Introducing the Rank-Biased Overlap as Similarity Measure for Feature Importance in Explainable Machine Learning: A Case Study on Parkinson's Disease

Article *in* Brain Informatics · August 2022

DOI: 10.1007/978-3-031-15037-1_11

Alessia Sarica
Universita' degli Studi "Magna Græcia" di Catanzaro
94 PUBLICATIONS **3,149** CITATIONS
SEE PROFILE

Andrea Quattrone
Universita' degli Studi "Magna Græcia" di Catanzaro
124 PUBLICATIONS **1,252** CITATIONS
SEE PROFILE

Aldo Quattrone
Catanzaro Neuroscience research center
814 PUBLICATIONS **28,943** CITATIONS
SEE PROFILE

Introducing the Rank-Biased Overlap as Similarity measure for Feature Importance in Explainable Machine Learning: a case study on Parkinson's disease

Alessia Sarica[1[0000-0003-1362-6718]], Andrea Quattrone[2[0000-0003-2071-2083]], Aldo Quattrone[1,3[0000-0003-2001-957X]]

[1] Neuroscience Research Center, Department of Medical and Surgical Sciences, Magna Graecia University, 88100 Catanzaro, Italy
[2] Institute of Neurology, Department of Medical and Surgical Sciences, Magna Graecia University, 88100 Catanzaro, Italy
[3] Neuroimaging Research Unit, Institute of Molecular Bioimaging and Physiology, National Research Council, 88100 Catanzaro, Italy
sarica@unicz.it

Abstract. Feature importance is one of the most common explanations provided by Machine Learning (ML). However, different classification algorithms or different training sets could produce different rankings of predictive features. Thus, the quantification of differences between feature importance is crucial for assessing model trustworthiness. Rank-biased Overlap (RBO) is a similarity measure between *incomplete*, *top-weighted* and *indefinite* rankings, which are all characteristics of feature importance. In RBO, tuning persistence *p* allows to truncate rankings at any arbitrary depth, so to evaluate their overlapping size at increasing number of features. Classification of Parkinson's disease (PD) with Explainable Boosting Machine (EBM) was chosen here as case study for introducing RBO in ML. An imbalanced dataset, 168 healthy controls (HC) and 396 PD patients, with 178 among clinical and imaging features was obtained from PPMI. Imbalanced, undersampled (K-Medoids) and oversampled (SMOTE) datasets were used for training EBMs, obtaining their respective feature importance. RBO score was calculated between ranking pairs incrementally increasing the depth by five features, from 1 to 178. All classifiers reached excellent AUC-ROC (~1) on test set, demonstrating the EBM prediction stability when trained on imbalanced datasets. RBO revealed that the maximum size of overlapping (80%) among rankings was obtained truncating at top 40 features, while their similarity decreased asymptotically to 50% when more than 45 features were considered. Thanks to RBO it was possible to demonstrate that, for the same accuracy, the more similar are the feature importance, the more stable is the model and the more reliable is the ML interpretability.

Keywords: explainable machine learning, feature importance, Parkinson's disease, Rank-Biased Overlap.

1 Introduction

Explainable Artificial Intelligence (XAI) and interpretable Machine Learning (ML) is a recently born field, which aim is to maximize the explainability and interpretability of ML findings [1]. One of the most common explanations provided by ML algorithms is the feature importance [2], that is the contribution of each feature in the classification. The ordered list of features by their individual contribution is a *top-weighted* ranking where the variables on the top are more predictive than the variables in the tail [2, 3]. In the medical and clinical field, the feature importance provides to the researcher an immediate overview of the biological measures involved in a specific disease [4].

The predictive contribution of each feature depends on the ML algorithm used for the classification. Indeed, different models produce different rankings of importance and one highly predictive feature in a classifier could be unimportant in another classifier [2]. Moreover, the same ML classifier could show different feature importance rankings when trained on different folds/subsets of the same dataset [1]. Another example is the prediction of a rare disease with an imbalanced dataset [5] and there is the need to balance the classes through undersampling or oversampling. The balance of classes could improve the ML performance but could also provide a different feature importance than the one obtained with an imbalanced training set, thus preventing an exhaustive interpretation of the findings. For these reasons, the comparison of feature importance rankings is fundamental for understanding how different ML approaches or different training sets influence the reliability and trustworthiness of the findings. In other words, the main questions are: how similar are the feature importance lists produced by different ML methods or by the same classifier trained on different datasets? What statistics, measure or metric should be used?

The quantification of the dissimilarity or similarity of two rankings is usually performed with correlation coefficients calculated with the Kendall's τ [6], Spearman's ρ [3] or their variants [7-9]. However, τ, ρ and their variants are unweighted measures and thus they are not able to emphasize the features on the top of the ranking [3]. Furthermore, these statistics are not applicable on indefinite and non-conjoint rankings, thus resulting not suitable for assessing the similarity of ML feature importance. On the contrary, the rank-biased overlap (RBO) [3] is a similarity measure that estimates the size of overlapping between indefinite ranked lists, representing a good candidate for comparing the classification feature importance. RBO score varies in a range from 0 to 1, where 1 indicates that the two rankings are identical, and zero indicates absence of similarity [3]. The weight given to the first d (depth) features in a ranking can be modified by tuning the persistence (p), a probability parameter in the range [0,1]. A lower value of p gives more importance to the top features, while a high value explores the ranking at a deeper depth [3].

The first aim of the present work is to introduce the RBO as a similarity measure for quantifying the differences between feature importance produced by explainable classification models. The Explainable Boosting Machine (EBM) [10] is a *glass*-box algorithm that showed high interpretability of ML findings, reaching excellent accuracies for example for the prediction of Alzheimer's disease [11] or for distinguishing between Parkinson's disease and SWEDD [12]. However, it has never been assessed whether and how EBM is able to deal with imbalanced datasets of neurodegenerative diseases. Thus, the second aim of the present study is to compare the performance of EBM models trained on imbalanced data and on balanced datasets obtained through undersampling with K-Medoids [13] and oversampling with Synthetic Minority Over-sampling Technique (SMOTE) [14]. The prediction of the Parkinson's disease (PD) was chosen here as case study, and for this purpose an imbalanced dataset with clinical and imaging features was obtained from the Parkinson's Progression Markers Initiative (PPMI). The third and last aim of this work is to use the RBO similarity measure for quantifying the differences among the three feature importance rankings produced by the EBM algorithm trained on the imbalanced, undersampled and oversampled dataset.

In summary, the three main contributions of the present study are: (i) introducing the RBO as measure for quantifying the similarity between feature importance rankings; (ii) building EBM classifiers on three different training sets - imbalanced, undersampled and oversampled datasets – and comparing their performance in predicting PD; (iii) assessing the similarity between feature importance rankings produced by the three EBM classifiers through the RBO score.

2 Materials and Methods

2.1 Participants

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. Table 1 reports the demographic, the clinical and imaging characteristics of the cohort, which consisted of 168 healthy controls (HC) and 396 PD. Only subjects without missing clinical and imaging features were considered and all data used for the analysis are acquired at the baseline visit.

2.2 Clinical and Imaging features

The number of items per clinical assessment and the total number of features (178) used for training the ML models are reported in Table 1, and consisted in: Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS) [15], part I, II and III, Montreal Cognitive Assessment (MoCA), State-Trait Anxiety Inventory (STAI), Geriatric Depression Scale (GDS), Scales for Outcomes in Parkinson's Disease - Autonomic Dysfunction (SCOPA-AUT), Judgment of Line Orientation (JLO), the University of Pennsylvania Smell Identification Test (UPSIT), Epworth Sleepiness Scale (ESS), Hoen and Yahr (H&Y) scale for assessing the stage of PD (not included in the training features since it is not for diagnosis). The neuroimaging technique commonly used for the diagnosis PD is the dopamine transporter single-photon emission computed tomography (DaT-SPECT) of the striatum, the region comprising caudate and putamen. The specific binding ratio (SBR) of these two regions of interest (ROI) is calculated for each hemisphere from the count densities and normalized by the occipital cortex uptake. More details of the imaging protocol can be found on www.ppmi-info.org.

Table 1. Demographic, clinical and imaging data of the imbalanced PPMI dataset.

	HC (168)	PD (396)	#[a]
Age	61.1±11.3	61.7±9.65	-
Gender (M/F)	109/59	260/136	-
H&Y	0.005±0.07	1.57±0.51	-
MDS-UPDRS-I	2.89±2.76	5.61±4.12	13
MDS-UPDRS-II	0.35±0.95	5.39±4.14	13
MDS-UPDRS-III	1.19±2.06	20.9±8.84	33
MoCA	28.1±1.09	26.9±2.38	26
STAI	47.7±4.97	47.3±5.32	40
GDS	5.17±1.39	5.26±1.45	15
SCOPA-AUT	5.11±3.38	8.58±6.51	21
JLO	13.1±1.95	12.8±2.1	1
UPSIT	34±4.75	22.3±8.34	4
ESS	5.66±3.38	5.81±3.42	8
Left Caudate SBR	3.0±0.63	1.99±0.59	1
Right Caudate SBR	2.9±0.61	1.98±0.59	1
Left Putamen SBR	2.14±0.56	0.812±0.35	1
Right Putamen SBR	2.16±0.58	0.843±0.36	1
		Tot:	178

[a] Number of items per test, i.e. number of features used for training EBM models. Age, gender, and H&Y not included in the features space.

2.3 Sampling of the dataset

The aim of sampling is to balance the dataset, thus, to obtain an equal sample size of the two classes. The original imbalanced dataset HC-PD$_{\text{imb}}$ (168-396) was randomly sampled by applying two different approaches, the first was an undersampling technique applied on the majority class (PD), the second one was an oversampling method applied on the minority class (HC), as described as follows.

Undersampling. The undersampling of the imbalanced dataset was done with the K-Medoids approach [13], which is an unsupervised method of clustering applied on the majority class (PD), where the number of clusters is equals the number of minority examples (HC=168). The final dataset HC-PD$_{\text{und}}$ (168-168) is a combination of all data from the minority set and the cluster centers from the majority set. The undersampling was conducted with the Python package *sklearn_extra.cluster.KMedoids* of scikit-learn (v. 0.23) (metric "euclidean" and method "pam").

Oversampling. SMOTE [14] was applied on the minority class (HC), for generating new synthetic data by randomly interpolating pairs of nearest neighbors. The final dataset HC-PD$_{\text{over}}$ (396-396) is a combination of all data from the majority (PD) and minority set (HC) and, additionally, the new synthetic minority data such that final dataset is balanced. The oversampling was conducted with the Python package *imblearn.over_sampling.SMOTE* (v. 0.9.0).

The original imbalanced dataset and the two sampled datasets – HC-PD$_{\mathrm{imb}}$, HC-PD$_{\mathrm{und}}$ and HC-PD$_{\mathrm{over}}$ – were then randomly split with a static seed into training and test sets with a percentage respectively of 80% and 20% by maintaining proportions between class distributions.

2.4 Machine Learning analysis

The EBM algorithm [10] is based on standard Generalized Additive Models (GAMs) [16], which accuracy is improved by adding pairwise interactions [17], taking the name of GA^{2}Ms with the form:

$$g(E[y]) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j), \qquad (1)$$

where E is the estimate of the additive model, $x_i = (x_{i1}, \dots, x_{ip})$ is the feature vector with p features, y_i the response, x_j denotes the jth variable in the feature space, g is the *link function* that adapts the GAMs to regression (g = identity) or classification (g = logistic), β_0 is the intercept that adjusts the prediction from the model, and f_j is the feature function, which could be plot for visualizing the contribution of each feature to the final prediction [17]. The feature importance is calculated after learning the best feature function f_j by training the model on one feature at a time, so to obtain its contribution to the prediction [17].

In this work, three EBM models were built on the three training sets – imbalanced, undersampled and oversampled - and the performance was evaluated on the test set with the Area under the Curve of the Receiver Operating Characteristic (AUC-ROC). Moreover, the AUC-ROC (mean±standard deviation) was calculated on the whole dataset with a 5-fold cross-validation (cv, *sklearn.model_selection.cross_val_score* of scikit-learn v. 0.23) for assessing overfitting. The pairwise interactions between features were not here considered to avoid complexity in the interpretation of the findings. The feature importance ranking of the three classifiers (FI[HC-PD$_{\mathrm{imb}}$], FI[HC-PD$_{\mathrm{und}}$], FI[HC-PD$_{\mathrm{over}}$]) was obtained by ordering the features by their mean absolute contribution in the prediction of the training data, calculated as logit of the probability (logarithm of the odds) from the logistic link function g (Eq. 1) [17]. Machine Learning analysis was conducted with the Python package *InterpretML* (v 0.2.7) [18] (implementation of EBM provided by Microsoft) on a MacOS 10.14.6 (2.9 GHz, 32GB of RAM). The Python package *seaborn* (v. 0.11.2) was used for plotting the feature importance rankings.

2.5 Rank-Biased Overlap (RBO)

The feature importance produced by explainable ML algorithms is a *top-weighted* ranking, that is an ordered list of items where the variables on the top are more important than the variables in the tail [2, 3]. Other two characteristics of a feature importance ranking is that it could be *incomplete*, that is it could not cover all variables in the domain, and it could be *indefinite*, since the user's decision to stop the ranking at a particular depth is arbitrary [3]. One of the most used measure of rank similarity is the correlation that quantifies the direction (positive or negative) and the magnitude of the association between a pair of lists. The Kendall's τ [6] and Spearman's ρ are two of the most widely used measures of correlation [3]. However, both τ and ρ require that the two rankings are conjoint and since they are unweighted measures, they are not able to place more emphasis on the items on the top of the rank [3]. Several variants of the correlation measures were proposed for considering the weight of items in a list and for comparing non-conjoint ranks, for example the top-weighted variant of the Kendall's τ, the τ_{AP} [7], the adaptations of Spearman's ρ [9] and Spearman's footrules [8], or the Kolmogorov-Smirnov's *D* [19]. However, all these variants do not fully satisfy the need to compare *indefinite* rankings, that is the need to truncate the feature importance at any particular and arbitrary depth [3]. To overcome this issue, a similarity measure was introduced, the rank-biased overlap (RBO), which is calculated as the expected average overlap between two indefinite rankings at incrementally increasing depths [3]. The depth of interest could be varied by tuning an input parameter of the RBO, called user's *persistence* (*p*). The persistence *p* is a probability (in the range [0,1]) of continuing to the next rank in the list, while on the contrary, 1-*p* is the probability that the user stops at a given depth *d* of the ranking [3]. A lower value of *p* gives more importance to top results, and when *p* = 0 only the first feature in the ranking is considered. Given two infinite rankings *S* and *T* to depth *d* and the persistence *p*, the RBO is calculated as follows [3]:

$$RBO(S,T,p) = (1-p)\sum p^{d-1} \cdot A_d, \qquad (2)$$

where, $d = 1\ to\ \infty$ is the depth of the ranking to be examined, $A_d = X_d/d$ is the agreement between *S* and *T*, i.e. the proportion of *S* and *T* that is overlapped, and $X_d = |S_{:d} \cap T_{:d}|$ is the size of overlap (intersection) between *S* and *T*. The RBO varies in the range [0,1], where 0 means disjoint rankings and 1 means identical rankings [3].

In this work, the RBO was used for assessing the similarity between pairs of feature importance rankings ($\mathrm{RBO_{imb\_und}}$, $\mathrm{RBO_{imb\_over}}$, $\mathrm{RBO_{und\_over}}$) that were obtained by training the EBM algorithm on the different datasets: imbalanced, undersampled and oversampled. Here, to investigate the similarity between feature importance rankings at different depths, the values of stopping depth *d*, i.e. the number of the top features in the ranking, were increased with a fixed step of 5 features in the range [1, 178]. Consequently, the value of persistence *p* was automatically increased and calculated as $p = \frac{d-1}{d}$, assuming the values in the range [0, 0.9944]. The Python package *rbo* (v.0.1.2) was used as implementation of the RBO by Webber et al. [3].

3 Results

3.1 Machine Learning analysis

The EBM models $\mathrm{HC\text{-}PD_{imb}}$ and $\mathrm{HC\text{-}PD_{over}}$ reached both an AUC-ROC of 1 (1±0 with 5-fold cv), while the classifier $\mathrm{HC\text{-}PD_{und}}$ had an AUC-ROC of 0.99 (0.998±0.004

with 5-fold cv). The rankings of the first twenty most important features in the models HC-PD$_{imb}$, HC-PD$_{und}$ and HC-PD$_{over}$ are reported in Figure 1A, B and C. Figure 1D reports the first fifty most important features ordered by their average importance across the three EBM models, where the first ten important features were NP2TRMR (MDS-UPRDS II item 2.10 Tremor), PUTAMEN_L (SBR of the left putamen), NP3FACXP (MDS-UPRDS III item 3.2 Facial expression), NP3BRADY (MDS-UPRDS III item 3.14 Global Spontaneity of movement) and NP3RTCON (MDS-UPRDS III item 3.18 Constancy of rest), PUTAMEN_R (SBR of the right putamen), NP2HWRT (MDS-UPRDS II item 2.7 Handwriting), NP3PRSPR (MDS-UPRDS III item 3.6a Pronation-Supination - Right Hand), NP3HMOVL (MDS-UPRDS III item 3.5b Hand movements - Left Hand) and NP3RIGRU (MDS-UPRDS III item 3.3b Rigidity - RUE).

3.2 RBO scores

The RBO scores calculated by tuning the value of depth *d* and consequently the persistence *p* in each pair of comparisons (RBO$_{imb\_und}$, RBO$_{imb\_over}$, RBO$_{und\_over}$) are reported in Table 2. The maximum similarity (~1) was obtained when only the first item (NP2TRMR, MDS-UPRDS II item 2.10 Tremor) in the ranking was compared between FI[HC-PD$_{imb}$] and FI[HC-PD$_{und}$]. The maximum values RBO$_{imb\_over}$ = 0.802 and RBO$_{und\_over}$ = 0.74 were reached both when the first 40 features in the rankings were compared (*p* = 0.975, Table 2).

Table 2. RBO of the pairwise comparisons of the feature importance rankings obtained by training the EBM models on the imbalanced, undersampled and oversampled datasets. Raising *p* increases the depth *d* of comparisons (number of features considered). In bold the maximum value.

p	*d*	**RBO$_{imb\_und}$**	**RBO$_{imb\_over}$**	**RBO$_{und\_over}$**
0	*1*	**~1**	~0	~0
0.800	*5*	0.755	0.548	0.423
0.900	*10*	0.780	0.695	0.603
0.950	*20*	0.803	0.778	0.705
0.967	*30*	0.806	0.799	0.733
0.975	*40*	0.800	**0.802**	**0.740**
0.980	*50*	0.788	0.795	0.735
0.990	*100*	0.679	0.692	0.646
0.993	*150*	0.571	0.583	0.546
0.994	*178*	0.520	0.531	0.499

Abbreviations: *d* = depth; *p* = persistence; imb = imbalanced dataset (HC-PD$_{imb}$); und = undersampled dataset (HC-PD$_{und}$); over = oversampled dataset (HC-PD$_{over}$).

Figure 2 depicts the RBO curves of the three ranking comparisons by raising the depth *d*, that is by considering a higher number of features as important. The RBO$_{imb\_over}$ and RBO$_{und\_over}$ curves show a similar increasing trend, moreover the three curves reach a plateau between *d* = 20 and *d* = 40, revealing that the maximum similarity among the three RBOs is obtained when the first 40 features are considered. For values *d* > 45 there is a decrease in the similarity among the three feature importance until the RBO curves asymptote to the final value of ~0.5.

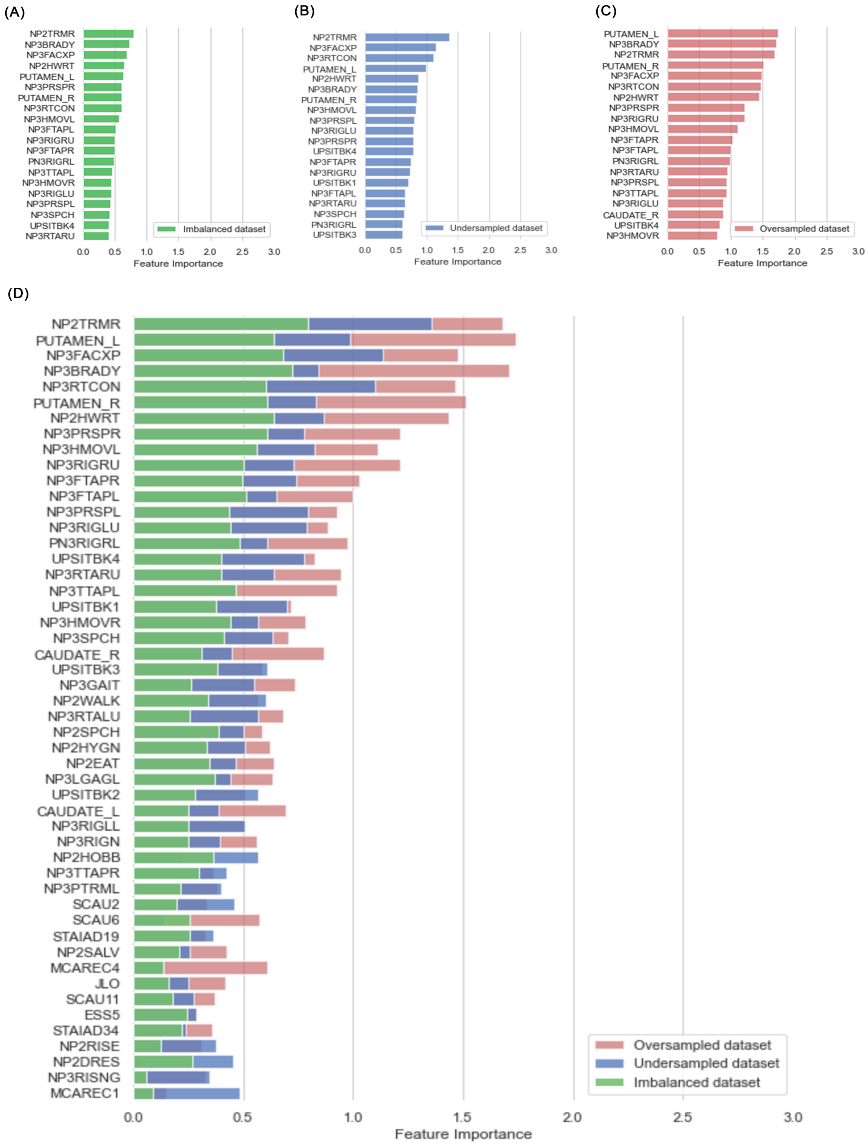

Fig. 1. Ranking of the first twenty most important features obtained by the EBM model trained (A) on the imbalanced dataset; (B) on the undersampled dataset; (C) on the oversampled dataset. (D) Feature importance (first fifty features) ordered by their average importance across the three classifiers trained on the imbalanced dataset (in green), on the undersampled dataset (in blue) and on the oversampled dataset (in red).

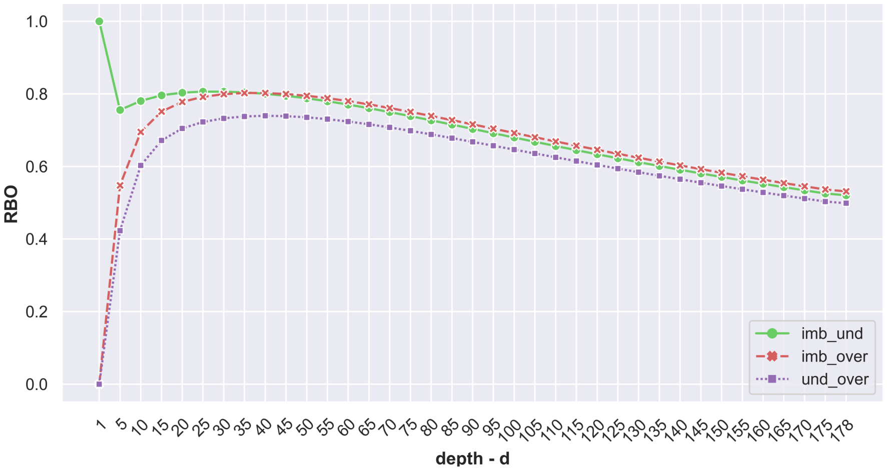

Fig. 2. RBO curves of the pairwise comparisons of feature importance rankings obtained by the EBM models HC-PD$_{\mathrm{imb}}$, HC-PD$_{\mathrm{und}}$ and HC-PD$_{\mathrm{over}}$ for increasing values of depth *d*, that is for increasing number of important features considered.

4 Discussion and Conclusions

The purpose of this work was to introduce the RBO [3] score as similarity measure for comparing feature importance rankings produced by explainable ML. The classification of Parkinson's disease from clinical and imaging features was chosen as case study and conducted with the Explainable Boosting Machine [10, 17, 20] algorithm on three datasets, imbalanced, undersampled and oversampled. EBM models reached excellent accuracies (~1), thus demonstrating the robustness of EBM in dealing with imbalanced datasets. Interestingly, RBO allowed to reveal that the three feature importance rankings had the highest size of overlapping (~80%) when the depth was truncated at 40 features.

The classification task has two main goals: to obtain good accuracy in distinguishing classes and to provide the feature contribution in the prediction [1, 2, 21]. The classifier performance could be evaluated through several metrics (e.g. accuracy, precision recall) that are easy to compare both quantitatively and statistically (e.g. McNemar's test) [22]. However, when a multiplicity of models reach excellent accuracies, it is difficult to decide which one is better and what Breiman calls the *Rashomon Effect* takes place [21]. Indeed, for the same performance, a classifier can consider a feature more or less important than the importance given by another classifier. For this reason, it is crucial to quantify the differences between ML rankings, because if different models produce similar feature importance, "*it is more likely that these reflect genuine aspect of the data*" as stated by Saarela and Jauhiainen [2]. The present study faces the Rashomon Effect [21], given that all the three EBM models trained on the imbalanced, undersampled and oversampled datasets reach the highest accuracy (AUC-ROC ~1). The stability of the EBM performance in presence of imbalanced data is an important finding for the automatic prediction of neurodegenerative diseases from clinical and imaging features. The rarity of some pathologies prevents having large enough samples as well as balance between classes [5], thus the ML struggles to provide reliable findings. On the contrary, EBM seems to be unaffected by the perturbations due to the imbalance between classes, probably thanks to the use of bagging, gradient boosting and additive modularity [10, 17, 20], which are all methods strongly suggested by the previous literature [21]. As further evidence of the stability of EBM algorithm, the RBO score found high similarity (80%) among the three feature importance at a depth of 40 features.

Another interesting finding is that the feature importance obtained with the oversampled dataset was slightly less similar than the other two rankings produced by the imbalanced and undersampled datasets. This is probably due to the nature of the SMOTE algorithm itself that could have altered the feature correlation of the original dataset by generating new synthetic minority data [23]. Indeed, it should be reported as limitation that EBM algorithm may consider important features that are on the contrary not predictive when correlation among features, heavy multicollinearity and/or nonlinearity around a prediction exist [10]. Another limitation of the present study is related to the percent split of training and test sets (80-20); future works might assess whether the use of different proportion could produce different accuracies and feature importance. Further research might explore the application of RBO to compare the explanations produced by different ML algorithms, such as Random Forest [24]. It would be also interesting to investigate how the tuning of EBM hyperparameters, such as the outer bags or the learning rate, could affect the feature importance and the accuracy in this specific case study.

In conclusion, the present work demonstrated that RBO is a suitable similarity measure, allowing to state that, for the same classification accuracy, the more similar are the feature importance produced with different training sets, the more stable is the model and the more reliable is the interpretability and explainability of the ML findings.

References

1 Molnar C: Interpretable machine learning. Lulu. com, 2020.

2 Saarela M, Jauhiainen S: Comparison of feature importance measures as explanations for classification models. SN Applied Sciences 2021;3:1-12.

3 Webber W, Moffat A, Zobel J: A similarity measure for indefinite rankings. ACM Transactions on Information Systems (TOIS) 2010;28:1-38.

4 Sarica A: Editorial for the Special Issue on “Machine Learning in Healthcare and Biomedical Application”, MDPI, 2022, 15, pp 97.

5 Dubey R, Zhou J, Wang Y, Thompson PM, Ye J, Initiative AsDN: Analysis of sampling techniques for imbalanced data: An n= 648 ADNI study. NeuroImage 2014;87:220-241.

6 Kendall MG: Rank correlation methods. 1948

7 Yilmaz E, Aslam JA, Robertson S: A new rank correlation coefficient for information retrieval: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008, pp 587-594.

8 Bar-Ilan J, Mat-Hassan M, Levene M: Methods for comparing rankings of search engine results. Computer networks 2006;50:1448-1463.

9 Bar-Ilan J: Comparing rankings of search results on the web. Information processing & management 2005;41:1511-1519.

10 Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp 1721-1730.

11 Sarica A, Quattrone A, Quattrone A: Explainable Boosting Machine for Predicting Alzheimer's Disease from MRI Hippocampal Subfields: International Conference on Brain Informatics, Springer, 2021, pp 341-350.

12 Sarica A, Quattrone A, Quattrone A: Explainable machine learning with pairwise interactions for the classification of Parkinson's disease and SWEDD from clinical and imaging features. Brain Imaging and Behavior 2022:1-11.

13 Park H-S, Jun C-H: A simple and fast algorithm for K-medoids clustering. Expert systems with applications 2009;36:3336-3341.

14 Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 2002;16:321-357.

15 Goetz CG, Tilley BC, Shaftman SR, Stebbins GT, Fahn S, Martinez-Martin P, Poewe W, Sampaio C, Stern MB, Dodel R, Dubois B, Holloway R, Jankovic J, Kulisevsky J, Lang AE, Lees A, Leurgans S, LeWitt PA, Nyenhuis D, Olanow CW, Rascol O, Schrag A, Teresi JA, van Hilten JJ, LaPelle N, Movement Disorder Society URTF: Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord 2008;23:2129-2170.

16 Hastie TJ, Tibshirani RJ: Generalized additive models. CRC press, 1990.

17 Lou Y, Caruana R, Gehrke J, Hooker G: Accurate intelligible models with pairwise interactions: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, 2013, pp 623-631.

18 Nori H, Jenkins S, Koch P, Caruana R: Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:190909223 2019

19 Melucci M: Weighted rank correlation in information retrieval evaluation: Asia Information Retrieval Symposium, Springer, 2009, pp 75-86.

20 Lou Y, Caruana R, Gehrke J: Intelligible models for classification and regression: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, pp 150-158.

21 Breiman L: Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science 2001;16:199-231.

22 Jollans L, Boyle R, Artiges E, Banaschewski T, Desrivieres S, Grigis A, Martinot JL, Paus T, Smolka MN, Walter H, Schumann G, Garavan H, Whelan R:

Quantifying performance of machine learning methods for neuroimaging data. Neuroimage 2019;199:351-365.

23 Patil A, Framewala A, Kazi F: Explainability of smote based oversampling for imbalanced dataset problems: 2020 3rd International Conference on Information and Computer Technologies (ICICT), IEEE, 2020, pp 41-45.

24 Sarica A, Cerasa A, Quattrone A: Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. Frontiers in aging neuroscience 2017;9:329.