

Article

Top-Weighted Structural Analysis of Ordered List Averaging for Medical Knowledge Extraction

Ricardo Sánchez-de-Madariaga ^{1,2,*} , Mario Pascual Carrasco ^{1,3} and Adolfo Muñoz Carrero ^{1,3} 

¹ Telemedicine and e-Health Research Unit, Carlos III Health Institute, Monforte de Lemos 5, 28029 Madrid, Spain; mario.pascual@isciii.es (M.P.C.); adolfo.munoz@isciii.es (A.M.C.)

² Research Network on Chronicity, Primary Care and Health Prevention and Promotion (RICAPPS), Carlos III Health Institute, 28029 Madrid, Spain

³ Infectious Diseases Thematic Area of the Networked Biomedical Research Center—CIBER (CIBERINFEC), Carlos III Health Institute, 28029 Madrid, Spain

* Correspondence: ricardo.sanchez@isciii.es

Abstract

Ordered List Averaging (OLA) was previously introduced as a methodology for evaluating machine learning algorithms according to their ability to prioritize medically relevant feature subsets. The original methodology demonstrated, through Friedman–Nemenyi statistical analysis applied to probability distributions of related datasets, that OLA can identify statistically supported machine learning algorithms for medical knowledge extraction. However, the internal structural organization of OLA-generated rankings has not previously been investigated. This work presents a structural analysis framework for OLA based on top-weighted ranking-agreement measures. Rather than proposing a new machine learning model-selection algorithm, the objective is to determine whether the ordering induced by OLA scores exhibits an intrinsic structural organization that provides complementary evidence supporting the interpretation of OLA rankings. To this end, ranking signatures generated by OLA are compared using Kendall distance, while the agreement between score-based and structure-based rankings is evaluated through Top-k overlap and Rank-Biased Overlap (RBO). Experiments conducted on nine biomedical datasets show that datasets previously exhibiting statistically validated OLA behavior consistently present stronger top-weighted structural agreement than datasets with weaker statistical support. Furthermore, the observed structural patterns remain stable across different values of the RBO persistence parameter and are robust to variations in dataset size. The results indicate that OLA rankings possess a meaningful and reproducible structural organization beyond their numerical scores. Consequently, top-weighted structural agreement provides complementary evidence supporting the interpretation of OLA rankings and strengthens the practical interpretation of OLA-based algorithm selection, particularly in situations where only a single working dataset is available and statistical validation through Friedman–Nemenyi cannot be performed.



Academic Editor: Hector Florez

Received: 21 May 2026

Revised: 15 July 2026

Accepted: 17 July 2026

Published: 18 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: Ordered List Averaging; machine learning; knowledge extraction; ranking analysis; Rank-Biased Overlap; Top-k overlap; Kendall distance; medical datasets

MSC: 68T09

1. Introduction

Machine learning (ML) algorithms are increasingly used not only for predictive modeling but also as tools for extracting domain knowledge from complex biomedical

datasets [1,2]. In many medical applications, the objective is not merely to maximize predictive performance but to identify meaningful relationships among variables that can support scientific interpretation and clinical decision making. Consequently, methodologies capable of evaluating how machine learning algorithms organize and prioritize relevant information have attracted increasing attention.

Ordered List Averaging (OLA), originally introduced as the second component of a previously published methodology for medical knowledge extraction [3], addresses this objective by ranking machine learning algorithms according to their ability to prioritize feature subsets that are consistent with expert medical knowledge. Unlike conventional model evaluation approaches, OLA does not seek the classifier with the highest predictive performance. Instead, it assigns each algorithm a score reflecting how effectively it ranks medically relevant feature subsets near the top of several ordered lists. The original methodology demonstrated that, when probability distributions of related datasets are available, Friedman and Nemenyi statistical analyses [4–6] can identify a statistically supported set of machine learning algorithms suitable for medical knowledge extraction.

Although this statistical validation establishes the practical usefulness of OLA, an important question remains unanswered. The original methodology evaluates algorithms through OLA scores but does not investigate the internal organization of the resulting rankings. In particular, it remains unclear whether OLA-generated rankings exhibit a meaningful structural organization or whether the observed ordering simply reflects numerical score differences. Understanding this internal structure is important because it may provide additional evidence supporting the interpretation of OLA rankings and the practical selection of machine learning algorithms when only a single working dataset is available and statistical validation through Friedman–Nemenyi cannot be performed.

Ranking analysis techniques provide a natural framework for investigating structural relationships between ordered lists [7,8]. Measures such as Kendall’s rank correlation [9,10], Top-k overlap [11] and Rank-Biased Overlap (RBO) [12] quantify different aspects of agreement between rankings, particularly emphasizing agreement among the highest-ranked elements. These measures have been widely applied in information retrieval, recommender systems, rank aggregation and voting theory, but, to our knowledge, they have not previously been used to analyze the internal structural organization of rankings generated by OLA.

The objective of this work is therefore not to propose a new model-selection algorithm, nor to replace the statistical validation framework previously established for OLA. Instead, we introduce a structural analysis framework that characterizes OLA-generated rankings through top-weighted ranking-agreement measures. By comparing the rankings induced by OLA scores with the rankings derived from the structural similarity of OLA signatures, we investigate whether OLA possesses an intrinsic structural organization beyond its numerical scoring mechanism.

Experimental results obtained from nine biomedical datasets show that datasets previously exhibiting statistically validated OLA behavior also present consistently stronger top-weighted structural agreement than datasets with weaker statistical support. Furthermore, the observed structural patterns remain stable across different ranking-agreement measures and parameter settings. These findings indicate that structural agreement provides complementary evidence supporting the interpretation of OLA rankings and contributes to a better understanding of how OLA identifies machine learning algorithms for medical knowledge extraction.

An important characteristic of the proposed framework is that it does not seek a universally optimal machine learning algorithm. Instead, OLA produces a ranking that depends on the statistical characteristics of each individual dataset, reflecting how different

algorithms organize medically relevant feature relationships. The structural analysis introduced in this work is intended to complement this process by investigating whether the organization induced by OLA is internally coherent. Consequently, the proposed framework provides additional insight into why a particular machine learning algorithm emerges as the highest-ranked candidate for a given dataset, rather than assuming that the same algorithm should be preferred across different biomedical problems.

The remainder of this paper is organized as follows. Section 2 reviews previous work on algorithm comparison and ranking analysis. Section 3 describes the proposed structural analysis framework. Section 4 presents the experimental methodology, Section 5 discusses the experimental results, and Section 6 concludes the paper.

2. Related Work

The evaluation of machine learning algorithms has traditionally focused on predictive performance measures such as accuracy, precision, recall, and F1-score [2]. Statistical comparison procedures including the Friedman test and the Nemenyi post hoc test are widely used for comparing machine learning algorithms across multiple datasets [1].

These approaches provide principled methodologies for identifying statistically competitive algorithms. However, they also present important limitations in practical knowledge extraction scenarios.

First, statistical comparison procedures typically require distributions of datasets or repeated experimental settings in order to estimate reliable significance measures [3].

Second, post hoc analyses such as Nemenyi usually identify groups of statistically equivalent algorithms rather than a unique model selection [6].

In many practical applications, however, researchers require a single working dataset, and a single machine learning model for extracting meaningful knowledge [13].

This limitation motivates the need for complementary methodologies capable of supporting single-model selection without requiring distributions of datasets.

Ranking-based methodologies provide a natural framework for analyzing how machine learning algorithms organize relationships between variables [10,11,14].

Unlike traditional feature-selection approaches focused on optimizing predictive performance [15,16], relevance-oriented ranking methodologies investigate the structural organization induced by predictive ordering behavior.

Measures of ranking similarity have been extensively studied in statistics, information retrieval, recommender systems, voting theory, and rank aggregation [9,11,12].

Among the most widely used measures are:

- Kendall tau correlation;
- Spearman correlation;
- Top-k overlap;
- Rank-Biased Overlap (RBO).

Kendall-based approaches evaluate global agreement between rankings through pairwise concordance analysis [7,8]. Formally, the Kendall distance counts the number of discordant ordering pairs:

$$K(\tau_1, \tau_2) = \#\{(i, j) : i < j, (\tau_{1(i)} - \tau_{1(j)}) (\tau_{2(i)} - \tau_{2(j)}) < 0\}$$

Although Kendall correlation provides a global measure of ranking similarity, it assigns comparable importance to disagreements occurring at all ranking positions.

In contrast, top-weighted agreement measures prioritize similarity near the highest-ranked positions [11].

Top-k overlap evaluates agreement among the first (k) positions of two rankings, whereas Rank-Biased Overlap introduces progressively decreasing weights for deeper ranking positions [12].

Top-weighted ranking measures were originally developed in information retrieval to reflect the observation that agreement among the highest-ranked items is generally more informative than agreement occurring at lower positions. In retrieval systems, users typically inspect only the first results returned by a search engine, making top-ranked agreement considerably more relevant than complete ranking similarity [17]. Rank-Biased Overlap (RBO) was specifically introduced to quantify agreement between indefinite rankings while assigning progressively decreasing weights to lower-ranked items through a persistence parameter [12]. Similarly, Top-(k) overlap focuses exclusively on agreement within the highest-ranked positions [18]. These measures have subsequently been adopted in recommendation systems, rank aggregation and comparative ranking analysis whenever emphasis on the leading positions is required. Despite the extensive literature on ranking comparison, relatively little work has focused on analyzing the structural organization induced by relevance-oriented ranking methodologies such as OLA.

The motivation for adopting top-weighted ranking measures in the present work is directly related to the construction of Ordered List Averaging (OLA). OLA assigns higher scores to machine learning algorithms that place medically relevant feature subsets near the beginning of the ordered lists. Consequently, the upper positions of OLA rankings contain the strongest evidence regarding the knowledge extraction capability of each algorithm. Top-weighted agreement measures are therefore particularly appropriate for investigating whether the structural organization induced by OLA is consistent with its scoring mechanism.

Most previous work in machine learning evaluation has focused primarily on predictive optimization, feature-selection performance, or statistical comparison of classifiers [4,15,16].

The objective of this work is to investigate whether the structural organization induced by OLA rankings and the predictive ordering induced by OLA scores exhibit meaningful top-weighted agreement associated with dataset quality.

More specifically, we investigate whether datasets previously identified as statistically reliable through Friedman–Nemenyi analysis also exhibit stronger structural coherence under top-weighted ranking-agreement measures.

In this work, structural coherence refers to the degree of agreement between the ordering induced by OLA scores and the ordering induced by the structural similarity of OLA ranking signatures, as quantified through Kendall distance, Top-k overlap and Rank-Biased Overlap.

Several studies have investigated the stability of rankings or the agreement between ranking methods. In these works, ranking-agreement measures are generally employed either to compare different ranking algorithms or to evaluate the consistency of ranking procedures under different experimental conditions. Their primary objective is therefore the comparison of rankings rather than the structural interpretation of a specific ranking methodology.

The present work addresses a different problem. Rather than proposing a new machine learning model-selection method or a new ranking-agreement measure, we investigate the internal structural organization of Ordered List Averaging (OLA). OLA was originally developed as a methodology for evaluating machine learning algorithms according to their ability to prioritize medically relevant feature subsets and was statistically validated using Friedman–Nemenyi analysis when probability distributions of related datasets were available. However, its structural properties have not previously been studied.

Our contribution is therefore complementary to the original OLA methodology. We introduce a structural analysis framework based on top-weighted ranking-agreement measures that characterizes the organization of OLA-generated rankings. By comparing score-based rankings with structure-based rankings derived from OLA signatures, we evaluate whether the internal organization of OLA is consistent with its scoring mechanism. This analysis provides additional insight into the interpretation of OLA rankings and offers complementary evidence supporting the practical selection of machine learning algorithms when only a single working dataset is available.

3. Methodology

3.1. Ordered List Averaging (OLA)

Ordered List Averaging (OLA) evaluates machine learning algorithms through ordered rankings of feature subsets according to predictive performance [3].

For each machine learning algorithm (A_i), feature subsets of sizes

$$s \in \{3, 4, 5, 6\}$$

are independently generated and evaluated using the F1 measure.

For each subset size (s), all candidate feature subsets are sorted in descending order according to predictive performance:

$$F1(FS_1) \geq F1(FS_2) \geq \dots \geq F1(FS_m)$$

where

- (FS_j) denotes a feature subset;
- (m) is the number of evaluated subsets.

Consequently, each machine learning algorithm produces four ordered lists corresponding to subset sizes 3, 4, 5, and 6.

These ordered lists define the ranking signature induced by the algorithm.

3.2. Expert Knowledge Encoding

The OLA methodology incorporates domain knowledge through the manual encoding of synergistic relationships between pairs of features and the corresponding medical outcome. Each feature pair is classified as either medically relevant or non-relevant according to established biomedical knowledge available prior to the machine learning analysis. Importantly, this encoding is performed independently of the machine learning algorithms evaluated and remains fixed throughout the complete OLA procedure.

Feature subsets of size 2 are manually encoded according to external medical knowledge regarding their relationship with the target outcome [3].

Each pairwise subset receives a binary relevance value:

$$r(FS) \in \{0, 1\}$$

where

- ($r(FS) = 1$) indicates a medically relevant subset;
- ($r(FS) = 0$) indicates a non-relevant subset.

Importantly, these manually encoded subsets are not used to generate the rankings themselves. Instead, they provide external semantic information for evaluating the ranking organization induced by OLA. Specific details of this expert knowledge encoding can be found in the original OLA methodology published in [3].

As with any expert-guided methodology, this encoding may be affected by incomplete or imperfect domain knowledge. Consequently, the quality of the resulting OLA scores depends on the quality of the available medical knowledge. To assess the robustness of the proposed structural analysis with respect to this potential source of uncertainty, the present work includes a sensitivity analysis based on controlled perturbations of the expert relevance encoding.

Although the expert relevance encoding inevitably introduces a degree of subjectivity, it reflects prior biomedical knowledge rather than information extracted from the analyzed datasets. Consequently, it represents an external source of domain knowledge rather than an optimization criterion fitted to the data.

3.3. OLA Score

The OLA score evaluates the extent to which a machine learning algorithm prioritizes medically relevant feature relationships near the top positions of the ordered lists [3].

Algorithms that place medically relevant subsets earlier in the rankings receive higher OLA scores.

Thus, the OLA score measures the alignment between the following:

- Predictive ordering;
- Externally guided medical relevance.

Unlike traditional feature-selection methodologies, OLA does not seek a single feature subset maximizing predictive performance. Instead, it evaluates how machine learning algorithms organize medically meaningful ranking structures.

3.4. Ranking Signatures

For each machine learning algorithm A_i , the collection of ordered lists generated by OLA defines a ranking signature:

$$S_i = \{L_3^{\{i\}}, L_4^{\{i\}}, L_5^{\{i\}}, L_6^{\{i\}}\}$$

These signatures characterize:

- Ranking organization;
- Predictive ordering behavior;
- Medically guided structural relationships.

The objective of this work is to analyze the structural relationships induced by these ranking signatures.

OLA generates four ordered lists corresponding to feature-subset sizes 3, 4, 5 and 6. The structural analysis proposed in this work is applied to the merged signature obtained from these four lists rather than to any individual subset size.

3.5. Kendall Distance Matrix

To analyze structural relationships between OLA-generated ranking signatures, pairwise Kendall distances [9,10] were computed between all machine learning algorithms.

For a set of (k) machine learning algorithms

$$A = \{A_1, A_2, \dots, A_k\}$$

the pairwise Kendall distances define a symmetric distance matrix:

$$D = [D_{ij}]_{k \times k}$$

where

$$D_{ij} = \text{Kendall}(S_i, S_j)$$

denotes the Kendall distance between the ranking signatures generated by algorithms:

$$A_i \text{ and } A_j$$

Since each algorithm produced four ranking signatures, one for each feature size 3, 4, 5 and 6, these four signatures were merged by weighing each feature with the order it occupies on each signature, summing up all these ordering numbers, and reordering with the total numbers. That is:

$$\text{Score}(f) = \sum_{s \in \{3, 4, 5, 6\}} \text{rs}(f)$$

The average structural distance of each algorithm was then computed as follows:

$$C_i = \left(\frac{1}{k-1} \right) \sum_{j \neq i} D_{ij}$$

where lower values indicate stronger structural agreement with the remaining algorithms.

3.6. Structural Ranking Analysis

To investigate the relationship between:

- Predictive ordering induced by OLA scores;
- Structural ordering induced by ranking geometry.

We analyze agreement between rankings using several ranking-comparison measures. Initially, global ranking agreement was evaluated through Kendall tau correlation.

However, because Kendall correlation assigns comparable importance to disagreements occurring at all ranking positions, additional top-weighted agreement measures were investigated.

3.7. Top-k Overlap

Top-k overlap evaluates the proportion of shared elements among the first (k) positions of two rankings [11].

Formally:

$$\text{Overlap}_{k(R_1, R_2)} = \frac{\left| \text{Top}_k(R_1) \cap \text{Top}_k(R_2) \right|}{k}$$

where $\text{Top}_k(R)$ denotes the first (k) elements of ranking (R).

This measure emphasizes agreement among the highest-ranked machine learning algorithms.

3.8. Rank-Biased Overlap (RBO)

To further emphasize agreement near the top positions of the rankings, we employ Rank-Biased Overlap (RBO) [12].

Unlike Kendall tau correlation, RBO assigns progressively lower importance to deeper ranking positions.

RBO is particularly suitable for the following:

- Short rankings;
- Partial agreement;
- Scenarios where top-ranked elements contain most of the informative structure.

The RBO score ranges between 0 (no agreement) and 1 (complete agreement).

The use of top-weighted measures is motivated by the construction of OLA itself. Since OLA scores increase when medically relevant feature subsets appear closer to the beginning of the ordered lists, the leading positions naturally contain the strongest evidence supporting the knowledge extraction capability of a machine learning algorithm. Consequently, structural agreement near the top of the rankings is expected to be more informative than agreement restricted to lower-ranked positions.

3.9. Single-Model Selection Perspective

The original OLA methodology evaluated machine learning algorithms through Friedman–Nemenyi analysis over distributions of datasets related to a working dataset.

Although this framework successfully identified statistically competitive algorithms for medical knowledge extraction, it presented two practical limitations:

- It required distributions of related datasets;
- It produced statistically equivalent sets of algorithms rather than a unique selection.

In contrast, OLA naturally produces a complete ranking of machine learning algorithms from a single dataset.

The objective of the present work is therefore not to replace statistical validation procedures, but rather to investigate whether structural ranking agreement provides additional evidence supporting the practical use of the highest-scoring OLA algorithm as a single-model selection criterion for knowledge extraction.

4. Experimental Setup

4.1. Machine Learning Algorithms

The experiments consider seven machine learning algorithms:

- Logistic Regression (LR);
- Naive Bayes (NB);
- Support Vector Machines (SVMs);
- Multilayer Perceptron (MLP);
- Decision Trees (DTs);
- k-Nearest Neighbors (KNs);
- Random Forests (RFs).

Each algorithm independently generates OLA ranking signatures. Appendix A resumes the principal characteristics of the seven selected ML algorithms.

4.2. Datasets

Four working datasets from the UCI Machine Learning Repository [19] were originally analyzed. These four datasets coincide with the datasets analyzed in [3]. Five additional working datasets (<http://www.openml.org>) were included in the research.

These nine biomedical datasets were selected to represent different medical classification problems. The datasets comprise a variety of clinical domains, including one veterinary application, and range from approximately 150 to some 3600 observations. Table 1 summarizes the basic structure of the nine datasets. Appendix B presents a short description of the nine datasets, including the medical outcome and features of each dataset.

In the original OLA methodology, each working dataset (the four original datasets D1–D4) generated a distribution of related datasets, allowing statistical evaluation through Friedman–Nemenyi analysis [3].

The original study [3] showed that datasets 1, 2, and 3 produced statistically significant Friedman results, whereas dataset 4 produced $p > 0.05$, and was therefore considered

as a dataset for which the Friedman test was not statistically significant in the original OLA validation.

Table 1. Summary of the nine working datasets used.

#	Name	Feature Types	Missing Data	Class Distribution	Instances
1	HP-UCI	1 CAT/11 NUM	NO	75/540	615
2	HF-UCI	5 CAT/7 NUM	NO	96/203	299
3	HD-UCI	4 CAT/8 NUM	NO	138/161	299
4	CKD-UCI	4 CAT/9 NUM	NO	43/115	158
5	COL-UCI	9 CAT/3 NUM	NO	53/92	145
6	THY-OML	11 CAT/1 NUM	NO	225/3621	3621
7	BT-OML	7 CAT/2 NUM	NO	117/160	277
8	HP-OML	7 CAT/5 NUM	NO	215/785	1000
9	LY-OML	12 CAT/0 NUM	NO	81/67	148

This previous statistical characterization provides an independent reference framework for evaluating structural ranking coherence in the present study.

4.3. Ranking Construction

For each dataset:

1. OLA scores were computed for all machine learning algorithms;
2. Algorithms were ranked according to OLA scores;
3. Ranking-geometry measures were independently computed from OLA signatures;
4. Algorithms were ranked according to structural agreement measures.

Agreement between predictive ordering and structural ordering was subsequently analyzed through Kendall tau correlation, Top-k overlap, and Rank-Biased Overlap.

5. Results and Discussion

5.1. Predictive Orderings

Table 2 shows the predictive ordering of the seven ML algorithms based on OLA scores for the four datasets D1–D4 and for each FS size. The orderings at the right of the dataset names are the merged orderings of the four FS sizes.

Table 2. The ordered OLA scores reached by each ML algorithm for each FS size on the four datasets.

HP-UCI	KN	DT	SVM	RF	MLP	LR	NB
size 3	303.5(KN)	302.7(RF)	302.0(DT)	294.1(SVM)	277.7(LR)	274.1(NB)	272.2(MLP)
size 4	1353.7(KN)	1345.8(DT)	1341.9(RF)	1337.4(SVM)	1269.2(MLP)	1269.2(LR)	1238.2(NB)
size 5	3595.2(KN)	3585.2(SVM)	3553.3(DT)	3549.7(RF)	3462.1(MLP)	3416.2(LR)	3306.0(NB)
size 6	6290.1(SVM)	6257.1(KN)	6194.6(DT)	6187.0(RF)	6073.4(MLP)	6019.4(LR)	5801.1(NB)
HF-UCI	DT	RF	KN	NB	SVM	MLP	LR
size 3	477.3(DT)	461.8(RF)	419.5(KN)	391.6(NB)	379.2(MLP)	370.8(SVM)	357.3(LR)
size 4	2152.2(DT)	2104.7(RF)	1854.6(KN)	1785.7(NB)	1741.3(SVM)	1682.6(MLP)	1657.3(LR)
size 5	5678.2(DT)	5528.1(RF)	4847.2(KN)	4791.5(SVM)	4767.4(NB)	4610.0(MLP)	4523.4(LR)
size 6	9812.8(DT)	9607.1(RF)	8570.6(SVM)	8505.6(KN)	8352.3(NB)	8211.3(MLP)	8078.6(LR)

Table 2. *Cont.*

HD-UCI	MLP	RF	KN	SVM	LR	NB	DT
size 3	451.0(KN)	445.7(RF)	442.8(MLP)	441.7(DT)	440.9(LR)	439.7(NB)	439.1(SVM)
size 4	2069.9(KN)	2065.1(RF)	2058.1(MLP)	2054.1(SVM)	2051.5(LR)	2040.3(NB)	2032.8(DT)
size 5	5627.1(MLP)	5615.2(RF)	5605.8(LR)	5603.4(SVM)	5601.4(KN)	5588.5(NB)	5503.4(DT)
size 6	10020.2(MLP)	9986.3(NB)	9980.7(RF)	9965.6(SVM)	9964.4(KN)	9950.1(LR)	9778.1(DT)
CKD-UCI	NB	MLP	KN	RF	DT	SVM	LR
size 3	432.8(NB)	423.9(RF)	416.0(MLP)	409.7(DT)	408.0(SVM)	404.9(KN)	362.2(LR)
size 4	1769.9(NB)	1735.9(MLP)	1734.4(DT)	1730.9(KN)	1726.1(RF)	1706.3(SVM)	1600.8(LR)
size 5	4804.8(MLP)	4781.2(KN)	4719.3(NB)	4687.9(RF)	4661.9(DT)	4657.2(SVM)	4492.8(LR)
size 6	9043.1(KN)	8914.4(MLP)	8858.8(RF)	8726.9(NB)	8708.2(SVM)	8668.8(LR)	8660.2(DT)

Table 3 summarizes the OLA score orderings of the seven ML algorithms for the nine working datasets. For datasets D1–D4, it can be considered a detail of Table 2.

Table 3. OLA score ordering of the seven ML algorithms for the nine datasets.

HP-UCI	KN	DT	SVM	RF	MLP	LR	NB
HF-UCI	DT	RF	KN	NB	SVM	MLP	LR
HD-UCI	MLP	RF	KN	SVM	LR	NB	DT
CKD-UCI	NB	MLP	KN	RF	DT	SVM	LR
COL-UCI	LR	MLP	KN	RF	NB	SVM	DT
THY-OML	LR	NB	DT	SVM	KN	MLP	RF
BT-OML	LR	MLP	NB	SVM	DT	RF	KN
HP-OML	RF	KN	DT	SVM	LR	MLP	NB
LY-OML	NB	KN	SVM	RF	DT	MLP	LR

Table 4 summarizes the Nemenyi sets of ML algorithms produced by the Nemenyi post hoc test in [3] for datasets D1–D4. It also shows the top-ranked OLA algorithm for these datasets using the OLA methodology.

Table 4. Nemenyi sets of the four UCI datasets in the original OLA methodology in [3].

	Nemenyi	Highest OLA alg.	Included
HP-UCI	RF, DT, SVM, KN	KN	YES
HF-UCI	DT, NB, LR, RF	DT	YES
HD-UCI	RF, DT, NB, SVM, KN, MLP	MLP	YES
CKD-UCI	DT, LR, SVM, KN, RF, MLP	NB	NO

Table 4 shows that the highest-ranked OLA algorithm is included inside the Nemenyi set in datasets D1, D2, and D3. However, this is not the case for dataset D4 (CKD-UCI). This result confirms the failure of that dataset in the p -value of the Friedman test as well as the p -values of the Kolmogorov–Smirnov test applied to its probability distribution generated in [3] using the Dataset Feature Splitting (DFS) methodology.

Sensitivity Analysis of Expert Knowledge Encoding

Table 5 shows the highest-ranked OLA algorithm in the nine working datasets. In order to establish the reliability and stability of these top-ranked OLA algorithms, we have performed sensitivity analysis of the expert knowledge used to compute the OLA score rankings.

Table 5. Sensitivity analysis of the expert knowledge of the nine datasets.

	Nemenyi	Highest OLA alg.	Included
HP-UCI	KN, DT	KN	YES
HF-UCI	DT, RF	DT	YES
HD-UCI	MLP, RF	MLP	YES
CKD-UCI	MLP	NB	NO
COL-UCI	LR, MLP	LR	YES
THY-OML	LR, NB, DT	LR	YES
BT-OML	LR, NB	LR	YES
HP-OML	RF	RF	YES
LY-OML	NB, KN	NB	YES

As introduced in the methodology in [3], medical knowledge is codified using lists of pairs of synergic features with respect to the medical outcome of the dataset. In the present research, sensitivity analysis has been performed by suppressing with 0.05 probability one or several of these feature pairs in a series of recurrent experiments and computing the resulting OLA score on each experiment.

Incidentally, this series of experiments has allowed us to generate a probability distribution of OLA scores where the Friedman–Nemenyi test may be applied. Table 5 also shows the Nemenyi sets produced by these tests in the nine working datasets.

Table 5 shows that the non-inclusion of the best OLA algorithm in the Nemenyi test of dataset D4 (CKD-UCI) is reconfirmed in the sensitivity analysis experiments results. In addition, the highest OLA algorithms of all five additional working datasets are indeed included within their respective Nemenyi tests.

5.2. Structural Orderings

Table 6 shows an example of the OLA signatures for the use case of the HD-UCI dataset (D3). The numbers correspond to the feature numbers of that dataset described in Appendix B.

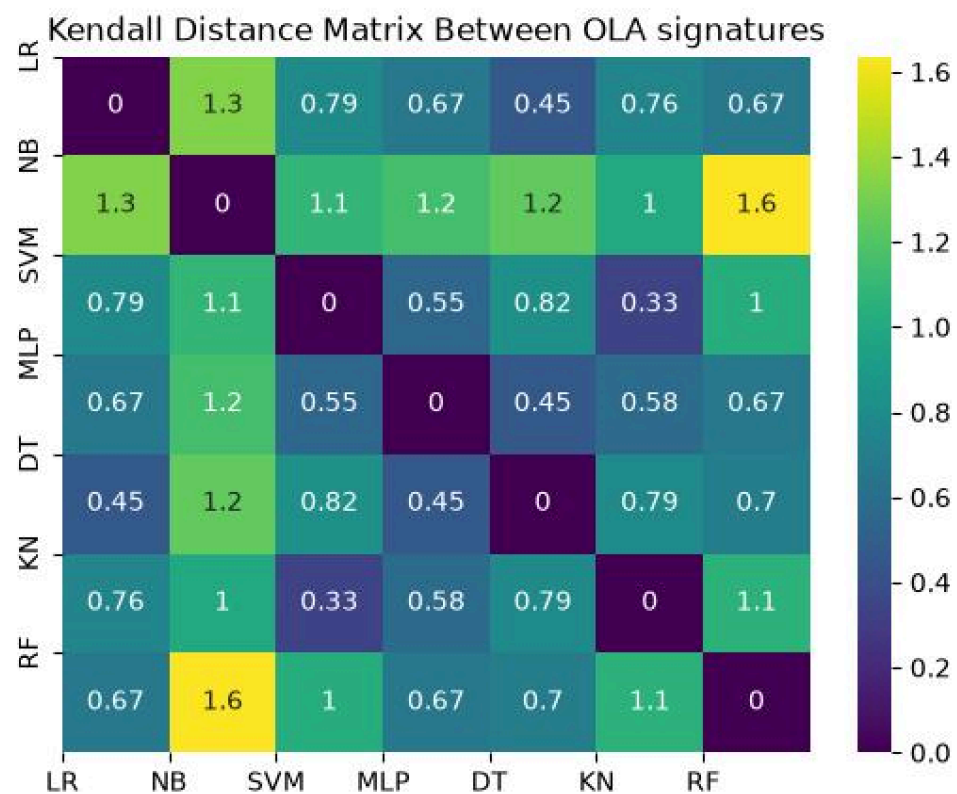
Figure 1 shows the pairwise Kendall distance matrix computed from the merged OLA signatures of the seven machine learning algorithms. Lower values indicate that two algorithms produce more similar structural organizations of the feature rankings, whereas higher values correspond to increasingly different ranking structures.

Several interesting patterns can be observed. First, the matrix is symmetric, as expected, and the diagonal contains values close to zero because each algorithm is identical to itself. More importantly, some pairs of algorithms exhibit substantially lower Kendall distances than others. For example, SVM and KN show the smallest structural distance (0.33), indicating that both algorithms generate highly similar OLA ranking signatures. Likewise, LR and DT (0.45) and MLP and DT (0.45) also present relatively strong structural similarity. In contrast, NB exhibits the largest distances with most of the remaining algorithms, particularly with RF (1.60), suggesting that Naïve Bayes induces a markedly different structural organization of feature relevance.

Table 6. The OLA signatures of the seven ML algorithms on the HD-UCI dataset.

LR	NB	SVM	MLP	DT	KN	RF
11	3	11	11	11	11	11
3	11	3	3	3	3	3
9	2	10	9	8	10	8
8	9	8	2	2	8	2
7	7	2	7	10	2	10
2	8	7	8	5	4	5
10	4	9	10	6	9	12
12	5	4	5	12	5	6
5	12	12	12	9	12	1
1	10	5	1	1	7	4
4	1	6	4	4	1	9
6	6	1	6	7	6	7

These observations indicate that OLA does not generate arbitrary rankings. Instead, machine learning algorithms naturally form groups exhibiting different degrees of structural similarity. This pairwise structural organization constitutes the basis for the subsequent ranking-geometry analysis. By aggregating the Kendall distances associated with each algorithm, a structure-based ranking can be constructed and subsequently compared with the OLA score ranking through Top-k overlap and Rank-Biased Overlap.

**Figure 1.** Kendall distance matrix generated by the OLA signatures of the HD-UCI dataset.

Tables 7 and 8 summarize the predictive and structural ML algorithm orderings produced by the OLA scores and the Kendall distances for datasets D1–D4 and datasets D5–D9, respectively.

Table 7. Predictive and structural ML algorithm orderings of datasets D1–D4. The numbers corresponding to the ML algorithms are those presented in Appendix A.

HP-UCI		HF-UCI		HD-UCI		CKD-UCI	
Predict.	Struct.	Predict.	Struct.	Predict.	Struct.	Predict.	Struct.
6	5	5	7	4	5	2	3
5	3	7	5	7	3	4	7
3	4	6	6	6	4	6	5
7	7	2	3	3	6	7	4
4	6	3	4	1	7	5	6
1	1	4	1	2	1	3	1
2	2	1	2	5	2	1	2

Table 8. Predictive and structural ML algorithm orderings of datasets D5–D9. The numbers corresponding to the ML algorithms are those presented in Appendix B.

COL-UCI		THY-OML		BT-OML		HP-OML		LY-OML	
Predict.	Struct.	Predict.	Struct.	Predict.	Struct.	Predict.	Struct.	Predict.	Struct.
1	3	1	2	1	2	7	7	2	3
4	1	2	1	4	4	6	1	6	6
6	7	5	3	2	3	5	6	3	1
7	4	3	7	3	1	3	3	7	2
2	6	6	5	5	6	1	2	5	7
3	2	4	6	7	5	4	4	4	4
5	5	7	4	6	7	2	5	1	5

5.3. Global Kendall Correlation

Initial experiments using global Kendall tau correlation revealed weak agreement between the following:

- Predictive ordering induced by OLA scores;
- Structural ordering induced by ranking geometry.

The obtained Kendall tau values were those shown in Table 9:

Table 9. Kendall tau values between predictive and structural orderings.

Dataset	Kendall Tau
D1	0.142857
D2	0.238095
D3	0.238095
D4	0.238095
D5	0.142857
D6	0.142857
D7	0.619047
D8	0.333333
D9	0.047619

The corresponding p -values were not statistically significant.

These results indicate that global agreement across the complete ranking space is weak.

5.4. Top-k Overlap and Rank-Biased Overlap (RBO) Analysis

Table 10 summarizes the RBO and Top-k analysis results for the nine working datasets.

Table 10. Summary of top-weight evaluation for the nine datasets.

	Friedman Quality	RBO	Top-1	Top-2	Top-3	Top-4	Top-5
HP-UCI	Significant	0.602	0.0	0.0	0.667	0.75	0.8
HF-UCI	Significant	0.650	0.0	0.5	0.667	0.75	0.8
HD-UCI	Significant	0.793	1.0	0.5	0.667	0.75	0.8
CKD-UCI	Non-significant	0.567	0.0	0.5	0.333	0.5	0.8
COL-UCI	N/A	0.626	0.0	0.5	0.333	0.75	0.8
THY-OML	N/A	0.721	0.0	1.0	0.667	0.75	0.8
BT-OML	N/A	0.686	0.0	0.5	0.667	1.0	0.8
HP-OML	N/A	0.793	1.0	0.5	0.667	0.75	0.8
LY-OML	N/A	0.650	0.0	0.5	0.667	0.75	0.8

Although global Kendall correlation produced weak agreement, Top-k overlap revealed substantially stronger agreement near the highest-ranking positions.

Datasets 3 and 8 exhibited the strongest top-ranked agreement. In contrast, dataset 4 exhibited noticeably weaker agreement.

These results suggest that the most informative structural relationships induced by OLA are concentrated near the highest-ranking positions.

RBO analysis reinforced the previous observations. Datasets 3 and 8 exhibited substantially stronger structural coherence than the remaining datasets.

Importantly, dataset 4—previously identified as a dataset for which the Friedman test was not statistically significant in the original OLA validation—produced the weakest structural agreement among the non-trivial datasets.

This observation is particularly relevant because the structural analysis performed in this work is independent from the original Friedman–Nemenyi analysis, yet both approaches consistently identify dataset 4 as structurally weaker.

Figure 2 shows that both Top-3 overlap and RBO follow a remarkably similar trend across the nine datasets. Datasets D3 and D8 consistently exhibit the highest structural agreement, whereas D4 and D5 present the weakest agreement. This consistency indicates that both measures capture similar aspects of the top-weighted organization induced by OLA despite being based on different mathematical principles.

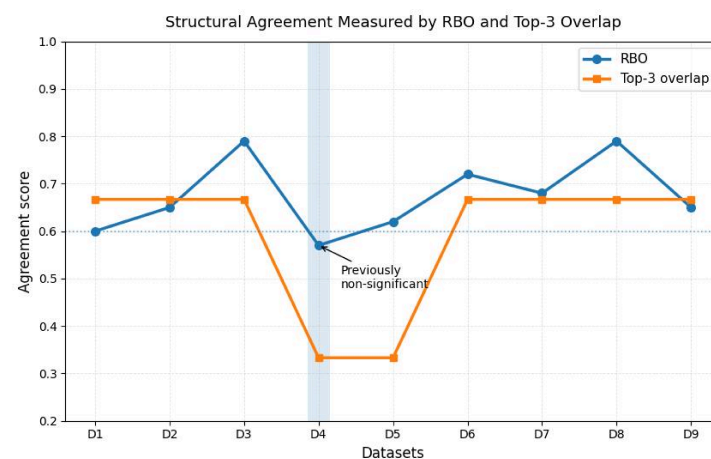


Figure 2. Top-weighted structural agreement across the nine datasets.

5.5. Sensitivity Analysis on the Persistence Parameter of RBO

Since the RBO measure is sensitive to the persistence parameter p , we have performed sensitivity analysis on the variation in this parameter from 0.80 to 0.95 and 0.98 in the algorithm orderings of the nine datasets.

The implementation of the RBO results obtained so far employed the default `rbo()` function of the software library (see Appendix A), whose default behavior corresponds to the unweighted overlap formulation ($p = 1.0$). To perform the sensitivity analysis, we adopted the extrapolated RBO formulation (`rbo_ext`), which explicitly incorporates the persistence parameter described by Webber et al. [12].

To evaluate the influence of the RBO persistence parameter, we repeated the analysis using ($p = 0.80, 0.85, 0.90, 0.95$) and (0.98). Table 11 summarizes the results obtained for the five parameter p values and the nine datasets. Figure 3 illustrates these results.

Table 11. Sensitivity analysis on the persistence parameter p of the RBO analysis for the nine working datasets.

P	80	85	90	95	98
HP-UCI	0.56	0.65	0.75	0.87	0.95
HF-UCI	0.62	0.70	0.79	0.89	0.95
HD-UCI	0.82	0.85	0.89	0.94	0.97
CKD-UCI	0.56	0.64	0.74	0.86	0.94
COL-UCI	0.59	0.68	0.77	0.88	0.95
THY-OML	0.70	0.76	0.83	0.91	0.96
BT-OML	0.65	0.72	0.81	0.90	0.96
HP-OML	0.82	0.85	0.89	0.94	0.97
LY-OML	0.62	0.70	0.79	0.89	0.95

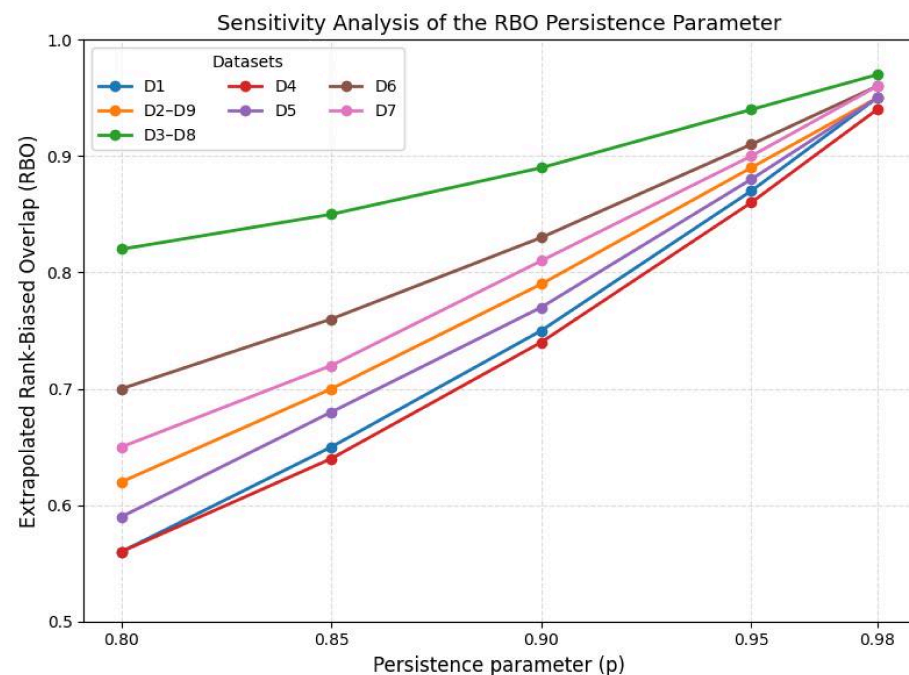


Figure 3. Sensitivity analysis of the extrapolated Rank-Biased Overlap (RBO) with respect to the persistence parameter (p). Although the absolute RBO values increase as (p) increases, the relative ordering of the datasets remains unchanged, demonstrating that the proposed structural analysis is robust to the choice of the persistence parameter.

Although the absolute values increased with the persistence parameter, the relative ordering of the datasets remained unchanged across all tested values. In particular, datasets D3 and D8 consistently exhibited the highest structural agreement, whereas D1 and D4 remained the lowest. These results indicate that the proposed structural analysis is robust with respect to the choice of the RBO persistence parameter.

5.6. Sensitivity Analysis on the Size of Dataset HP-OML

In order to further investigate the stability of the proposed single-model selection criterion, we performed sensitivity analysis on the size of one of the largest datasets, namely, the Hepatitis dataset HP-OML.

Table 12 shows that, by reducing the size from 1000 down to 500 observations, RBO and Top-k overlap decreased, indicating that the degree of structural agreement is influenced by the amount of available data. Nevertheless, the highest-scoring OLA algorithm (Random Forest) remained unchanged throughout the analysis. This result provides additional empirical evidence that the proposed OLA-based selection criterion is stable under substantial reduction in dataset size.

Table 12. Sensitivity analysis of the dataset Hepatitis of OpenML.

	Highest OLA alg.	RBO	Top-1	Top-3
HP-OML-1000	RF	0.793	1.0	0.667
HP-OML-750	RF	0.567	0.0	0.333
HP-OML-500	RF	0.579	0.0	0.667

5.7. Structural Agreement vs. Medical Relevance

An important observation emerging from the experiments is that structural centrality and medical relevance should not be interpreted as equivalent concepts.

Ranking-geometry measures such as:

- Kendall correlation;
- Top-k overlap;
- RBO

characterize internal structural organization and agreement properties of OLA-generated rankings.

However, they do not directly incorporate external medical knowledge.

In contrast, OLA scores explicitly depend on manually encoded medically relevant feature relationships.

Consequently, if the objective is identifying medically meaningful feature relationships, the highest-scoring OLA algorithm remains the primary criterion for practical model selection.

Structural agreement measures instead provide complementary information regarding:

- Ranking organization;
- Internal coherence;
- Top-weighted structural consistency.

Thus, structural centrality should not be interpreted as optimality, and consensus does not necessarily imply medical relevance.

5.8. Single-Model Selection

The original OLA methodology relied on Friedman–Nemenyi analysis over distributions of related datasets in order to identify statistically competitive machine learning algorithms.

However,

- Friedman–Nemenyi analysis requires multiple related datasets;
- Typically, it produces sets of statistically equivalent algorithms rather than a unique model selection.

In contrast, OLA naturally produces a complete ranking from a single working dataset.

The observed structural coherence revealed by Top-k overlap and Rank-Biased Overlap suggests that the highest-scoring OLA algorithm provides complementary structural evidence supporting the interpretation of the machine learning algorithm ranked highest by OLA for selecting a single machine learning model for medical knowledge extraction when only one dataset is available.

Importantly, this interpretation should not be viewed as replacing statistical validation procedures, but rather as providing complementary structural evidence supporting practical single-model selection.

5.9. Limitations

The present work has several limitations that should be considered when interpreting the reported results.

First, although the experimental study has been substantially expanded to include nine biomedical datasets covering different medical applications and one veterinary dataset, the overall number of datasets remains relatively limited. Consequently, additional evaluations involving larger collections of heterogeneous datasets will be required to establish the full generality of the observed structural patterns.

Second, OLA relies on expert-guided encoding of medically relevant synergistic feature pairs. While this external knowledge constitutes one of the strengths of the methodology, it also introduces a degree of subjectivity because the encoded relationships depend on the completeness and accuracy of current biomedical knowledge. To assess the robustness of the proposed framework with respect to this potential source of uncertainty, the present research includes perturbation experiments in which the expert knowledge is randomly modified. These experiments indicate that the highest-ranked OLA algorithm remains consistent with the corresponding Friedman–Nemenyi validation.

Third, the structural analysis proposed in this work has been developed and evaluated exclusively within the OLA framework. Although the ranking-agreement measures employed are general, further research will be required to determine whether similar structural behavior can be observed in other knowledge-oriented ranking methodologies.

Although top-weighted agreement measures reveal consistent structural patterns, the present work establishes association rather than causation. The observed structural coherence should therefore be interpreted as complementary evidence supporting the interpretation of OLA rankings rather than as proof of the superiority of a particular machine learning algorithm.

Finally, the present study should not be interpreted as proposing a universally applicable model-selection strategy. Rather, the proposed framework provides a structural interpretation of OLA-generated rankings and complementary evidence supporting the interpretation of OLA-based algorithm selection in medical knowledge extraction scenarios.

6. Conclusions and Future Work

In this work, we investigated the internal structural organization of Ordered List Averaging (OLA) through a ranking analysis framework based on Kendall distance, Top-(k) overlap, and Rank-Biased Overlap (RBO). Unlike the original OLA methodology, which evaluates machine learning algorithms according to medically guided relevance and validates its results through Friedman–Nemenyi analysis when probability distributions of

related datasets are available, the present study focuses on understanding the structural properties of the rankings generated by OLA.

The experimental evaluation performed on nine biomedical datasets consistently showed that datasets previously exhibiting statistically validated OLA behavior also present stronger top-weighted structural agreement than datasets with weaker statistical support. Furthermore, the observed structural patterns remained stable across different values of the RBO persistence parameter and under dataset size perturbations, providing additional evidence of the robustness of the proposed structural analysis.

An important outcome of this work is the distinction between two complementary sources of information provided by OLA. On the one hand, OLA scores incorporate expert medical knowledge through manually encoded synergistic feature relationships and therefore remain the primary criterion for identifying machine learning algorithms suitable for medical knowledge extraction. On the other hand, the structural analysis introduced in this work characterizes the internal organization of OLA rankings independently of the scoring procedure, providing complementary evidence regarding their coherence and reproducibility.

The results therefore suggest that OLA-generated rankings possess a meaningful structural organization beyond their numerical scores. Rather than constituting an alternative model-selection strategy, the proposed framework complements the original OLA methodology by providing an additional structural interpretation of its rankings. This complementary perspective contributes to understanding why OLA identifies particular machine learning algorithms and provides complementary structural evidence that is consistent with the practical interpretation of OLA-based algorithm selection established in the original methodology, especially in situations where only a single working dataset is available.

Furthermore, perturbation experiments based on uncertainty in the expert knowledge demonstrated that the highest-ranked OLA algorithm consistently belonged to the corresponding Friedman–Nemenyi set of statistically supported machine learning algorithms. This observation provides additional empirical support for the practical interpretation of OLA rankings.

Future work will investigate alternative ranking-geometry measures, graph-based representations of ranking structure, weighted agreement measures, and theoretical models relating structural organization to medically guided relevance. In addition, larger collections of biomedical datasets and broader experimental studies will further clarify the generality of the observed structural patterns.

Author Contributions: Conceptualization, R.S.-d.-M.; Methodology, R.S.-d.-M.; Software, R.S.-d.-M.; Validation, R.S.-d.-M.; Formal analysis, R.S.-d.-M.; Investigation, R.S.-d.-M.; Writing—original draft, R.S.-d.-M.; Writing—review and editing, R.S.-d.-M., M.P.C. and A.M.C.; Supervision, R.S.-d.-M., M.P.C. and A.M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in [UC Irvine Machine Learning Repository] at [<https://archive.ics.uci.edu/>] (accessed on 20 May 2026).

Acknowledgments: This work has been partially administratively supported by projects IMPACT DATA 2 (TMPY 187/25, MCI/AEI/FEDER, UE). (Infraestructura de Datos para la Cohorte de IMPaCT alineada con los Desarrollos Europeos), RICAPPS (Network for Research on Chronicity, Primary Care and Health Promotion) research network of the Spanish Ministry of Science and Innovation and CIBERINFEC (Consorcio Centro de Investigación Biomédica en Red—Enfermedades Infecciosas CB21/13/00091).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. The Seven Families of Supervised ML Algorithms Used

- LR. *Logistic Regression* has been implemented using the Scikit-learn library (version 1.6.1) [20] in IDLE Python 3.11, Wilmington, Delaware, USA.
- This algorithm belongs to the linear classification learning scientific family. It implements the linear regression algorithm, which is based on classical statistics, adapted to a categorical, i.e., non-numeric, target feature using a non-linear transformation. The maximum number of iterations has been set to 1000.
- NB. *Naïve Bayes* has been implemented using the Scikit-learn library written in Python.
- This algorithm belongs to the Bayesian statistical learning scientific family. A non-linear algorithm that sometimes may behave linearly.
- SVM. *Support Vector Machine* has been implemented using the Scikit-learn library written in Python.
- The hyperparameters have been adjusted to C from 0 to 100 at intervals of 10 and the 10 resulting scores have been averaged. An ‘rbf’ kernel has been used as mapping transformation.
- This algorithm belongs at the same time to the linear classification and instance-based learning scientific families. It uses a non-linear mapping transformation in order to treat non-linear problems.
- MLP. *Multilayer Perceptron* has been implemented using the Scikit-learn library written in Python. An architecture of [100 × 100] neurons has been used. An activation function ‘tanh’ has been performed for the hidden layer. An initial learning rate of 0.001 and a maximum number of iterations of 100,000 have been used. A ‘random_state’ 0 argument has been performed in order to have reproducible results across multiple function calls.
- This algorithm belongs to the linear classification learning scientific family. It implements the perceptron learning rule, which is the standard neural network architecture. Due to its multilayer structure, it is able to learn non-linear data concepts.
- DT. *Decision Tree* has been implemented using the Scikit-learn library written in Python.
- This algorithm belongs to the information theory learning scientific family. It creates step-function-like decision boundaries to learn from non-linear relationships in data.
- KN. *k-Nearest Neighbors* has been implemented using the Scikit-learn library written in Python.
- This algorithm belongs to the instance-based learning scientific family. It is considered to create decision boundaries that are often non-linear.
- RF. *Random Forest* has been implemented using the Scikit-learn library written in Python.

This algorithm belongs to the family of ensemble learning algorithms, specifically to bootstrap aggregation (bagging) methods based on Decision Trees (DTs).

Appendix B. The Nine Datasets Used

D1. HCV data. UCI Machine Learning Repository.

Hepatitis C prediction: 12 features, 615 instances.

List of numbered features.

Hepatitis C. 1 AGE, 2 SEX, 3 ALB albumin, 4 ALP alkaline phosphatase, 5 ALT alanine amino-transferase, 6 AST aspartate amino-transferase, 7 BIL bilirubin, 8 CHE choline esterase, 9 CHO cholesterol, 10 CRE creatinine, 11 GGT gamma glutamyl transpeptidase, 12 PRO protein.

D2. Heart Failure. UCI Machine Learning Repository.

Heart failure clinical records: 12 features, 299 instances.

List of numbered features.

Heart Failure. 1 AGE, 2 ANA anemia, 3 CPH creatinine phosphokinase, 4 DIA diabetes, 5 EIJ ejection fraction, 6 HBP high blood pressure, 7 PLA platelets, 8 SCR serum creatinine, 9 SSO serum sodium, 10 SEX, 11 SMO smoking, 12 TIM time.

D3. Heart Disease. UCI Machine Learning Repository.

Heart-h: 12 features, 299 instances.

List of numbered features.

Heart disease. 1 AGE, 2 SEX, 3 CP chest pain, 4 BPS blood pressure, 5 FBS fasting blood sugar, 6 ECG electrocardiographic results, 7 MHR maximum heart rate, 8 ANG exercise induced angina, 9 STD depression induced by exercise relative to rest, 10 SLO slope of the peak exercise ST segment, 11 CA number of major vessels colored by fluoroscopy, 12 CHO serum cholesterol.

D4. Chronic Kidney Disease. UCI Machine Learning Repository.

Chronic Kidney Disease: 13 features, 158 instances.

List of numbered features.

Chronic Kidney Disease. 1 AGE, 2 ALB albumin, 3 URE blood urea, 4 CRE serum creatinine, 5 SOD sodium, 6 POT potassium, 7 HEM hemoglobin, 8 WHI white blood cell count, 9 RED red blood cell count, 10 HTN hypertension, 11 DM diabetes mellitus, 12 CAD coronary artery disease, 13 ANE anemia.

D5. Horse Colic database. openML.org

Horse Colic: 12 features, 145 instances.

List of numbered features.

Horse Colic. 1 SUR surgery, 2 AGE, 3 REC rectal temperature, 4 PUL pulse, 5 RAT respiratory rate, 6 EXT temperature of extremities, 7 PER peripheral pulse, 8 MUC color of mucous membranes, 9 TIM capillary refill time, 10 PAI pain level, 11 PERI peristalsis, 12 DIS abdominal distension.

D6. Thyroid Disease records. openML.org

Thyroid Disease: 12 features, 3621 instances.

List of numbered features.

Thyroid Disease. 1 AGE, 2 SEX, 3 THY thyroxine, 4 ANT antithyroid medication, 5 SIC sick, 6 PRE pregnant, 7 SUR thyroid surgery, 8 I131 I131 treatment, 9 HYPO hypothyroidism, 10 HYPE hyperthyroidism, 11 LIT lithium, 12 GOI goiter.

D7. Breast Tumor. openML.org

Breast Tumor: 9 features, 277 instances.

List of numbered features.

Breast Tumor. 1 AGE, 2 MEN menopause, 3 NOD positive axillary lymph nodes, 4 CAP extracapsular extension, 5 GRA tumor grade, 6 BRE breast, 7 QUA breast quadrant, 8 RAD radiotherapy, 9 REC recurrence.

D8. Hepatitis. openML.org

Hepatitis: 12 features, 1000 instances.

List of numbered features.

Hepatitis. 1 AGE, 2 STE steroid, 3 ANT antivirals, 4 ANO anorexia, 5 SPI spiders, 6 ASC ascites, 7 VAR varices, 8 BIL bilirubin, 9 ALK alkaline phosphatase, 10 SGOT serum glutamic oxaloacetic transaminase, 11 ALB serum albumin, 12 PROT prothrombin time.

D9. Lymphography Dataset. openML.org

Lymphography: 12 features, 148 instances.

List of numbered features.

Lymphography. 1 LYM lymphatic vessels, 2 BYP collateral lymphatic circulation bypassing, 3 LYMD lymph nodes showing decreased size, 4 LYME enlarged lymph nodes, 5 SHA shape of the lymph nodes, 6 DEF filling defects, 7 MOR morphological changes,

8 STR alterations in the internal structure, 9 PAT characteristic radiological patterns, 10 DIS displacement of lymph nodes, 11 EXC exclusion due to obstruction, 12 NUM number of lymph nodes.

References

1. Mitchell, T. *Machine Learning*; McGraw-Hill: New York, NY, USA, 1997.
2. Witten, I.H.; Frank, E.; Hall, M.A. *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd ed.; Morgan Kaufmann: San Francisco, CA, USA, 2011.
3. Sánchez-de-Madariaga, R.; Pascual Carrasco, M.; Muñoz Carrero, A. A Methodology to Extract Knowledge from Datasets Using ML. *Mathematics* **2025**, *13*, 1807. [\[CrossRef\]](#)
4. Demšar, J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
5. Friedman, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675–701. [\[CrossRef\]](#)
6. Nemenyi, P. *Distribution-Free Multiple Comparisons*; Princeton University: Princeton, NJ, USA, 1963.
7. Wang, S.; Deng, Q.; Feng, S.; Zhang, H.; Liang, C. A Survey on Rank Aggregation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*; IJCAI: Bremen, Germany, 2024; pp. 8281–8289. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Hu, S.; Wang, X.; Lyu, S. Rank-based Decomposable Losses in Machine Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13240–13260.
9. Kendall, M.G. A New Measure of Rank Correlation. *Biometrika* **1938**, *30*, 81–93. [\[CrossRef\]](#)
10. Kendall, M.G.; Gibbons, J.D. *Rank Correlation Methods*, 5th ed.; Oxford University Press: Oxford, UK, 1990.
11. Dwork, C.; Kumar, R.; Naor, M.; Sivakumar, D. Rank Aggregation Methods for the Web. In *Proceedings of the 10th International Conference on World Wide Web*; Association for Computing Machinery: New York, NY, USA, 2001; pp. 613–622.
12. Webber, W.; Moffat, A.; Zobel, J. A Similarity Measure for Indefinite Rankings. *ACM Trans. Inf. Syst.* **2010**, *28*, 20. [\[CrossRef\]](#)
13. Raschka, S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *arXiv* **2018**, arXiv:1811.12808.
14. Sałabun, W.; Urbaniak, K. How strongly do rank similarity coefficients differ used in decision making problems? *Procedia Comput. Sci.* **2021**, *192*, 4570–4577. [\[CrossRef\]](#)
15. Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
16. Liu, H.; Motoda, H. *Feature Selection for Knowledge Discovery and Data Mining*; Springer: Berlin/Heidelberg, Germany, 1998.
17. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008.
18. Fagin, R.; Kumar, R.; Sivakumar, D. Comparing Top k Lists. *SIAM J. Discret. Math.* **2003**, *17*, 134–160. [\[CrossRef\]](#)
19. UC Irvine Machine Learning Repository. Available online: <https://archive.ics.uci.edu/> (accessed on 21 May 2026).
20. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.