Explanations as a New Metric for Feature Selection: A Systematic Approach

Haomiao Wang, Emmanuel Doumard, Chantal Soulé-Dupuy, Philippe Kemoun, Julien Aligon, Paul Monsarrat

Explanations as a new metric for feature selection: a systematic approach

Haomiao Wang, Emmanuel Doumard, Chantal Soulé-Dupuy, Philippe Kémoun, Julien Aligon †,*, Paul Monsarrat †

***Abstract*— With the extensive use of Machine Learning (ML) in the biomedical field, there was an increasing need for Explainable Artificial Intelligence (XAI) to improve transparency and reveal complex hidden relationships between variables for medical practitioners, while meeting regulatory requirements. Feature Selection (FS) is widely used as a part of a biomedical ML pipeline to significantly reduce the number of variables while preserving as much information as possible. However, the choice of FS methods affects the entire pipeline including the final prediction explanations, whereas very few works investigate the relationship between FS and model explanations. Through a systematic workflow performed on 145 datasets and an illustration on medical data, the present work demonstrated the promising complementarity of two metrics based on explanations (using ranking and influence changes) in addition to accuracy and retention rate to select the most appropriate FS/ML models. Measuring how much explanations differ with/without FS are particularly promising for FS methods recommendation. While *reliefF* generally performs the best on average, the optimal choice may vary for each dataset. Positioning FS methods in a tridimensional space, integrating explanations-based metrics, accuracy and retention rate, would allow the user to choose the priorities to be given on each of the dimensions. In biomedical applications, where each medical condition may have its own preferences, this framework will make it possible to offer the healthcare professional the appropriate FS technique, to select the variables that have an important explainable impact, even if this comes at the expense of a limited drop of accuracy.**

***Index Terms*— Biomedical, Explainability, Feature selection, Machine learning, Metrics**

I. Introduction

Although it is undeniable that artificial intelligence (AI) can be beneficial to the biomedical field, the major challenge lies in combining and hybridizing human intelligence with AI [1]. Except for a few intrinsically explainable models (*i.e.* glass-boxes), one of the barriers to this hybrid alliance is the "black-box" effect of most AI/machine learning (ML) systems with a lack of explainability of predictions [2].

In the biomedical field, explainability plays an important part of strengthening the relationship between patients and medical practitioners [3], [4]. Physicians should reasonably explain and decipher the decision-making process of AI and consequently be able to communicate with the patient appropriately. This is the basis for respecting patient autonomy and obtaining informed consent [5]. Despite the high performance of these "black-box" models, empower the medical practitioner to assess the quality of model inputs, to verify the absence of bias or discrimination and to understand the influence of the different variables in prediction is a matter of medical ethics [5]. Consequently, in recent years, various regulations and guidance [6], [7] have included explainability as an essential principle of AI in biomedical and Explainable Artificial Intelligence (XAI) has become an emerging topic in biomedical AI [8].

Understanding pathophysiology often requires a holistic approach, biomedical data are often high-dimensional, but contain extensive irrelevant and/or redundant information [9]. There is a real risk that useless information interferes with ML models, resulting in the sparsification of data and a series of problems (known as the *curse of dimensionality* [10]), and the time required to explain a model increases with the number of features [9]. A frequent solution to address the *curse of dimensionality* is the Feature selection (FS), which refers to the process of selecting relevant features (*i.e.*, variables) from the original features [11]. As a dimensionality reduction technique, FS is considered to be a more desirable process

This work was supported by the Occitanie Region, the Federal University of Toulouse Midi-Pyrénées (grant ADI 2021, N°ALDOCT89533), the Programme d'Investissements d'Avenir and the Agence Nationale pour la Recherche (grant EUR CARe N°ANR-18-EURE-0003) and the national infrastructure "ECELLFrance: Development of mesenchymal stem cell based therapies" (PIA-ANR-11-INBS-005).

The symbol * indicates corresponding author, and † authors that contributed equally to this work.

Haomiao Wang is with the RESTORE Research Center, Université de Toulouse, INSERM 1301, CNRS 5070, EFS, ENVT, Toulouse, France (e-mail: haomiao.wang@inserm.fr)

Emmanuel Doumard is with the Institut de Recherche en Informatique de Toulouse (IRIT), Université de Toulouse, CNRS/UMR 5505, Toulouse, France, and also with the RESTORE Research Center, Université de Toulouse, INSERM 1301, CNRS 5070, EFS, ENVT, Toulouse, France (e-mail: emmanuel.doumard@irit.fr)

Chantal Soulé-Dupuy and Julien Aligon are with the Institut de Recherche en Informatique de Toulouse (IRIT), Université de Toulouse, CNRS/UMR 5505, Toulouse, France (e-mail: chantal.soule-dupuy@irit.fr, julien.aligon@irit.fr)

Philippe Kémoun is with the RESTORE Research Center, Université de Toulouse, INSERM 1301, CNRS 5070, EFS, ENVT, Toulouse, France, and also with the Oral Medicine Department and CHU de Toulouse-Toulouse Institute of Oral Medicine and Science (e-mail: philippe.kemoun@univ-tlse3.fr)

Paul Monsarrat is with the RESTORE Research Center, Université de Toulouse, INSERM 1301, CNRS 5070, EFS, ENVT, Toulouse, France, the Oral Medicine Department and CHU de Toulouse-Toulouse Institute of Oral Medicine and Science, as well as the Toulouse Artificial and Natural Intelligence Institute (ANITI), (e-mail: paul.monsarrat@univ-tlse3.fr)

(compared to feature extraction which creates new features) in the biomedical context, since the transformation undermines the understanding of original features [12].

Nevertheless, little is known about the consequences of adding FS to an ML pipeline on the final explanations. In this work, a systematic workflow performed on 145 datasets investigated the impact of the combination of several FS/ML models on the final prediction explanations obtained. Results showed that incorporating explanation-based metrics to perform FS would provide better ML explanations for users, suggesting their interest to improve ML-based decision support in the biomedical field.

A. Feature selection

According to the "no free lunch" (NFL) theorem [13], there is no omnipotent feature selection method. Many methods have been developed over the past decades for various purposes. FS methods can be categorized into three main types according to their dependency on an ML model, *i.e.*, filter, wrapper and embedded. Filter methods are independent of the learning model and focus on the evaluation metric computed solely from the data (features and target). Filter methods can be further subdivided into similarity-based, information theory-based, sparse-learning-based, and statistic-based methods [14]. The wrapper methods are "wrapped" into the learning model; they directly use the final model in the evaluation of the subset. The wrapper method must then be considered as an optimization problem. Current research emphasizes the search strategy, using either Genetic Algorithm (GA) [15]–[18], Swarm intelligence [19]–[21], Simulated Annealing [22]–[24] or other meta-heuristic algorithms [25]–[27]. Embedded methods mean that the ML algorithms integrate the FS process intrinsically. For example, since the tree and rule-based models incorporate splitting steps during training, part of the features can consequently be discarded. Another example is the regularization-based model, especially the *Lasso* [28] which penalizes small values and makes regression coefficients sparse, hence playing the role of FS.

A critical issue in comparing FS methods is defining the quality. Using synthetically generated data allows a calibrated field of experimentation with clear ground truth, which makes it easier to evaluate FS methods by creating an accurate indicator of success rate [29], [30]. However using real-world data is more complex since defining a clear ground truth is seldom possible, and the most common practice has been to use an indirect indicator [11], typically a performance metric (*e.g.*, accuracy [30]–[35], balanced accuracy [31], [36], [37], classification error [33], [38], AUC [35], [37], [39]). Nevertheless, researchers have demonstrated that accuracy is not sufficient to identify the fitness of a ML model [40], [41]. As a result, a substitution of accuracy must be found for the evaluation of FS methods.

B. Additive methods in XAI

To address the lack of transparency of traditional AI in the biomedical field, the usage of XAI has been expanding in recent years [42]–[45]. XAI methods can be categorized as intrinsically explainable models and *post-hoc* methods [46]. The former refers to the ML model itself, *i.e.* glass-boxes, self-explainable given its structural simplicity (*e.g.* decision trees, rule-based models [47], fuzzy systems [48], [49]). *Post-hoc* methods need additional calculation after model training, since they are in fact model-agnostic, and therefore more broadly applicable. An alternative taxonomy classifies model explanations into *local* and *global*. Local explanation refers to the individual explanation of each instance, while global explanation refers to the explanation of the entire model behavior. The local explanations may be converted into a global explanation, through statistical concepts [49] or additive methods.

Additive methods, popular for ML in the biomedical field [42], [43], explain a model by assessing the contribution of each feature for each instance, and the global explanation is derived by averaging all the local explanations of the instances. *LIME (Local Interpretable Model-agnostic Explanations)* [50] generates neighborhoods for each instance to be explained and trains an interpretable linear model with the neighboring data. The prediction of the instance is explained by this linear model as a vector of feature weight. Another well-known concept was inspired by the game theory, *i.e.*, Shapley values [51]. The influence on the prediction of a feature is computed based on the margin contribution to all possible coalitions of this feature. To avoid the exponential time complexity, the Shapley value was approximated by Monte Carlo Sampling [52]. The *Coalitional-based method* [53] proposed to create groups of features in advance, approximating the influence of features based only on the pre-computed groups, rather than all coalitions. Another solution, *SHAP (SHapley Additive exPlanations)* [54] adopts the same concept of perturbation as *LIME* to estimate the feature importance. Some model-specific explainers, such as *TreeSHAP*, *LinearSHAP* and *DeepSHAP*, were proposed to overcome the computational burden of *KernelSHAP*, which is model-agnostic.

C. Feature Selection and XAI

Before its use in XAI, the game theory and Shapley values were used as feature selection evaluation metrics [55]–[57]. Shapley values have also been integrated into other FS techniques such as Borutashap [58], [59]. Indeed, since the additive methods assign the contribution values to each feature, some studies use direct feature importance of the global explanation as an FS method [60], [61]. In an alternative approach, *SCI-XAI* [62] integrated FS and XAI concepts into a single pipeline, but this work only took Tree-based Ensemble models into account, and used a fully interpretable model to quantify the impact of FS methods. Therefore, it is difficult to generalize to all real-world cases due to model and explainer limitation.

II. Methodology

A. Experimental workflow

The following steps were implemented to investigate the impact of the feature selection methods on the explanations obtained (Figure 1): dataset selection, feature selection, model

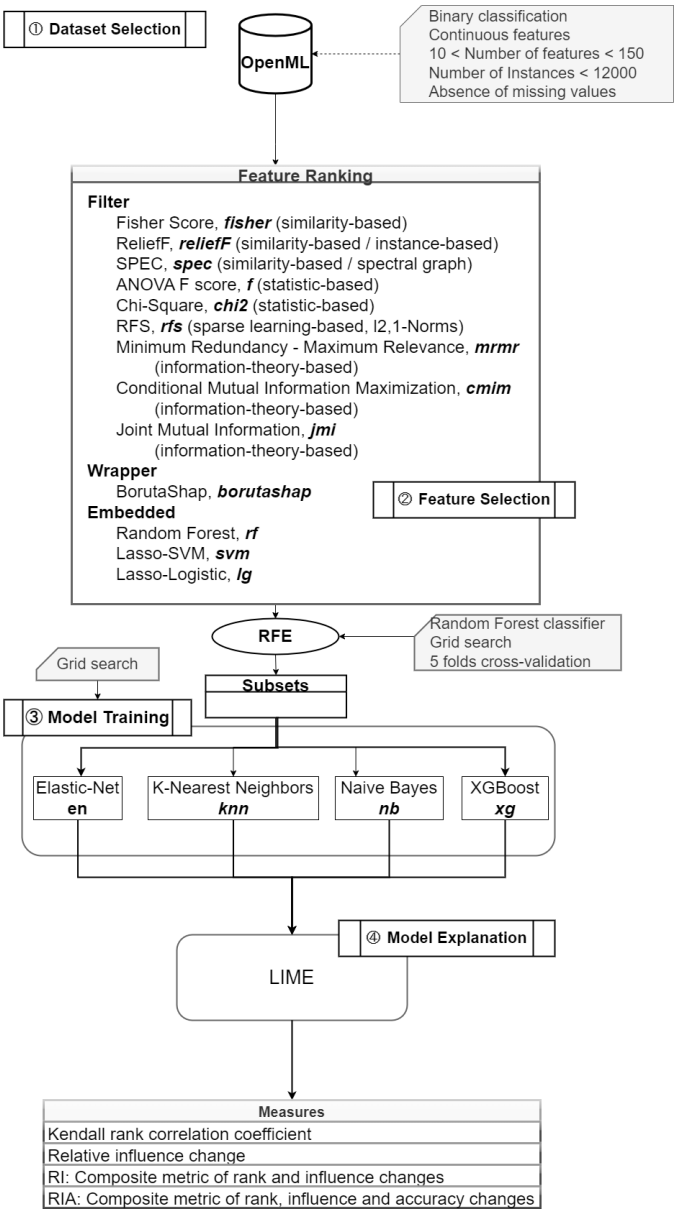

Fig. 1. **Experimental workflow.**

training, and model explanation with the subsequent calculation of suitable metrics. The results of the experiments together with supplementary materials are accessible on Github[1].

1) Dataset selection from the OpenML repository: OpenML (Open Machine Learning) [63] is an open science data repository for ML, including thousands of datasets available from a variety of application fields. This work was restricted to datasets with 1) binary classification tasks, 2) continuous explanatory features only, 3) no missing data, 4) 10 to 150 features, 5) less than 12,000 instances.

229 datasets met the previous conditions. A manual sorting step allowed the exclusion of twenty-four duplicate datasets. Sixty-four time-series datasets were detected. Since they could not be handled by traditional classifiers [64], and because they were all historical exchange rate datasets, only 3 were randomly chosen, removing the timestamp feature. Finally, 144 datasets were taken into consideration for the experiment.

2) Feature Selection: A wide range of FS methods was implemented to cover both the different families of FS (filter, wrapper and embedded) and different computational approaches within the same family. Among filter-based family, similarity-based methods (*fisher*, *reliefF* [65] and *spec* [66]), statistic-based methods (*f* and *chi2*), sparse-learning-based methods (*rfs* [67]) and information-theory-based methods (*mrmr* [68], *cmim* [69] and *jmi* [70]) were undertaken. For embedded methods, feature importance (Random Forest, *rf*) or coefficient (Linear Support Vector - *svm*, Logistic Regression - *lg*) were used as feature ranking. For the wrapper method, *borutaShap*, feature ranking was obtained by combining the Boruta technique with SHAP values [58]. The workflow included a recursive feature elimination step with 5-fold cross-validation (RFECV) [71], to determine a suitable threshold in a reproducible way. A RandomForest with hyperparameters tuned (*max_depth, min_sample*) beforehand through grid search was used in the RFECV.

3) Model Training: Four classification algorithms were chosen to reflect a certain diversity of algorithmic strategies in ML: Elastic-Net (*i.e.*, *en*, penalized linear model) [72], K-Nearest Neighbors (*i.e.*, *knn*, distance-based model), Naive Bayes (*i.e.*, *nb*, probabilistic model), XGBoost (*i.e.*, *xg*, tree-based ensemble model) [73]. The hyperparameters of *en* (*l1_ratio*), *knn* (*n_neighbors*) and *xg* (*max_depth, min_child_weight, gamma, eta*) were tuned by grid search with 5-fold cross-validation. Each classifier was trained separately with each feature subset generated by each FS method (as previously described). All instances were used for training to avoid any bias caused by sampling. An accuracy score was also computed as the performance metric.

4) Model Explanation: As a model-agnostic explanation method, LIME can be applied to any model, regardless of the ML model employed. As an additive method, it assigns an influence value to each feature of each instance, which represents its contribution to the prediction. Moreover, an advantage of LIME over other additive methods such as KernelSHAP and coalitional-based methods is less computational complexity when the number of features increases [74], which is critical for the feasibility of this study.

B. Used Metrics for explanations comparison

The goal of the experiments was to identify the FS method that achieves the highest level of accuracy with the fewest number of features while making the smallest alterations to the original explanation. The first two can be easily measured by the accuracy value and retention rate. For the third criterion, four metrics were computed to compare the explanations obtained after FS for each ML model, with the complete model (without FS) used as a reference.

1) Kendall rank correlation coefficient: As a local method, LIME computes a global explanation with feature importance ranking using Equation 1, where $M_{i,j}$ denotes the j^{th} feature explanation of a given instance i in a dataset with n instances and f features, Function $argsort(v_f)$ returns the descending

[1]https://github.com/haomiaow/XAI_feature_selection

order of an f-element vector (*i.e.*, feature importance vector) :

$$ranking = argsort(\sum_{i=1}^{n}|M_{i,j}|) \tag{1}$$

The Kendall rank correlation coefficient [75], *i.e.*, Kendall's τ, is a non-parametric statistic that measures the similarity between two rankings. This method compares the position of each element pair in the two rankings to determine whether the pair is concordant or discordant. Equation 2 defines the τ using the number of concordant and discordant pairs:

$$\tau = \frac{\#\ of\ concordant\ pairs - \#\ of\ discordant\ pairs}{\frac{1}{2} \times n \times (n-1)} \tag{2}$$

where n is the number of elements in a ranking.

A τ of 1, 0 or -1 means the two rankings are identical, independent or inverse, respectively. Since the same features are required for comparison, only the intersection of two subsets was considered.

2) Relative influence change: This metric is complementary to τ since it is possible to observe changes in feature ranking with small changes in influence. A feature's contribution eliminated by feature selection was considered as zero. In order to make the influence metric independent of the number of features, normalization was performed relative to the total influence.

$$inf'_f = \frac{inf_f}{\sum_{j=1}^{f}|inf_j|} \tag{3}$$

In Equation 3, for a given instance described by n features, inf'_f denotes the normalized value of the f^{th} feature's influence value inf_i.

$$diff = \sum_{j=1}^{f} \overline{|M_{i,j}^{FS\prime} - M_{i,j}^{O\prime}|}\Big|_{i=1}^{n} \tag{4}$$

Equation 4 shows the *Relative influence change* between two explanation value matrices of a dataset with n features and m instances, $M^{O\prime}$ denotes the normalized original explanation matrix, $M^{FS\prime}$ represents the normalized explanation obtained after applying a feature selection method.

3) Composite metric of rank and influence changes: the RI metric: The intuition behind this metric is to limit the penalty of rank change between different features if their influences are close, and conversely. In fact, the metric combines the relative influence change and the ranking change. C denotes the normalized explanation value matrices M' reordered by feature importance in descending order; l stands for the size of the selected subset of features; function $PR(M, f)$ returns the percentile rank of a feature f in the ranking calculated from the given explanation matrix M using Formula 1. The root of four serves to adjust the two penalties to the same scale.

$$RI = \overline{(|PR(M^{FS}, j) - PR(M^{O}, j)| + \epsilon)} \times \overline{((\overline{|C_{i,j}^{FS} - C_{i,j}^{O}|}\Big|_{i=1}^{n})^{\frac{1}{4}} + \epsilon) - \epsilon^2}\Big|_{j=1}^{l} \tag{5}$$

4) Composite metric of rank, influence and accuracy changes: the RIA metric: Based on the previous metric, the RIA penalizes a model with highly degraded accuracy compared to the original model. Function $Acc(M)$ returns the model accuracy associated with a given explanation M.

$$RIA = \overline{(|PR(M^{FS}, j) - PR(M^{O}, j)| + \epsilon)} \times \overline{((\overline{|C_{i,j}^{FS} - C_{i,j}^{O}|}\Big|_{i=1}^{n})^{\frac{1}{4}} + \epsilon)} \times \overline{(Acc(M^{O}) - Acc(M^{FS})) + \epsilon) - \epsilon^3}\Big|_{j=1}^{l} \tag{6}$$

An analysis of variance (ANOVA) has been performed to test for a putative significant difference between FS methods for a given metric and a given ML model, adjusted on the accuracy of the model, the dataset, the number of features and instances and the retention rate. The level of significance was considered at p-value $\leq .05$ after considering multiple comparisons using Tukey's HSD test.

C. Running environment

The experiments were performed on a workstation with an Intel® Xeon® Gold 6230 processor and 125GB of RAM under Python 3.9.7. *fisher*, *reliefF*, *spec*, *f*, *chi2*, *mrmr*, *cmim*, *jmi* and *rfs* were found in the Python library scikit-feature version 1.0.2. *borutashap* implementation was found in the library BorutaShap version 1.0.16. *rf*, *svm* and *lg*, as well as the ML models (*en*, *knn* and *nb*) were used from scikit-learn [76] version 1.0.2. *xg* classifier was found in the library XGBoost version 1.5.0, and the LIME library version 0.2.0.1 was used as the explanation method.

III. Result

A. Statistical analysis

Figure 2-A illustrates *Kendall's* τ of the feature contribution rankings generated for each FS method, using the explanation without FS (*all*) as a reference. Similar results were observed for a given FS technique according to the different ML models, with a tendency toward better τ for the *xg* ML model. The highest τ was achieved with the *xg* ML model with the *borutashap* FS method (.58), but in other models, the highest τ were found with *reliefF*. The lowest τ was observed in the *knn* ML model (*spec* FS model, .21). The *spec* FS model exhibited the lowest coefficients regardless of the ML model, with a τ between .21 and .29.

Figure 2-B describes the *Relative influence change* between the explanations generated by each FS method, and the original explanations (*all*). The most significant changes were achieved for the *rfs* and *rf* FS technique while the least significant

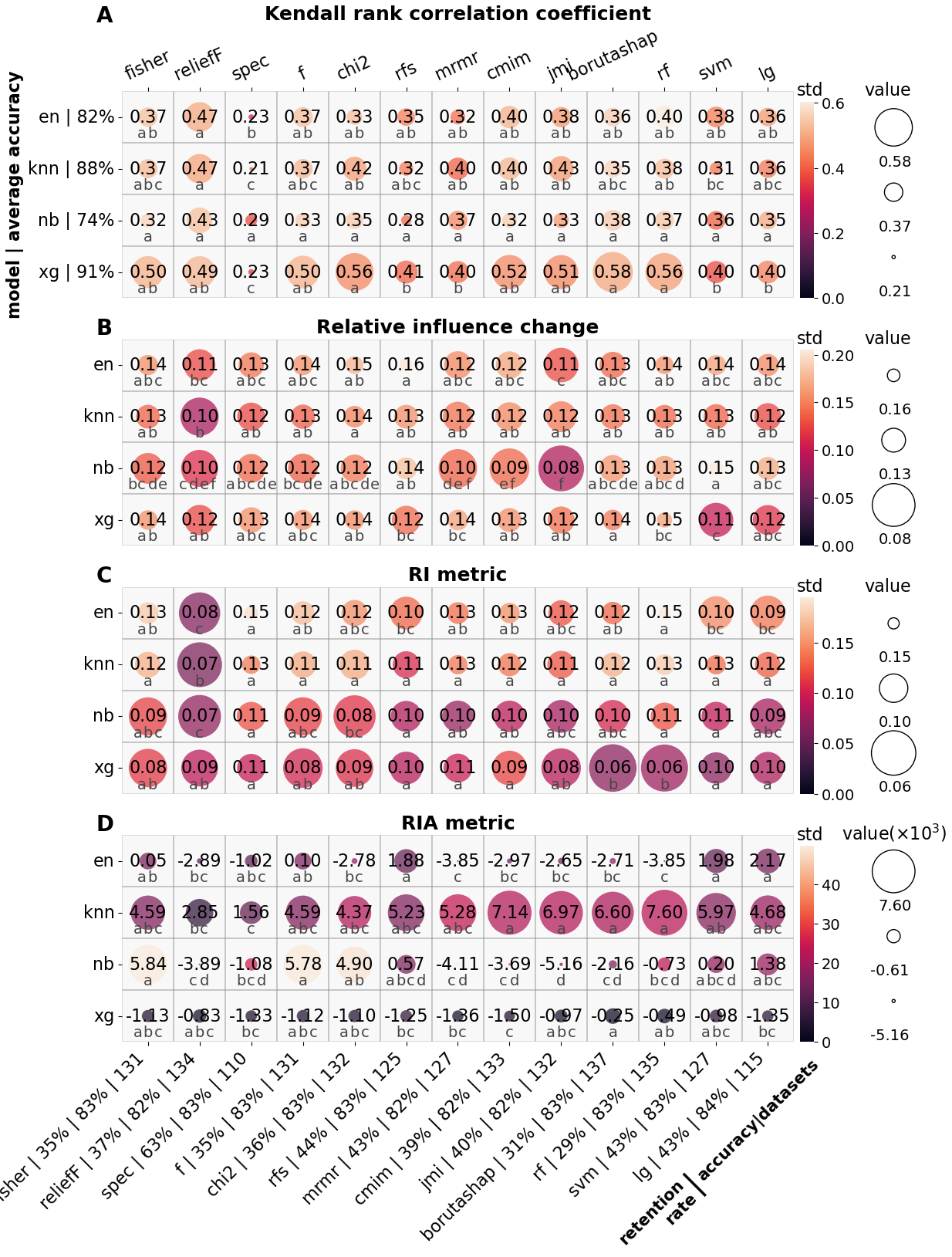

Fig. 2. **Kendall rank correlation coefficient (A), Relative influence change (B), RI (C) and RIA (D) metrics.** The circle size and value represent the differences between explanations generated by the feature selection method on the x-axis and the original feature set, for the classifier on the y-axis. The circle color indicates the standard deviation. For a given model (in row), identical letters indicate no significant difference of metric between FS methods after adjustment on the accuracy of the model, the dataset id, the number of features and instances and the retention rate, and using the Tukey's HSD test as multiple test correction at a significance level of 0.05.
The x-axis shows the feature selection methods and the selection rate, average accuracy, and the number of datasets involved; the y-axis shows classifiers and the average accuracy.

changes were achieved for *reliefF* consistently across all ML models.

The lowest *RI* was encountered for *reliefF*, with both reduced mean metrics and standard deviations regardless of the ML model and conversely for *spec* (Figure 2-C). Weighted by the accuracy change (Figure 2-D), the positive value of RIA for the *knn* model could be related to an improvement in accuracy for all FS methods. On the contrary, for the *xg* model, accuracy was slightly degraded, although the differences between FS methods were very minor. The *RIA* of an FS method was largely dependent on the choice of model.

Table I summarizes the average retention rate and accuracy across the datasets according to the FS and ML methods, highlighting high variability in retention rates across the FS methods. Table II presents the pair-wise Kendall's τ of feature rankings between FS methods. Supposing that some techniques of the same family provide fairly consistent results (*e.g.*, the *fisher* score and the *chi2* with *f* score), in that case, the τ would be generally low between techniques, which would tend to show specificities for each FS method. The retention rate of *spec* (63.1%) was significantly higher than with other FS methods (37.9 % on average). *spec* was the most discordant technique compared to the others, with negative τ for almost all comparisons (Table II).

B. Use-cases

Two biomedical datasets were chosen to illustrate how explanations change according to the FS and ML model, especially to clarify the relationship between model accuracy and model explanation similarity. The summary plot of the explanations, inspired by SHAP [54], was used below for an overview of the model explanation.

1) Oxford Parkinson's Disease Detection dataset: Oxford Parkinson's Disease Detection dataset (*parkinsons*, OpenML ID 1488), contains 197 voice recordings (instances) from 23 Parkinson's disease patients and 8 healthy individuals. Each voice record consists of a decomposition of 21 variables (*i.e.*, feature engineering) [77]. Figure 3 shows the summary plots of each FS method for the *en* model. Several FS methods exhibited very low retention rates, with only 1 or 2 features (*jmi*, *fischer*, *f*, *mrmr*, *borutashap*, *cmim* and *rf*) and resulting in τ that cannot be estimated, or a value of 1 for Kendall's τ. On the contrary, *spec* was the most conservative with a high accuracy rate. However, the presence of a non-monotonic correlation between data and influence values in numerous features has created challenges for users attempting to extract meaningful insights. As the objective of the experiments was to attain the highest level of accuracy and the minimum number of features, with explanations as close as possible to the reference (*i.e.*, the original explanation). The *lg* FS technique had the best behavior: only 8 features were chosen and the order of importance of the variables was closely respected compared to the full model. The explanation of most features (7 out of 8) was monotonically related to the data values, allowing the users to easily understand the contribution of the features to the model.

2) Indian Liver Patient Dataset: The Indian Liver Patient dataset (*ilpd*, OpenML ID 41945, [78]) contains 583 instances, including 416 subjects with liver damage and 167 healthy subjects. The dataset was used to evaluate liver disease prediction algorithms and 10 features were recorded: age, gender, TB (Total Bilirubin), DB (Direct Bilirubin), Alkphos (Alkaline Phosphatase), SGPT (Alanine Aminotransferase), SGOT (Aspartate Aminotransferase), TP (Total Proteins), ALB (Albumin), AG (Albumin and Globulin Ratio).

Opposite explanations were found between models. *Age* and *Alkaline Phosphatase* were explained oppositely in the *nb* model (47.9% of accuracy) compared to other models (71.4 to 79.1% accuracy, Figure 4). In terms of the *xg* model, all FS models provided similar accuracy (Figure 5). However, their explanations differed significantly. *spec* and *mrmr*, achieved similar accuracy, but the explanation of *mrmr* was meaningless. The best subset was generated by *borutashap*, having the most similar explanations to the full set of variables with the highest accuracy, although this subset contains more variables than the others FS methods.

IV. Discussion

The main objective of this study was to explore the impact of FS methods on the final prediction explanations obtained. Although retention rate and accuracy are fairly well-established metrics for FS method evaluation, using explanations and related metrics is particularly promising for recommending FS methods.

A. Experiment design

ML pipelines can be complex in their implementation, i.e., being able to integrate stages of standardization, feature engineering, encoding for multi-category classification or regression tasks. To reduce this complexity, all datasets available on OpenML were filtered. Only datasets containing exclusively numerical features were chosen to avoid adding a categorical feature encoding and datasets with missing values were excluded to avoid an additional data imputation step, which may have itself influenced the final explanations. Size of datasets was limited to ensure a sufficient number of features for meaningful FS within acceptable computation times.

Only FS methods providing feature ranking were tested in this study. Apart from these methods, *subset selection* is another category of FS methods with a specific subset as the output; these outputs could not be compared. Moreover, subset selection methods integrate complex and sophisticated strategies to simplify the exponential search for the optimal subset. The computational expensiveness would have undermined the feasibility of the experiment. Nonetheless, the framework offers extensibility that enables users to conveniently integrate any FS method (such as domain-specific FS) for comparison with existing methods.

A crucial challenge in feature ranking is to determine a suitable threshold [79], to select the minimal and necessary number of features (*i.e.*, the retention rate), hence the implementation of the RFECV step. The optimal minimum number of features was determined as the subset of features with the

TABLE I

NUMBER OF DATASETS INVOLVED, RETENTION RATE AND ACCURACY OF EACH FEATURE SELECTION METHOD

FS method	# of datasets involved	Retention rate		ML Models							
				en		knn		nb		xg	
		avg.	std.	avg.	std.	avg.	std.	avg.	std.	avg.	std.
all	145	100.0%	-%	81.75%	11.95%	83.89%	9.15%	73.02%	14.53%	92.01%	7.08%
fisher	131	34.7%	28.0%	80.12%	12.66%	88.78%	7.28%	72.16%	13.59%	91.07%	7.30%
reliefF	134	36.7%	27.4%	79.03%	12.70%	89.17%	7.09%	70.26%	15.00%	91.20%	7.24%
spec	110	63.1%	29.9%	82.27%	**11.88%**	84.23%	11.30%	**73.84%**	14.79%	91.08%	7.46%
f	131	34.9%	27.6%	80.08%	12.79%	88.79%	7.30%	72.08%	13.51%	91.12%	7.27%
chi2	132	35.6%	28.2%	79.10%	14.29%	88.33%	9.90%	71.76%	13.37%	91.13%	7.40%
rfs	125	44.3%	29.0%	81.65%	11.93%	87.22%	7.98%	72.72%	**13.10%**	90.88%	7.65%
mrmr	127	42.7%	30.4%	79.16%	13.81%	86.38%	8.86%	70.54%	15.25%	**91.48%**	**6.61%**
cmim	133	38.7%	31.1%	79.23%	12.76%	88.26%	7.41%	70.48%	15.03%	90.90%	7.43%
jmi	132	40.0%	31.5%	79.45%	12.93%	88.28%	7.62%	69.83%	15.56%	91.11%	7.37%
borutashap	**137**	31.5%	**25.5%**	79.18%	13.12%	**90.08%**	**6.55%**	69.80%	14.73%	91.45%	7.22%
rf	135	**29.3%**	25.7%	78.70%	13.16%	89.84%	6.88%	70.44%	14.28%	91.34%	7.42%
svm	127	42.9%	27.1%	81.69%	11.97%	87.50%	7.66%	72.47%	13.87%	91.07%	7.82%
lg	115	42.8%	29.6%	**82.71%**	12.32%	87.46%	8.38%	73.73%	13.97%	90.24%	7.85%

The number in bold indicates the best value for each column (i.e., lowest retention rate, highest accuracy, lowest standard deviation) and underlined number indicates the worst value.

TABLE II

PAIR-WISE KENDALL'S τ OF THE FEATURE RANKINGS

	fisher	**reliefF**	**spec**	**f**	**chi2**	**rfs**	**mrmr**	**cmim**	**jmi**	**borutashap**	**rf**	**svm**	**lg**
fisher	1	0.125	-0.048	**0.937**	0.393	0.079	0.030	0.092	0.086	0.165	0.166	0.068	0.054
reliefF	0.125	1	-0.047	0.118	0.128	0.061	0.053	0.100	0.103	0.169	**0.175**	0.094	0.069
spec	-0.048	-0.047	1	-0.046	-0.030	-0.014	-0.014	-0.034	-0.037	-0.036	-0.039	-0.017	**0.013**
f	**0.937**	0.118	-0.046	1	0.396	0.076	0.028	0.089	0.084	0.160	0.158	0.069	0.051
chi2	0.393	0.128	-0.030	**0.396**	1	0.064	0.059	0.078	0.086	0.139	0.132	0.111	0.051
rfs	0.079	0.061	-0.014	0.076	0.064	1	0.057	0.093	0.076	0.087	0.095	**0.130**	0.124
mrmr	0.030	0.053	-0.014	0.028	0.059	0.057	1	0.208	**0.210**	0.047	0.050	0.066	0.052
cmim	0.092	0.100	-0.034	0.089	0.078	0.093	0.208	1	**0.347**	0.097	0.105	0.073	0.057
jmi	0.086	0.103	-0.037	0.084	0.086	0.076	0.210	**0.347**	1	0.103	0.117	0.080	0.054
borutashap	0.165	0.169	-0.036	0.160	0.139	0.087	0.047	0.097	0.103	1	**0.229**	0.104	0.072
rf	0.166	0.175	-0.039	0.158	0.132	0.095	0.050	0.105	0.117	**0.229**	1	0.099	0.077
svm	0.068	0.094	-0.017	0.069	0.111	0.130	0.066	0.073	0.080	0.104	0.099	1	**0.225**
lg	0.054	0.069	0.013	0.051	0.051	0.124	0.052	0.057	0.054	0.072	0.077	**0.225**	1

For each row, the number in bold indicates the highest value of τ (most similar) and the underlined number indicates the minimum value (the inverse likely).

highest accuracy score while features were eliminated one-by-one. RFE algorithm is frequently combined with Support Vector Machine (SVM-RFE, [80]–[85]) and Random Forest (RF-RFE, [81], [82], [85]–[87]), achieving powerful performance in many studies. The size of the subset was determined by cross-validation (CV). This paper pioneers the use of RFECV to determine the threshold of selection, enabling automatic comparison between ML models and FS methods. Since some studies have concluded that RF-RFE outperforms SVM-RFE [81], [86] and that accuracy is the best-known evaluation metric, they were chosen for this experiment associated with the tuning of hyperparameters. This step is of course user definable and can be substituted by any other evaluation model and metric.

Finally, this study deliberately focused on "classical" ML models (*en*, *knn*, *nb*, *xg*) for classification tasks, and did not consider deep learning models or regression tasks. However, since the explanation method is model-agnostic, this framework can be effortlessly incorporated for other ML models.

B. Retention rate and model accuracy

The consensual goal of FS is to choose the relevant and non-redundant features. *A priori*, with a real dataset this information is unknown. Therefore a common solution was to use indirect indicators [88] such as the retention rate and accuracy [30]–[35] to assess whether a method is able to select as few features as possible without harming the model performance.

One of the most influential factors of accuracy is model selection, a fundamental question in the ML task [89]. In the *ilpd* dataset, the accuracy of model *nb* was significantly lower than in other models, and the explanation was also clearly different from others. The model choice has a significant impact on the explanation profile [74], and adding FS to a model pipeline provides reinforcement.

Existing reviews have compared diverse combinations of FS/ML in several application fields. The performance of FS techniques highly depends on the context [32]. Mean accuracy between FS methods showed no significant differences while the retention rates were dissimilar (Table I). The most notable example was for *spec*. As demonstrated in Table II, *spec* provided the most different rankings compared to the other FS methods including *reliefF*, although *reliefF* belongs to the same family (similarity-based) and can even be considered as a special case of the *spec* framework. In the original paper [66], *spec* was designed to be a unified supervised and unsupervised feature selection framework that combines ranking, similarity and penalty functions using the graph

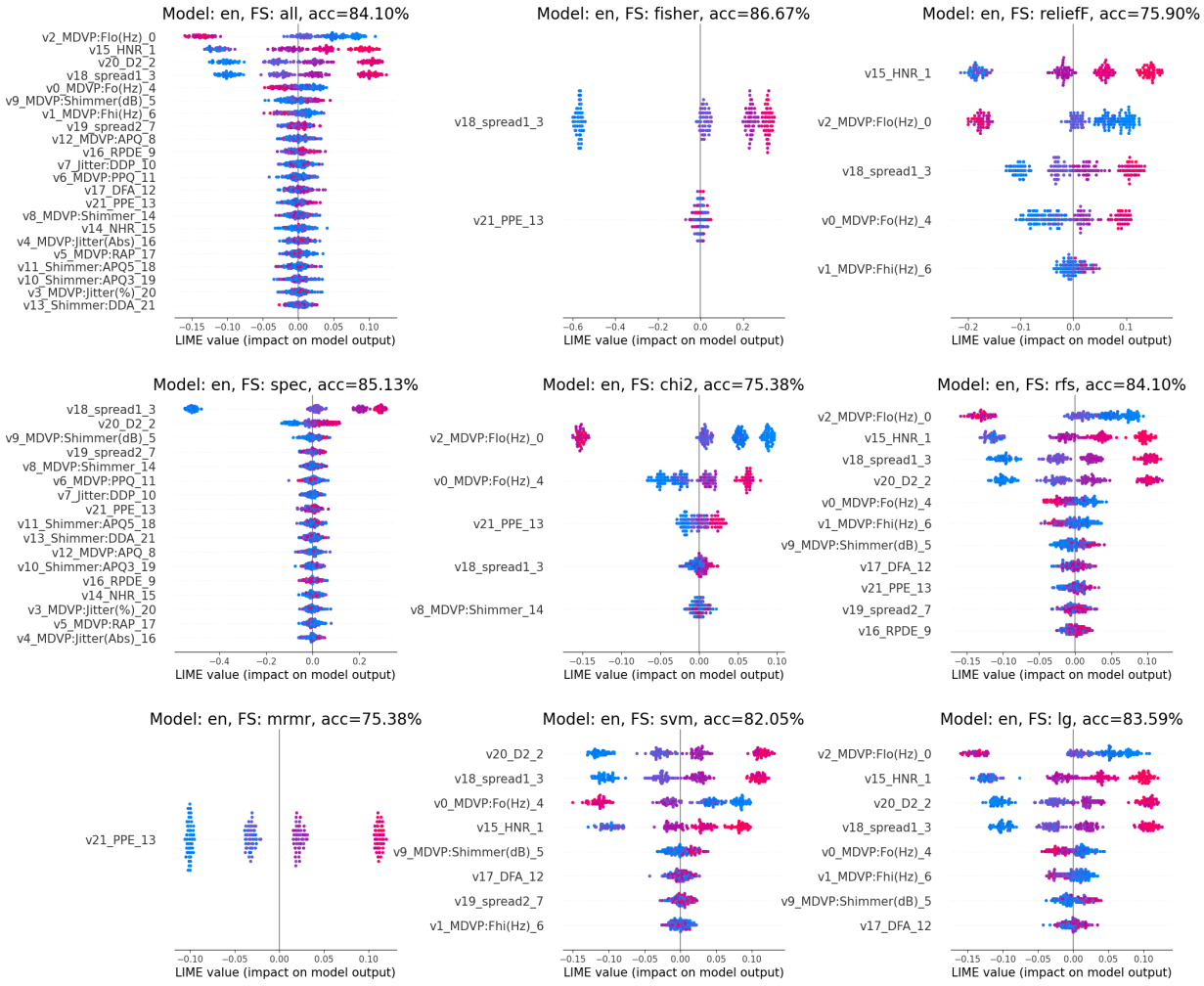

Fig. 3. **Summary plots for the Oxford Parkinson's Disease Detection dataset with different FS methods in the *en* ML model.** The features are sorted in descending order of their contribution to the model. Each feature is renamed according to its original order in the dataset, the feature name, and the order of importance in the explanation for the full feature set ($v[OriginalOrder]\_[FeatureName]\_[ImportanceOrder]$). For each feature, a dot represents an instance of the dataset (red for a high feature value and blue for a low value). A positive value on the x-axis indicates feature contributes positively to the prediction for this instance, and conversely for a negative value.

spectrum. The adaptation of *spec* to unsupervised FS can cause the ranking of generated features to be out of alignment with other FS methods, which varied the retention rate but had little effect on accuracy.

For quite identical accuracies, explanations can lack meaning, as was the case for the *mrmr* method in the *xg* model of the *ilpd* dataset (Figure 5). This issue was due to the extremely low retention rate and an imbalance in the target class (a frequent characteristic of biomedical datasets [90]). The model almost used the ZeroR rule [91], which directly returns the majority category as the prediction. The *parkinsons* use-case (Figure 3), demonstrated that the method with the highest accuracy does not guarantee that the explanations will be similar to the original explanations. That means, if physicians expect relevant explanations, an accuracy-focused practice can lead to inappropriate decisions. Thus, an alternative metric is required to supplement accuracy.

C. Retention rate and model explanation

The main difficulty in assessing feature selection methods is the lack of a ground truth [11]. The complete model (*i.e.*, model trained with all features) contains all available information captured by the ML model and reflected in the explanations (*i.e.*, the full explanation). The hypothesis of this work was that explanations could serve as ground truth, containing the influence of each feature in the model and all confounding factors between features. The aim of FS in the biomedical field may be not only to improve or maintain accuracy, but also to preserve as much information [14] as possible from the original model with a minimum number of features. In fact, the closer the explanations are to the full explanation, the better is FS method, although there may be a slight degradation in accuracy.

The explanations contain two types of information about the features: 1) the influence of the feature on the prediction, and 2) the importance of the feature, which represents the position of a feature in relation to other features. In XAI studies, the former has been used to measure differences between

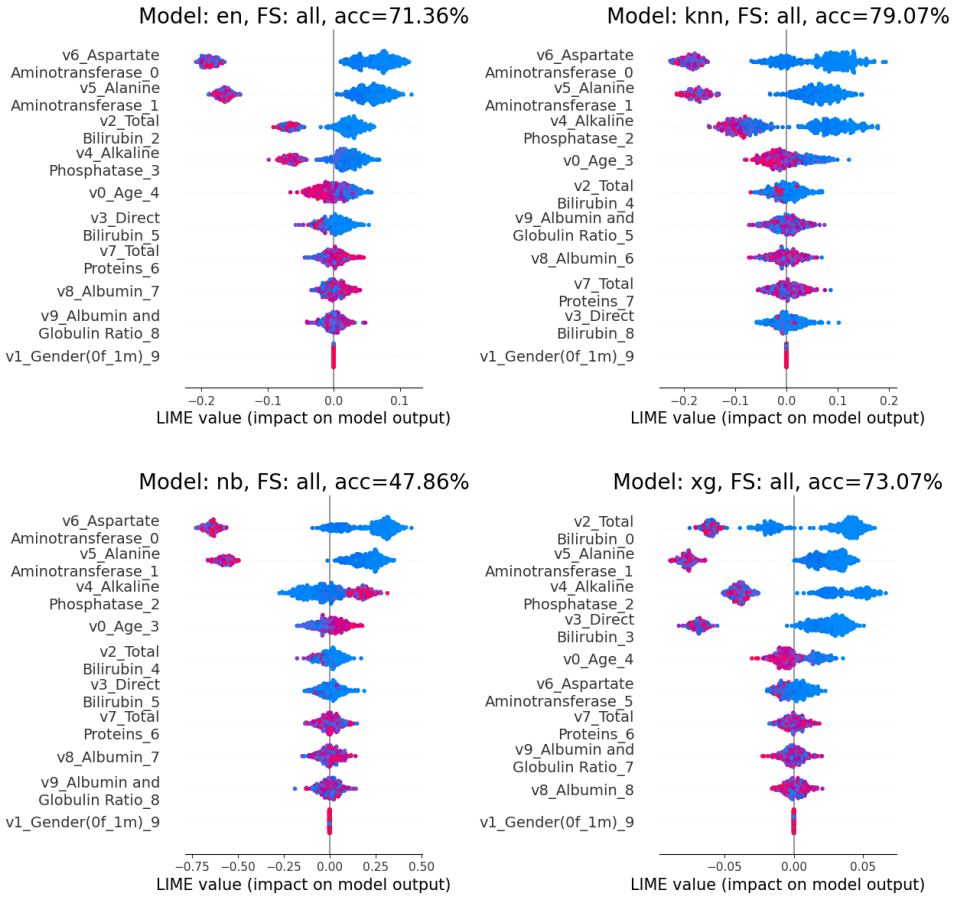

Fig. 4. **Summary plots for the Indian Liver Patient dataset with a full feature set in four models.**

explanations [92]. *Relative influence change* thus allowed to consider influence changes of each feature. Feature importance by ranking was rarely investigated. *Kendall's* τ [75], was used in this work to measure the correlation between two rankings.

The present work demonstrated that the information provided by ranking and influence differed significantly and were therefore complementary. Consequently, the originality of *RI* is to combine both metrics to provide a comprehensive view of explanation differences. It has been demonstrated that accuracy plays an essential role in FS evaluation. Therefore, *RIA* was designed to account for accuracy changes. Furthermore, the threshold for selecting features was determined using accuracy via RFECV. It would also be interesting to use metrics based on explanations to determine the suitable number of features to select.

D. Retention rate, model accuracy and model explanation

Classically for FS, the minimal subset of features to be selected was determined using the accuracy. However, in biomedical applications, the stakeholders may have their own preferences and priorities for FS. For example, physicians can prioritize understandable and reliable explanations to communicate effectively with patients. A loss of accuracy, within reasonable limits, can thus be assumed if the explanations better correspond to reality.

The three-way relationship between the retention rate, model accuracy and model explanation was explored (Figure 6). In most cases, these three dimensions are difficult to conciliate: the optimal FS method was different for each dimension. *reliefF* and *spec* had completely different behaviors, with high accuracy/high retention rate/high variation of explanations and lower accuracy/low retention rate/low variation of explanations, respectively. *lg* and *rf* providing the best accuracy and retention rate, respectively.

A trade-off must then be made between the three dimensions, to be considered in a future FS recommender system and matching with user priorities.

V. Conclusion

In brief, FS must take into account the characteristics and complexity of the dataset and select the appropriate evaluation measures. As stated in the NFL theorem, no single method is the best in every dimension. Each metric also has its own properties. Accuracy of a model is not always trustworthy, while the explanations of the model seems more reliable but much more resource-consuming. The accuracy could be used as a precondition before calculating the explanations as the balance between the three factors made sense only when accuracy was within an acceptable range.

Each decision maker and medical issue has its own preferences, it might be better to let users decide according to

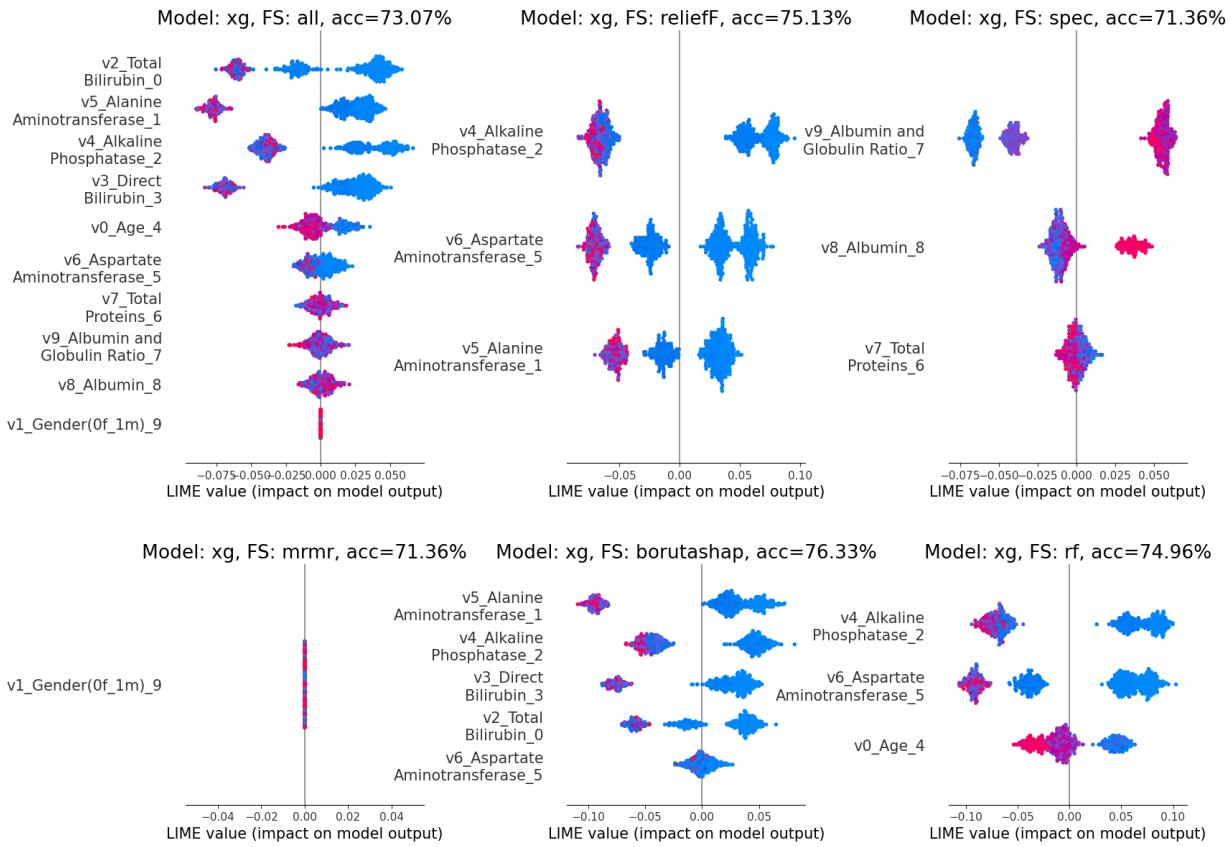

Fig. 5. **Summary plots for the Indian Liver Patient dataset with different feature selection methods in the *xg* model.**

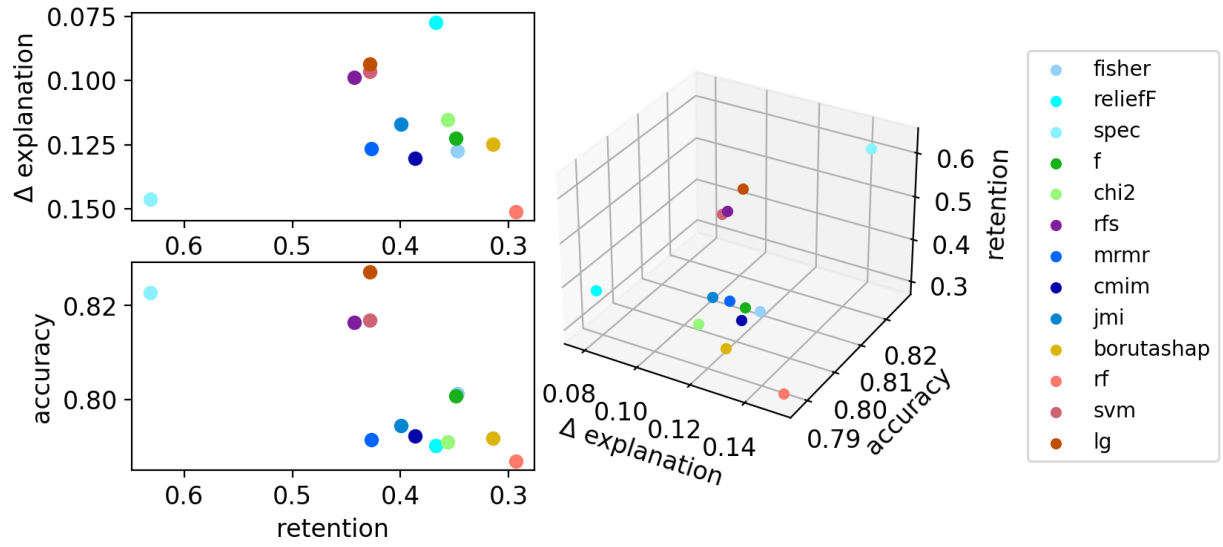

Fig. 6. **Positioning the feature selection methods for *en* model in a tridimensional space with axes that represent respectively explanation (RI), accuracy and retention rate.**

their own needs and create a ”Human-in-the-loop” iteration for feature selection included in the entire ML pipeline. In future work, the study will concentrate on the application of this FS solution on others biomedical datasets with the help of domain experts. New metrics and workflows based on user experience can be developed and integrated into existing systems. The ultimate FS solution should be hybrid, both data and knowledge-driven.

References

[1] J. H. Chen, G. Dhaliwal, and D. Yang, “Decoding artificial intelligence to achieve diagnostic excellence: Learning from experts, examples, and experience,” *JAMA*, vol. 328, pp. 709–710, Aug. 2022.

[2] E. H. Shortliffe and M. J. Sepúlveda, "Clinical Decision Support in the Era of Artificial Intelligence," *JAMA*, vol. 320, pp. 2199–2200, 12 2018.

[3] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies," *Journal of Biomedical Informatics*, vol. 113, p. 103655, 2021.

[4] B. Mittelstadt, "The impact of artificial intelligence on the doctor-patient relationship," tech. rep., Council of Europe, 2022.

[5] S. Reddy, "Explainability and artificial intelligence in medicine," *The Lancet Digital Health*, vol. 4, no. 4, pp. e214–e215, 2022.

[6] P. Ala-Pietilä, Y. Bonnet, U. Bergmann, M. Bielikova, C. Bonefeld-Dahl, W. Bauer, L. Bouarfa, R. Chatila, M. Coeckelbergh, V. Dignum, *et al.*, *The assessment list for trustworthy artificial intelligence (ALTAI)*. European Commission, 2020.

[7] K. Lekadir, R. Osuala, C. Gallin, N. Lazrak, K. Kushibar, G. Tsakou, S. Aussó, L. C. Alberich, K. Marias, M. Tsiknakis, S. Colantonio, N. Papanikolaou, Z. Salahuddin, H. C. Woodruff, P. Lambin, and L. Martí-Bonmatí, "Future-ai: Guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging," 2021.

[8] Y. Zhang, Y. Weng, and J. Lund, "Applications of explainable artificial intelligence in diagnosis and surgery," *Diagnostics*, vol. 12, no. 2, p. 237, 2022.

[9] V. Berisha, C. Krantsevich, P. R. Hahn, S. Hahn, G. Dasarathy, P. Turaga, and J. Liss, "Digital medicine and the curse of dimensionality," *NPJ digital medicine*, vol. 4, no. 1, pp. 1–8, 2021.

[10] R. Bellman and R. Bellman, *Adaptive Control Processes: A Guided Tour*. Princeton Legacy Library, Princeton University Press, 1961.

[11] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 1-4, pp. 131–156, 1997.

[12] B. S. Wade, S. H. Joshi, B. A. Gutman, and P. M. Thompson, "Machine learning on high dimensional shape data from subcortical brain surfaces: A comparison of feature selection and classification methods," *Pattern Recognition*, vol. 63, pp. 731–739, 2017.

[13] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE transactions on evolutionary computation*, vol. 1, no. 1, pp. 67–82, 1997.

[14] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, p. 94, 2018.

[15] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," in *Feature extraction, construction and selection*, pp. 117–136, Springer, 1998.

[16] F. Tan, X. Fu, Y. Zhang, and A. G. Bourgeois, "Improving feature subset selection using a genetic algorithm for microarray gene expression data," in *2006 IEEE International Conference on Evolutionary Computation*, pp. 2529–2534, IEEE, 2006.

[17] O. Babatunde, L. Armstrong, J. Leng, and D. Diepeveen, "A genetic algorithm-based feature selection," *International Journal of Electronics Communication and Computer Engineering*, vol. 5, pp. 889–905, 07 2014.

[18] S. Oreski and G. Oreski, "Genetic algorithm-based heuristic for feature selection in credit risk assessment," *Expert systems with applications*, vol. 41, no. 4, pp. 2052–2064, 2014.

[19] E. Emary, H. M. Zawbaa, and A. E. Hassanien, "Binary grey wolf optimization approaches for feature selection," *Neurocomputing*, vol. 172, pp. 371–381, 2016.

[20] B. H. Nguyen, B. Xue, and M. Zhang, "A survey on swarm intelligence approaches to feature selection in data mining," *Swarm and Evolutionary Computation*, vol. 54, p. 100663, 2020.

[21] M. Rostami, K. Berahmand, E. Nasiri, and S. Forouzandeh, "Review of swarm intelligence-based feature selection methods," *Engineering Applications of Artificial Intelligence*, vol. 100, p. 104210, 2021.

[22] J. C. Debuse and V. J. Rayward-Smith, "Feature subset selection within a simulated annealing data mining algorithm," *Journal of Intelligent Information Systems*, vol. 9, no. 1, pp. 57–81, 1997.

[23] R. Meiri and J. Zahavi, "Using simulated annealing to optimize the feature selection problem in marketing applications," *European journal of operational research*, vol. 171, no. 3, pp. 842–858, 2006.

[24] Y. Li and Y. Liu, "A wrapper feature selection method based on simulated annealing algorithm for prostate protein mass spectrometry data," in *2008 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp. 195–200, IEEE, 2008.

[25] E.-G. Talbi, L. Jourdan, J. Garcia-Nieto, and E. Alba, "Comparison of population based metaheuristics for feature selection: Application to microarray data classification," in *2008 IEEE/ACS International Conference on Computer Systems and Applications*, pp. 45–52, IEEE, 2008.

[26] I. Fister Jr, X.-S. Yang, I. Fister, J. Brest, and D. Fister, "A brief review of nature-inspired algorithms for optimization," *arXiv preprint arXiv:1307.4186*, 2013.

[27] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606–626, 2015.

[28] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[29] L. C. Molina, L. Belanche, and À. Nebot, "Feature selection algorithms: A survey and experimental evaluation," in *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pp. 306–313, IEEE, 2002.

[30] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and information systems*, vol. 34, no. 3, pp. 483–519, 2013.

[31] P. Sun, D. Wang, V. C. Mok, and L. Shi, "Comparison of feature selection methods and machine learning classifiers for radiomics analysis in glioma grading," *IEEE Access*, vol. 7, pp. 102010–102020, 2019.

[32] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, "Benchmark for filter methods for feature selection in high-dimensional classification data," *Computational Statistics & Data Analysis*, vol. 143, p. 106839, 2020.

[33] C.-F. Tsai, "Feature selection in bankruptcy prediction," *Knowledge-Based Systems*, vol. 22, no. 2, pp. 120–127, 2009.

[34] Y. Liu and M. Schumann, "Data mining feature selection for credit scoring models," *Journal of the Operational Research Society*, vol. 56, no. 9, pp. 1099–1108, 2005.

[35] S. Shilaskar and A. Ghatol, "Feature selection for medical diagnosis: Evaluation for cardiovascular diseases," *Expert Systems with Applications*, vol. 40, no. 10, pp. 4146–4153, 2013.

[36] H.-C. Liu, P.-C. Peng, T.-C. Hsieh, T.-C. Yeh, C.-J. Lin, C.-Y. Chen, J.-Y. Hou, L.-Y. Shih, and D.-C. Liang, "Comparison of feature selection methods for cross-laboratory microarray analysis," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 10, no. 3, pp. 593–604, 2013.

[37] G. Kou, P. Yang, Y. Peng, F. Xiao, Y. Chen, and F. E. Alsaadi, "Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods," *Applied Soft Computing*, vol. 86, p. 105836, 2020.

[38] B. Remeseiro and V. Bolon-Canedo, "A review of feature selection methods in medical applications," *Computers in biology and medicine*, vol. 112, p. 103375, 2019.

[39] Y. Aphinyanaphongs, L. D. Fu, Z. Li, E. R. Peskin, E. Efstathiadis, C. F. Aliferis, and A. Statnikov, "A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization," *Journal of the Association for Information Science and Technology*, vol. 65, no. 10, pp. 1964–1987, 2014.

[40] S. M. McNee, J. Riedl, and J. A. Konstan, "Being accurate is not enough: How accuracy metrics have hurt recommender systems," in *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, (New York, NY, USA), p. 1097–1101, Association for Computing Machinery, 2006.

[41] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.

[42] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 11, pp. 4793–4813, 2020.

[43] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies," *Journal of Biomedical Informatics*, vol. 113, p. 103655, 2021.

[44] T. H. Vo, N. T. K. Nguyen, Q. H. Kha, and N. Q. K. Le, "On the road to explainable ai in drug-drug interactions prediction: A systematic review," *Computational and Structural Biotechnology Journal*, 2022.

[45] Q.-H. Kha, T.-O. Tran, V.-N. Nguyen, K. Than, N. Q. K. Le, *et al.*, "An interpretable deep learning model for classifying adaptor protein complexes from sequence information," *Methods*, vol. 207, pp. 90–96, 2022.

[46] C. Molnar, *Interpretable machine learning: A guide for making Black Box models explainable*. Christoph Molnar, 2022.

[47] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pp. 0210–0215, IEEE, 2018.

[48] E. Lughofer, "Model explanation and interpretation concepts for stimulating advanced human-machine interaction with "expert-in-the-loop"," *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, pp. 177–221, 2018.

[49] E. Lughofer and M. Pratama, "Evolving multi-user fuzzy classifier system with advanced explainability and interpretability aspects," *Information Fusion*, vol. 91, pp. 458–476, 2023.

[50] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), p. 1135–1144, Association for Computing Machinery, 2016.

[51] L. S. Shapley, *17. A Value for n-Person Games*, pp. 307–318. Princeton: Princeton University Press, 1953.

[52] E. Štrumbelj and I. Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and information systems*, vol. 41, no. 3, pp. 647–665, 2014.

[53] G. Ferrettini, E. Escriva, J. Aligon, J.-B. Excoffier, and C. Soulé-Dupuy, "Coalitional strategies for efficient individual prediction explanation," *Information Systems Frontiers*, vol. 24, no. 1, pp. 49–75, 2022.

[54] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 4765–4774, Curran Associates, Inc., 2017.

[55] S. Cohen, E. Ruppin, and G. Dror, "Feature selection based on the shapley value," *other words*, vol. 1, p. 98Eqr, 2005.

[56] S. Cohen, G. Dror, and E. Ruppin, "Feature selection via coalitional game theory," *Neural Computation*, vol. 19, no. 7, pp. 1939–1961, 2007.

[57] X. Sun, Y. Liu, J. Li, J. Zhu, H. Chen, and X. Liu, "Feature evaluation and selection with cooperative game theory," *Pattern recognition*, vol. 45, no. 8, pp. 2992–3002, 2012.

[58] E. Keany, "Is this the best feature selection algorithm "borutashap"?." `https://medium.com/analytics-vidhya/is-this-the-best-feature-selection-algorithm-borutashap-8bc238aa1677`, 2020. [Online; accessed 2022-09-08].

[59] E. Keany, "Borutashap package," 2022. `https://github.com/Ekeany/Boruta-Shap/`, Last accessed on 2022-10-22.

[60] X. Man and E. P. Chan, "The best way to select features? comparing mda, lime, and shap," *The Journal of Financial Data Science*, vol. 3, no. 1, pp. 127–139, 2021.

[61] Y. Liu, Z. Liu, X. Luo, and H. Zhao, "Diagnosis of parkinson's disease based on shap value feature selection," *Biocybernetics and Biomedical Engineering*, vol. 42, no. 3, pp. 856–869, 2022.

[62] P. A. Moreno-Sanchez, "An automated feature selection and classification pipeline to improve explainability of clinical prediction models," in *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pp. 527–534, IEEE, 2021.

[63] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo, "Openml: networked science in machine learning," *SIGKDD Explorations*, vol. 15, no. 2, pp. 49–60, 2013.

[64] H. Yoon, K. Yang, and C. Shahabi, "Feature subset selection and feature ranking for multivariate time series," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 9, pp. 1186–1198, 2005.

[65] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Machine learning proceedings 1992*, pp. 249–256, Elsevier, 1992.

[66] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in *Proceedings of the 24th international conference on Machine learning*, pp. 1151–1157, 2007.

[67] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $\ell 2$, 1-norms minimization," *Advances in neural information processing systems*, vol. 23, 2010.

[68] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[69] F. Fleuret, "Fast binary feature selection with conditional mutual information.," *Journal of Machine learning research*, vol. 5, no. 9, 2004.

[70] H. Yang and J. Moody, "Feature selection based on joint mutual information," in *Proceedings of international ICSC symposium on advances in intelligent data analysis*, vol. 1999, pp. 22–25, Citeseer, 1999.

[71] I. Guyon, J. Weston, S. D. Barnhill, and V. N. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, pp. 389–422, 2004.

[72] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the royal statistical society: series B (statistical methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[73] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

[74] E. Doumard, J. Aligon, E. Escriva, J.-B. Excoffier, P. Monsarrat, and C. Soulé-Dupuy, "A comparative study of additive local explanation methods based on feature influences," in *24th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data ((DOLAP 2022)*, vol. 3130, (Edinburgh, United Kingdom), pp. 31–40, CEUR-WS.org, Mar. 2022.

[75] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[76] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[77] M. Little, P. McSharry, E. Hunter, J. Spielman, and L. Ramig, "Suitability of dysphonia measurements for telemonitoring of parkinson's disease," *Nature Precedings*, pp. 1–1, 2008.

[78] B. V. Ramana, M. S. P. Babu, N. Venkateswarlu, *et al.*, "A critical study of selected classification algorithms for liver disease diagnosis," *International Journal of Database Management Systems*, vol. 3, no. 2, pp. 101–114, 2011.

[79] B. Seijo-Pardo, V. Bolón-Canedo, and A. Alonso-Betanzos, "On developing an automatic threshold applied to feature selection ensembles," *Information Fusion*, vol. 45, pp. 227–245, 2019.

[80] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1, pp. 389–422, 2002.

[81] P. M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, "Recursive feature elimination with random forest for ptr-ms analysis of agroindustrial products," *Chemometrics and intelligent laboratory systems*, vol. 83, no. 2, pp. 83–90, 2006.

[82] Y. Xu, Z. Li, and L. Luo, "A study on feature selection for trend prediction of stock trading price," in *2013 International Conference on Computational and Information Sciences*, pp. 579–582, IEEE, 2013.

[83] K. Yan and D. Zhang, "Feature selection and analysis on correlated gas sensor data with recursive feature elimination," *Sensors and Actuators B: Chemical*, vol. 212, pp. 353–363, 2015.

[84] N. Rtayli and N. Enneya, "Enhanced credit card fraud detection based on svm-recursive feature elimination and hyper-parameters optimization," *Journal of Information Security and Applications*, vol. 55, p. 102596, 2020.

[85] M. Lee, J.-H. Lee, and D.-H. Kim, "Gender recognition using optimal gait feature based on recursive feature elimination in normal walking," *Expert Systems with Applications*, vol. 189, p. 116040, 2022.

[86] Q. Chen, Z. Meng, X. Liu, Q. Jin, and R. Su, "Decision variants for the automatic determination of optimal feature subset in rf-rfe," *Genes*, vol. 9, no. 6, p. 301, 2018.

[87] B. F. Darst, K. C. Malecki, and C. D. Engelman, "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data," *BMC genetics*, vol. 19, no. 1, pp. 1–6, 2018.

[88] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *The Journal of Machine Learning Research*, vol. 5, pp. 1205–1224, 2004.

[89] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," *arXiv preprint arXiv:1811.12808*, 2018.

[90] M. M. Rahman and D. N. Davis, "Addressing the class imbalance problem in medical datasets," *International Journal of Machine Learning and Computing*, pp. 224–228, 2013.

[91] R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald, and D. Scuse, "Weka manual for version 3-9-1," *University of Waikato, Hamilton, New Zealand*, 2016.

[92] E. Doumard, J. Aligon, E. Escriva, J.-B. Excoffier, P. Monsarrat, and C. Soulé-Dupuy, "A quantitative approach for the comparison of additive local explanation methods," *Information Systems*, vol. 114, p. 102162, 2023.