Decision Making and Analysis
https://ojs.luminescience.cn/DMA

LUMINESCIENCE
https://luminescience.cn

Original Research

Framework for fair model comparison as standard for multicriteria decision making

Károly Héberger

Plasma Chemistry Research Group, Institute of Materials and Environmental Chemistry, HUN-REN Research Centre for Natural Sciences, Centre of Excellence, Hungarian Academy of Sciences, Magyar tudósok krt. 2, Budapest H-1117, Hungary

Corresponding to: Károly Héberger, E-mail: heberger.karoly@ttk.hu

Abstract: Choosing an appropriate or best multi-criteria decision-making (MCDM) method remains challenging as more than 200 are available. Different data structures and disciplinary practices complicate selection, because methods and weighting schemes can produce widely divergent rankings. A consensus framework is proposed guided by parsimony, the bias–variance trade-off, complementary theoretical foundations, statistical testing, and balanced ranking characteristics. Seven methods are selected: sum of ranking differences (SRD); sums of Euclidean and Manhattan distance differences (DnE and DnM); summed SRD scores from, with each alternative serving once as the benchmark and a heatmap (HM) is produced; the Generalized Pair Correlation Method coupled with conditional Fisher's exact test and Williams' *t*-test, both in probability-weighted ordering form (CEpW and WtpW); and the benchmark-agnostic Deep Ranking Analysis by Power Eigenvectors (DRAPE). Three case studies with different data structures illustrate the framework. Mean ranks determine the consensus ordering of alternatives, while standard deviations describe variation across methods. Wilcoxon matched-pairs tests and Kendall's tau distances assess pairwise differences, and Friedman Analysis of Variance (ANOVA) evaluates overall differences among alternatives. The seven MCDM methods are ranked using three separate procedures: the sum of Kendall's distances, adjacent absolute rank differences, and leave-one-method-out validations. These procedures produce highly different orderings in all three case studies, revealing no consistent hierarchy among the seven methods. The pattern is reassuring: none is consistently identified as a candidate for exclusion from the consensus or from the standardization. Although the present work focuses on ranking, the framework can also support clustering, sorting, resilience assessment, and choice. The ensemble combines complementary approaches and statistical testing, offering a basis for standardization. Other MCDM methods, including Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), VlseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR), Analytic Hierarchy Process (AHP), etc. may be included cautiously, provided the guiding principles are preserved. Using more than 13 methods is not recommended because of rank degeneration.

Keywors: Method comparison, Ranking, Uncertainty estimation of ranks, Randomization test, Cross-validation, Wilcoxon matched pair test, Williams *t*-test

1. Introduction

Currently, more than 200 methods have been developed for Multi-Criteria Decision-Making (MCDM), also called Multi-Criteria Decision Analysis (MCDA) or even Multi-Attribute Decision-Making (MADM), and selection from

Received: Jul. 21, 2026; Revised: Sept. 15, 2026; Accepted: Nov. 11, 2026; Published: Nov. 26, 2026
Copyright © 2026 Károly Héberger.
DOI: https://doi.org/10.55976/dma.420261654257-283
This is an open-access article distributed under a CC BY license (Creative Commons Attribution 4.0 International License)
https://creativecommons.org/licenses/by/4.0/

among Pareto optimal solutions. Some authors use a different term that essentially refers to the same concept: Multi-Criteria Decision Analysis (MCDA).

Taherdoost and Madanchian [1] have reviewed the Decision-Making (DM) process, decomposed it into several steps, differentiated models between their bases: *i.e*. DM based on number of criteria and number of decision makers. They also ordered different types of MCDM methods according to their usage frequency in SCOPUS (Elsevier).

Cinelli *et al*. [2] elaborated a powerful taxonomy and made distinctions of *ranking, sorting, clustering and choice*. They produced a Multiple Criteria Decision Analysis Method Selection Software (MCDA-MSS), which is a Decision Support System (DSS). *Clustering and sorting* were reviewed by Barrera *et al*. [3] "… to assess the quality of preference-ordered customer groups." The "parameters have been adapted for contexts with thousands of alternatives. … The multicriteria system proposed allows to segment thousands of customers into ordered categories by preferences according to company strategies. The segments generated are more homogeneous, robust and understandable" [3]. Wątróbski *et al*. [4] outlined a framework how to select the best MCDM from among 49 frequently used ones. Baydas *et al.* restricted their success analysis of maximum normalization (COmbinative Distance-based ASssessment (CODAS)) and fuzzy approaches to economic data [5]. Although various frameworks and decision-support systems have been developed to facilitate MCDM method selection [1-5], selecting an appropriate method remains challenging because the characteristics of the decision problem and the available data may not always be fully understood or specified in practice.

The other crucial problem is the subjective character of users' preference. Namely, the selection of weights is necessarily subjective even without knowing it. Defining objective weights seemed to be a successful way to overcome subjectivism. Objective weighting means repeatable, algorithm-based approaches producing numerical data in the decision matrix. Wang and Lee [6] have compared subjective and objective weightings. Mukhametzyanov [7] has discussed the peculiarities of various objective methods and questioned the rationality of all objective methods for evaluating the criteria weights. A comprehensive analysis has been carried out on the most frequently used MCDM weighting methods [8]. Hybrid methods were attempted to unite objective and subjective weighting schemes with all their advantages and disadvantages [9-11]. Apart from the non-trivial selection of a proper weighting scheme, it should be emphasized that there are so many "objective" approaches with different rankings and various best alternatives that the selection from among them would necessarily be subjective.

As a solution, some authors suggest choosing weight definition methods according to characteristics and information structure in the data to reduce the possibility of subjective judgment [12, 13].

Another way is to apply more methods and find their consensus. Vinogradova [14] proposed a new approach for merging the results of all MCDM methods into a single value while considering the suitability of the methods for the task to be solved; *i.e*. she suggested a weighted consensus approach. In image analysis, each raw classifier score was converted to a probability value with the help of statistical Extreme Value Theory. These probabilities were fused to perform more accurate multiattribute searches, as well as enabled attribute-based similarity searches. The normalization of raw scores into a multi-attribute space makes their comparisons and combinations commensurate [15].

Generally, the rank average of several MCDM methods is suggested for the consensus. However, rank aggregation introduces a novel way of ambiguities. Namely, many aggregation methods exist and the disagreement about which method is to be chosen still seems to be a chronic problem. Dwork *et al*. examined the aggregation methods from a historical perspective [16]. Rank-aggregation approaches can roughly be classified into distribution-based, heuristic, and stochastic optimization methods, where Thurstone's scaling and its extensions constitute the classical, distribution-based approaches [17].

Knowing that:

i) classical aggregation methods do not take into account the direction of alternatives,

ii) restrictions on economic or financial problems cannot cover large number of deviating problems,

iii) a formal selection cannot be (Pareto) optimal for data of deviating structure,

iv) statistical testing is forgotten to determine significance almost always:

The next question remains yet:

What should be done if the different MCDM methods provide various rankings of alternatives? (By the way, this happens always except in trivial cases.)

Hence, the objectives of this study are to kill three birds in one stone: i) make selection of complementary MCDM methods for consensus while avoiding the weighting problem ii) provide a statistically justified framework for rank aggregation, and iii) assign reliable uncertainties to rank of alternatives.

2. Methods, justification and results

First an input matrix is formed, alternatives arranged in the columns and criteria in the rows.

Table 1 summarizes the recommended MCDM techniques for consensus.

Table 1. Advantages, disadvantages, bases and statistical background of the selected MCDM methods

Methods	Advantages	Disadvantages	Based on	Backed by
Sum of ranking difference (SRD) [18, 19]	Robust, stable, conservative, scale invariant, ties allowed, avoids the rank reversal problem (fool proof)	Rank degeneration (Not all alternatives are distinguished always), constant benchmark is not rankable.	One benchmark, Manhattan distance of rank differences	Randomization test [20] (*y*-scrambling, permutation test), *X*-randomization Wilcoxon and Alpaydin tests, cross-validation [21]
Sum of Euclidean distance differences, DnE, [22]	Finer distinction than SRD, ties possible, but seldom	Non-robust, scale sensitive, unidirectional alternatives needed	One benchmark, Euclidean distance differences	By analogy of SRD [22]
Sum of Manhattan distance differences, DnM, [22]	Finer distinction than by SRD, ties possible, but seldom	"robust", but sensitive to outliers, scale sensitive, unidirectional alternatives needed	One benchmark, Sum of absolute Manhattan distance differences	By analogy of SRD [22]
Generalized pair correlation method, GPCM [23] ordering by number of probability weighted wins minus probability weighted losses, CepW	Independent train of thought, easily understandable, general, robust but sensitive to y-outliers Insensitive to direction of alternatives	No decision can be made, if there are no significant differences by Conditional Fisher's exact test (CE), Prediction is not examined	One benchmark, A series of 2 × 2 contingency tables, pairwise associations	Conditional Fisher's exact test (CE) [24]
Generalized pair correlation method, GPCM [23] ordering by number of probability weighted wins minus probability weighted losses, WtpW	Independent train of thought, easily understandable, general, robust but sensitive to *y*-outliers	No decision can be made, if there are no significant differences by Wt, Prediction is not examined. Sensitive to direction of alternatives	One benchmark, pairwise Pearson correlation coefficients on the condition that the alternatives themselves are also correlated	Williams *t*-test (Wt) [24]
Heatmap like distance matrix, arranged in increasing sums of alternatives sum, HM	See at SRD, but finer distinction than Pearson and Spearman correlation coefficients [25]	See at SRD, many benchmarks may provide biased distances	All alternatives serve as benchmark once and only once, Comparison with One VAriable (COVA) type [25]	Randomization test preserved, Pairwise testing of sum of alternatives possible
Deep Ranking Analysis by Power Eigenvectors (DRAPE) [26]	Unsupervised, no benchmark is needed	Unstable ranking by thresholding, zero (very small) eigenvectors require a correction factor	Tournament table (dominance matrix), power-weakness ratio Power-Weakness Ratio (PWR) [26]	No statistical testing, but PWR values are subjected to principal component analysis (PCA)

The techniques in Table 1 are generally applicable and correspond to principle of parsimony. The sum of Euclidean distance differences (DnE) and the sum of Manhattan distance differences (DnM) usually provide highly similar ordering enhancing rank stability. The Conditional Fisher's exact test with probability-weighted ordering (CEpW) and William's t-test with probability-weighted ordering (WtpW) are based on totally different ways of thinking, and they produce explanatory alternatives *i.e.*, prediction is not considered. HM uses different benchmarks and ensures high variance in the ranking. Deep Ranking Analysis by Power Eigenvectors (DRAPE) is unsupervised, may produce more orderings by thresholding. All MCDM methods have their own bias. However, the biases follow normal distribution and cancel each other out. They are well-balanced and complementary to each other. Different bases and correct statistical tests justify their application.

The entire framework is visualized in Figure 1. The MCDM methods are distributed into four groups: A (SRD, DnE and DnM), B (CEpW, WtpW), C (HM) and D (DRAPE). Groups A, B and D need benchmarks which are not always readily available. In those cases, row-averages, *e.g.* means of criteria, are valid choices. Maximum Likelihood Principle (MLP), which is a naturalistic principle of inference. MLP requires no contrived assumptions; it is an intuitive law of informational compatibility. The way of thinking here is similar to that for proficiency testing (round-robin tests): The systematic errors (biases) of

different laboratories or different methods follow normal distribution with zero mean and constant variance. The base of proficiency testing is a *well-substantiated empirical finding* in analytical chemistry. Variances of random errors and that of systematic errors (biases) are additive [27], *i.e.* the laboratories and methods are compared using *z*-scores with the assumption of normality. The analogy is obvious: MCDM methods can be evaluated in such a way as labs and different methods (see the Eq. A1_4 in the Appendix 1).

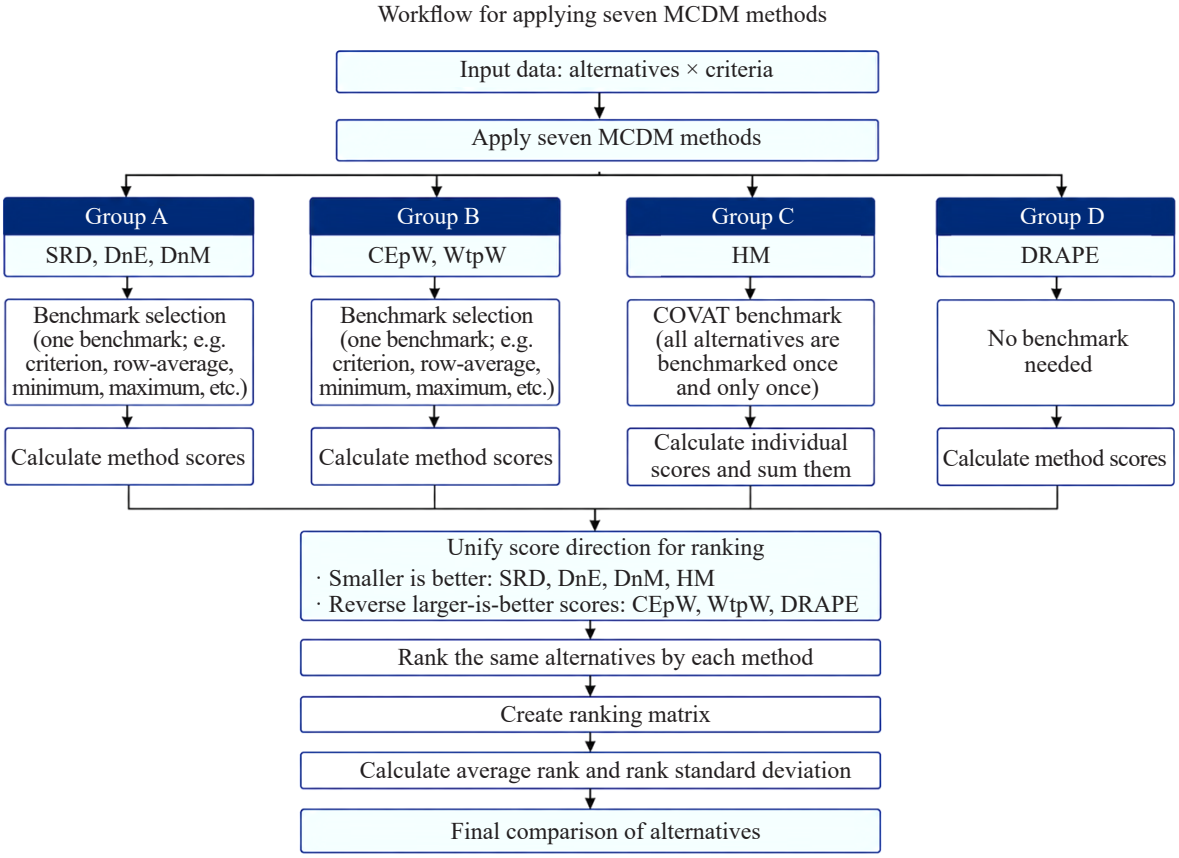


Figure 1. General scheme of the selected seven MCDM methods

There are other data fusion options to be completed according to the data structure: Row-minima is the best choice for error-like terms whereas row-maxima is suitable for correlation, similarity and accuracy like terms. Hybrid fusion is also possible, if error and correlation like terms are present in the input matrix for the Sum of Ranking Differences (SRD). As DnE and DnM do not apply rank transformation, they need to be put on the same scale, but calculation of Euclidean and Manhattan distance from a constant benchmark is also feasible. No data fusion is meaningful for HM, where the benchmarks are to be changed and the inner data structure is revealed. Neither benchmark nor scaling or data fusion is needed for DRAPE. After calculation of means and standard deviations of ranks for alternatives, the consensus step remains only, the final comparison can be done by pairwise statistical testing.

All statistical tests, box and whisker and parallel coordinate plots have been completed and drawn by the accredited Statistica™ program package (version 14.01.25 TIBCO software Inc. Palo Alto, CA, USA.).

2.1 Case study No 1

Elaiwat *et al*. [28] have compared three algorithms, namely, Neural Networks (NN), Support Vector Machines (SVM), and Naïve Bayes (NB), coupled to three performance indicators (Acc, Cohen K and F1-score), *i.e*. the alternatives for three experiments (Exp1, Exp2, and Exp3, this experimental setup was designed to systematically test three different data structures to determine the most effective way to predict activity cliffs) and for three different settings i) random split with tokenized sequences, T5, ii) random split with numeric encodings (Morgan fingerprint), T6, and target-based split with numeric encodings (Morgan fingerprint).

Figure 2 shows the SRD ordering of fused algorithms & performance indicators.

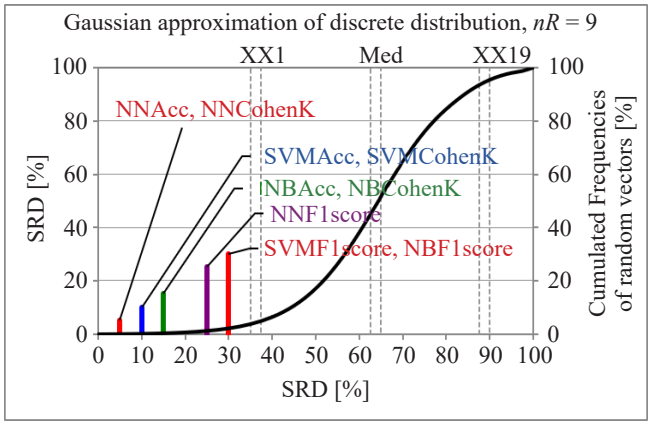


Figure 2. SRD analysis of Elaiwat *et al.*'s tables 5, 6 and 7. Sum of ranking differences (SRD values, scaled between 0 and 100) ordered the combinations of algorithms and performance indicators. Scaled SRD values are plotted on *x* axis and left *y* axis, right *y* axis shows the relative frequencies of random ranking distribution: black curve is the Gaussian approximation of discrete frequencies. The probability values are as follows (right *y* axis): XX1 and XX19 are 5% and 95% of first kind errors, respectively; Med = Median

The regularities can easily be seen: i) accuracy and Cohen's kappa are in the first group (SRD < 20), F1-score in the second one (SRD > 20); ii) Neural network < Support Vector Machine < Naïve Bayes ordering of algorithms correspond to best to worst. (The smaller the better in case of SRD). This corresponds to empirical investigations, SVM is better than Naïve Bayes *c.f., e.g.* ref. [29].

All combinations are better than random ranking (all lines are located left to XX1 dotted lines).

Figure 3 shows much finer distinction for the combination of algorithms & performance indicators.

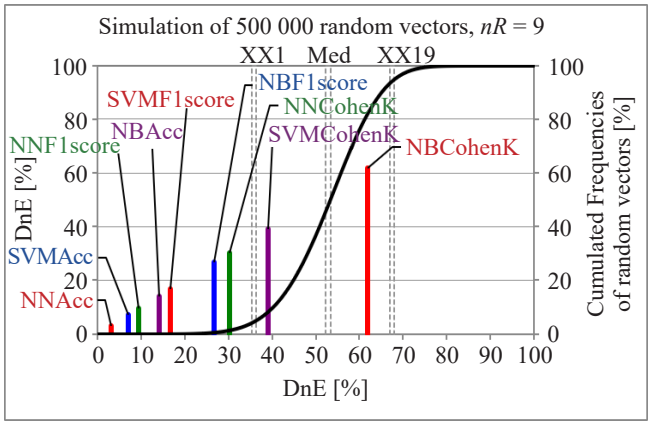


Figure 3. DnE analysis of Elaiwat *et al.*'s tables 5, 6 and 7. Sum of Euclidean distance differences (DnE values, scaled between 0 and 100) ordered the combinations of algorithms and performance indicators. Scaled DnE values are plotted on *x* axis and left *y* axis, right *y* axis shows the relative frequencies of random ranking distribution: black curve is the Gaussian approximation of discrete frequencies. The probability values are as follows (right *y* axis): XX1 and XX19 are 5% and 95% of first kind errors, respectively; Med = Median

The pattern is not so unambiguous as for Figure 2, but the degeneracies (equal SRD scores, *i.e.* tied ranks) are resolved, and Cohen's kappa is placed in the last (worst) three positions. Two combinations (SVMCohenK and NBCohenK) are indistinguishable from random ranking. Their ordering, however, makes sense.

Figure 4 the DnM analysis exhibits different patterns than the DnE analysis (Figure 3), *e.g.* The DnM values are hardly different for NBAcc and SVMF1score, and not large distance is between NNF1score and SVMAcc. (In the DnE plot there are well distinguished). However, the ordering is almost the same. One rank reversal can be observed between NNF1score and SVMAcc. The same two combinations (SVMCohenK and NBCohenK) are indistinguishable from random ranking as in Figure 3.

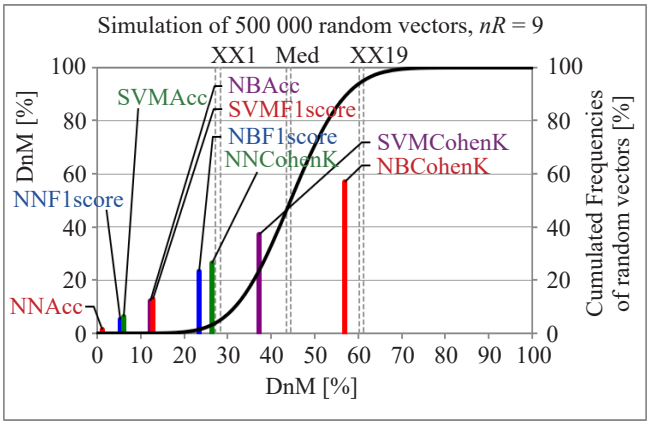


Figure 4. DnM analysis of Elaiwat *et al.*'s tables 5, 6 and 7. Sum of absolute Manhattan distance differences (DnM values, scaled between 0 and 100) ordered the combinations of algorithms and performance indicators. Scaled DnM values are plotted on *x* axis and left *y* axis, right *y* axis shows the relative frequencies of random ranking distribution: black curve is the Gaussian approximation of discrete frequencies. The probability values are as follows (right *y* axis): XX1 and XX19 are 5% and 95% of first kind errors, respectively; Med = Median

The results of generalized pairwise correlation methods (CEpW and WtpW) can be found in the sheets of CEpW and WtpW, respectively of the supplementary material (CaseStudyNo1T5_T7all). The ranking was enforced by the difference of lines 21 and 22: Number of probability weighted wins (pWinner) minus number of probability weighted losses (pLoser) for both of CEpW and WtpW, whereas column E contains the theoretical probabilities (α (theor)). If their values are smaller than the predefined 0.05, the comparison is to be considered as significant. All 2 × 2 contingency tables are given below line 27 for CEpW and all Pearson correlation coefficients are given below line 27 (column D) for WtpW to complete the Williams *t*-test.

Relative HeatMap									
colours:	x < =	4.0	12.0	20.0	28.0	36.0			
		8.0	16.0	24.0	32.0	40.0			
	NNAcc	NNCohenK	SVMAcc	SVMCohenK	NBCohenK	NNF1score	NBAcc	SVMF1score	NBF1score
NNAcc	0.00	0.00	5.00	15.00	10.00	25.00	20.00	30.00	35.00
NNCohenK	0.00	0.00	5.00	15.00	10.00	25.00	20.00	30.00	35.00
SVMAcc	5.00	5.00	0.00	15.00	5.00	25.00	25.00	30.00	30.00
SVMCohenK	15.00	15.00	15.00	0.00	20.00	15.00	25.00	20.00	25.00
NBCohenK	10.00	10.00	5.00	20.00	0.00	30.00	20.00	35.00	25.00
NNF1score	25.00	25.00	25.00	15.00	30.00	0.00	40.00	10.00	35.00
NBAcc	20.00	20.00	25.00	25.00	20.00	40.00	0.00	40.00	30.00
SVMF1score	30.00	30.00	30.00	20.00	35.00	10.00	40.00	0.00	30.00
NBF1score	35.00	35.00	30.00	25.00	25.00	35.00	30.00	30.00	0.00
Sum	140.00	140.00	140.00	150.00	155.00	205.00	220.00	225.00	245.00

Figure 5. Heatmap ordering of the combinations for algorithms and performance indicators (alternatives) for Elaiwat *et al.*'s data. Relative color coding was used, see header. The inner, pairwise-distance structure is not the same as the overall sum ordering

Figure 5 visualizes the SRD scores in a COVAT manner, HM. All alternatives served as benchmark once and only once. The alternatives are arranged in sum of (rows and) columns. The matrix is symmetric. The best alternative (NNAcc) corresponds to the previous rankings (Figures 2-5 and CEpW sheet (there is only tiny the difference between NNCohenK and NNAcc, WtpW sheet) in the supplementary material, CaseStudyNo1T5_T7all), as expected. Inner rearrangement can be observed according to the data structure. There is no doubt, NNAcc and NNcohenK are the most similar (smallest, *i.e.* zero rank distance from each other. SVMAcc is not far from them, but the same distance is observed between SVMAcc and NBCohenK, SRD = 5). The largest individual difference is in between NNF1score and NBAcc.

The calculation of the tournament table and power-

weakness ratio (PWR) for DRAPE analysis is reproducibly described in part 2.1 and 2.2 of ref. [29]. The ranking of alternatives can be found with some explanations in the DRAPE sheet in the supplementary material, CaseStudyNo1T5_T7all).

The overall ranking of alternatives by all seven MCDM methods can be found in the 7MCDMranks sheet in the supplementary material, CaseStudyNo1T5_T7all, together with raw values. As is indicated in Figure 1, the direction of ranking should be unified: SRD, DnE, DnM and HM, the smaller the better, whereas CEpW, WtpW and DRAPE, the larger the better; *i.e.* the last three MCDM methods should be reversed.

In the final step, rank averages and standard deviations are calculated. Both the two sample *t* test, and the Mann-Whitney test provide the same result exactly. It is visualized in Figure 6.

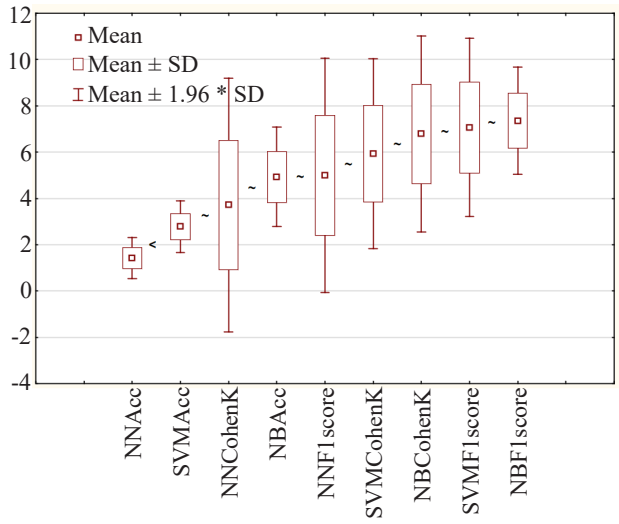


Figure 6. Box and whisker plot of rank averages and uncertainties for nine alternatives. Pairwise statistical testing suggests one significant difference (<) at the 5% level; all others are not significant (~). Friedman ANOVA, Chi Squares statistics (n = 7, degree of freedom = 8) = 31.34, limit probability of significance is p = 0.00012. The overall comparison of all alternatives is significant

Pairwise comparison of alternatives has found only one significant difference (between NNAcc and SVMAcc), all other pairwise comparisons are not significant. Overall comparison of alternatives by Friedmann Analysis of Variance (ANOVA) suggests significance at the p = .00012 level. (χ^2 statistics (n = 7, df = 8) = 31.3, whereas Coefficient of Concordance is 0.560, and the average rank is 0.486).

Alternatively, a one-way ANOVA can be carried out (data not shown). The variances are inhomogeneous (see also Figure 6). *Post hoc* tests provide three to five overlapping groups (Scheffé and Fisher's Least Significant Difference (LSD) test, respectively), whereas Bonferroni and Tukey's Honestly Significant Difference (HSD) test founds four overlapping groups.

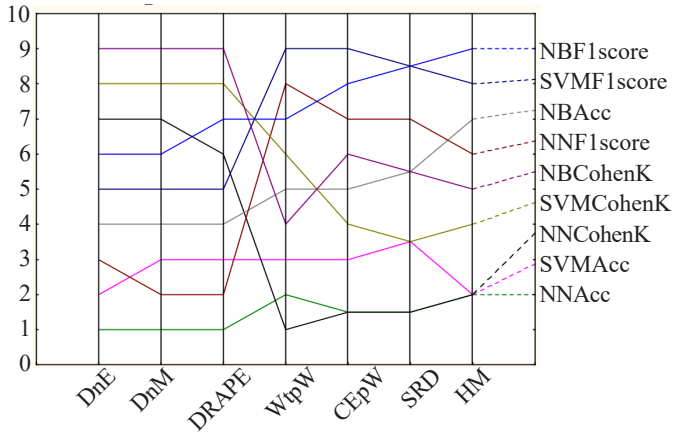


Figure 7. Parallel coordinate plot, the MCDM methods are rearranged according to the minimal sum of adjacent absolute rank differences across all alternatives. The largest turbulence is between DRAPE and WtpW

The conclusions for Elaiwat *et al.*'s tables (5, 6 and 7) are moderate. Sensitive non-parametric tests show some interplay of algorithms and performance indicators. Without doubt the combination of NNAcc is the best, the rank averages are increasing but their grouping is not unambiguous.

A parallel coordinate plot is suitable for the ordering of MCDM methods (Figure 7), if they are rearranged according to the minimal sum of adjacent absolute rank differences across all alternatives. It is interesting to note that there is no "superiority" ordering them, which is reassuring. (Note that the reverse ordering is equally optimal). No left to right increase of number of crossings can be detected. The timely evolution of methods does not meet the ordering, either: (CEpW and WtpW have been elaborated first, then, SRD, later HM, DRAPE and finally DnE and DnM).

2.2 Case study No 2

Bernal *et al.* [30] have compared four different types of tree-based algorithms, Extra Trees (ET), Random Forest (RF), Gradient Boosting Machines (GBM), and XGBoost (XGB) coupled to three fragments (ECFP4, MACCS, custom) and descriptors (RDKit), *i.e.* the alternatives (denoted by ET_ECFP4, ET_MACCS, ET_custom, ET_RDKit, RF_ECFP4, RF_MACCS, RF_custom, RF_RDKit, GBM_ECFP4, GBM_MACCS, GBM_custom, GBM_RDKit, XGB_ECFP4, XGB_MACCS, XGB_custom and XGB_RDKit), on the basis of six performance indicators (Accuracy, Matthews correlation coefficient (MCC), Precision, Recall, ROC-AUC and F1-score), and for three data sets (DU-145, LNCaP, and PC3), for details and explanations see ref. [30].

Figure 8 shows the SRD ordering of combination of algorithms with fingerprints and descriptors.

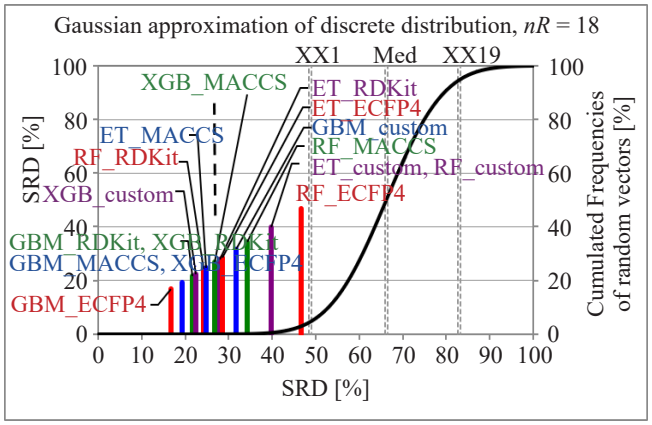


Figure 8. SRD analysis of Bernal *et al.*'s table 2. Sum of ranking differences (SRD values, scaled between 0 and 100) ordered the combinations of algorithms with fragments and (RDkit) descriptor. Scaled SRD values are plotted on *x* axis and left *y* axis, right *y* axis shows the relative frequencies of random ranking distribution: black curve is the Gaussian approximation of discrete frequencies. The probability values are as follows (right *y* axis): XX1 and XX19 are 5% and 95% of first kind errors, respectively; Med = Median

Some regularities can be seen: Higher density of lines (20 < SRD < 30) can be observed; RF_ECFP4 seems to be an outlier visually. The fingerprint-descriptor pattern is not unambiguous: all of them occur in the first and in the second part of the set (see the occurrence in the left and right parts to the dotted vertical line in Figure 8). However, gradient boosting (GBM and XGB) dominates the first part of algorithms (GBM and XGB are placed in the six first positions), whereas RF and ET dominate the second part (right to the dotted line). This hierarchy of algorithms is in agreement of empirical findings [29, 31]. Remember: The smaller the better in case of SRD.

The frequency of algorithms corresponds best to worst. Although all combinations are better than random ranking (all lines are located left to XX1 dotted lines), the large empty range up to SRD = 16 reveals that the selected maximum benchmark is not the best. Ties and outliers (biased estimations) should be present.

Figure 9 shows finer distinction and grouping for the combination of algorithms with fingerprints and descriptor.

Here, the degeneracies (equal SRD scores, *i.e.* tied ranks) are resolved, some Euclidean distance differences are, indeed, very small hence the grouping tendency. The custom fingerprints dominate the last part (three times out of four), whereas RDKit descriptor can be found in the first six positions. Gradient boosting remains the best algorithm (GBM and XGB); they are located in the best three positions, though RF_RDK goes to position 4. The dotted lines would be placed between ET_ECFP4 and ET_MACCS, but their DnE values are almost the same; hence, it is removed. All combinations are better than random ranking (all lines are located left to XX1 dotted lines).

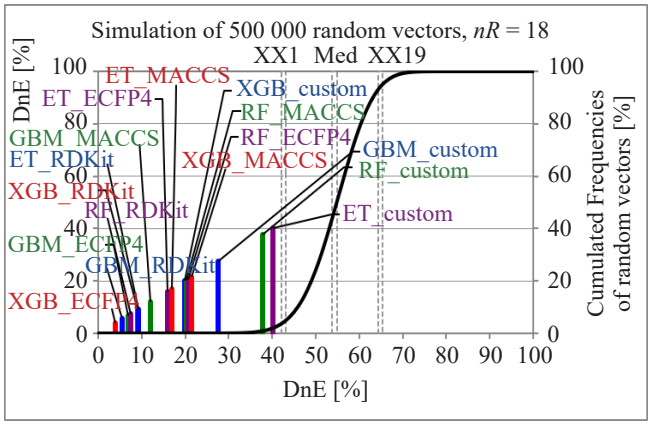


Figure 9. DnE analysis of Bernal *et al.*'s table 2. Sum of Euclidean distance differences (DnE values, scaled between 0 and 100) ordered the combinations of algorithms with fragments and (RDkit) descriptor. Scaled DnE values are plotted on *x* axis and left *y* axis, right *y* axis shows the relative frequencies of random ranking distribution: black curve is the Gaussian approximation of discrete frequencies. The probability values are as follows (right *y* axis): XX1 and XX19 are 5% and 95% of first kind errors, respectively; Med = Median

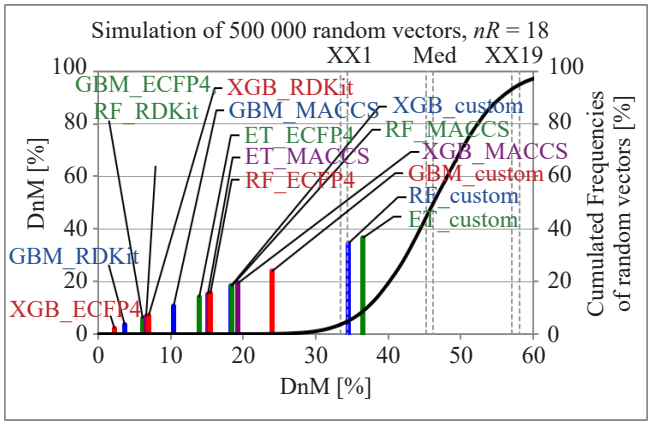


Figure 10. DnM analysis of Bernal *et al.*'s table 2. Sum of absolute Manhattan distance differences (DnM values, scaled between 0 and 100) ordered the combinations of algorithms with fragments and (RDkit) descriptor. Scaled DnM values are plotted on *x* axis and left *y* axis, right *y* axis shows the relative frequencies of random ranking distribution: black curve is the Gaussian approximation of discrete frequencies. The probability values are as follows (right *y* axis): XX1 and XX19 are 5% and 95% of first kind errors, respectively; Med = Median

Figure 10 shows that the DnM analysis is characterized by better distinguishing ability than the DnE one. The pattern is also different: two lines (RF_custom and ET_custom) are indistinguishable from random ranking (they are located right to the XX1 lines, *i.e.* from the 5% error limit. (Figure 10). Their ordering, however, may make sense. XGB and GBM algorithms are placed in the first three positions.

The results of generalized pairwise correlation methods (CEpW and WtpW) can be found in the sheets of CEpW and WtpW, respectively of the supplementary material (CaseStudyNo2all). The ranking was enforced in the difference of lines 21 and 22, again, accidentally: Number of probability weighted wins (pWinner) minus number of probability weighted losses (pLoser) for both of CEpW and WtpW, whereas column E contains theoretical probabilities (α (theor)). If their values are smaller than the predefined 0.05, the comparison is to be considered as significant. All 2 × 2 contingency tables are given below line 27 for CEpW and all Pearson correlation coefficients are given below line 27 (column D) for WtpW to complete the Williams *t*-test. CEpW results emphasize earlier observations: gradient boosting (GBM and XGB) techniques dominate the first part of algorithms (GBM and XGB algorithms are placed in the six first positions), whereas custom fingerprint is fairly in the worst places (three out of four cases) right part of lines.

Figure 11 visualizes the SRD scores in a COVAT manner, abbreviated as HM. All alternatives served as benchmark once and only once. The alternatives are arranged in sum of (rows and) columns. The matrix is symmetric. The best alternative is without doubt the one

denoted by GBM_ECFP4. It corresponds to the previous rankings (Figures 8-10 as well as CEpW and WtpW sheets in the supplementary material, CaseStudyNo2all). The worst alternative changes (in most cases ET_custom, but custom fingerprint dominates the last part of ordering. Inner rearrangement can be observed according to the data structure. Interestingly, the smallest pairwise SRD distance is RF_Custom and ET_custom (*i.e.* they are the most similar) but the distance between XGB_Custom and XGB_RDKit or that of GBM_RDKit and XGM_RDKit are not much higher (SRD < 10). The largest individual difference is in between RF_ECFP4 and GBM_custom.

Relative HeatMap																
colours:	x < =	6.0	17.9	29.8	41.7	53.7										
		11.9	23.9	35.8	47.7	59.6										
	GBM_ECFP	RF_RDKit	XGB_ECFP4	ET_MACCS	XGB_RDKit	GBM_RDKit	XGB_custom	ET_RDKit	ET_ECFP4	GBM_MACCS	RF_MACCS	RF_custom	ET_custom	XGB_MACCS	GBM_custom	RF_ECFP4
GBM_ECFP4	0.00	24.22	8.70	30.43	13.66	12.42	16.77	29.19	36.65	22.36	39.75	46.58	47.20	28.57	27.33	49.69
RF_RDKit	24.22	0.00	30.43	26.09	30.43	29.19	31.68	16.15	21.74	34.16	27.33	24.22	24.84	39.75	41.61	33.54
XGB ECFP4	8.70	30.43	0.00	29.19	18.01	17.39	18.63	35.40	34.16	16.77	37.89	48.45	48.45	22.36	30.43	51.55
ET_MACCS	30.43	26.09	29.19	0.00	34.78	34.16	37.27	24.84	14.91	29.19	22.36	28.57	28.57	34.78	44.72	40.37
XGB_RDKit	13.66	30.43	18.01	34.78	0.00	8.07	6.21	31.06	42.24	29.81	47.20	44.72	44.10	37.89	17.39	55.28
GBM_RDKit	12.42	29.19	17.39	34.16	8.07	0.00	9.32	29.81	41.61	30.43	47.20	45.96	46.58	37.89	21.12	54.04
XGB_custom	16.77	31.68	18.63	37.27	6.21	9.32	0.00	34.16	41.61	27.33	47.20	47.20	47.20	36.02	17.39	56.52
ET_RDKit	29.19	16.15	35.40	24.84	31.06	29.81	34.16	0.00	29.81	42.86	34.78	26.71	27.95	47.83	36.65	29.81
ET_ECFP4	36.65	21.74	34.16	14.91	42.24	41.61	41.61	29.81	0.00	28.57	13.04	22.36	21.74	37.89	54.04	38.51
GBM_MACCS	22.36	34.16	16.77	29.19	29.81	30.43	27.33	42.86	28.57	0.00	34.78	48.45	47.83	14.91	37.89	59.63
RF_MACCS	39.75	27.33	37.89	22.36	47.20	47.20	47.20	34.78	13.04	34.78	0.00	24.84	24.84	36.65	57.14	31.06
RF_custom	46.58	24.22	48.45	28.57	44.72	45.96	47.20	26.71	22.36	48.45	24.84	0.00	3.73	55.90	50.93	27.33
ET_custom	47.20	24.84	48.45	28.57	44.10	46.58	47.20	27.95	21.74	47.83	24.84	3.73	0.00	56.52	51.55	27.95
XGB_MACCS	28.57	39.75	22.36	34.78	37.89	37.89	36.02	47.83	37.89	14.91	36.65	55.90	56.52	0.00	44.72	53.42
GBM custom	27.33	41.61	30.43	44.72	17.39	21.12	17.39	36.65	54.04	37.89	57.14	50.93	51.55	44.72	0.00	59.63
RF_ECFP4	49.69	33.54	51.55	40.37	55.28	54.04	56.52	29.81	38.51	59.63	31.06	27.33	27.95	53.42	59.63	0.00
Sum	433.54	435.40	447.83	460.25	460.87	465.22	474.53	477.02	478.88	504.97	526.09	545.96	549.07	585.09	592.55	668.32

Figure 11. Heatmap ordering of the combinations for algorithms with fragments and (RDkit) descriptor (alternatives) for Bernal *et al.*'s table 2. Relative color coding was used, see header. The inner, pairwise-distance structure is not the same as the overall sum ordering

The Deep Ranking Analysis by Power Eigenvectors (DRAPE), *i.e.* ranking of alternatives (Bernal *et al.*'s data) can be found with some explanations in the DRAPE sheet in the supplementary material, CaseStudyNo2all).

The overall ranking of alternatives by all seven MCDM methods can be found in the 7MCDMranks sheet in the supplementary material, CaseStudyNo2all, together with raw values. As is indicated in Figure 1, the direction of ranking should be unified: SRD, DnE, DnM and HM, the smaller the better, whereas CEpW, WtpW and DRAPE, the larger the better; *i.e.* the last three MCDM methods should be reversed.

In the final step, rank averages and standard deviations are calculated. Both the Wilcoxon test and the Mann-Whitney test provide the same result with one exception (Figure 12). It is visualized in Figure 12.

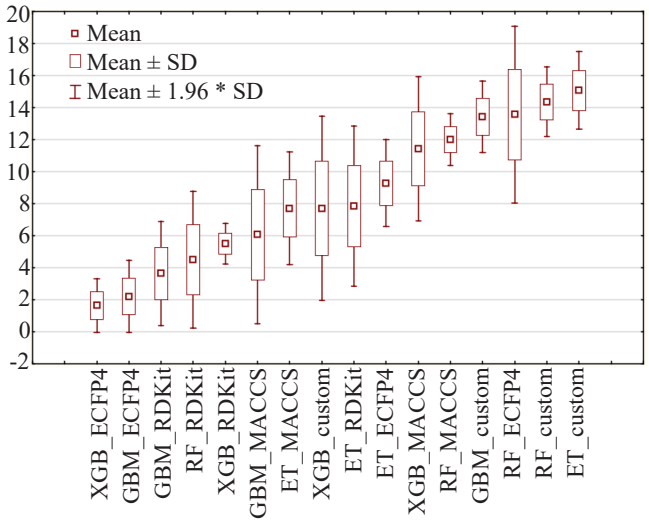


Figure 12. Box and whisker plot of rank averages and uncertainties for 16 alternatives. Pairwise statistical testing suggests no significant differences at the 5% level; However, there is a borderline comparison: ET_RDKit and ET_ECFP4, $p = 0.052$ according to Wilcoxon matched pair test (not significant just). Friedman ANOVA Chi Squares statistics ($n = 7$, degree of freedom = 15) = 88.56, limit probability of significance is $p = 0.00000$. The overall comparison of all alternatives is highly significant

None of the pairwise comparison of alternatives provided significant differences (one borderline case, see Caption to Figure 12. Overall comparison of alternatives by Friedmann ANOVA suggests significance at the p = .00000 level. (χ^2 statistics (n = 7, df = 15) = 90.1, whereas Coefficient of Concordance is 0.858, and the average rank is 0.835). A one-way ANOVA was also carried out (data not shown). The variances are inhomogeneous (see also Figure 12). *Post hoc* tests provide six to nine overlapping groups.

The conclusions for Bernal *et al.*'s table 2 are straightforward. Quasi-continuous ranking resulted with individual variances. The best alternatives can be selected unambiguously. The custom fingerprint is among the worst representations. Sensitive non-parametric tests show some interplay of algorithms and fingerprints and RDKit descriptor. The large number of *post hoc* groups suggests more significant ordering than for Case study No. 1, despite non-significance of adjacent pairwise comparisons.

A parallel coordinate plot was also drawn to order the MCDM methods (Figure 13), when they are rearranged according to the minimal sum of adjacent absolute rank differences across all alternatives. It is interesting to note that present ordering is not better than that of Case study No. 1, but completely different, which is reassuring (no superior MCDM methods appeared). (Note that the reverse ordering is equally optimal). Again, no left to right increase of number of crossings can be detected. This ordering does not coincide with the timely evolution of MCDM methods, either.

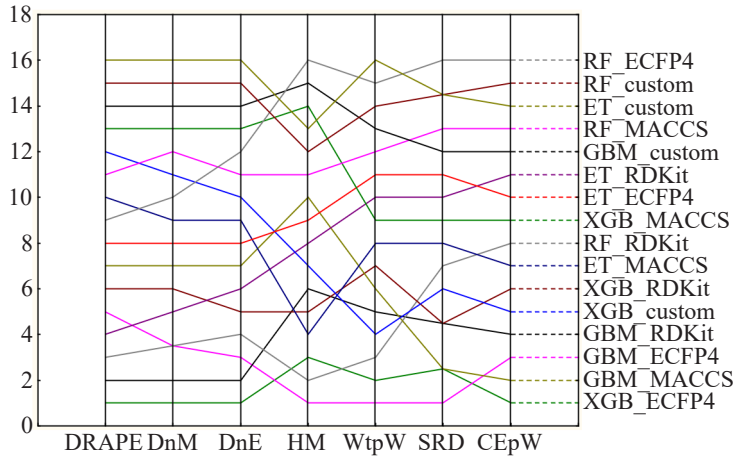


Figure 13. Parallel coordinate plot, the MCDM methods are rearranged according to the minimal sum of adjacent absolute rank differences across all alternatives. HM caused the largest turbulence; the most similar pairs are DRAPE and DnM as well as WtpW and SRD

2.3 Case study No 3

Shili Lin [32] summarized rank aggregation studies in a detailed and straightforward manner. His Table 2 compares data fusion methods for 40 items and definite ranking. The fusion possibilities were Arithmetic Mean (ARM), Median (MED) Geometric Mean (GEM), L_2-norm (Euclidean norm, L2N), Monte Carlo methods (MC1, MC2, MC3), Kendall's distance (CEK) and the Superman's distance (obviously Spearman distance, CES), details can be found in ref. [32]. In fact, the data fusion methods can hardly be discriminated against each other, especially in the first 25 ranked positions. Therefore, this data set is particularly suitable for comparison and statistical analysis (see in the next part).

Figure 14 visualized the SRD ordering of data fusion methods.

All data fusion methods surpass the randomization test. As SRD the smaller the better MC3 and CES can be a bit better, all other methods produce quasi-continuous ordering. The benchmark was the known (true) ranking; the ordering is unambiguous no ties can be observed.

The DnE analysis (Figure 15) shows somewhat finer distinction of data fusion methods.

Many of the Euclidean distance differences are, indeed, very small; no grouping tendency can be observed. What is remarkable: MED, and CEK are distant and they are the first two fusion methods—just the opposite, what can be seen in Figure 14. In fact, this is a logical consequence, as the methods can hardly be differentiated (*c.f.* sensitivity analysis of Vinogradova's data in ref. [22].

All combinations are better than random ranking (all lines are located left to XX1 dotted lines).

The DnM analysis (Figure 16) is characterized as having the same distinguishing ability as the DnE one. The pattern is somewhat different: while two lines (MED and CEK) are similar to the DnE analysis—the first two. However, MC1-MC3 and CES moved from middle position to the last ones. All data fusion methods are clearly distinguished from random ranking.

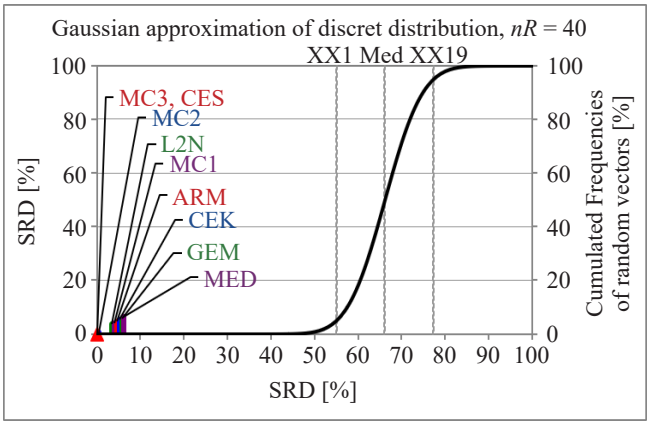


Figure 14. SRD analysis of Lin's table 2. Sum of ranking differences (SRD values, scaled between 0 and 100) ordered the data fusion methods. Scaled SRD values are plotted on *x* axis and left *y* axis, right *y* axis shows the relative frequencies of random ranking distribution: black curve is the Gaussian approximation of discrete frequencies. The probability values are as follows (right *y* axis): XX1 and XX19 are 5% and 95% of first kind errors, respectively; Med = Median

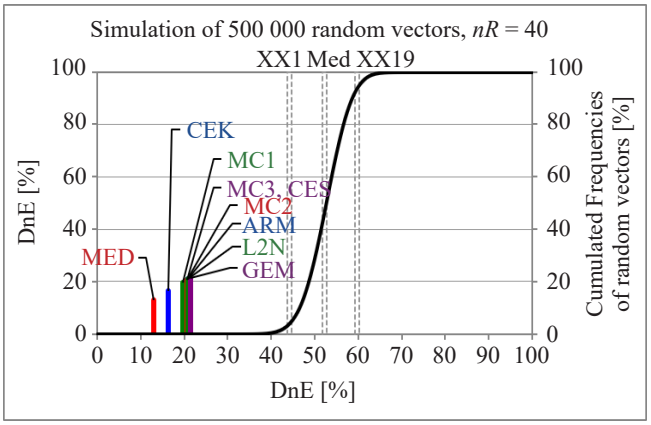


Figure 15. DnE analysis of Lin's table 2. Sum of Euclidean distance differences (DnE values, scaled between 0 and 100) ordered the data fusion methods. Scaled DnE values are plotted on *x* axis and left *y* axis, right *y* axis shows the relative frequencies of random ranking distribution: black curve is the Gaussian approximation of discrete frequencies. The probability values are as follows (right *y* axis): XX1 and XX19 are 5% and 95% of first kind errors, respectively; Med = Median

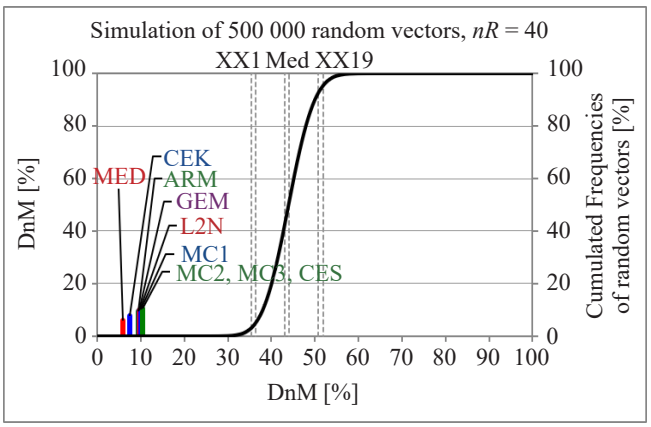


Figure 16. DnM analysis of Lin's table 2. Sum of absolute Manhattan distance differences (DnM values, scaled between 0 and 100) ordered the data fusion methods. Scaled DnM values are plotted on *x* axis and left *y* axis, right *y* axis shows the relative frequencies of random ranking distribution: black curve is the Gaussian approximation of discrete frequencies. The probability values are as follows (right *y* axis): XX1 and XX19 are 5% and 95% of first kind errors, respectively; Med = Median

The results of generalized pairwise correlation methods (CEpW and WtpW) can be found in the CEpW and WtpW sheets respectively of the supplementary material (CaseStudyNo3LinST240x9rnk). The ranking was enforced in the difference of lines 43 and 44, *i.e.*, the number of probability weighted wins (pWinner) minus number of probability weighted losses (pLoser) for both of CEpW and WtpW, whereas column D contains theoretical probabilities (α (theor)). If their values are smaller than the predefined 0.05, the pairwise comparison is to be considered as significant. All 2 × 2 contingency tables are given below line 49 for CEpW and all Pearson correlation coefficients are given below line 49 (column C) for WtpW to complete the Williams *t*-test. CEpW results favor MC3 and CES whereas CEK, MC3 and GEM are located. WtpW favors MC3 and CES, but ARM, MED and GEM are located at the end of the ordering.

Figure 17 visualizes the SRD scores in a COVAT manner, abbreviated as HM. All alternatives served as benchmark once and only once. The alternatives are arranged in sum of (rows and) columns. The matrix is symmetric. Four alternatives are indistinguishable as best ones (L2N, MC2, MC3, CES whereas the worst is MED.

Relative HeatMap									
colours:	x < =	0.8	2.3	3.8	5.3	6.8			
		1.5	3.0	4.5	6.0	7.5			
	L2N	MC2	MC3	CES	ARM	CEK	MC1	GEM	MED
L2N	0.00	3.50	3.50	3.50	2.75	3.00	3.25	4.75	5.50
MC2	3.50	0.00	0.50	0.50	4.25	5.25	3.50	6.00	6.25
MC3	3.50	0.50	0.00	0.00	4.50	5.25	4.00	5.75	6.25
CES	3.50	0.50	0.00	0.00	4.50	5.25	4.00	5.75	6.25
ARM	2.75	4.25	4.50	4.50	0.00	2.75	4.75	4.50	4.50
CEK	3.00	5.25	5.25	5.25	2.75	0.00	5.50	3.50	4.00
MC1	3.25	3.50	4.00	4.00	4.75	5.50	0.00	7.50	7.25
GEM	4.75	6.00	5.75	5.75	4.50	3.50	7.50	0.00	4.50
MED	5.50	6.25	6.25	6.25	4.50	4.00	7.25	4.50	0.00
Sum	29.75	29.75	29.75	29.75	32.50	34.50	39.75	42.25	44.50

Figure 17. Heatmap ordering of the data fusion methods. for Lin's table 2. Relative color coding was used, see header. The inner, pairwise-distance structure is not the same as the overall sum ordering

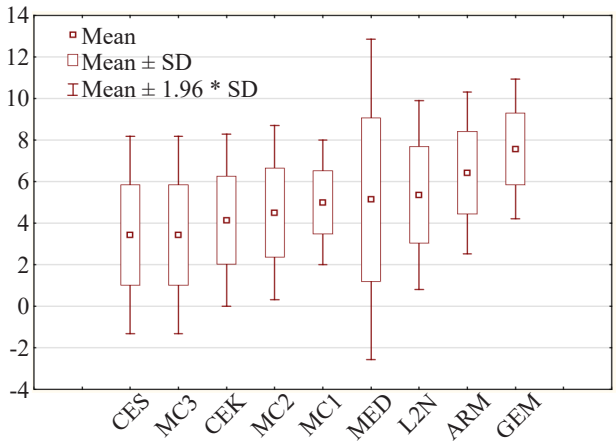


Figure 18. Box and whisker plot of rank averages and uncertainties for 9 alternatives. Pairwise statistical testing suggests no significant differences at the 5% level of the Mann Whitney U test and Wilcoxon matched pair test. Friedman ANOVA Chi Squares statistics (n = 7, degree of freedom = 8) = 88.56, limit probability of significance is p = 0.07845. The overall comparison of all alternatives is Not significant

The Deep Ranking Analysis by Power Eigenvectors (DRAPE), *i.e.* ranking of alternatives (Lin's data) can be found with some explanations in the DRAPE sheet in the supplementary material, CaseStudyNo3LinST240x9rnk). Here MED is ranked first, MC3 and CES are indistinguishable (tie: 4.5) and L2N is the last.

The overall ranking of alternatives by all seven MCDM methods can be found in the 7MCDM sheet in the supplementary material, CaseStudyNo3LinST240x9rnk, together with raw values. As is indicated in Figure 1, the

direction of ranking should be unified: SRD, DnE, DnM and HM, the smaller the better, whereas CEpW, WtpW and DRAPE, the larger the better; *i.e.* the last three MCDM methods should be reversed.

In the final step, rank averages and standard deviations are calculated. Mann-Whitney U test and Wilcoxon Matched pair test provide the same result with non-significant pairwise differences at the 5% error level. It is visualized in Figure 18.

Overall comparison of alternatives by Friedmann ANOVA does not suggest significance (p = 0.07845 level. (χ^2 statistics (n = 7, df = 8) = 14.1, whereas Coefficient of Concordance is 0.252, and the average rank is 0.128).

Although the alternatives are tantamount to each other; the MC1 bears the smallest uncertainty whereas MED has the largest one.

Parallel coordinate plot (Figure 19) rearranges the MCDM methods to the minimal sum of adjacent absolute rank differences across all alternatives. The largest turbulence is between WtpW and DnE, whereas the ordering of DnE and DRAPE is interchangeable.

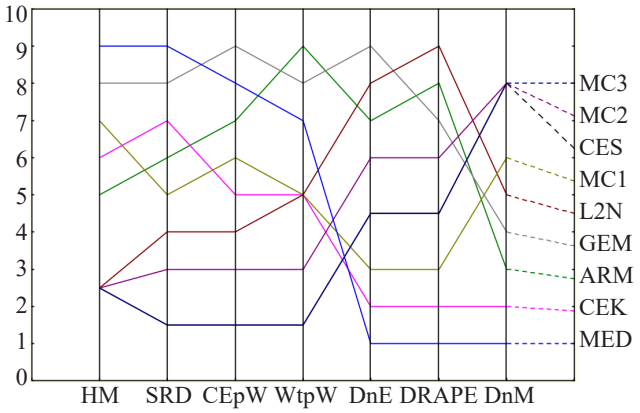


Figure 19. Parallel coordinate plot, the MCDM methods are rearranged according to the minimal sum of adjacent absolute rank differences across all alternatives. The largest turbulence is between WtpW and DnE, whereas ordering of DnE and DRAPE is interchangeable

2.4 Justification 1 (correlation analysis)

Perhaps, it is not unambiguous why just these three case studies were selected, why they are just complementary. First, they are different in input sizes of matrices: 9 × 9, 18 × 16 and 40 × 9, respectively. All inputs were subjected to the seven MCDM techniques, which behave differently on these data sets.

Suitability for MCDM methods can be disclosed by their correlation structure. A comparison requires the same correlation measure: Kendall's τ coefficient was selected because of its beneficial properties. Its distance (d_τ) has symmetric distribution and Kendall τ distance impart metric properties (with criteria of non-negativity, symmetry, identity and triangular inequality). Kendall τ distance is scaled between 0 and 1 with the formula $d_\tau = (1 - \tau)/2$.

The correlation structure of the MCDM methods (CEpW, DnE, DnM, DRAPE, HM, SRD, and WtpW) for the three case studies is summarized in Table 2.

Table 2. Kendall's τ correlations coefficients arranged in their increasing sums for the three case studies. Significant correlations are indicated by boldface setting, whereas red color signalizes positive, blue color negative correlations

Sum	EliawtNo1	WtpW	DnM	DnE	DRAPE	CEpW	HM	SRD
3.305	WtpW	1.000	0.000	0.056	0.056	**0.761**	**0.667**	**0.766**
3.351	DnM	0.000	1.000	**0.944**	**0.944**	0.141	0.203	0.118
3.524	DnE	0.056	**0.944**	1.000	**0.889**	0.197	0.261	0.177
3.524	DRAPE	0.056	**0.944**	**0.889**	1.000	0.197	0.261	0.177
4.047	CEpW	**0.761**	0.141	0.197	0.197	1.000	**0.794**	**0.956**
4.049	HM	**0.667**	0.203	0.261	0.261	**0.794**	1.000	**0.862**
4.055	SRD	**0.766**	0.118	0.177	0.177	**0.956**	**0.862**	1.000

Sum	BernalNo2	HM	CEpW	DRAPE	SRD	WtpW	DnM	DnE
4.543	HM	1.000	**0.517**	**0.517**	**0.616**	**0.683**	**0.577**	**0.633**
4.764	CEpW	**0.517**	1.000	**0.467**	**0.903**	**0.767**	**0.527**	**0.583**
4.894	DRAPE	**0.517**	**0.467**	1.000	**0.515**	**0.567**	**0.946**	**0.883**
5.062	SRD	**0.616**	**0.903**	**0.515**	1.000	**0.819**	**0.576**	**0.633**
5.146	WtpW	**0.683**	**0.767**	**0.567**	**0.819**	1.000	**0.628**	**0.683**
5.200	DnM	**0.577**	**0.527**	**0.946**	**0.576**	**0.628**	1.000	**0.946**
5.362	DnE	**0.633**	**0.583**	**0.883**	**0.633**	**0.683**	**0.946**	1.000
Sum	**LinNo3**	**DnM**	**DRAPE**	**HM**	**DnE**	**SRD**	**WtpW**	**CEpW**
–1.313	DnM	1.000	0.206	**–0.636**	0.206	**–0.794**	**–0.677**	**–0.618**
1.612	DRAPE	0.206	1.000	–0.370	**0.886**	–0.200	0.120	–0.029
1.839	HM	**–0.636**	–0.370	1.000	–0.247	**0.802**	**0.549**	**0.741**
1.964	DnE	0.206	**0.886**	–0.247	1.000	–0.086	0.120	0.086
2.268	SRD	**–0.794**	–0.200	**0.802**	–0.086	1.000	**0.717**	**0.829**
2.665	WtpW	**–0.677**	0.120	**0.549**	0.120	**0.717**	1.000	**0.837**
2.845	CEpW	**–0.618**	–0.029	**0.741**	0.086	**0.829**	**0.837**	1.000

Although the arrangement is systematic, the significance pattern is random. Moreover, the ordering of MCDM methods does not follow any other ordering pattern than random, which is reassuring: no systematic information was introduced by the selection of MCDM methods. The τ coefficients cover a broad range from –0.794 to +0.946; separately 0 to 0.956, 0.467 to 0.883 and –0.794 to 0.886 for case studies No 1, 2 and 3, respectively. Therefore, a formal criterion (threshold) cannot reasonably be defined. Out of the broad range above no intelligent (smart) ranking is possible: a selection of MCDM tools is impossible by ensuring straight and reverse ranking at the same time. Neither too large nor too small similarity (diversity) is acceptable.

Kendall's correlation coefficient (τ) can be recalculated into Kendall τ distance. The Kendall τ distance also ranges broadly from 0.02 to 0.9 (a distance can only be positive, or zero); separately: 0.0219 to 0.500, 0.0272 to 0.267 and 0.0817 to 0.897 for case studies No 1, 2 and 3, respectively. These data sets are of large, medium, and small diversity. Although case studies are designed for illustrative purposes, the coverage shows little playground to find a threshold or better combination of MCDM techniques. Anyone attempting to construct an out-of-dataset above faces a steep uphill battle.

Friedman test of overall significance can be categorized as significant, highly significant and non-significant for case studies No 1, 2 and 3, respectively (see captions to Figures 6, 12 and 18).

Sum ordering of Kendall' τ-s leads to the following MCDM method orderings:

Case study No 1: WtpW < DnM < DnE < DRAPE < CEpW < HM < SRD

Case study No 2: HM < CEpW < DRAPE < SRD < WtpW < DnM < DnE

Case study No 3: DnM < DRAPE < HM < DnE < SRD < WtpW < CEpW

It is worth considering another legitim criterion for ordering MCDM methods. Namely, they can be ordered according to the minimal sum of adjacent absolute rank differences across all alternatives (*c.f.* parallel coordinate plots, Figures 7, 13, and 19). The reverse ranking is also feasible (equivalent).

Case study No 1: DnE < DnM < DRAPE < WtpW < CEpW < SRD < HM

Case study No 2: DRAPE < DnM < DnE < HM < WtpW < SRD < CEpW

Case study No 3: HM < SRD < CEpW < WtpW < DnE < DRAPE < DnM

Again, one can safely say the pattern is random, no hierarchy can be detected. Formulated differently: no superior (or inferior) MCDM methods exist; The ordering corresponds to the data structure of the individual data sets.

Matrix plots of correlations visualize the correlation pattern (Figures 20-22, for case studies 1-3, respectively). The diagonal elements contain histograms (for a survey of allocation), the off-diagonal elements show the correlation or lack of it. The horizontal arrangements of MCDM methods (indicated in the diagonal elements) are plotted on the *x* axis, vertical ones on the *y* axis. The matrix is symmetric; it is sufficient to consider the upper off-diagonal part.

Figure 20 (Case study No. 1) plots the matrix of correlations (Kendall's τ correlation): out of 21 off-diagonal elements 11 combinations are not correlated, six are highly correlated ($\tau_i > 0.8$), but even the remaining four are significant at the 5% level.

Figure 21 (Case study No. 2) shows the matrix of correlations (Kendall's τ correlation): all 21 off-diagonal elements are significantly correlated; four are highly correlated ($\tau_i > 0.8$), but the remaining 17 are also significant at the 5% level.

Figure 22 (Case study No. 3) shows the matrix of correlations (Kendall's τ correlation): four are highly and positively correlated ($\tau_i > 0.8$), four are significantly and negatively correlated ($\tau_i < 0.6$), three are significantly and positively correlated but the remaining 10 are not significantly correlated at the 5% level, either positively or negatively.

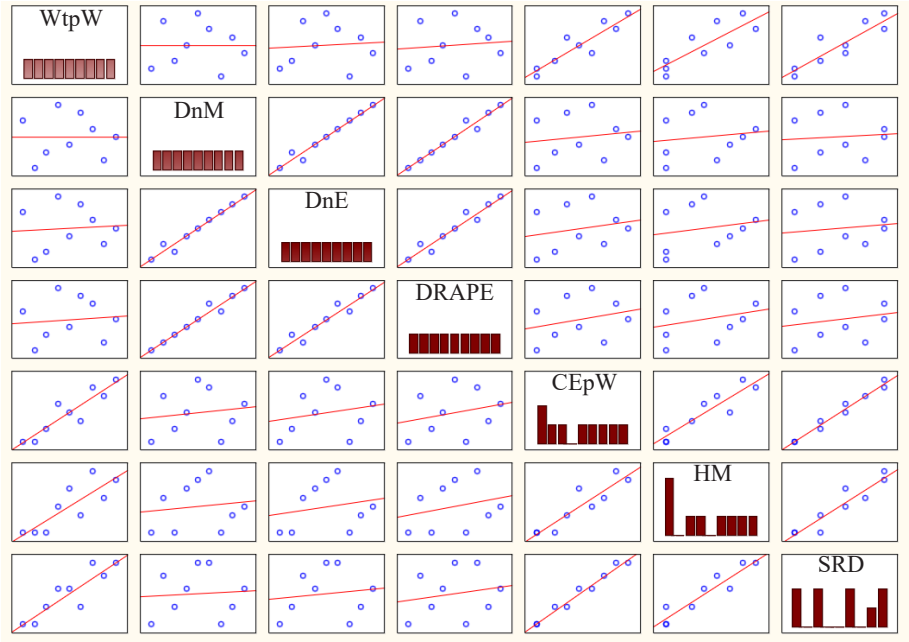


Figure 20. Matrix plot of correlations for Case study No. 1

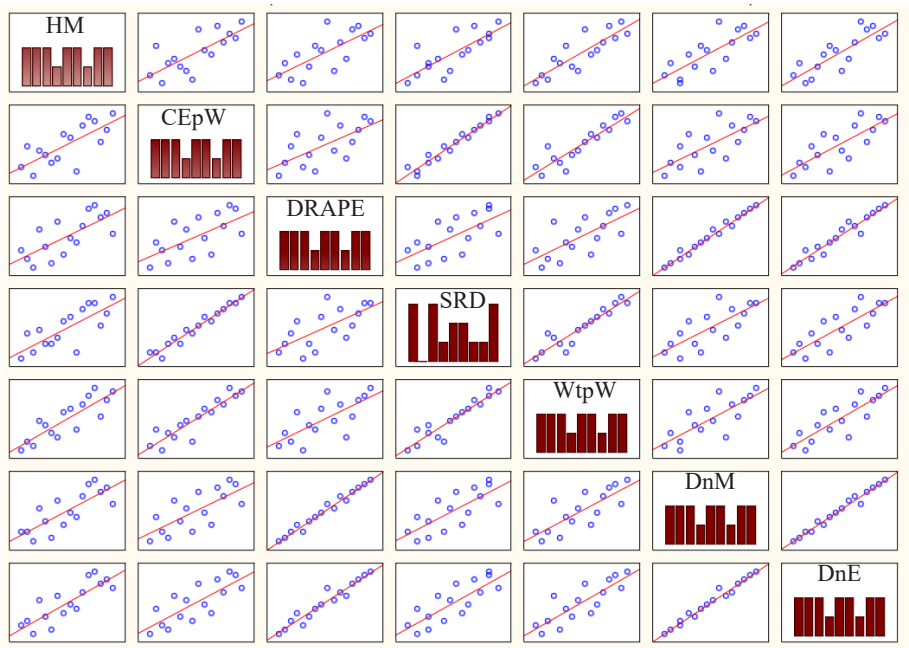


Figure 21. Matrix plot of correlations for Case study No. 2

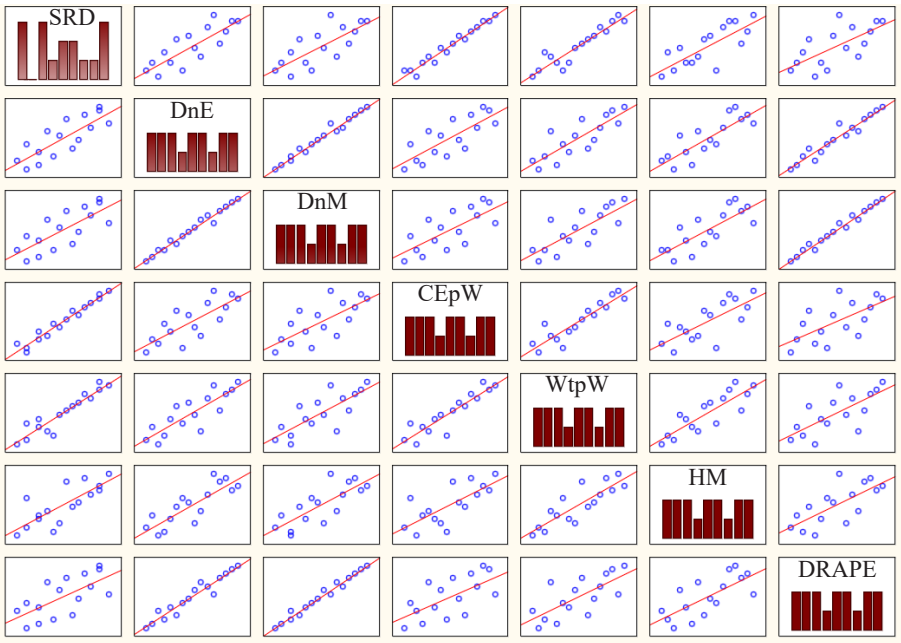


Figure 22. Matrix plot of correlations for Case study No. 3

The matrix plots make it easily perceivable that the case studies cover all reliable situations. The only missing highly negative correlations are irrational as the rank directions calculated to be unidirectional (*c.f.* Figure 1), the smaller the better; highly significant reverse rankings are out of picture. However, such a highly and negatively correlated ($\tau_i < -0.8$) case can serve diagnostic purposes. Its appearance indicates wrong or not fully unidirectional ranking.

2.5 Justification 2 (statistical testing)

Legendre's detailed paper [33] discusses small numbers of rankings, permutation testing and the relationship between Kendall's W test and Friedman's ANOVA. Moreover, it refers to Kendall's W calculations [34] as proper statistical testing of ranked lists. Here, the rows renamed as alternatives and MCDM methods are enumerated in the columns (7MCDMranks sheets in all cases).

Kendall's W provides an overall concordance for the seven MCDM methods to be compared. It is scaled: $0 \leq W \leq 1$, where 1 means high concordance and 0 corresponds to random one.

Kendall's W test and Friedman's test are directly related but their null hypotheses to be accepted (or rejected) are different: Kendall's H_0: There is no systematic concordance among the rankings whereas Friedman's H_0: The alternatives (rows) have equivalent rank distributions.

Table 3. Comparison of the statistical significance by χ^2 test for the three case studies

	Kendall's W [a]	χ^2 (degree of freedom) [b]	Theoretical limit of significance, p
Case Study No 1	0.570	31.34 (8)	**1.22×10^{-4}** [c]
Case Study No 2	0.843	88.56 (15)	**1.84×10^{-12}** [d]
Case Study No 3	0.252	14.13	0.0784 [e]

[a] Using standard ties correction.
[b] χ^2 test for Kendall's W is related to Friedman's test; however, their interpretation is different.
[c] Significant at the 5% level.
[d] Highly significant at the 5% level. (The very small p-value indicates that the observed concordance is highly unlikely as compared to random rankings).
[e] Not significant at the 5% level.

2.6 Justification 3 (Leave-one-method-out)

It would be interesting to know which method strengthens or weakens global concordance. A kind of sensitivity analysis can answer that. Leave-one-out (LOO) technique was selected: one MCDM method was left out once and only once seven times.

Table 4 summarizes the LOO Kendall's W values, their difference from the overall Kendall's W values, χ^2 statistics and theoretical limit probability of significance for Case Study No 1.

Table 4. The LOO Kendall's W values, their difference from the overall Kendall's W values, χ^2 statistics and theoretical limit probability of significance for Case Study No 1

Omitted MCDM method	Leave-one-method-out (LOO), Kendall's W [a]	Difference of LLO-W and overall-W	χ^2 (degree of freedom)	Theoretical limit of significance, p in 10^{-4} units [b]
DnM	0.6041	+0.0445	29.00	**3.18**
DnE	0.5895	+0.0300	28.30	**4.21**
DRAPE	0.5802	+0.0206	27.85	**5.04**
WtpW	0.5783	+0.0187	27.76	**5.23**
SRD	0.5555	−0.0041	26.66	**8.08**
HM	0.5522	−0.0074	26.51	**8.60**
CEpW	0.5431	−0.0165	26.07	**1.02**

[a] Using standard ties correction.
[b] Significant at the 5% level.

The significance (at the 5% level) was preserved in all cases, if left out any of the MCDM methods.

The first four differences (column 3 in Table 4) are positive; *i.e.* leaving out the corresponding methods enhances the concordance, and the remaining three are negative, *i.e.* their removal decreases the concordance. Formulated the DnM example in another way around: DnM contributes most strongly to disagreement with the remaining methods in Case Study No. 1. Similarly, Removing CEpW produces the smallest LOO-W, *i.e.*, its removal supports the overall ranking structure. Ordering the seven MCDM methods according to the differences:

Case Study No. 1: DnM < DnE < DRAPE < WtpW < SRD < HM < CEpW

Comparing this ordering with the sum ordering and rank difference ordering (Justification 1) a totally different ordering is produced. This is reassuring, no MCDM methods should be rightly removed out of seven selected.

Table 5 summarizes the LOO Kendall's W values, their difference from the overall Kendall's W values, χ^2 statistics and theoretical limit probability of significance for Case Study No 2.

Table 5. LOO Kendall's W values, their difference from the overall Kendall's W values, χ^2 statistics and theoretical limit probability of significance for Case Study No 2

Omitted MCDM method	Leave-one-method-out (LOO), Kendall's W [a]	Difference of LLO-W and overall-W	χ^2 (degree of freedom)	Theoretical limit of significance, p in 10^{-10} units [b]
HM	0.8598	+0.0164	77.39	**2.09**
DRAPE	0.8598	+0.0163	77.38	**2.10**
CEpW	0.8558	+0.0124	77.02	**2.43**
SRD	0.8470	+0.00353	76.23	**3.39**
DnM	0.8415	−0.00194	75.74	**4.17**
WtpW	0.8405	−0.00292	75.65	**4.32**
DnE	0.8301	−0.0133	74.71	**6.38**

[a] Using standard ties correction.
[b] Highly significant at the 5% level.

Any of the MCDM methods left out the highly significant character has not been changed (at 5% level). The first four differences (column 3 in Table 5) are positive, *i.e.* leaving out the corresponding methods enhances the concordance, and the remaining three are negative, *i.e.* their removal decreases the concordance. Formulated the first two items in Table 5 differently Both HM and DRAPE reduce the overall concordance almost equivalently (≈ and slightly): HM and DRAPE contribute mostly to disagreement with the remaining methods in Case Study No. 2. Similarly, Removing DnE produces the smallest LOO-W, *i.e.*, its removal supports the overall ranking structure. Ordering the seven MCDM methods according to the differences:

Case Study No. 2: HM ≈ DRAPE < CEpW < SRD < DnM < WtpW < DnE

Comparing this ordering with the sum ordering and rank difference ordering (Justification 1) a totally different ordering (with one tie) is produced; again, this is reassuring, no MCDM methods should be rightly removed out of seven ones selected.

Table 6 summarizes the LOO Kendall's W values, their difference from the overall Kendall's W values, χ^2 statistics and theoretical limit probability of significance for Case Study No 3.

Table 6. LOO Kendall's W values, their difference from the overall Kendall's W values, χ^2 statistics and theoretical limit probability of significance for Case Study No 2

Omitted MCDM method	Leave-one-method-out (LOO), Kendall's W [a]	Difference of LLO-W and overall-W	χ^2 (degree of freedom)	Theoretical limit of significance, *p*
DnM	0.4662	+0.2139	22.38	**0.00426** [b]
DRAPE	0.2799	+0.0275	13.43	0.09781
HM	0.2629	+0.0105	12.62	0.12574
DnE	0.2574	+0.00509	12.36	0.13604
SRD	0.2416	−0.0107	11.60	0.17002
CEpW	0.2044	−0.0479	9.811	0.27855
WtpW	0.1990	−0.0534	9.550	0.29805

[a] Using standard ties correction.
[b] significant at the 5% level.

The striking result is that the overall non-significant behavior changes if DnM is left out. Leaving out any of the other MCDM methods does not alter the non-significant character (at 5% level). The first four differences (column 3 in Table 6) are positive; *i.e.* leaving out the corresponding methods enhances the concordance, and the remaining three are negative; *i.e.* their removal decreases the concordance. DnM is without doubt the foremost source of discordance (contributes mostly to disagreement with the remaining methods in Case Study No. 3. Similarly, removing WtpW produces the smallest LOO-W, *i.e.*, its removal supports the overall ranking structure at most. Ordering the seven MCDM methods according to the differences:

Case Study No. 3: DnM << DRAPE < HM < DnE < SRD < CEpW < WtpW

Comparing this ordering with the sum ordering and rank difference ordering (Justification 1) a totally different ordering (without ties) is produced; again, this is reassuring. Removal of one MCDM methods can be considered as accidental. In any case, the remaining six methods exhibit a statistically significant global concordance (and show their similarity) in a way of non-significant comparisons.

Comparing all three cases studies one can safely state the seven MCDM methods selected are complementary. It would be difficult to remove any of them, which were valid for all situations. In fact, a smaller number of methods were sufficient, see (Appendix 1). If a method must be omitted, DnM were an acceptable candidate because of its high similarity to DnE and for a special data structure as for Case Study No 3.

3. Discussion

SRD play a main role in the seven selected MCDM methods because of i) it is easily perceivable (smaller rank differences to a benchmark the better) and fair [18]; ii) it has a solid statistical basis (SRD without ties and a fixed benchmark vector equals Spearman footrule (as a measure for disarray [35]); iii) it has two validation steps: randomization (permutation) test (where the scores of alternatives are compared to the SRD scores of random rankings [18, 20]) and the final SRD ordering is cross-validated by re-sampling [18, 21] and can be tested pairwise by Wilcoxon matched pair or Alpaydin test [21].

The only disadvantage of SRD might be if the benchmark selection is not appropriate. Not proper benchmark selection leads to unacceptable results. To eliminate this discrepancy, all alternatives were used as benchmarks providing fine SRD-distance structure. The symmetric HM matrix is a kind of reversed correlation matrix (Group C in Figure 1). Moreover, instead of L_1 norm of ranks (SRD), sum of Euclidean distance differences and sum of absolute Manhattan distance differences (DnE and DnM) utilize the same scheme. Finer distinction is possible while the randomization test and giving uncertainties for rankings and statistical tests remain. The scheme is the same, but the distance measures are different for DnE and DnM (Group A in Figure 1).

Group B in Figure 1 realizes two methods selecting explanatory variables in a supervised way (CEpW and WtpW). Conditional Fisher's exact test of hypergeometric distribution tests significances of 2 × 2 contingency tables [24] whereas Wiliams' *t* test establishes whether significant difference exist between correlation coefficients [24] both in a generalized pair-correlation manner (GPCM) [23]. The robust (nonparametric, CEpW) and parametric (WtpW) testing complement each other. What is left, the unsupervised ordering of alternatives by DRAPE [26].

There is highly improbable, if not impossible, that the different biases, different variances, different trains of thoughts were shifted in one direction, as they complement each other in many ways. Empirical case studies support these considerations largely See case studies and Justification 1 parts).

All of them are backed by statistical tests (DRAPE can be subjected to a leave-one out cross-validation and assigning uncertainties to DRAPE ordering and testing them). A smoothed version of power/weakness ratio has also been elaborated [36], again, it may cause multiple solutions (unnecessary subjectivism) like weighting; hence, smoothing is not recommended for standardization.

Naturally, not only the above MCDM methods can be used in rank averaging and aggregation. Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), VlseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR), Analytic Hierarchy Process (AHP), *et*c. can be recommended, just to name a few of the most frequently used ones. Any MCDM methods with weighting and those like ELimination Et Choix Traduisant la REalité (ELECTRE) and Ranking Organization METHod for Enrichment Evaluations (PROMETHEE) are not recommended because of many versions, various utility functions, *etc*.; *i.e*. They are easy to manipulate and do not level with the principle to parsimony. In any case, benchmarking the MCDM techniques has proven to be unsuccessful but the attempt raised attention that the data pretreatment should also be carefully selected according to the data structure [37]. The suggested seven MCDM methods are *well balanced and complimentary* to each other. Many other techniques can leverage the equilibrium and bias the final results. For example, SHapley Additive exPlanations (SHAP) technique [38] can substitute DRAPE as an unsupervised one. Calculation of SHAP values has become a kind of standard to find plausible explanation of Artificial Intelligence (AI) black box techniques. In fact, it may strengthen (overweight) the role of explanatory variable selection (besides CEpW and WtpW).

Why does SRD play central role in standardization?

First of all, SRD realizes a multicriteria optimization [39, 40]. SRD is entirely general, fair, robust [41] and it would be expedient to be considered as a *de facto* standard for model (algorithm) comparison. SRD's power and usefulness have been proven in many issues: Its scientific reliability was emphasized by data compliant with the test guidelines of the Organisation for Economic Co-operation and Development (OECD) [42]. SRD coupled with the variable selection method called Uninformative Variable Elimination (UVE) can select *more useful variables* than the UVE algorithm itself [43]. The introduction of DRAPE was *validated* by SRD procedure [44]. A new rank-order validation metric (relative to other QSAR/QSPR performance measures) has been *supported and validated* by SRD analysis [45]. Coupling Partial Least Squares Discriminant Analysis (PLS-DA) with SRD "not only avoids the complicated parameter selection process but *also improves the specificity, sensitivity, and accuracy* of the classification model [46]. "… the Sum of Ranking Differences (SRD) offers a *fairer and more robust framework* for comparative model assessment" [47]. "Heberger and coworkers … provided an elegant and unbiased procedure to rank the models based on multiple criteria, called sum of rank differences (SRD) procedure." [48]. There is no need to enumerate many hundred successful application examples. Let's end the eulogy of SRD with a contemporary quote: "SRD technique has proven to be effective… to identify the most favorable optimal solution. … the identified Pareto solution should yield the most favorable value for bias, quality of prediction, … resilience and robustness, … this way to ensure better reproducibility for the generated Pareto solutions." [49].

Though present work is focused on ranking; however, the suggested techniques are suitable for grouping (clustering), sorting, choice and determination of rank variances.

4. Summary and conclusions

A framework easy to overview is suggested to solve the ambiguous rankings of various multicriteria decision making (MCDM, Pareto optimization) methods. After careful selection, a consensus of seven well-balanced (complementary) techniques is recommended. Three case studies illustrate and visualize the working of the suggested MCDM methods.

Different validation methods justify the claims: Sum of binary comparisons of ranks (Kendall's τ correlation matrices), minimum adjacent rank differences (parallel

coordinate plots), leave-one-out sensitivity analyses: Row(alternative)-wise and column (MCDM method)-wise—all prove the complementary character of the seven methods selected.

The final majority view for MCDM methods is calculated by mean ranks of alternatives and their standard deviations. Then, a Wilcoxon matched pair test will return the significance between alternatives (at a given error level). Kendall's W concordance and Freidman ANOVA demonstrate the overall significance of alternatives.

As the above methods cover various trains of thought, they are well-balanced, complementary, backed by suitable and different statistical tests, obey the principle of parsimony, *etc*.—the suggested ensemble of MCDM methods is ready for standardization.

The framework provides an elegant proof-of concept of rank prediction with uncertainties in addressing questions of basic interest for decision makers.

5. Appendix 1

Application of just seven MCDM methods is not mandatory. Five is the minimum, more than seven produces not much gain. Calculation details are given below. The other reason is the limited short-term memory of human beings to process information: Miller's paper (best cited ever in psychology) divulges the limit of the human brain for processing information [50].

5.1 Justification of exactly seven MCDM tools

1) If the MCDM methods are uncorrelated and the number of methods is small (7) no Gaussian distribution can be assumed. No exact tail probability can be calculated without distributional assumption. However, Chebyshev inequality is distribution-free:

$$P\left(\left|\bar{r}-\mu\right| \geq d\right) \leq \frac{\sigma^2}{md^2} \qquad \text{(A1_1)}$$

Where $\bar{r}$ is the rank average (row-wise), d is the rank difference allowed, m is the number of MCDM methods. For simplicity, d = 1 in rank units, and Bernal's data were used (CaseStudyNo2T2all, sheet "7MCDMranks" Columns J-Q) (Table A_1).

Table A_1. Probability of differing ±1 rank units as a function of numbers for ranked lists (m)

Number of ranks, m	SD of mean ranks: $SD = \frac{\sigma}{\sqrt{m}}$	Upper bound for Eq. A1_1 in [%]	Probability of being ±1 in [%]
2	0.32	10.1	89.9
3	0.26	6.8	93.3
4	0.23	5.1	94.9
5	**0.20**	**4.1**	**96.0**
6	0.18	3.4	96.6
7	0.17	2.9	97.1
8	0.16	2.5	97.5
9	0.15	2.3	97.7

Reaching the 5% error limit is indicated by bold font.

The approximation by Chebyshev inequality suggests that five MCDM methods are sufficient (at a predefined 5% error level).

Seven were selected to be sure; larger number of methods than seven produces only slight changes, *i.e*. they cannot produce much gain.

2) If the correlation between MCDM method is considerable; then, Spearman–Brown prediction formula is suitable to estimate reliability (ρ_i) [51].

$$\rho_i = \frac{m\bar{r}}{1+(m-1)\bar{r}} \qquad \text{(A1_2)}$$

Where m is the number of MCDM methods to be compared and $\bar{r}$ is the mean of Spearman correlation coefficient without the diagonal elements of the correlation matrix. Using Bernal's data (CaseStudyNo2T2all, sheet "7MCDMranks" Columns J-Q) as earlier, Table A_2 is obtained:

That is, as the mean correlation between MCDM methods increases; then, reliability also increases. The entries in the table are "reliabilities" (composite rank correlations coefficients, Spearman or Pearson, they are identical for ranking data) and not probabilities. Comparing seven correlated MCDM methods is reliable unquestionably.

Table A_2. Reliability (ρ_i) of seven MCDM methods as a function of numbers for ranked lists (m)

m	DRAPE	DnM	DnE	HM	WtpW	SRD	CEpW
2	0.8773	0.9108	0.9306	0.8772	0.9126	0.9010	0.8848
3	0.9147	0.9387	**0.9527**	0.9146	0.9399	0.9317	0.9201
4	0.9347	**0.9533**	0.9641	0.9346	**0.9543**	0.9479	0.9389
5	0.9470	0.9623	0.9711	0.9470	0.9631	**0.9579**	**0.9505**
6	**0.9555**	0.9684	0.9758	**0.9554**	0.9690	0.9647	0.9584
7	0.9616	0.9728	0.9791	0.9615	0.9733	0.9696	0.9641

(Surpassing the predefined 5% limit is indicated by bold).

3) There is no need to assume any distribution and any approximation is superfluous, if we have only seven methods. Then, exact distribution can be calculated with all possible subsets:

$$P\left(\left|\overline{R}_S - \overline{R}_7\right| \geq d\right) = \frac{\text{Number of subsets: } \left|\overline{R}_S - \overline{R}_7\right| \geq d}{\binom{7}{m}} \quad \text{(A1_3)}$$

Where d is the tolerated difference, *e.g.* ±1 rank units, $\overline{R}_S$ is the row average of any subsets, $\overline{R}_7$ is the seven-method mean, the denominator contains all possible subsets.

Then, exact (theoretical) probabilities of comparing 2, 3, …6 MCDM methods (m = 7 is identically equal to zero; hence, it is omitted from Table A_3) using Bernal's data (CaseStudyNo2T2all, sheet "7MCDMranks" Columns J-Q) as earlier.

Table A_3. Exact probability of an error for at least 1 rank unit or higher in percentage [%]

	Subsets (21) m = 2	Subsets (35) m = 3	Subsets (35) m = 4	Subsets (21) m = 5	Subsets (7) m = 6
XGB_ECFP4	4.8	0	0	0	0
GBM_ECFP4	28.6	11.4	8.6	0	0
GBM_RDKit	33.3	11.4	8.6	0	0
RF_RDKit	61.9	37.1	25.7	4.8	0
XGB_RDKit	0	0	0	0	0
GBM_MACCS	76.2	34.3	28.6	19.0	0
ET_RDKit	52.4	42.9	28.6	19.0	0
ET_MACCS	47.6	28.6	8.6	0	0
XGB_custom	57.1	48.6	34.3	23.8	0
ET_ECFP4	28.6	8.6	2.9	0	0
XGB_MACCS	42.9	40.0	14.3	0	0
RF_MACCS	4.8	0	0	0	0
GBM_custom	19.0	2.9	0	0	0
RF_ECFP4	47.6	48.6	28.6	23.8	0
RF_custom	9.5	0	0	0	0
ET_custom	9.5	2.9	0	0	0

The inference is obvious: comparison of five MCDM methods is sufficient in the overwhelming majority of cases (11); sometimes, a comparison of a pair would do (RF_MACCS, XGB_ECFP4) as well. Only four cases need at least comparison of six MCDM methods (GBM_MACCS, ET_RDKit, XGB_custom, RF_ECFP4).

5.2 Sensitivity analysis (alternative-wise)

The above Table A_3 can also be considered as a sensitivity analysis. Each alternative (XGB_ECFP4, *etc*.) produces different patterns.

However, other ways can also be considered.

Leave-one-row-out for an MCDM comparison matrix (here alternatives are in the rows) are presented using Lin's data (CaseStudyNo3LinST240x9rnk, sheet "7MCDMranks" Columns J-Q) (Table A_4).

As the column-wise mean of alternatives is exactly five, the largest deviation is ±0.5 rank units but generally much smaller: zero for *e.g*. SRD, if MC1 is omitted, 0.0625 for *e.g*. DRAPE, if CES omitted, *etc*.

The smaller perturbation than ±0.5 rank unit underlines the stability of MCDM methods selected and confirms the correctness of the MCDM method selection.

Table A_4. Mean rank of each MCDM method over the remaining 8 alternatives

Omitted alternative	HM	SRD	CEpW	WtpW	DnE	DRAPE	DnM
CES	5.3125	5.4375	5.4375	5.4375	5.0625	5.0625	4.6250
MC3	5.3125	5.4375	5.4375	5.4375	5.0625	5.0625	4.6250
CEK	4.8750	4.7500	5.0000	5.0000	5.3750	5.3750	5.3750
MC2	5.3125	5.2500	5.2500	5.2500	4.8750	4.8750	4.6250
MC1	4.7500	5.0000	4.8750	5.0000	5.2500	5.2500	4.8750
MED	4.5000	4.5000	4.6250	4.7500	5.5000	5.5000	5.5000
L2N	5.3125	5.1250	5.1250	5.0000	4.6250	4.5000	5.0000
ARM	5.0000	4.8750	4.7500	4.5000	4.7500	4.6250	5.2500
GEM	4.6250	4.6250	4.5000	4.6250	4.5000	4.7500	5.1250

5.3 Theoretical considerations

Already Lin has emphasized the limitations of theoretical models [32]. Even the simplest assumptions are not fulfilled either for short-ranked lists in Thurstone model: some parameters are unknown (unidentifiable); hence, a variance is to be set unity, and the correlation is assumed to be zero for pairwise comparison of ranked lists. Such assumptions are never fulfilled in practice [32].

Jan and Shieh postulated a model for block design [52] which is analogous here:

$$X_{i,j} = \mu + \theta_i + b_j + \varepsilon_{i,j} \qquad \text{(A1_4)}$$

Where $X_{i,j}$ is the input matrix for comparison of seven MCDM methods selected (7MCDM sheets), θ_i is the bias of alternatives i (1, …, n, n is the number of rows), b_j is the bias of the MCDM methods j (1, …, m, m is the number of columns), and $\varepsilon_{i,j}$ is the random error independently and identically distributed (*i.i.d*.). They, in ref. [52], also recommend a modification of Friedman's test. This equation serves as a basis for theoretical treatment, even if it does not take into account the correlation of biases. Such problems have not been resolved in up-to-date research.

Not only does the correlation structure disturb the theoretical approaches but it is a prototype of "square circle". Namely, an index should be developed (and optimized), which contains relevance and diversity at the same time. An ideal candidate for such an index would be the Tversky coefficient [53]. It contains two regularization parameters (α, β). However, they are moving in the same direction, *i.e*. Increasing (α and β) penalizes distinctive features, in other words, prioritize the similarity alone. Decreasing (α and β) minimizes the penalty for unique features. That is, highly diverse items are to be ranked well. How to optimize them for correlated rank lists have remained unresolved yet.

Such a "nonsense" index was also developed for re-ranked lists (recommender systems) where a regularization parameter (λ) is introduced to balance accuracy (relevance, novelty) and diversity using the marginal relevance (MR) recently [54]:

$$MR(R) = 1 - \lambda \frac{1}{|R|} \sum_{i \in R} \text{Relevance}(i) + \lambda \text{Diversity}(N) \quad \text{(A1_5)}$$

Where a set cutoff N is defined to only consider the first positions in re-ranking.

Difficulties are to be caused not only by the optimization of regularization parameter(s). Pairwise comparisons are the base of correlation analyses. However, multiple comparisons should provide a more sophisticated picture [55]. Even this approach should be regularized (balanced) between similarity and diversity and applied for ranked data.

Supplementary materials

Supplementary materials associated with this article are available online and can be accessed via the journal website: https://ojs.luminescience.cn/DMA/article/view/654.

Declaration of generative AI use

The author declares that generative artificial intelligence (AI) tools or AI-assisted technologies were used during the preparation of this manuscript.

AI tools were limited to the following purposes: i) language correction and translation (Deep L and Google translate (GEMINI)); ii) formatting assistance. (The references have been transferred to Vancouver format by Copilot Chat (basic)); iii) programming assistance (ChatGPT 5.2-5.5 Thinking was used to understand the DRAPE algorithm and Kendall's W calculations); iv) figure preparation (Figure 1 boxes and coloring); v) specified purposes: (ChatGPT 5.2-5.5 Thinking was used to locate a few literature sources (refs [12, 17, 34, 44, 50, 52])).

All AI-assisted outputs were critically reviewed, verified, and revised by the author. Generative AI tools were not used to replace the authors' scientific judgment, generate unsupported scientific claims, interpret results independently, or assume responsibility for any part of the research. The author remains fully responsible for the originality, accuracy, integrity, and ethical compliance of the manuscript.

Code availability

Detailed discussion and codes for SRD, DnE and DnM can be found in supplementary material of ref. [22], its DOI link is https://doi.org/10.1016/j.aca.2025.344649

Computer codes (MS Excel macros) are also available for SRD, DnE, DnM, PCM, GPCM as links with input and output examples (last access date October 06 /2026):

http://aki.ttk.hu/srd/

https://www.ttk.hun-ren.hu/rolunk/adatbazisok-programok

https://cran.r-project.org/web/packages/rSRD/index.html detailed explanations can be found in ref. [56] with easy-to-understand examples.

Funding

Financial support was granted by the Ministry of Innovation and Technology of Hungary from the National Research, Development and Innovation Fund, financed under the K-type funding scheme, K 134260, title: Development of soft sensor models based on in vitro experiments and creation of related metrics. The founder had absolutely NO any role in the study design, data collection, analysis, interpretation, manuscript preparation, or decision to submit the manuscript for publication.

Data availability

All input and output data are available for case studies No. 1, 2 and 3 as supplementary material.

Conflicts of interest

There are no conflicts to declare to the best of my knowledge.

References

[1] Taherdoost H, Madanchian M. Decision making: models, processes, techniques. *Cloud Computing and Data Science*. 2024; 5(1): 1-14. doi:10.37256/ccds.5120243284.

[2] Cinelli M, Kadziński M, Miebs G, Gonzalez M, Słowiński R. Recommending multiple criteria decision analysis methods with a new taxonomy-based decision support system. *European Journal of Operational Research*. 2022; 302(2): 633-651. doi:10.1016/j.ejor.2022.01.011.

[3] Barrera F, Segura M, Maroto C. Multiple criteria decision support system for customer segmentation using a sorting outranking method. *Expert Systems with Applications*. 2024; 238: 122310. doi:10.1016/j.eswa.2023.122310.

[4] Wątróbski J, Jankowski J, Ziemba P, Karczmarczyk A, Zioło M. Generalised framework for multi-criteria method selection. *Omega*. 2019; 86: 107-124.

doi:10.1016/j.omega.2018.07.004.
[5] Baydaş M, Yılmaz M, Jović Ž, Stević Ž, Özuyar SEG, Özçil A. A comprehensive MCDM assessment for economic data: success analysis of maximum normalization, CODAS, and fuzzy approaches. *Financial Innovation*. 2024; 10(1): 105. doi:10.1186/s40854-023-00588-x.
[6] Wang TC, Lee HD. Developing a fuzzy TOPSIS approach based on subjective weights and objective weights. *Expert Systems with Applications*. 2009; 36(5): 8980-8985. doi:10.1016/j.eswa.2008.11.035.
[7] Mukhametzyanov IZ. Specific character of objective methods for determining weights of criteria in MCDM problems: entropy, CRITIC and SD. *Decision Making: Applications in Management and Engineering*. 2021; 4(2): 76-105. doi:10.31181/dmame210402076i.
[8] Ayan B, Abacıoğlu S, Basilio MP. A comprehensive review of the novel weighting methods for multi-criteria decision-making. *Information*. 2023; 14: 285. doi:10.3390/info14050285.
[9] Erbey A, Fidan Ü, Gündüz C. A robust hybrid weighting scheme based on IQRBOW and entropy for MCDM: stability and advantage criteria in the VIKOR framework. *Entropy*. 2025; 27: 867. doi:10.3390/e27080867.
[10] Amusan OT, Nwulu NI, Gbadamosi SL. Multi-criteria decision-based hybrid energy selection system using CRITIC weighted CODAS approach. *Scientific African*. 2024; 26: e02372. doi:10.1016/j.sciaf.2024.e02372.
[11] Paramanik AR, Sarkar S, Sarkar B. OSWMI: an objective-subjective weighted method for minimizing inconsistency in multi-criteria decision making. *Computers & Industrial Engineering*. 2022; 169: 108138. doi:10.1016/j.cie.2022.108138.
[12] Gere A, Szakál D, Héberger K. Multiobject optimization of National Football League drafts: comparison of teams and experts. *Applied Sciences*. 2022; 12(13): 6303. doi:10.3390/app12136303.
[13] Cinelli M, Kadziński M, Gonzalez M, Słowiński R. How to support the application of multiple criteria decision analysis? Let us start with a comprehensive taxonomy. *Omega*. 2020; 96: 102261. doi:10.1016/j.omega.2020.102261.
[14] Vinogradova I. Multi-attribute decision-making methods as a part of mathematical optimization. *Mathematics*. 2019; 7: 915. doi:10.3390/math7100915.
[15] Scheirer WJ, Kumar N, Belhumeur PN, Boult TE. Multi-attribute spaces: calibration for attribute fusion and similarity search. In: *IEEE Conference on Computer Vision and Pattern Recognition*. Providence, RI, USA: IEEE; 2012. p.2933-2940. doi:10.1109/CVPR.2012.6248021.
[16] Dwork C, Kumar R, Naor M, Sivakumar D. *Rank aggregation revisited*. 2002. Available from: https://www.researchgate.net/publication/2869423_Rank_Aggregation_Revisited [Accessed 10th November 2026].
[17] Boix-Cots D, Pardo-Bosch F, Pujadas P. A systematic review on multi-criteria group decision-making methods based on weights: analysis and classification scheme. *Information Fusion*. 2023; 96: 16-36. doi:10.1016/j.inffus.2023.03.004.
[18] Héberger K. Sum of ranking differences compares methods or models fairly. *TrAC Trends in Analytical Chemistry*. 2010; 29(1): 101-109. doi:10.1016/j.trac.2009.09.009.
[19] Kollár-Hunek K, Héberger K. Method and model comparison by sum of ranking differences in cases of repeated observations (ties). *Chemometrics and Intelligent Laboratory Systems*. 2013; 127: 139-146. doi:10.1016/j.chemolab.2013.06.007.
[20] Héberger K, Kollár-Hunek K. Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers. *Journal of Chemometrics*. 2011; 25(4): 151-158. doi:10.1002/cem.1320.
[21] Sziklai BR, Baranyi M, Héberger K. Does cross-validation work in telling rankings apart? *Central European Journal of Operations Research*. 2025; 33(4): 1503-1528. doi:10.1007/s10100-024-00932-1.
[22] Héberger K. Sum of Euclidean distance differences and sum of absolute Manhattan distance differences: multicriteria decision making tools for small data tables. *Analytica Chimica Acta*. 2026; 1381: 344649. doi:10.1016/j.aca.2025.344649.
[23] Héberger K, Rajkó R. Generalization of pair-correlation method (PCM) for nonparametric variable selection. *Journal of Chemometrics*. 2002; 16(8-10): 436-443. doi:10.1002/cem.748.
[24] Rajkó R, Héberger K. Conditional Fisher's exact test as a selection criterion for pair-correlation method. Type I and type II errors. *Chemometrics and Intelligent Laboratory Systems*. 2001; 57(1): 1-14. doi:10.1016/S0169-7439(01)00101-0.
[25] Andrić F, Bajusz D, Rácz A, Šegan S, Héberger K. Multivariate assessment of lipophilicity scales—computational and reversed phase thin-layer chromatographic indices. *Journal of Pharmaceutical and Biomedical Analysis*. 2016; 127: 81-93. doi:10.1016/j.jpba.2016.04.001.
[26] Todeschini R, Grisoni F, Ballabio D. Deep ranking analysis by power eigenvectors (DRAPE): a wizard for ranking and multi-criteria decision making. *Chemometrics and Intelligent Laboratory Systems*. 2019; 191: 129-137. doi:10.1016/j.chemolab.2019.06.005.
[27] Martín J, Velázquez N, Asuero AG. Youden two-sample method. In: Kounis LD. (ed.) *Quality Control and Assurance-An Ancient Greek Term Re-Mastered*. London: IntechOpen; 2017. doi:10.5772/66397.
[28] Elaiwat S, Daoud S, Jaradat NJ, Taha M. Deep learning model for activity cliffs prediction: a

comprehensive approach to protein kinase inhibitors. *Journal of Computer-Aided Molecular Design*. 2026; 40: 9. doi:10.1007/s10822-025-00721-1.

[29] Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*. 2014; 15: 3133-3181.

[30] Bernal L, Rastelli G, Pinzi L. Improving machine learning classification predictions through SHAP and features analysis interpretation. *Journal of Chemical Information and Modeling*. 2025; 65(21): 11716-11732. doi:10.1021/acs.jcim.5c02015.

[31] Torres LF. *XGBoost: the king of machine learning algorithms, a state-of-the-art algorithm in data science*. 2023. Available from: https://medium.com/latinxinai/xgboost-the-king-of-machine-learning-algorithms-6b5c0d4acd87 [Accessed 10th November 2026].

[32] Lin S. Rank aggregation methods. *WIREs Computational Statistics*. 2010; 2(5): 555-570. doi:10.1002/wics.111.

[33] Legendre P. Species associations: the Kendall coefficient of concordance revisited. *Journal of Agricultural, Biological, and Environmental Statistics*. 2005; 10: 226-245. doi:10.1198/108571105X46642.

[34] Kendall MG, Babington Smith B. The problem of *m* rankings. *Annals of Mathematical Statistics*. 1939; 10: 275-287. doi:10.1214/aoms/1177732186.

[35] Diaconis P, Graham RL. Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society Series B (Methodological)*. 1977; 39(2): 262-268. doi:10.1111/j.2517-6161.1977.tb01624.x.

[36] Consonni V, Ballabio D, Cruz Muñoz E, Termopoli V, Todeschini R. Smoothed power-weakness ratio (sPWR): a new informative system for multi-criteria decision making. *Chemometrics and Intelligent Laboratory Systems*. 2026; 270: 105624. doi:10.1016/j.chemolab.2025.105624.

[37] Sałabun W, Wątróbski J, Shekhovtsov A. Are MCDA methods benchmarkable? A comparative study of TOPSIS, VIKOR, COPRAS, and PROMETHEE II methods. *Symmetry*. 2020; 12: 1549. doi:10.3390/sym12091549.

[38] Shapley LS. A value for *n*-person games. In: Kuhn HW, Tucker AW. (eds.) *Contributions to the Theory of Games*. Vol. 2. Princeton, NJ: Princeton University Press; 1953. p.307-317. doi:10.1515/9781400881970-018.

[39] Lourenço JM, Lebensztajn L. Post-Pareto optimality analysis with sum of ranking differences. *IEEE Transactions on Magnetics*. 2018; 54(8): 8202810. doi:10.1109/TMAG.2018.2836327.

[40] Rácz A, Bajusz D, Héberger K. Consistency of QSAR models: correct split of training and test sets, ranking of models and performance parameters. *SAR and QSAR in Environmental Research*. 2015; 26(7-9): 683-700. doi:10.1080/1062936X.2015.1084647.

[41] Jin X, Li S, Zhang W, Zhu J, Sun J. Prediction of soil-available potassium content with visible near-infrared ray spectroscopy of different pretreatment transformations by the boosting algorithms. *Applied Sciences*. 2020; 10(4): 1520. doi:10.3390/app10041520.

[42] Pore S, Szepesi Z, Roy K. Machine learning-based quantitative structure activity relationship modeling of repeated dose toxicity: a data-driven approach following Organisation for Economic Co-operation and Development test guidelines 407, 408, and 422 supported by experimental validation. *Chemical Research in Toxicology*. 2026; 39(4): 544-565. doi:10.1021/acs.chemrestox.5c00459.

[43] Nie M, Meng L, Chen X, Hu X, Li L, Yuan L, et al. Tuning parameter identification for variable selection algorithm using the sum of ranking differences algorithm. *Journal of Chemometrics*. 2019; 33(4): e3113. doi:10.1002/cem.3113.

[44] Valsecchi C, Ballabio D, Consonni V, Todeschini R. Deep ranking analysis by power eigenvectors (DRAPE): a polypharmacology case study. *Chemometrics and Intelligent Laboratory Systems*. 2020; 203: 104001. doi:10.1016/j.chemolab.2020.104001.

[45] Roy K, Mitra I, Ojha PK, Kar S, Das RN, Kabir H. Introduction of r_m^2 (rank) metric incorporating rank-order predictions as an additional tool for validation of QSAR/QSPR models. *Chemometrics and Intelligent Laboratory Systems*. 2012; 118: 200-210. doi:10.1016/j.chemolab.2012.06.004.

[46] Chen X, Xu Y, Meng L, Chen X, Yuan L, Cai Q, et al. Non-parametric partial least squares-discriminant analysis model based on sum of ranking difference algorithm for tea grade identification using electronic tongue data. *Sensors and Actuators B: Chemical*. 2020; 311: 127924. doi:10.1016/j.snb.2020.127924.

[47] Li S, Sun P. Toward more trustworthy QSAR: a systematic discussion on data set partitioning. *Journal of Chemical Information and Modeling*. 2026; 66(4): 2199-2210. doi:10.1021/acs.jcim.5c02465.

[48] Pavurala N, Madhavarao CN, Lee J, Das J, Ashraf M, O'Connor T. Cell culture media and Raman spectra preprocessing procedures impact glucose chemometrics. *Journal of Chemometrics*. 2025; 39: e70005. doi:10.1002/cem.70005.

[49] Costa N, Lourenço J. A proposal for selecting a Pareto solution with desirable properties. *Sustainability*. 2026; 18: 5843. doi:10.3390/su18125843.

[50] Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*. 1956; 63(2): 81-97. doi:10.1037/h0043158.

[51] Kingston NM, Tiemann G. Spearman-Brown prophecy formula. In: *The SAGE Encyclopedia of Research Design*. 2nd ed. Thousand Oaks (CA): SAGE Publications, Inc.; 2022. p.1571. doi:10.4135/9781071812082.n584.

[52] Jan SL, Shieh G. An improved nonparametric test and sample size procedures for the randomized complete block designs. *Sankhya B, The Indian Journal of Statistics*. 2025; 87: 686-711. doi:10.1007/s13571-025-00362-2.
[53] Tversky A. Features of similarity. *Psychological Review*. 1977; 84(4): 327-352. doi:10.1037/0033-295X.84.4.327.
[54] Carnovalini F, Rodà A, Wiggins GA. Popularity bias in recommender systems: the search for fairness in the long tail. *Information*. 2025; 16(2): 151. doi:10.3390/info16020151.
[55] Miranda-Quintana RA, Bajusz D, Rácz A, Héberger K. Extended similarity indices: the benefits of comparing more than two objects simultaneously. Part 1: theory and characteristics. *Journal of Cheminformatics*. 2021; 13: 32. doi:10.1186/s13321-021-00505-3.
[56] Sziklai BR, Gere A, Héberger K, Staudacher J. rSRD: an R package for the sum of ranking differences statistical procedure. *The R Journal*. 2026; 18(3): 97-119. doi:10.32614/RJ-2026-045.