Benchmarking report for Prognosis Task, significance ranking

created by challengeR v1.0.1

29 April, 2021

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

- Visualization of assessment data: Dot- and boxplot, podium plot and ranking heatmap
- Visualization of ranking stability: Blob plot, violin plot and significance map, line plot

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

1 Ranking

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.

The analysis is based on 0 algorithms and 120 cases. 0 missing cases have been found in the data set.

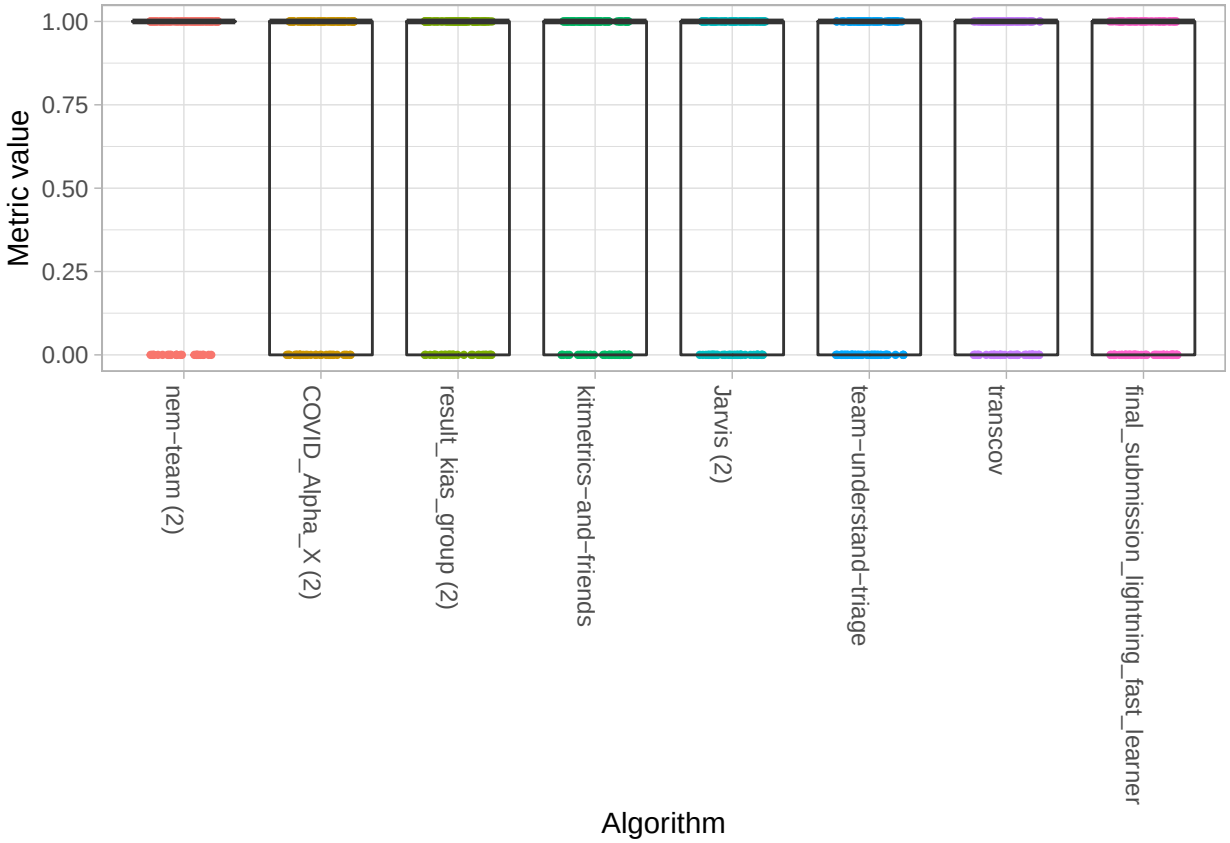
Ranking:

	prop_significance	rank
nem-team (2)	0.8571429	1
COVID_Alpha_X (2)	0.5714286	2
result_kias_group (2)	0.5714286	2
kitmetrics-and-friends	0.4285714	4
Jarvis (2)	0.1428571	5
team-understand-triage	0.1428571	5
transcov	0.1428571	5
final_submission_lightning_fast_learner	0.0000000	8

2 Visualization of raw assessment data

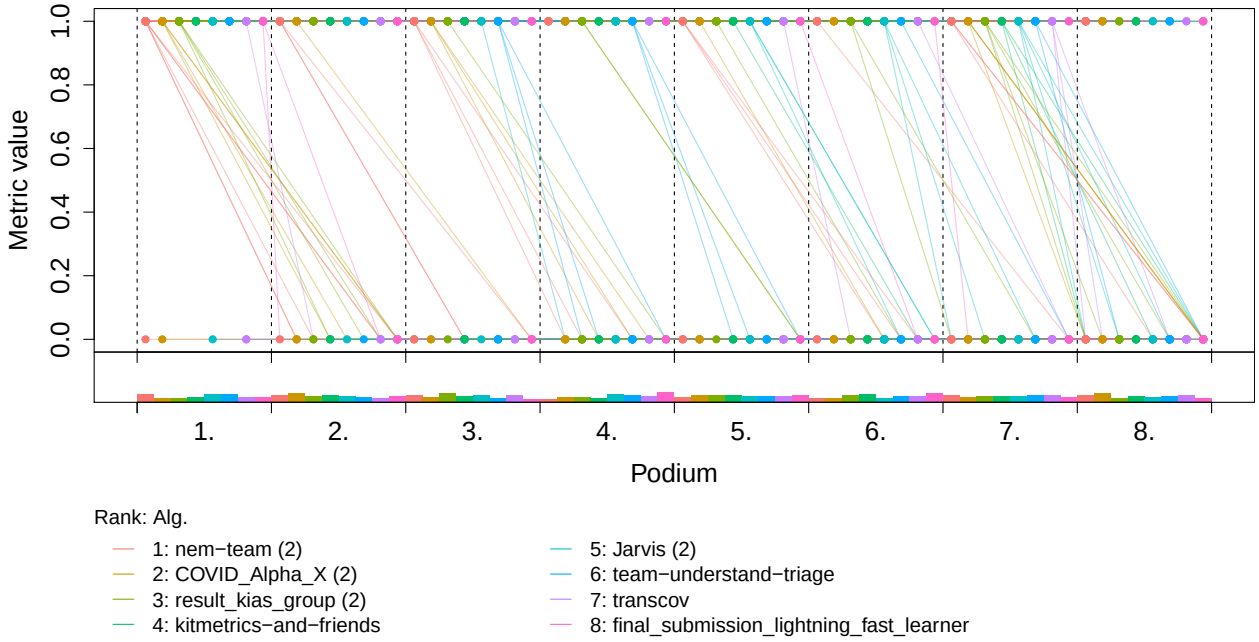
2.1 Dot- and boxplot

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.



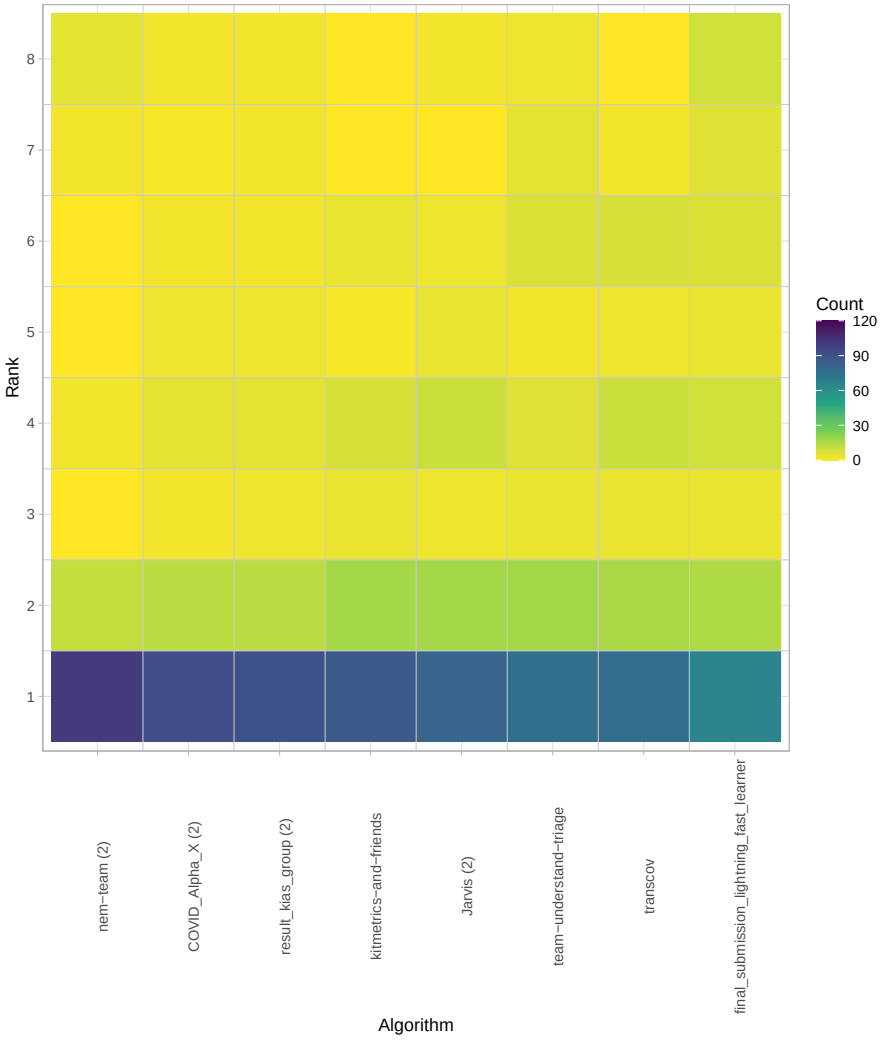
2.2 Podium plot

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=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 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 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.



2.3 Ranking heatmap

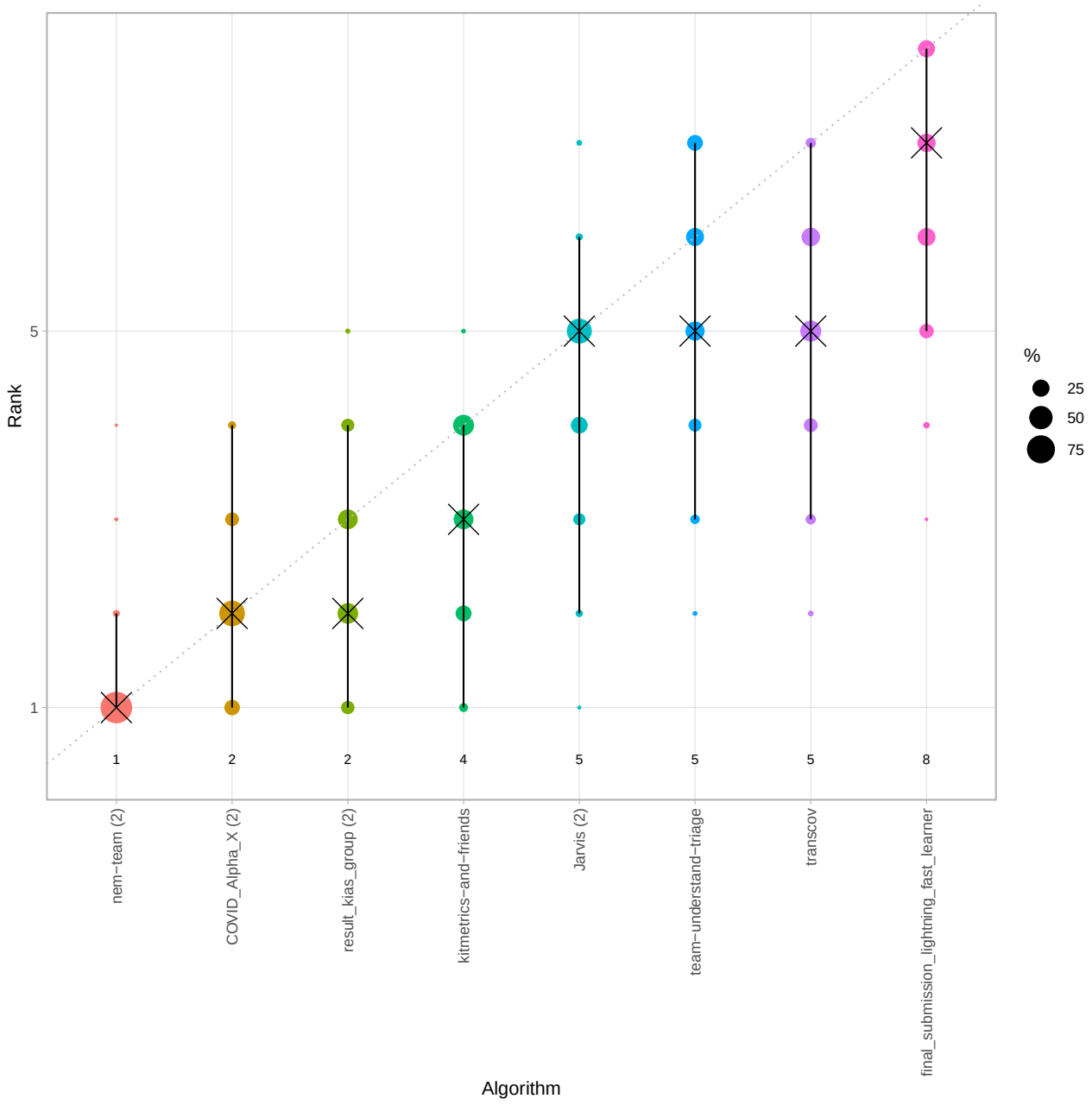
Ranking heatmaps for visualizing raw assessment data. Each cell (i, A_j) shows the absolute frequency of cases in which algorithm A_j achieved rank i.



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.

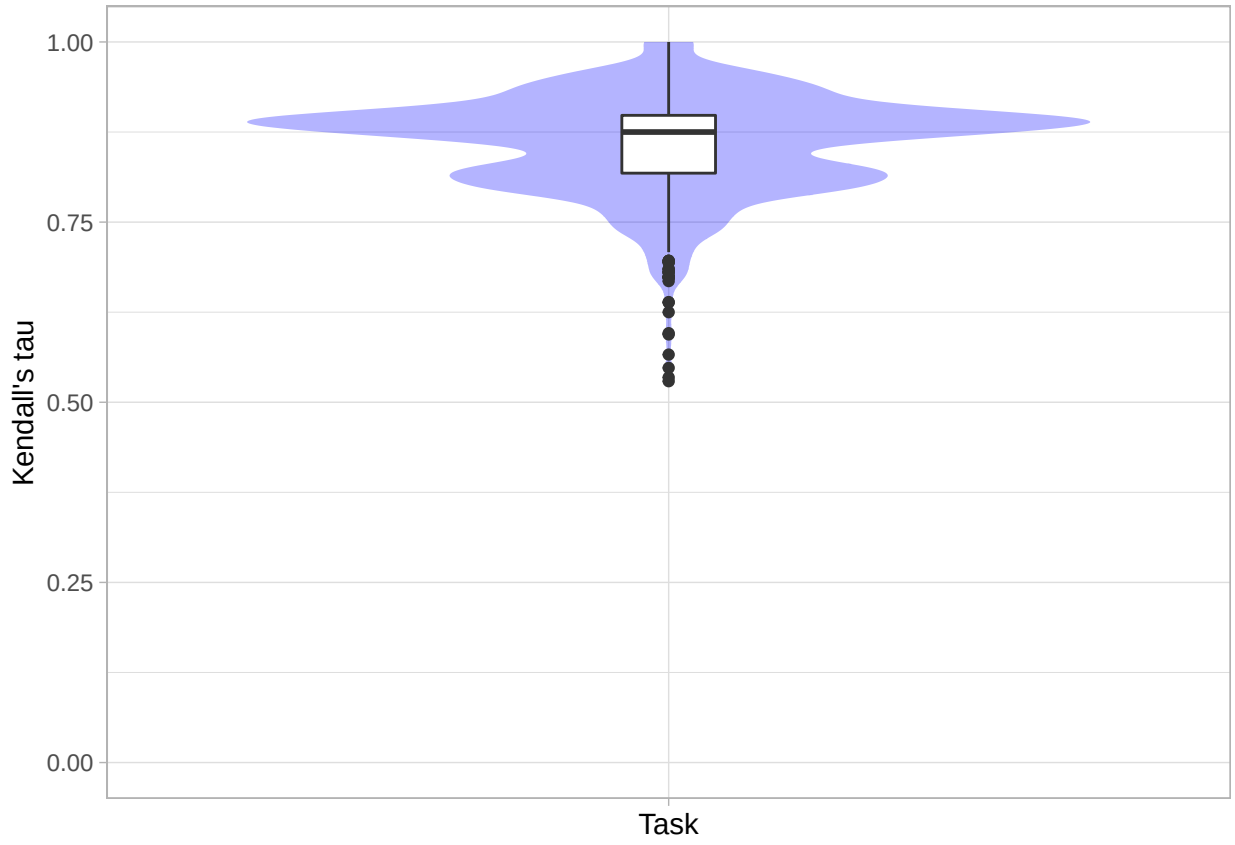


3.2 *Violin plot* for visualizing ranking stability based on bootstrapping

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.

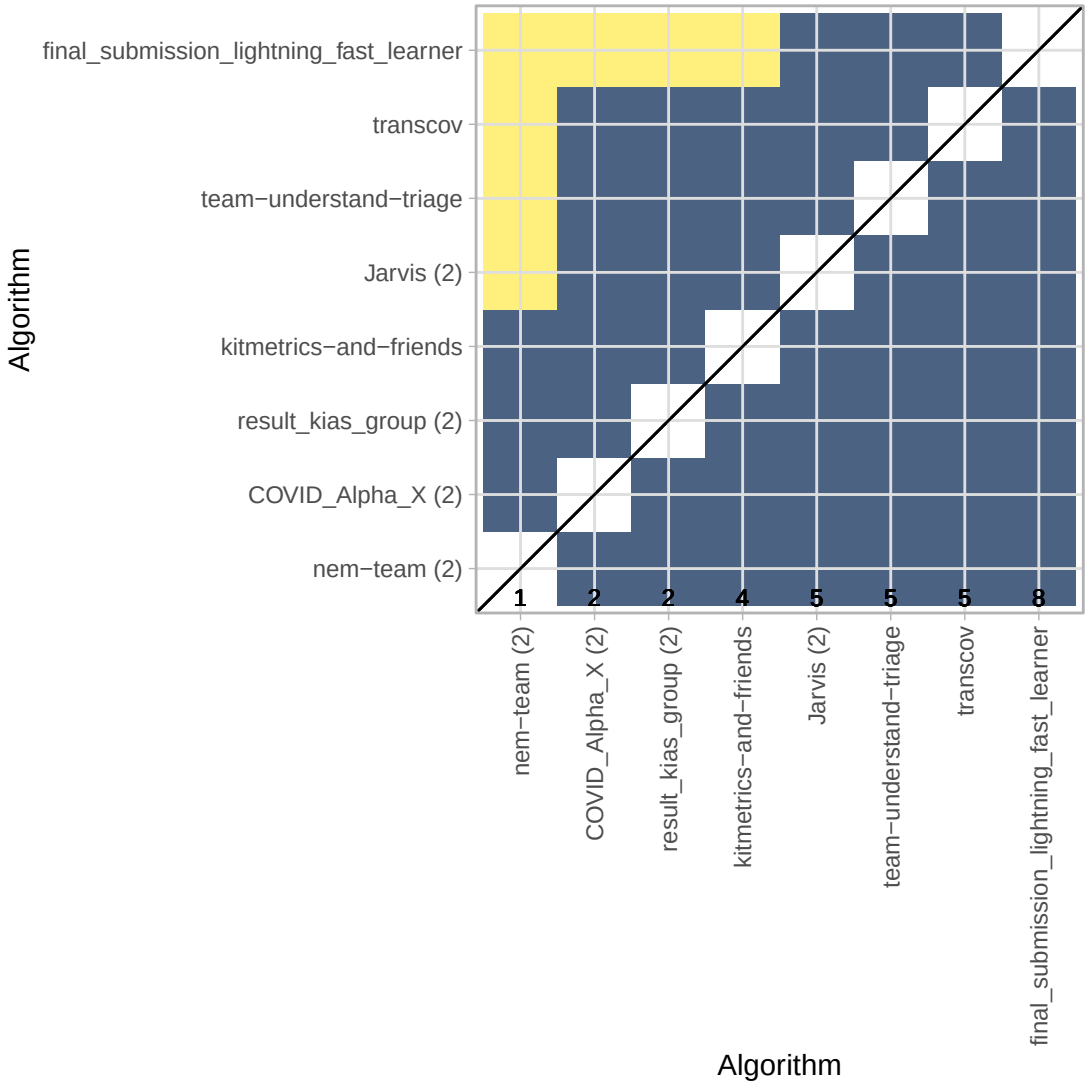
Summary Kendall's tau:

Task	mean	median	q25	q75
dummyTask	0.8599467	0.875	0.8179129	0.8981462



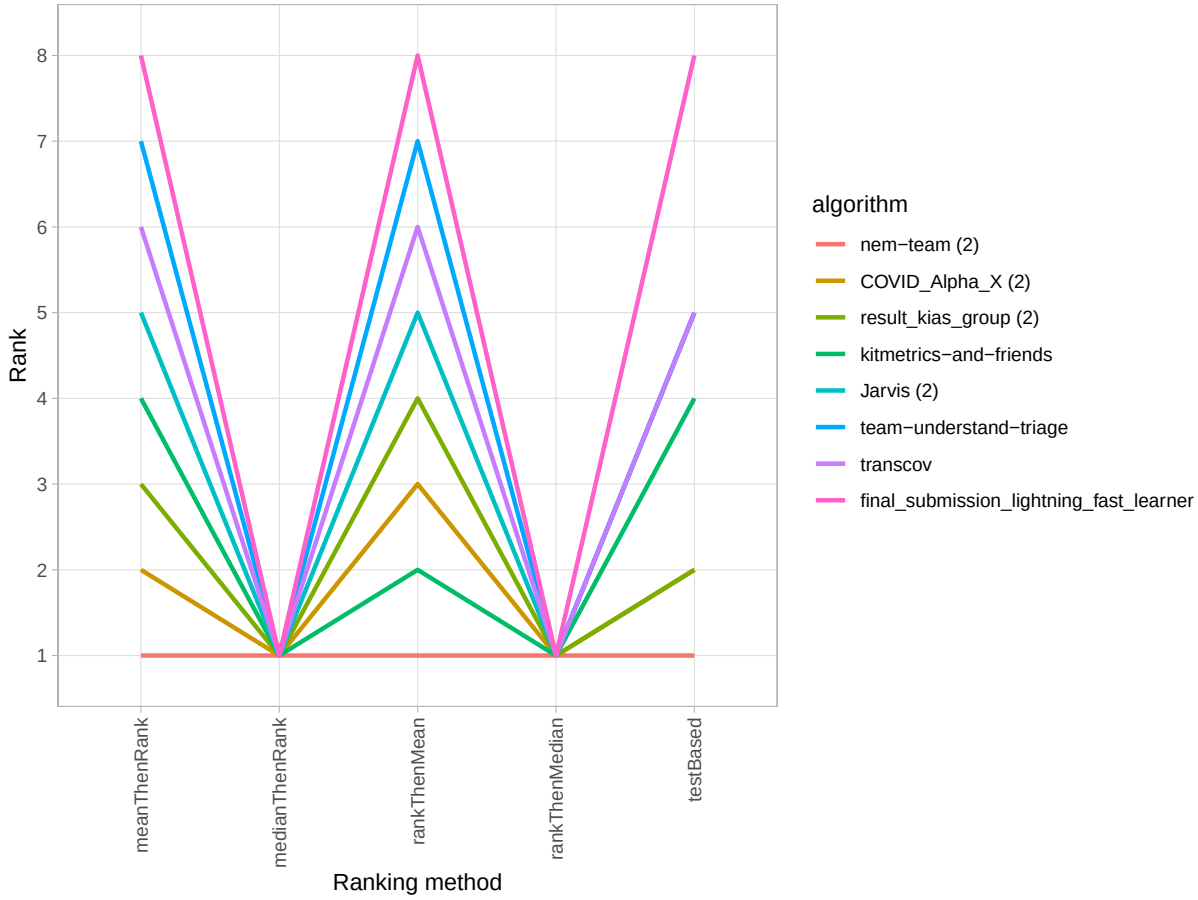
3.3 *Significance maps* for visualizing ranking stability based on statistical significance

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.



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 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/.