Ranking Scheme for MSD 2018

(Lena Maier-Hein, Björn Menze, Annika Reinke, Annette Kopp-Schneider)

There will be one ranking per metric (DSC: Dice Similarity Coefficient; NSD: Normalized Surface Distance) separately for both the training tasks and the mystery tasks:

1. A so-called *significance score* is determined for each algorithm a separately for each task (i.e. sub-challenge) c_i and metric $m_j \in$ {DSC, NSD} and referred to as $s_{ij}(a)$. It is computed as follows:
    - Performance assessment per case: Determine performance $m_j(a_l, t_{ik})$ of all algorithms a_l, $(l = 1, \dots, N_A)$ for all test cases t_{ik}, $(k = 1, \dots, N_i)$ where N_i is the number of test cases in competition c_i and N_A is the number of competing algorithms.
        - In case of N/A value: set $m_j(a_l, t_{ik})$ to worst possible value (i.e. 0 for DSC and NSD).
    - Statistical tests: perform all pairwise comparisons between algorithms $(a_l, a_{l'})$ with the values $m_j(a_l, t_{ik}) - m_j(a_{l'}, t_{ik})$ $(k = 1, \dots, N_i)$ using Wilcoxon signed rank test.
    - Significance scoring: $s_{ij}(a_l)$ then equals the number of algorithms performing significantly worse than a_l according to the test ($\alpha = 0.05$).
    - Significance ranking: The ranking is computed from the scores $s_{ij}(a_l)$ (shared places possible) with the highest score corresponding to the best algorithm(s) (rank 1).
        - If a task has multiple sub-tasks, the ranking scheme is applied separately to each sub-task, and the final ranking is computed from the mean significance ranks.
2. The final ranking for the training tasks and mystery tasks is computed from the mean significance ranks of all algorithms, where the mean is determined from the seven training tasks and three mystery tasks, respectively.

Further reading:

[1] Lena Maier-Hein*, Matthias Eisenmann*, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P. Bradley, Aaron Carass, Carolin Feldmann, Alejandro F. Frangi, Peter M. Full, Bram van Ginneken, Allan Hanbury, Katrin Honauer, Michal Kozubek, Bennett A. Landman, Keno März, Oskar Maier, Klaus Maier-Hein, Bjoern H. Menze, Henning Müller, Peter F. Neher, Wiro Niessen, Nasir Rajpoot, Gregory C. Sharp, Korsuk Sirinukunwattana, Stefanie Speidel, Christian Stock, Danail Stoyanov, Abdel Aziz Taha, Fons van der Sommen, Ching-Wei Wang, Marc-André Weber, Guoyan Zheng, Pierre Jannin*, Annette Kopp-Schneider* (*: shared first/senior authors), “Is the winner really the best? A critical analysis of common research practice in biomedical image analysis competitions”, arXiv:1806.02051, 2018