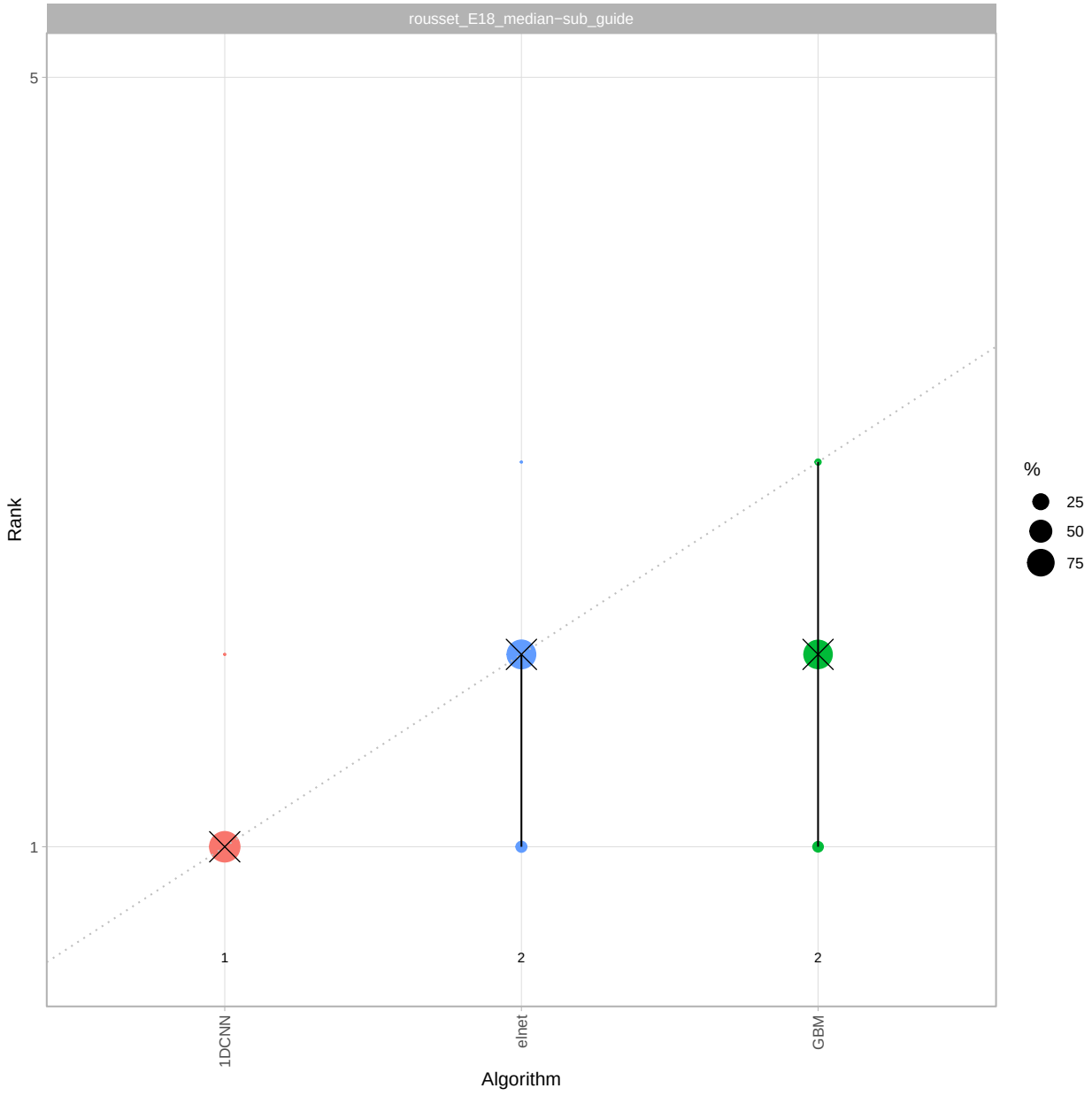


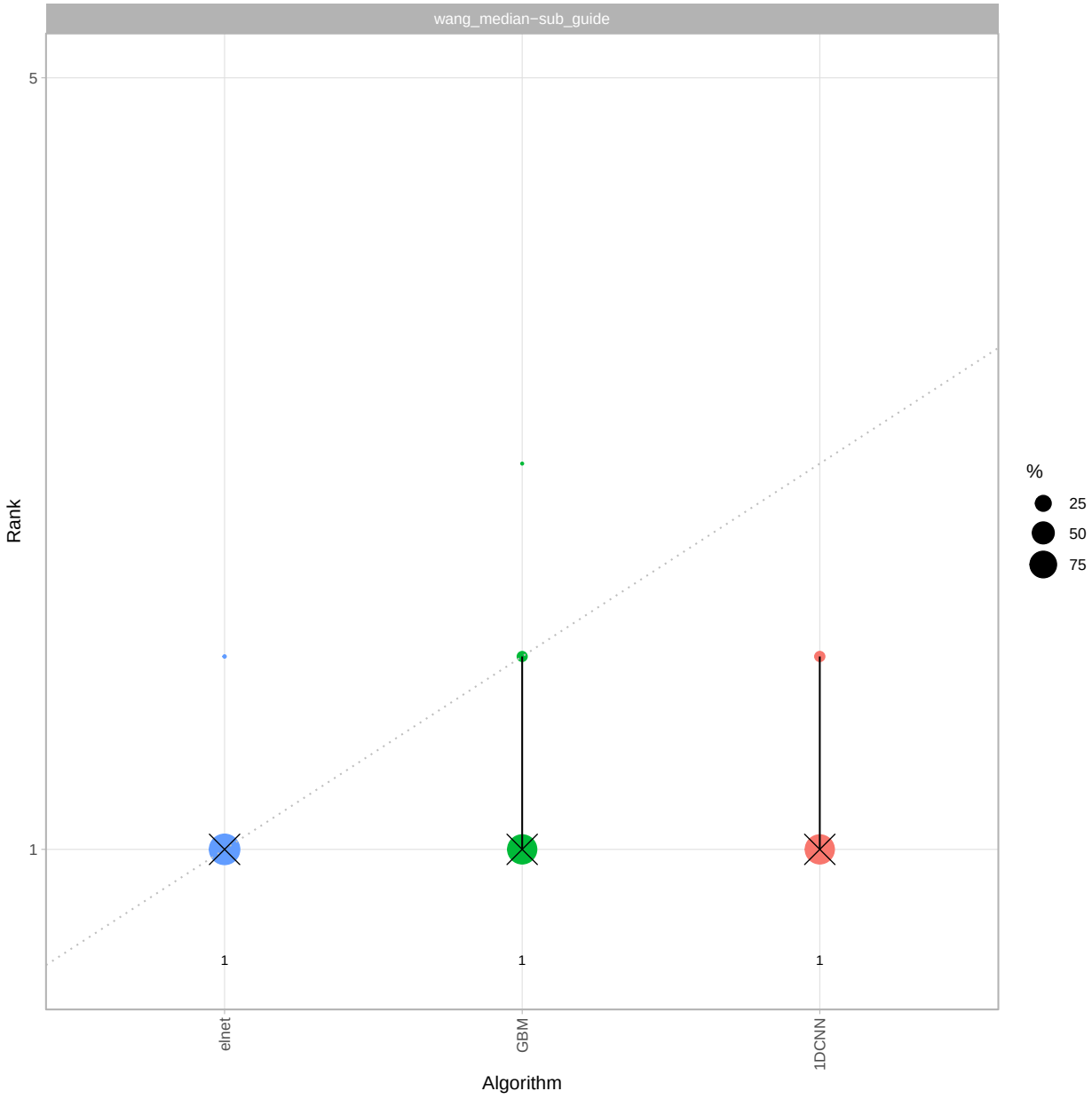

wang_median−sub_guide
Rank
3
2
1
elnet
GBM
1DCNN
Algorithm
Count
200
150
100
50
0

3 Visualization of ranking stability

3.1 *Blob plot* for visualizing ranking stability based on bootstrap sampling

Algorithms are color-coded, and the area of each blob at position $(A_i, \text{rank } j)$ is proportional to the relative frequency A_i achieved rank j across $b = 1000$ bootstrap samples. The median rank for each algorithm is indicated by a black cross. 95% bootstrap intervals across bootstrap samples are indicated by black lines.




wang_median−sub_guide
5
1
Rank
1
1
1
%
25
50
75
elnet
GBM
1DCNN
Algorithm

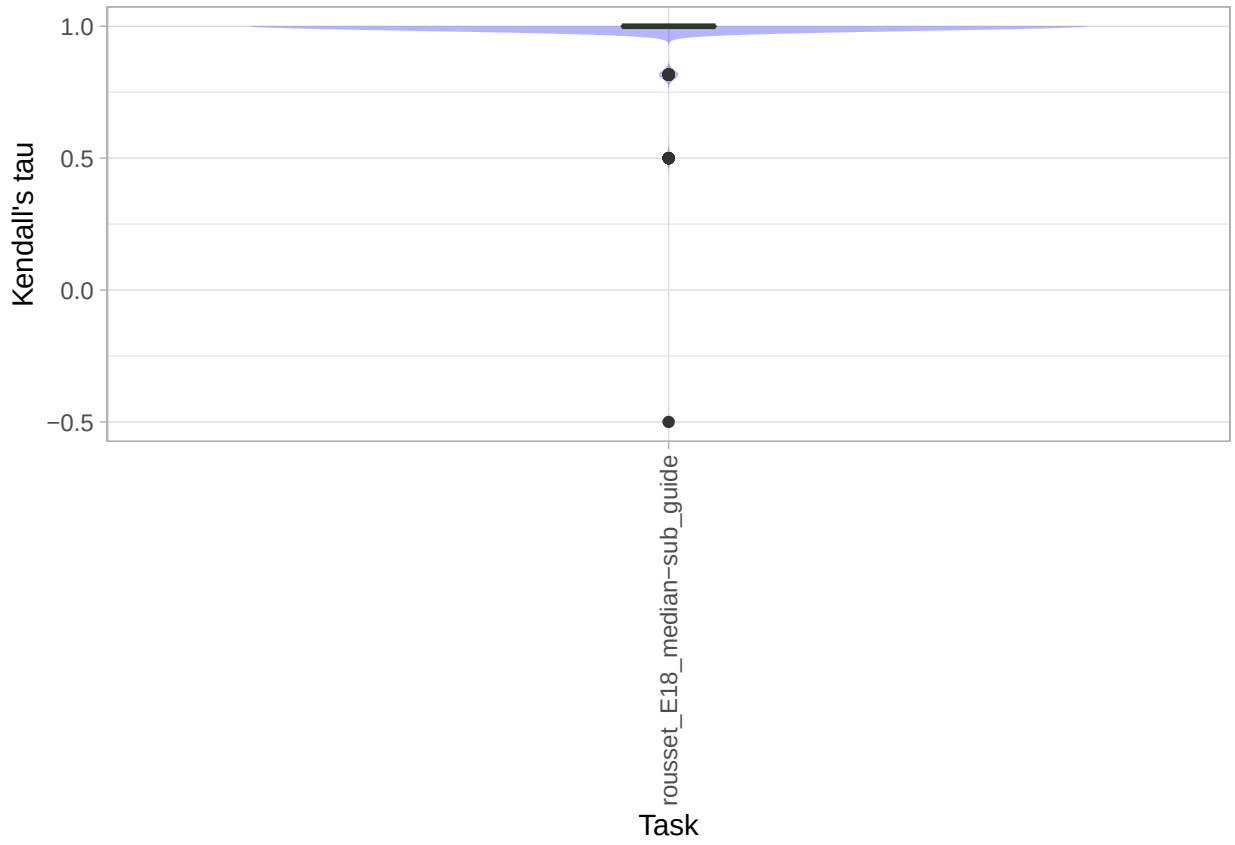
3.2 *Violin plot* for visualizing ranking stability based on bootstrapping

The ranking list based on the full assessment data is pairwise compared with the ranking lists based on the individual bootstrap samples (here $b = 1000$ samples). For each pair of rankings, Kendall's τ correlation is computed. Kendall's τ is a scaled index determining the correlation between the lists. It is computed by evaluating the number of pairwise concordances and discordances between ranking lists and produces values between -1 (for inverted order) and 1 (for identical order). A violin plot, which simultaneously depicts a boxplot and a density plot, is generated from the results.

Summary Kendall's tau:

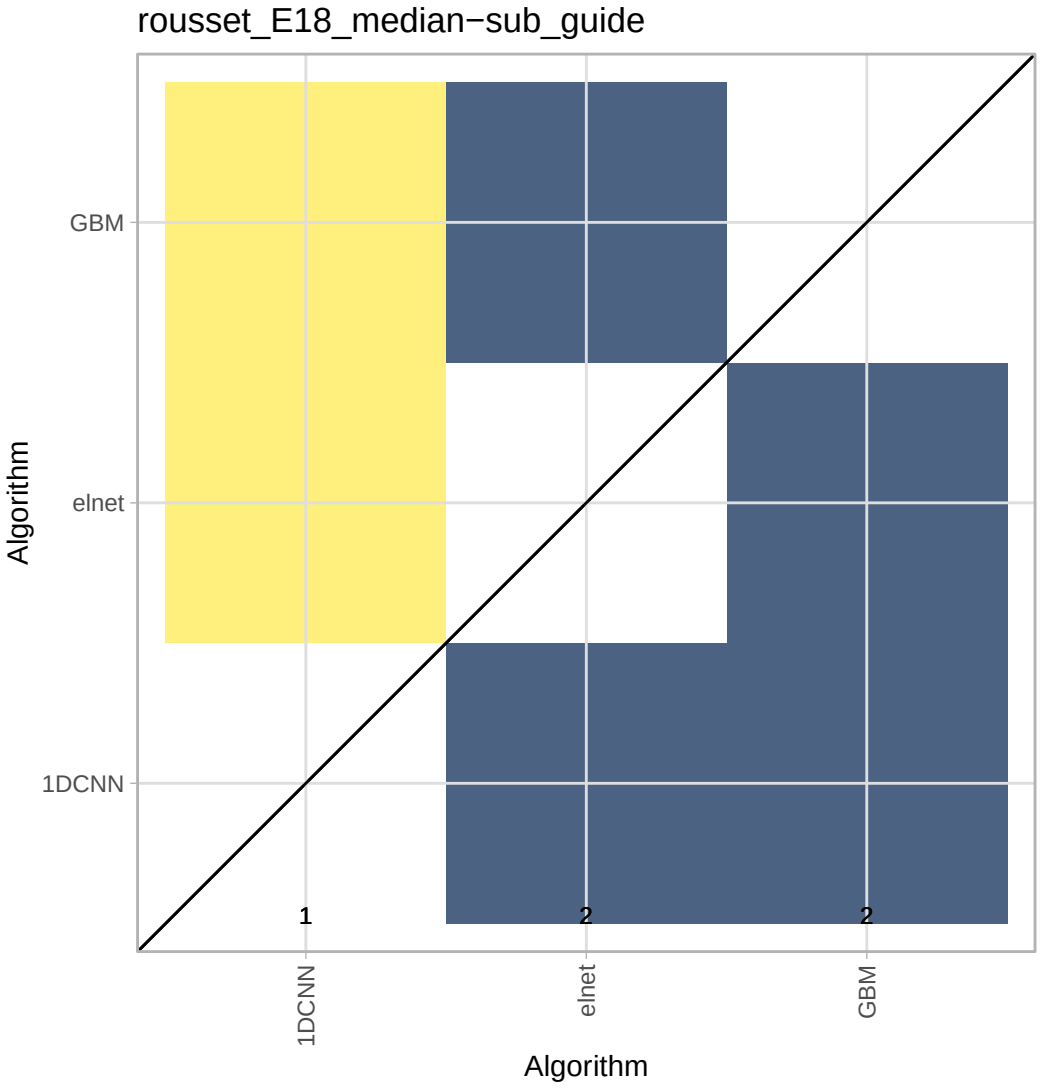
Task	mean	median	q25	q75
rousset_E18_median-sub_guide	0.9892435	1	1	1
wang_median-sub_guide	NaN	NA	NA	NA

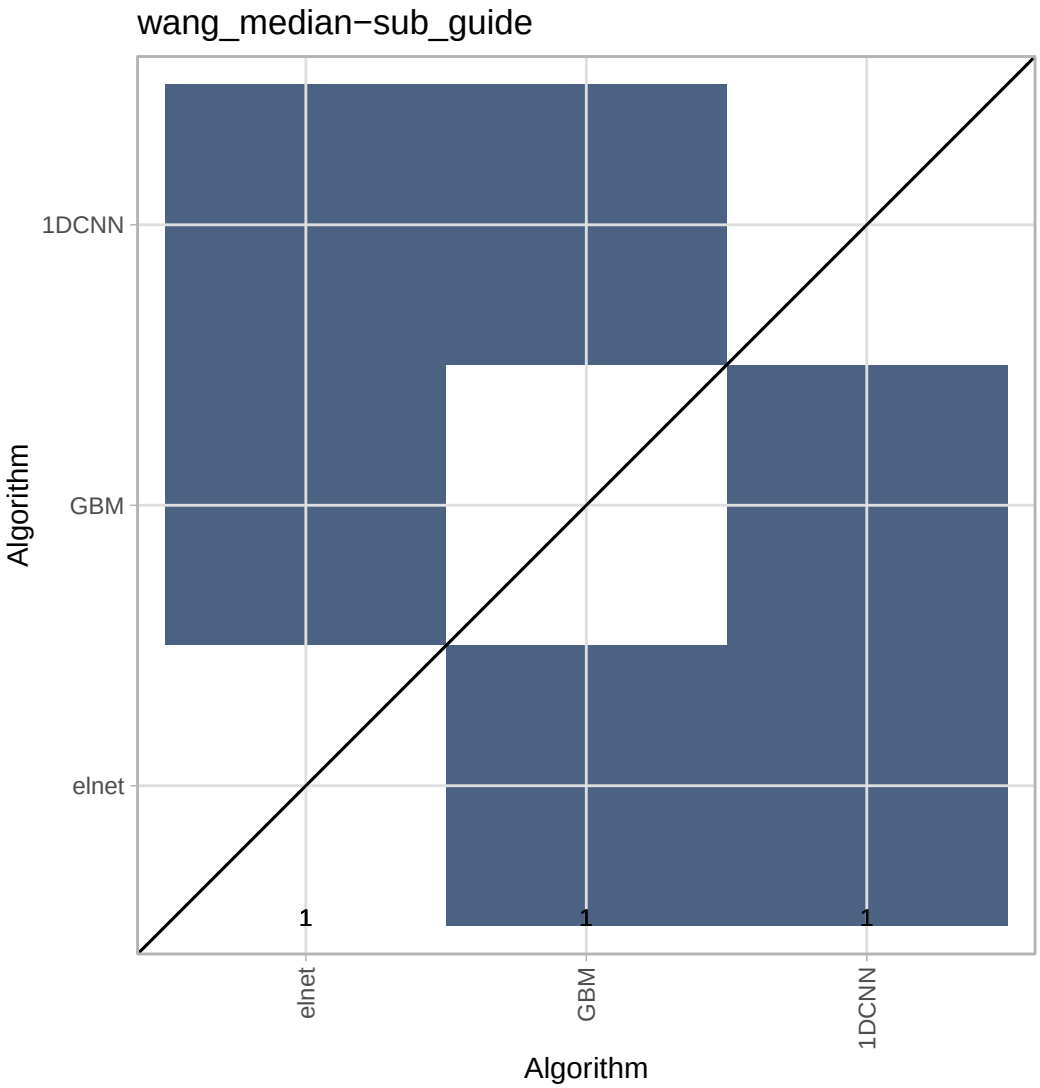
No Kendall's tau could be calculated for any bootstrap sample in task wang_median-sub_guide because of missing variability. Task dropped from figure.



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

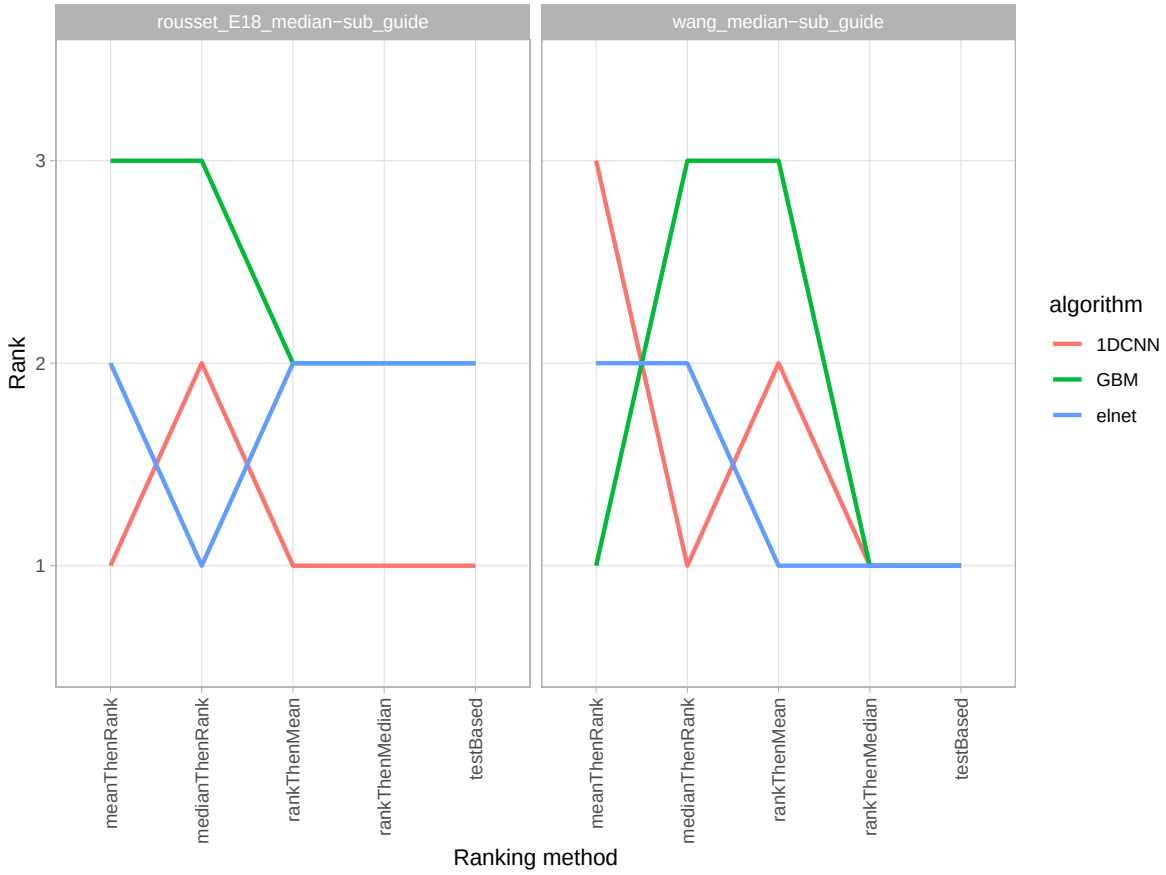
Significance maps depict incidence matrices of pairwise significant test results for the one-sided Wilcoxon signed rank test at a 5% significance level with adjustment for multiple testing according to Holm. Yellow shading indicates that metric values from the algorithm on the x-axis were significantly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




wang_median−sub_guide
1DCNN
GBM
elnet
Algorithm
1
1
1
elnet
GBM
1DCNN
Algorithm

3.4 Ranking robustness to ranking methods

Line plots for visualizing ranking 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.



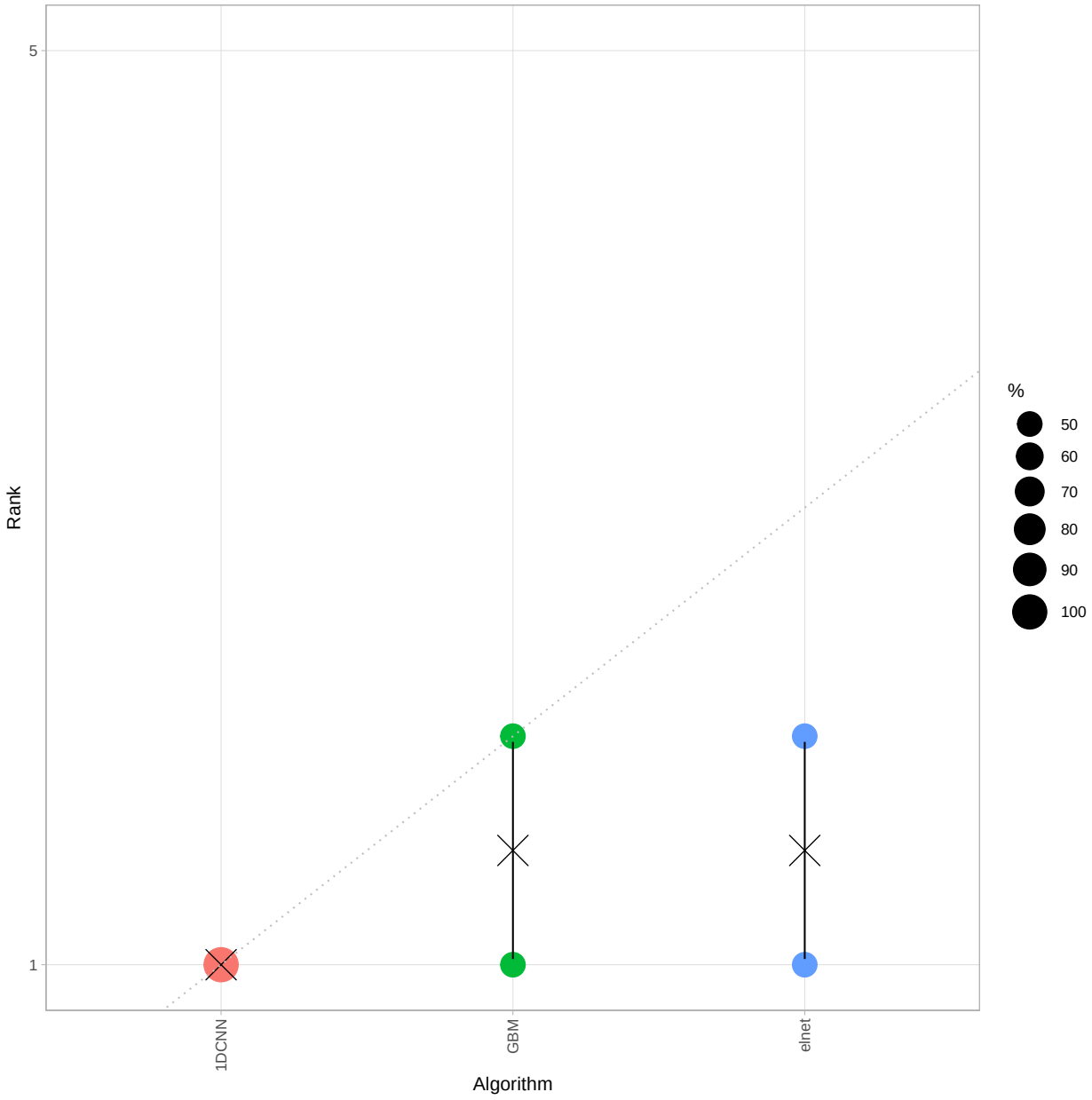
4 Visualization of cross-task insights

The algorithms are ordered according to consensus ranking.

4.1 Characterization of algorithms

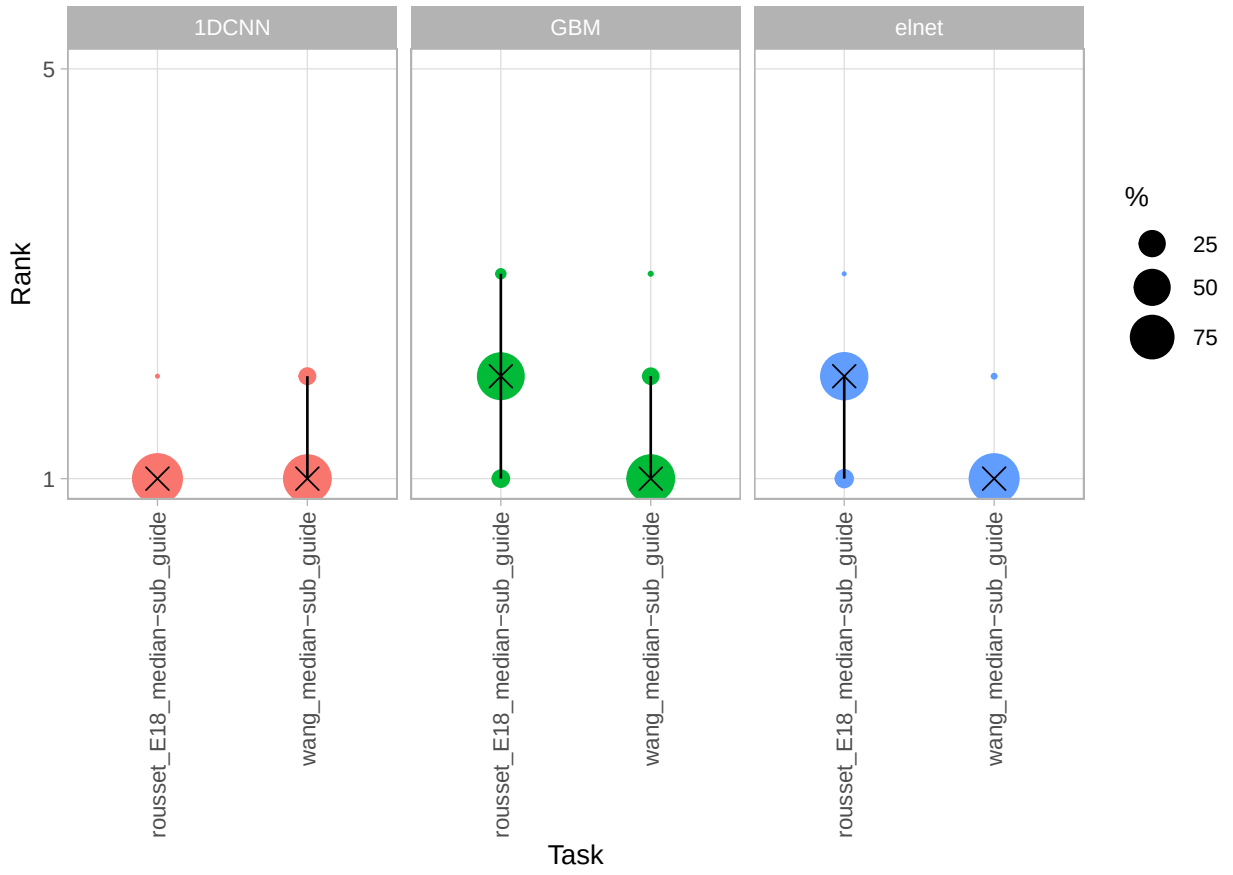
4.1.1 Ranking stability: Variability of achieved rankings across tasks

Algorithms are color-coded, and the area of each blob at position (A_i, rank j) is proportional to the relative frequency A_i achieved rank j across multiple tasks. The median rank for each algorithm is indicated by a black cross. This way, the distribution of ranks across tasks can be intuitively visualized.

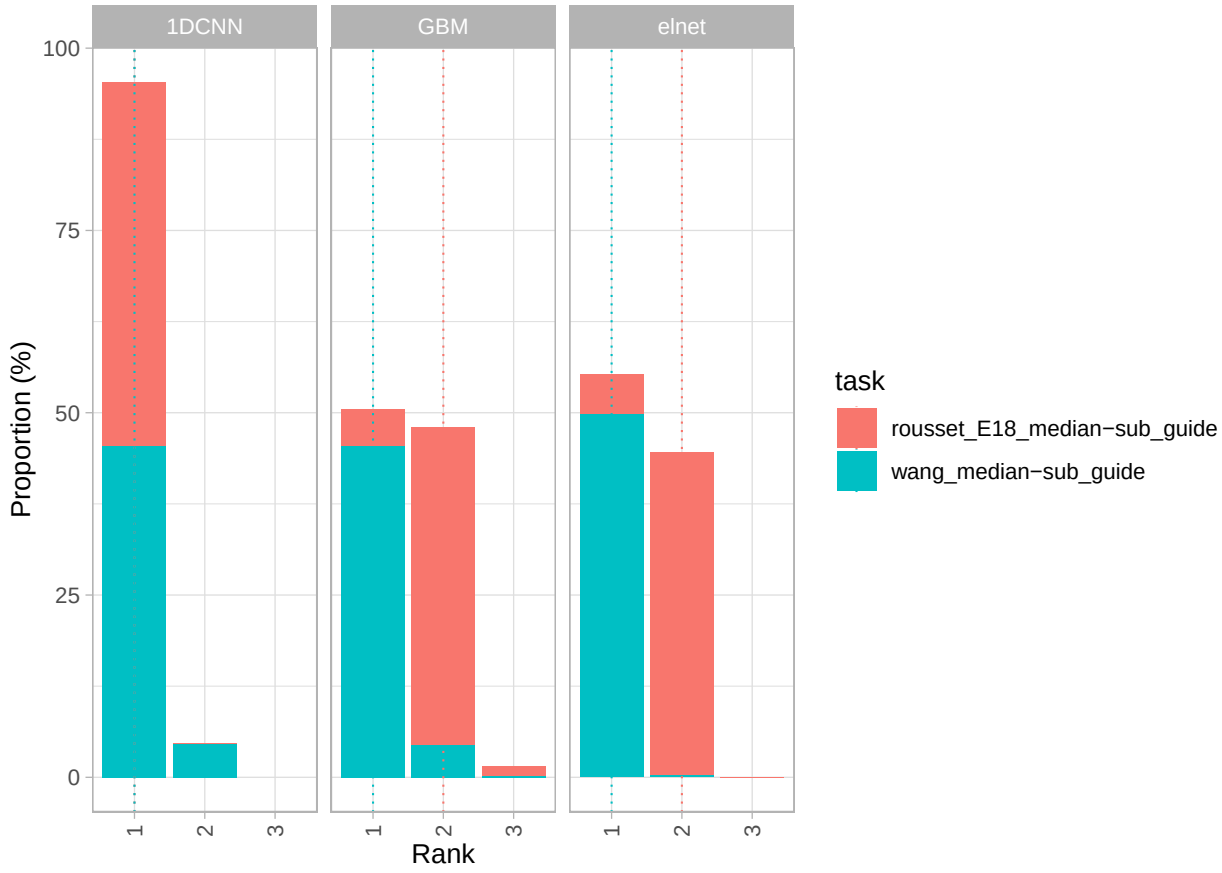


4.1.2 Ranking stability: Ranking variability via bootstrap approach

A blob plot of bootstrap results over the different tasks separated by algorithm allows another perspective on the assessment data. This gives deeper insights into the characteristics of tasks and the ranking uncertainty of the algorithms in each task.



An alternative representation is provided by a stacked frequency plot of the observed ranks, separated by algorithm. Observed ranks across bootstrap samples are displayed with coloring according to the task. For algorithms that achieve the same rank in different tasks for the full assessment data set, vertical lines are on top of each other. Vertical lines allow to compare the achieved rank of each algorithm over different tasks.



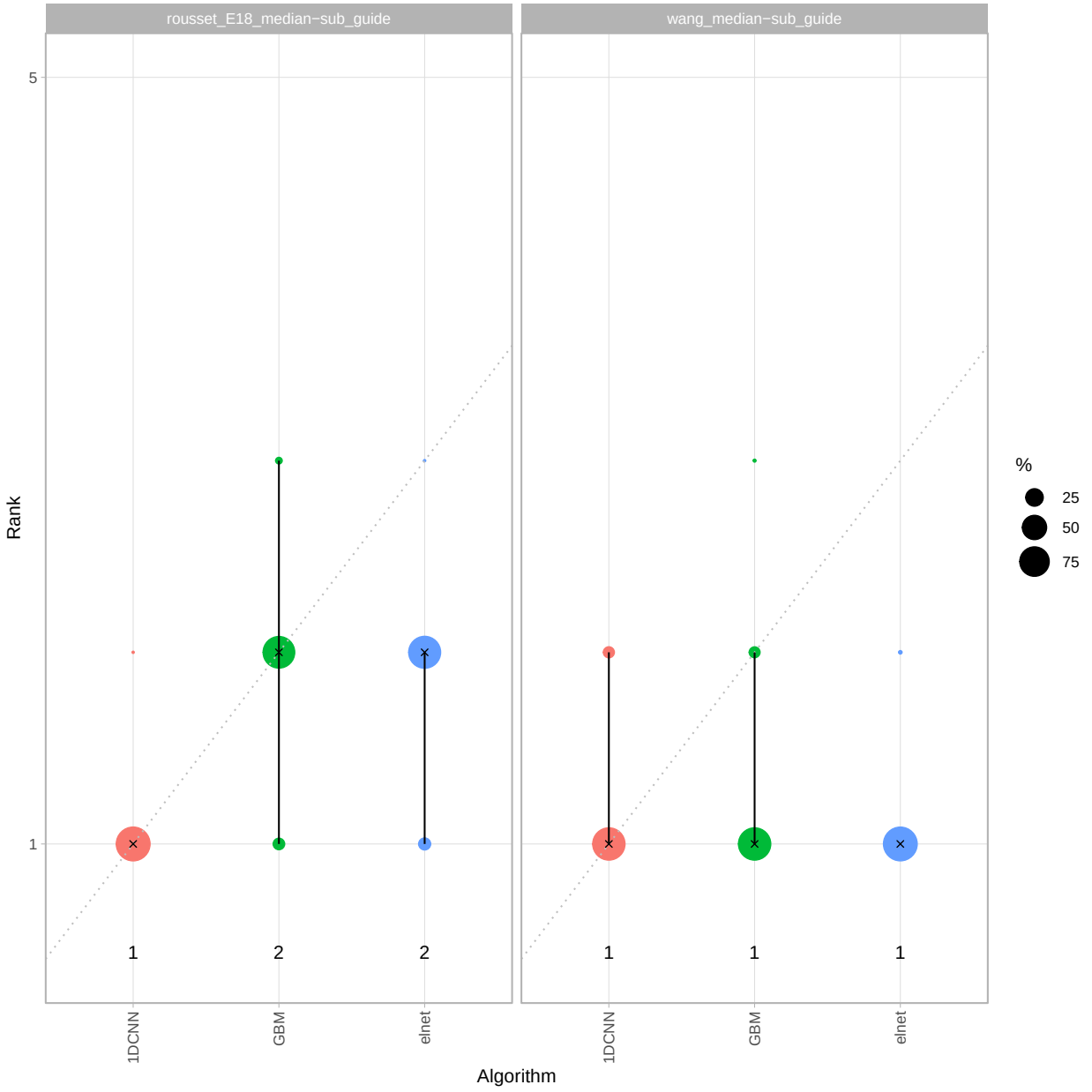
4.2 Characterization of tasks

4.2.1 Visualizing bootstrap results

To investigate which tasks separate algorithms well (i.e., lead to a stable ranking), a blob plot is recommended.

Bootstrap results can be shown in a blob plot showing one plot for each task. In this view, the spread of the blobs for each algorithm can be compared across tasks. Deviations from the diagonal indicate deviations from the consensus ranking (over tasks). Specifically, if rank distribution of an algorithm is consistently below the diagonal, the algorithm performed better in this task than on average across tasks, while if the rank distribution of an algorithm is consistently above the diagonal, the algorithm performed worse in this task than on average across tasks. At the bottom of each panel, ranks for each algorithm in the tasks are provided.

Same as in Section 3.1 but now ordered according to consensus.



4.2.2 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.
```

5 References

Wiesenfarth, M., Reinke, A., Landman, B.A., Eisenmann, M., Aguilera Saiz, L., Cardoso, M.J., Maier-Hein, L. and Kopp-Schneider, A. Methods and open-source toolkit for analyzing and visualizing challenge results. *Sci Rep* **11**, 2369 (2021). https://doi.org/10.1038/s41598-021-82017-6

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/.