# Variance-based importance measures for machine learning model interpretability

Iooss Bertrand, Vincent Chabridon, Thouvenot Vincent

Institut pour la Maîtrise des Risques
Sûreté de Fonctionnement - Management - Cindyniques

# Variance-based importance measures for machine learning model interpretability

IOOSS Bertrand
*SINCLAIR AI Laboratory*
*EDF R&D*
Chatou, France
bertrand.iooss@edf.fr

CHABRIDON Vincent
*SINCLAIR AI Laboratory*
*EDF R&D*
Chatou, France
vincent.chabridon@edf.fr

THOUVENOT Vincent
*SINCLAIR AI Laboratory*
*THALES SIX GTS*
Palaiseau, France
vincent.thouvenot@thalesgroup.com

***Résumé***—Les algorithmes statistiques d'apprentissage automatique (ou *machine learning*) connaissent un essor sans précédent dans le monde industriel, notamment pour l'aide à la décision en ingénierie des systèmes critiques. Toutefois, leur manque d'"interprétabilité" est un verrou à lever afin de rendre ces outils intelligibles et auditables. Ce papier vise à dresser une cartographie de certaines métriques d'interprétabilité (appelées "mesures d'importance") dont le but est de quantifier l'impact de chaque prédicteur sur la variance de la sortie du modèle statistique. Il est montré que le choix d'une métrique pertinente doit être guidé par les contraintes inhérentes aux données et au modèle considéré (caractère linéaire ou non du phénomène d'intérêt, dimension du problème, dépendance des prédicteurs) et par le type d'étude que l'utilisateur souhaite mener (détecter les variables influentes, les hiérarchiser, etc.). Enfin, ces métriques sont estimées et analysées sur un jeu de données public afin d'illustrer certaines de leurs propriétés théoriques et empiriques.

***Keywords***—apprentissage statistique, interprétabilité, analyse de sensibilité, effets de Shapley, indices de Sobol'

***Abstract***—Machine learning algorithms benefit from an unprecedented boost in the industrial world, in particular in support of decision-making for critical systems. However, their lack of "interpretability" remains a challenge to leverage in order to make these tools fully intelligible and auditable. This paper aims to track and synthesize of a panel of interpretability metrics (called "importance measures") whose aim is to quantify the impact of each predictor on the statistical model's output variance. It is shown that the choice of a relevant metric has to be guided by proper constraints imposed by the data and the considered model (linear vs. nonlinear phenomenon of interest, input dimension, input dependency) together with taking the type of study the user wants to perform into consideration (detect influential variables, rank them, etc.). Finally, these metrics are estimated and analyzed on a public dataset so as to illustrate some of their theoretical and empirical properties.

***Keywords***—statistical learning, interpretability, sensitivity analysis, Shapley effects, Sobol' indices

## I. Introduction

Machine learning (ML) is one of the sound and substantial branches of artificial intelligence technology and provides a large panel of algorithmic tools to learn from data (e.g., numerical data, images, sounds, texts). However, ML algorithms are often considered as black-box models, linking features (also called "inputs") to variables of interest (also called "outputs"), and may provide predictions which turn out to be difficult to explain or interpret. Therefore, the industrial deployment of these solutions requires tools together with a panel of best practices to perform explainable and interpretable ML [1]–[3]. For example, in the industrial nondestructive testing field (e.g., for aeronautics or nuclear industry), generalized automated inspections (that will allow a large gain in terms of efficiency and economy) are planned to be used intensively. However, for these analyses (and the underlying algorithms that are used), some transparency guarantees are required to ensure strong confidence in the predictions [4].

ML interpretability is linked to the ability, for a human mind, to understand representations of the ML model such as resulting predictions and associated decisions. ML models are intensively used to make predictions on output quantities of interest and these quantities are used to make decisions (e.g., regarding safety criteria or economic efficiency and profitability ones). In this view, the strong connections that exist between ML interpretability [5] and sensitivity analysis (SA) of model outputs [6], [7] have been recently recognized [8], [9]. Indeed, in its broadest sense, SA aims at studying how the uncertainty in a model output can be apportioned to different sources of uncertainty in its inputs. As in SA, ML interpretability can be based either on visualization tools, global metrics (to interpret the global model behavior) or on local ones (to interpret the model output for a given instance).

To interpret ML models, a large panel of methods are available. Some of them are original contributions from the ML community, while others have been purely reinvented [10]. However, these methods are often empirical or approximate and the description of their underlying assumptions and conditions of applicability is often omitted (see, e.g., [5]). Therefore, developing synthetic and methodological overviews, as done in SA (see, e.g., [11]), would be useful to develop a global understanding of the different methods.

A taxonomy of interpretability methods can be built with various criteria [10]. A first classification could be proposed following the type of results provided by the method (e.g.,

visualization tools or summary statistics). Such an approach is useful from a user-friendly perspective but does not really help, neither to understand the underlying similarities and differences between the methods, nor to take quantitatively informed decisions. Moreover, the literature on the subject often distinguishes intrinsically interpretable models and model interpretability techniques. In the first case, models are called "transparent", while in the second case, the techniques are referred to as "post-hoc interpretability" [2], [10]. By post-hoc techniques, one considers that a supplementary layer of statistical quantities (e.g., sensitivity indices, feature importance measures) is required to achieve a proper interpretation of the model results and a full understanding of the model behavior. As an example, linear models (such as linear and logistic regressions) are often assumed to be the prototype of intrinsic transparent models. However, even for this simple class of models, direct interpretability might be tricky for many reasons (as, e.g., strong dependencies between inputs). Thus, from an industrial perspective where generic tools are often required, post-hoc and model-agnostic techniques, that are independent of the type of ML model, seem inevitable.

Inspired by what has been achieved in SA and taking a generic application viewpoint based on variance-based metrics, this work discusses three main challenges which compose the outline of the paper: complexity of the input-output relationship (and, consequently, the complexity of the ML model), dependence among inputs and large input dimension. Section II aims at primarily describing some practical ML settings. Then, Section III distinguishes the ML models along with the underlying complexity they try to capture. Several popular global metrics, gathered under the common term "importance measures" (IM) in the rest of this paper, are investigated. In particular, the variance-based IM which are able to measure the nonlinearity degree (as well as the degree of interaction between inputs) in a ML model, are reviewed. Section IV focuses on the dependence between inputs which appears to be underrepresented in the ML interpretability literature, even if such a problem is ubiquitous in daily ML-based applications. Section V discusses the high-dimensional input space issue. Indeed, whether in SA or in ML, the computation of IM (such as, for instance, the popular Shapley effects) is known to be subject of the curse of dimensionality. In Section VI, the different IM are applied to a public dataset, while Section VII gives synthesizes this work.

To help the reader, Table I provides a table of acronyms used all along the paper.

Table I
ACRONYMS.

| | |
|---|---|
| HSIC | Hilbert-Schmidt Independence Criterion |
| IM | Importance Measure |
| LMG | Lindeman-Merenda-Gold measure |
| MDA | Mean Decrease Accuracy |
| MDI | Mean Decrease Impurity |
| ML | Machine Learning |
| OOB | Out Of Bag |
| PME | Proportional Marginal Effects |
| PMVD | Proportional Marginal Variance Decomposition |
| RF | Random Forest |
| RKHS | Reproducing Kernel Hilbert Space |
| RWA | Relative Weight Analysis |
| SA | Sensitivity Analysis |
| SAGE | Shapley Additive Global Importance |
| SHAFF | SHApley eFfects via random Forests |
| SHAP | SHapley Additive exPlanations |
| SPVIM | Shapley Population Variable Importance Measure |
| SRC | Standard Regression Coefficient |
| SVD | Singular Value Decomposition |
| VIF | Variance Inflation Factor |

## II. A METHODOLOGICAL VIEW OF MACHINE LEARNING

### A. Machine learning methodology

Supervised learning consists in building a statistical model from a set of labeled (i.e., input-output) samples. Such a model is then used in order to predict, for a new set of input values, the corresponding output value [12]. Mathematically, one assumes that the phenomenon under study can be expressed by a relationship given by :

$$Y = f_{\text{true}}(\mathbf{X}) \ , \tag{1}$$

where $Y$ is the output (supposed to be scalar here, for the sake of simplicity) and $\mathbf{X} = (X_1, \ldots, X_d)$ a vector of $d$ inputs, usually called "predictors" or "features" in ML. The idea is that the true relationship $f_{\text{true}}(\cdot)$ is unknown, but one can build a surrogate statistical model $f(\cdot)$ based on the available data gathered in a $n$-size sample denoted by $\left(X_1^{(i)}, \ldots, X_d^{(i)}, Y^{(i)}\right)_{i=1,\ldots n}$. Note that, if $Y$ is *quantitative*, one lies in the *regression* framework, while if $Y$ is *qualitative*, one lies in the *classification* framework.

Figure 1 synthesizes the general scheme of building and using (for prediction and/or for interpretation) a ML model (denoted by $f(\cdot)$). Different steps are required, from the data and features extraction/selection (Step A), to the ML model building process (Step B) and its validation (Step B'). The ML model is then used to make predictions from new inputs (Step C), associated with interpretative elements (Step C').

### B. Settings for machine learning interpretability

Step C' of the ML scheme (Fig. 1), which mainly consists in calculation/estimation of IM, is closely related to the SA step of the uncertainty quantification of numerical experiments methodology [13]. In the SA community, four main settings have been recently recognized [7] in real-world studies: the *model exploration* which aims to understand the behavior of a model by trying to investigate the input-output relationship, the *factor fixing* (or *screening*) which aims to reduce the number of uncertain inputs by setting unimportant factors as constants, the *factor prioritization* (or *ranking*) which aims to give a quantitative ranking among the most important factors, and the *input distribution robustness* which aims to analyze variation of the quantity of interest with respect to uncertainty in inputs' distributions.

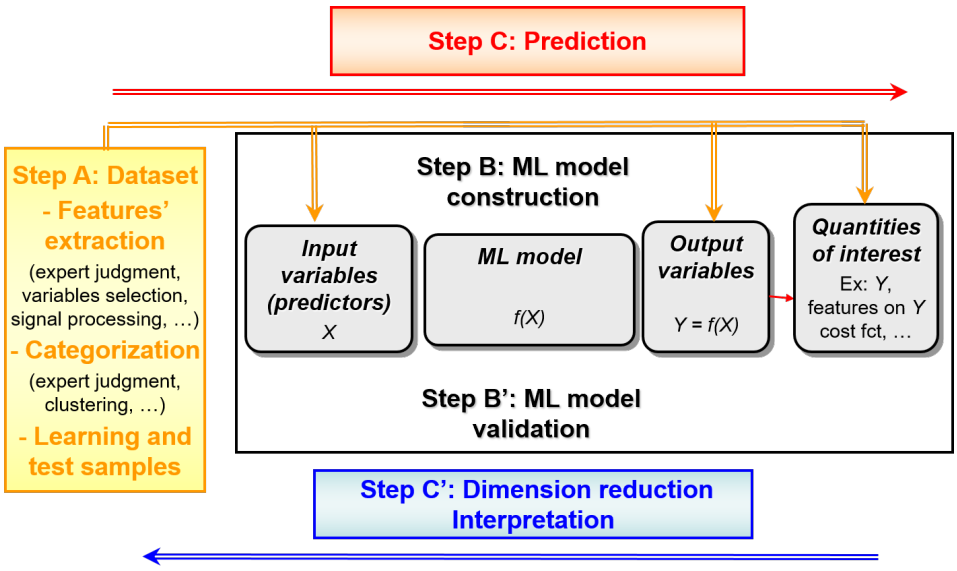


Figure 1. Machine learning global methodology.

Therefore, by analogy with the SA methodology, and inspired from the analyses of [2], [3], [10], [14], four main settings of ML interpretability can be defined: visualization, features identification, measures of importance, robustness of the decision boundary. They are briefly explained in the following subsections. Note that the objectives and results could be different if a tool is applied to the data or to ML model predictions.

*1) Visualization:* The visualization of the relationship between the output (label) and the features can be made via the use of well-known plots in ML interpretability (partial dependence plots, individual conditional expectation, accumulated local effects) [10] and in SA (scatterplots or conditional expectation plots, parallel coordinate plot) [7]. A fruitful discussion about interpretation of these graphs is given in [15].

Figures 2 and 3 provide illustrations of a scatterplot and a parallel coordinate plot applied to the Boston housing dataset (see Section VI). The parallel coordinate plot (usually called "cobweb plot" in SA [16]) consists in visualizing the data as a set of trajectories. By interactively selecting a small number of trajectories corresponding to a specific event of interest for the user (e.g., the 10% largest output values, as shown in Fig. 3), specific values of other variables are automatically emphasized, which leads to visually identify the influential input values regarding the occurrence of this event.

*2) Features identification:* Identifying features in the data (or with a ML model) aims to find the minimal number of inputs that reach a maximized accuracy of the ML model (see, e.g., [17], [18]). It involves the use of statistical techniques such as correlation measures (e.g., Pearson coefficient, Spearman's rank correlation coefficient, Kendall's tau, copula, Hoeffding's D, distance correlation) and kernel-based metrics [19]. Even if the context is slightly different, it mainly corresponds to the "factor fixing" setting in SA that is achieved using the so-called "screening" techniques [7].

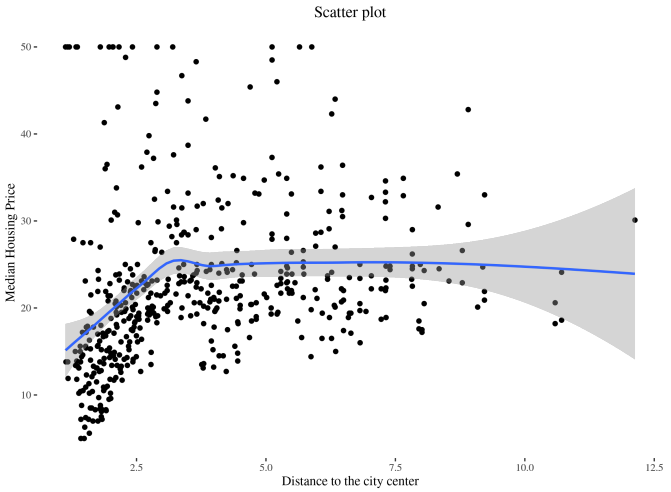


Figure 2. Scatterplot between one input and the output of the Boston housing dataset.

*3) Measures of importance:* This setting consists in measuring, in a quantitative manner, the impact of the inputs on the output (see, e.g., [20], [21]). It corresponds to the "factor prioritization" setting in SA, where the goal is to detect and rank influential inputs to explain the model (e.g., to experts or authorities). One can be interested in two types of impact. The first one, which is global, consists in explaining how a feature or a set of features impact the output distribution. The second one, which is more local, consists in explaining how the features impact the model's output for one specific instance. Once again, IM can be computed on the ML model predictions or directly on the data. Note that a confidence on

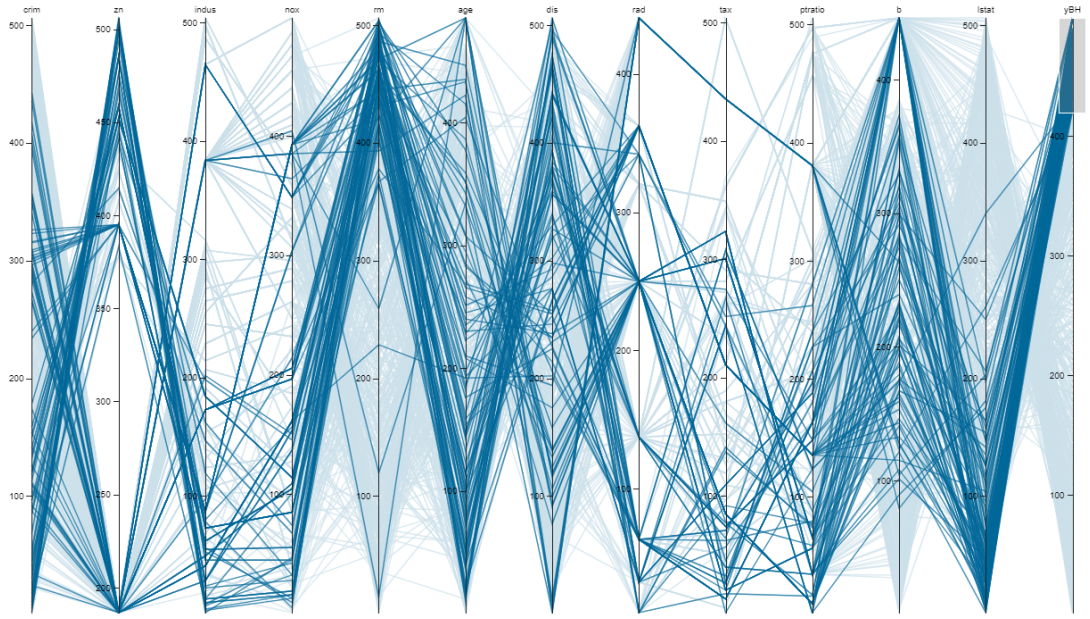


Figure 3. Parallel coordinate plot of the Boston housing dataset. The last column corresponds to the variable of interest (median value of owner-occupied homes).

the ML model needs the equality between the IM of the ML model with the IM computed on the data.

*4) Robustness of the decision boundary:* This last setting looks at the variability of some results obtained from the ML model as a function of perturbations in the data (inputs or output). It concerns two major topics: $(i)$ the counterfactual examples which consist in explaining the prediction related to one individual by another close individual with an opposite target prediction [10]; $(ii)$ the robustness with respect to the input data, i.e., to tackle the way the output label changes when the distribution of one input (or a group of inputs) change due to a perturbation in the data [22]. It corresponds to the "input distribution robustness" setting in SA.

All in all, this paper is mainly devoted to the "Measure of importance" setting, and discusses several adapted interpretability metrics in the following sections.

## III. Machine learning model complexity

In this section, the goal is to discuss how ML model complexity might affect the interpretability step. The term "complexity" denotes here the linearity or nonlinearity of the ML model (assuming that the complexity of the model is adapted and validated regarding the linearity / nonlinearity of the underlying phenomenon of interest, e.g., by using some diagnostics tool). The occurrence of interaction effects between inputs is also symptomatic of complexity [23]. It is considered in this paper as nonlinearity for the sake of simplicity.

As a consequence, choosing a specific ML model leads to a specific complexity which may induce some relevant choices for interpretability. Thus, in this section, we first discuss existing "post-hoc" interpretability methods dedicated to two classes of models with different complexity: on the one hand, the linear regression model and on the other hand, the random forest algorithm (which can easily capture nonlinearities). These basic algorithms have been chosen, not only because of their simplicity, but also as they remain some of the most versatile and popular algorithms among the ML practitioners (see, e.g., [12], [24]). Finally, we present more general and "model-agnostic" interpretability metrics.

Note that, in the remaining of this section, the predictors are assumed to be independent.

### A. IM for independent inputs in the linear context

From the $n$-size input-output sample, it is possible to fit the following linear model:

$$Y = Y(\mathbf{X}) = \sum_{j=0}^{d} \beta_j X_j + \varepsilon, \tag{2}$$

where $X_0 = 1$, $\beta = (\beta_0, \ldots, \beta_d)^t \in \mathbb{R}^{d+1}$ is the vector of regression parameters, and $\varepsilon \in \mathbb{R}$ the model's error of constant variance $\sigma^2$. Since there is an intercept term in Eq. (2), we consider that all inputs are centered (i.e., $\mathbb{E}X_j = 0$ for $j = 1, \ldots, d$) without any loss of generality.

A global IM of the model given by Eq. (2) can be understood as a "relative importance" or a "relative contribution"

of a given input variable $X_j$ (with $j \in \{1, \ldots, d\}$) on the dispersion of another variable, typically the output variable $Y$ [25], [26]. A usual and useful dispersion metric can be the *coefficient of determination* $R^2$ (of the linear regression model) given by:

$$R^2 = R^2_{Y(\mathbf{X})} = 1 - \frac{\sum_{i=1}^{n} \left[ Y^{(i)} - \widehat{Y}(\mathbf{X}^{(i)}) \right]^2}{\sum_{i=1}^{n} \left[ Y^{(i)} - \mathbb{E}[Y] \right]^2}, \quad (3)$$

where $\widehat{Y}(\cdot)$ is the predictor of the model built from the data. This quantity represents the percentage of output variance explained by the model.

Thus, in the context of linear models, a formal definition of an IM has been proposed by [26]: *"The proportionate contribution each variable makes to $R^2$, considering both its direct effect (i.e., its correlation with the response) and its effect when combined with the other variables in the model."* In addition to that, four desirability criteria for the $R^2$ decomposition into shares (due to each input) have been proposed by [27]:

1) *Proper decomposition*: the sum of all shares should be equal to the model variance;
2) *Non-negativity*: all shares should be non-negative;
3) *Exclusion*: if $\beta_j = 0$ (with $\beta_j$ being the regression coefficient associated to $X_j$), the share of $X_j$ should be zero;
4) *Inclusion*: if $\beta_j \neq 0$, the share of $X_j$ should be nonzero.

The associated IM of the linear model are called the *standard regression coefficients* (SRC):

$$\mathrm{SRC}_j = \beta_j \sqrt{\frac{\mathrm{Var} X_j}{\mathrm{Var} Y}}. \quad (4)$$

This metric is used for relative importance [28] as it indicates "the effect of moving an input away from its expected value by a fixed fraction of its standard deviation while holding all other variables fixed at their expected values" [29]. Each $\mathrm{SRC}_j^2$, also known as the "betasq" [28], are variance-based IM which expresses the part of the $R^2$ explained by the input $X_j$. However, this interpretation is only valid when all the inputs are mutually independent.

### B. IM for independent inputs in the RF context

For features ranking in ML, importance score methods associated to permutation and resampling techniques are widely used, for example via the random forest (RF) model [29], [30]. The RF model is a substantial modification of boostrap aggregating (also known as "bagging") that builds a large collection of decorrelated trees, and then averages them. In particular, bagging [31] is a ML ensemble meta-algorithm designed to improve robustness and accuracy of ML algorithms used for classification or regression. It consists in generating $B$ bootstrap samples from the original dataset, train $B$ ML models and aggregate their predictions (e.g., by averaging in the regression case, or by voting for a supervised classification). An important feature of RF is the use of out-of-bag (OOB) samples. For each observation, the RF predictor is constructed by averaging only the trees of the RF corresponding to bootstrap samples in which the observation did not appear.

A well-known RF-based IM is the *Mean Decreasing Impurity* (MDI) [32]. For this criterion, at each split in each tree, the improvement in the split-criterion is the IM attributed to the splitting feature, and is accumulated over all the trees in the forest separately for each feature. However, it is known that MDI is biased, in particular because it tends to inflate the importance of continuous or high-cardinality categorical variables [33]. RF also uses the OOB samples to construct a different IM, apparently to measure the prediction strength of each feature. When the $b$-th tree is grown, the OOB samples are passed down the tree, and the prediction accuracy is recorded. Then the values for the $j$-th feature are randomly permuted, i.e., the values of the features are shuffled, in the OOB samples, and the accuracy is again computed. The decrease in accuracy as a result of this permuting is averaged over all trees, and is used as a measure of the importance of feature $j$ in the RF. This criterion is called *Mean Decreasing Accuracy* (MDA). It is a variance-based IM that can be generalized to any ML model as explained in the next section.

### C. Model-agnostic IM for independent inputs

In the general ML setting, [34] introduced the so-called *permutation feature importance*, based on a similar idea to MDA. It consists in computing the increase in the model's prediction error after permuting the feature. Suppose one has a trained model, denoted by $\widehat{f}(\cdot)$, with $\mathbf{X}$ a feature matrix, $\mathbf{y}$ a target vector and $L(\mathbf{y}, f(\cdot))$ an error measure. Algorithm 1 provides the structure of the permutation feature importance algorithm.

**Algorithm 1** Permutation feature importance.

1) Estimate the original error $e^o = L(\mathbf{y}, f(\mathbf{X}))$;
2) For each feature $j$:
   a) Generate a feature matrix $\mathbf{X}^{\text{perm}}$ by permuting feature $j$ in the data $\mathbf{X}$;
   b) Estimate the permutation error $e^p = L(\mathbf{y}, f(\mathbf{X}^{\text{perm}}))$;
   c) Compute the permutation feature IM for $j$: $e^p/e^o$ or $e^p - e^o$.
3) Sort the features by descending permutation feature IM.

For MDA in RF, the choice of the feature matrix is immediate: it is the OOB samples. Such samples do not exist for a large majority of ML techniques. Instead of OOB, one can use a validation set (different from the training set). Permutation feature IM represents the increase in model error when a single feature value is randomly shuffled. Feature IM provides a highly compressed, global insight into the model's behavior. Back to Algorithm 1 (see step (c)), note that when the error ratio is used instead of the difference-based one, the result is unitless. Thus, results obtained for two different problems are comparable. The features IM takes into account both

the main feature effect and the interaction effects on model performance, because by permuting the feature, its interaction with the others features is broken. Another advantage is that permutation feature IM does not require to retrain the model.

As explained in [35], [36], and reviewed in [29], the MDA (and then the permutation feature IM) is equivalent (up to a multiplicative constant) to the *total Sobol' index* [37] which is one of the most widely used tool in SA and that has been proven to be robust and easily interpretable [38]. In the case of independent inputs and a deterministic relationship $Y = f(\mathbf{X})$, the variance-based sensitivity measures, also called *Sobol' indices* [7], [39], write:

$$S_j = \frac{\text{Var}\left(\mathbb{E}\left[Y|X_j\right]\right)}{\text{Var}(Y)}, \quad (5)$$

$$S_{j,k} = \frac{\text{Var}\left(\mathbb{E}\left[Y|X_j, X_k\right]\right)}{\text{Var}(Y)} - S_j - S_k, \ \ldots \quad (6)$$

The terms $S_j$ ($j \in \{1, \ldots, d\}$) are called "first-order Sobol' indices" (measures of the individual effect of each input), $S_{j,k}$ ($j \in \{1, \ldots, d\}, k \in \{1, \ldots, d\} \setminus j$) "second-order" Sobol' indices (measures of the second-order interactions) and so on. Moreover, the total Sobol' index for any component $j$ is defined by:

$$T_j = S_j + \sum_{k=1, k\neq j}^{d} S_{j,k} + \cdots + S_{1,\ldots,d} = \frac{\mathbb{E}\left[\text{Var}(Y|\mathbf{X}_{-j})\right]}{\text{Var}(Y)}, \quad (7)$$

where $\mathbf{X}_{-j}$ is the vector $\mathbf{X}$ without $X_j$. This total index measures the effect of $X_j$ through its individual contribution and all its interactions with other inputs. Thus, this interpretation allows to better understand the MDA metric. Finally, let us note that, in the linear regression model (2), the total Sobol' indices are equal to the first-order Sobol' indices and to the $\text{SRC}^2$.

Therefore, Sobol' indices can be used to identifying the most influential features learned by ML models, as well as to detect interactions between features. They have been used to prune redundant neurons in artificial neural networks model [40] and to capture high-order interactions between image regions and their contributions to a neural network [41]. Thanks to SA-based efficient computation schemes of total Sobol' indices, [42] use the notion of "mean dimension" (which represents the mean interaction degree between inputs that acts on the output of a model) to characterize the internal structure of complex modern neural network architectures. The mean dimension allows to summarize to which extent the ML model is dominated by high or low-order interactions. This tool is then useful to analyze architectures or neural networks-based models regarding some accuracy metrics. The mean dimension also allows to identify, from a deep analysis, interactions between layers of the neural network.

## IV. Problems with dependent inputs in machine learning interpretability

Since the early works of [43] and [44] in the statistical community, it is well-known that dealing with correlated or multicollinear explanatory variables may lead to a misinterpretation of the regression analysis results. As an example, the regression weights lose their meaning and become tricky to interpret. Therefore, various diagnostic tools (e.g., visualization tools, dependence and multicollinearity statistics, stepwise procedures) have been proposed to identify multicollinearity in the regressors and to allow redundant features removal.

Among the panel of literature reviews or recent state-of-the-art reports (see, e.g., [45], [46]) which have been proposed on the topic of ML interpretability (or more generally about explainable artificial intelligence), the issue of the dependent features appears to be globally underrepresented, even if such a problem is ubiquitous in daily applications of ML. A recent work from [47] clearly points out this pitfall and stresses the need for better characterization of input dependence. Therefore, in a preliminary step of ML building, knowing if there are some dependencies between inputs is essential. The next section gives a short literature review on this topic.

### A. Dependency diagnostics

Linear statistical dependence is called *multicollinearity*. Identifying it is of great importance since it affects statistical estimation of the regression coefficients and influences the efficiency of least-squares estimation [48]. Two variables are collinear if they lie almost on the same line, i.e., if they have a high linear correlation between them. This notion is generalized to more than two variables by saying that collinearity exists if there is a high multiple correlation when one of the variables is regressed on the others. A standard and simple measure of multicollinearity is the so-called *variance inflation factor* (VIF) which writes [43], [49]:

$$\text{VIF}_j = \frac{1}{1 - R^2_{X_j(\mathbf{X}_{-j})}}, \quad (8)$$

where $\mathbf{X}_{-j}$ is the vector of all the inputs except $X_j$. $R^2_{X_j(\mathbf{X}_{-j})}$ then represents the $R^2$ from the regression of the input $X_j$ on the remaining inputs. The smallest value of VIF is 1 and indicates the absence of collinearity, while a value above 5 if often suspicious and might indicate multicollinearity. For large values of VIF, removing one of the incriminated variable would eliminate redundancy in the model and avoid numerical issues in the regression stage.

Nonlinear statistical dependence is far more complex to detect [19]. General nonlinear measures exist to detect bivariate statistical dependence, as the use of the Hoeffdings'D [50] or the HSIC criteria [51]. Statistical tests based on copulas can also be used to detect multivariate dependencies [52].

### B. IM for dependent inputs in the linear context

When the inputs are dependent, it is difficult to understand what is the root cause for this correlation. For example, the coefficients of the linear regression model Eq. (2) are no longer interpretable because the $\text{SRC}^2$ do not sum to one anymore (see, e.g., [26], [53]).

In the linear regression analysis literature, other measures have been developed, most often by finding ways to partition

the $R^2$ among the $d$ inputs [26], [27]. A particularly interesting IM is the so-called LMG (for "Lindeman-Merenda-Gold", see [28], [54]) which uses sequential sums of squares from the linear model and obtains an overall measure by averaging over all orderings of inputs. Mathematically, let $u$ be a subset of indices in the set of all subsets of $\{1,\dots,d\}$ and $\mathbf{X}_u = (X_j : j \in u)$ a group of inputs. The underlying idea is to measure the elementary contribution of any given variable $X_j$ to a given subset model $Y(\mathbf{X}_u)$ by the increase in $R^2$ that results from adding that predictive variable to the regression model:

$$\text{LMG}_j = \frac{1}{d!} \sum_{\substack{\pi \in \text{permutations} \\ \text{of } \{1,\dots,d\}}} r^2_{Y,(X_j|\mathbf{X}_\pi)} \ , \tag{9}$$

where $r^2_{Y,(X_j|\mathbf{X}_\pi)} = R^2_{Y(\mathbf{X}_{v\cup\{j\}})} - R^2_{Y(\mathbf{X}_v)}$ with $v$ the indices entered before $j$ in the order $\pi$. By recombinations (and recalling that $\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$), we have:

$$\begin{aligned}
\text{LMG}_j &= \frac{1}{d} \sum_{i=0}^{d-1} \sum_{\substack{u \subseteq -\{j\} \\ |u|=i}} \binom{d-1}{i}^{-1} r^2_{Y,(X_j|\mathbf{X}_u)} && (10)\\
&= \frac{1}{d} \sum_{u \subseteq -\{j\}} \binom{d-1}{|u|}^{-1} r^2_{Y,(X_j|\mathbf{X}_u)} && (11)\\
&= \frac{1}{d!} \sum_{u \subseteq -\{j\}} (d-1-|u|)!\,|u|!\, r^2_{Y,(X_j|\mathbf{X}_u)}. && (12)
\end{aligned}$$

In Eqs. (11) and (12) (resp. (9)), this averaging process over all combinations (resp. permutations) is carried out in the absence of order between the inputs. Such IM has been underused in practice due to the large amount of required calculations for moderate size $d$. Indeed, the number of linear regressions that is required in computing the summands in Eq. (11) dramatically increases with $d$.

A weighted analog of LMG, called "PMVD" for *proportional marginal variance decomposition* has been introduced by [55] and [53]. It is also described in details in [27]. PMVD aims to favor orderings for which the early inputs have a large contribution (see also [20]):

$$\begin{aligned}
\text{PMVD}_j &= \sum_{\substack{\pi \in \text{permutations} \\ \text{of } \{1,\dots,d\}}} \frac{L(\pi)}{\sum_\pi L(\pi)} r^2_{Y,(X_j|\mathbf{X}_\pi)} \ , && (13)\\
L(\pi) &= \prod_{i=1}^{d-1} \left[ r^2_{Y,(\mathbf{X}_{\pi_{i+1},\dots,\pi_d}|\mathbf{X}_{\pi_1,\dots,\pi_i})} \right]^{-1} . && (14)
\end{aligned}$$

One of the interest of PMVD is that it forces the IM to be zero if $\beta_j$ is zero.

### C. IM for dependent inputs considering RF model and similar tree ensemble based model

In the context of RF algorithm, MDA is a standard widely feature IM. However, it suffers some bias when the inputs are dependent. To avoid this bias, some authors proposed to introduce "conditionality" [56] in the MDA estimation. To illustrate this, the *conditional-MDA* is presented in Algorithm 2. Note that the dependence of the criteria to the correlation computation can be an issue since the Pearson's correlation coefficient only measures linear relationship.

**Algorithm 2** Conditional MDA.

1) For each tree:
   a) Compute the OOB-prediction accuracy before the permutation;
   b) Determine a certain number of variables $Z$ to be conditioned on. It is suggested to include only variables whose correlation with the variable of interest $X_j$ exceeds a given threshold (e.g., 0.2);
   c) For all variables $Z$ to be conditioned on:
      i) Take the cutpoints that split this variable in the current tree;
      ii) Create a grid by bisecting the sample space at each cutpoint;
      iii) Within this grid, permute the values of $X_j$ and compute the OOB-prediction accuracy after permutation.
   d) Compute the difference between the prediction accuracy before and after the permutation for one tree;
2) Compute the average of the difference over all the tree to obtain the importance of $X_j$ for the forest.

Recently, using a SA formalism, [14], [57] proposed the first convergence result for Breiman's MDA without any simplification. The study of existing implementations of RF shows that there are several definitions of MDA. These versions do not converge to the same theoretical quantity, and therefore lead to different IM. It has also been demonstrated that these different MDA can be decomposed as the sum of Sobol' indices and a third unknown term. This last term does not correspond to an IM, and strongly biases the MDA (as empirically observed) when the inputs are dependent. Therefore, [14], [57] introduced the *Sobol-MDA* metric, a new IM for RF. The general principle is to project the partitions of the trees according to the input of interest in order to eliminate it from the prediction mechanism. This principle makes it possible to define the Sobol-MDA in a consistent way with respect to the total Sobol' index, which gives the proportion of variance lost when the variable considered is removed from the model. This IM is, in particular, very efficient for the features identification setting.

### D. Model-agnostic IM for dependent inputs

In the general context, IM adapted to dependent inputs have been extensively studied in SA (see, e.g., [7]). An interesting (but practically complex) solution is based on the definition of four Sobol' indices [58]: the first two ones are the "independent" first-order and total Sobol' indices (which measure the effects of an input that is not due to its dependence with other inputs); the two others are the "full" first-order and

total Sobol' indices (which measure the effects of an input including the effects due to its dependence with other inputs).

Another solution is based on the use of the *Shapley value* concept coming from cooperative game theory [59] where it consists in fairly distributing both gains and costs to several actors working in coalition. The Shapley value applies primarily in situations when the contributions of each actor are unequal, ensuring that each actor gains as much (or more) as they would have from acting independently. Now, one assumes the actors are identified with a set of inputs and the value assigned to each coalition is identified to the explanatory power of the subset of model inputs composing the coalition. Then, these Shapley values can be interpreted as IM of model inputs. In the linear context, Shapley values (using a $R^2$-based cost function) turn out to be the LMG index introduced previously in Eq. (11). In the SA community, the so-called *Shapley effects* have been proposed by [60], considering the Sobol' indices as the cost function. These Shapley effects are valid IM for any finite-variance model. Similarly to the LMG (Eq. (12)), the Shapley effect writes:

$$Sh_j = \sum_{u \subseteq -\{j\}} \frac{(d-1-|u|)!\,|u|!}{d!} \left[c(u \cup \{j\} - c(u)\right]. \quad (15)$$

where $c(u) = \text{Var}\left(\mathbb{E}\left[Y|\mathbf{X}_u\right]\right)/\text{Var}(Y)$ corresponds to the so-called "closed Sobol' index".

For the ranking setting, both Shapley effects [60] and "SHAP" (*SHapley Additive exPlanations*) metric [61] are based on the Shapley value concept and have been developed for global SA and ML interpretability, respectively. In SA, the main advantage of Shapley effects over Sobol' indices lies on the correct treatment of the dependent inputs' case. For the same reason, SHAP is intensively used in ML interpretability but remains a local metric (contrarily to the Shapley effects which are global) and explains each individual prediction.

Using the Shapley formulation, several authors have proposed some alternative IM. [21] introduced "SAGE" (*Shapley Additive Global Importance*), to propose a general view of global feature importance for different types of learning models and loss functions. Shapley effects correspond to a notable specific case of SAGE. [62] proposed a computationally efficient procedure for estimating the "SPVIM" (*Shapley Population Variable Importance Measure*) metric, which is actually a Shapley effect with any value function.

Let us also mention that [14], [63] recently introduced "SHAFF" (*SHApley eFfects via random Forests*), a fast and accurate estimator of Shapley effects, based on the use of a RF algorithm. The estimation of Shapley effects induces two major difficulties (see [7] for a review of different techniques): first, the algorithmic complexity is exponential with respect to the dimension of the inputs; second, it is necessary to be able to efficiently estimate the expectation of the output conditional on a subset of inputs. Because of these two difficulties, existing algorithms for estimating Shapley effects are either computationally heavy or biased when the inputs are dependent. By generalizing the principle of Sobol-MDA (see Section IV-C), SHAFF solves these problems using importance sampling and projected random forests. This approach allows an accurate and fast estimation of conditional expectations, and thus significantly improves the accuracy of the estimation of Shapley effects.

It has also been recently observed that Shapley effects exhibit a particular undesirable feature called the "Shapley's joke": an exogenous input (i.e., not included explicitly in the model) can be associated to a strictly positive index if it is correlated to another endogenous input [64]. More generally, Shapley effects tend to perform an equitable share of the dependence effect between inputs. However, one could argue that having a better discrimination between inputs (e.g., for the features identification setting) might be of better interest. Therefore, another game-theoretic allocation rule, called the "*proportional value*", has been investigated in [65], [66]. A first contribution consists in an extension of this solution concept to the case of null players' game, i.e., games involving players for which the cost function is zero, which is of particular interest considering the Shapley's joke issue. A second contribution consists in providing a set of new IM, called the *proportional marginal effects* (PME). PME indices allow to avoid the Shapley's joke and tend to have much power of discrimination between inputs than Shapley effects.

## V. Curse of dimensionality

In addition to the problem of correlated features, dealing with a high dimensional input vector can be a challenge too. Indeed, building a relevant ML model while avoiding any overfitting requires to find the set of the most influential features. In other words, the idea is to find among a possible large number of features, the *effective dimension* [67] which is the dimension of the most important factors carrying the most information in the model. Such a notion which is a cornerstone of the global SA practice, echoes perfectly the definition of an IM and seems appropriate for the current context of interpretability.

Formally defining what is a"large dimension" $d$ seems difficult. One could first give an order of magnitude in terms of number of predictors (typically, when $d \approx 100$). However, such a definition might not be sufficient. One could also argue that "large" should be defined by combining both $d$ and $n$. Considering our applications of interest, one will consider that the problem becomes "high-dimensional" when $d$ is above $100$ and that algorithms such as RF do not perform well. Then, one would need to apply a more complex ML model (e.g., an autoencoder) or to use a dimension reduction technique (e.g., principal component analysis). At this stage and prior to that, using IM which enable to capture and analyze the most influential predictors can be of high interest. Moreover, taking advantage of the structure of the ML model can help to derive relevant IM in this context.

### A. IM for high-dimensional problems in the linear regression context

An alternative to the LMG indices has been proposed to deal with the high-dimensional context, where computation of the indices becomes intractable due to the combinatorial

complexity. The idea is to first perform a singular value decomposition (SVD) of the vector of inputs, in order to transform the correlated inputs into uncorrelated variables. Then, an adequate reweighing process, using the SRC of different linear regressions, leads to the formulation of the so-called "*relative weight measures*" [25], that has been called later "*Johnson indices*" or "*relative weights analysis*" (RWA). These IM are known to be adapted to large input dimension as well as providing similar results to those obtained via LMG indices (see, e.g., [26], [68]), at a highly reduced computational cost.

### *B. Model-agnostic methods adapted to high-dimensional problems*

Several model-agnostic metrics exist and can be used in order to identify influential features in a large dimension context. Among several methods, one can mention information-theoretic methods mostly related to the Shannon entropy (e.g., mutual information, variation of information, squared-loss mutual information), dissimilarity measures (e.g., f-divergences) and dependence measures (e.g., distance correlation). As developed in [69], all these metrics have theoretical connections. Note that, most of these metrics go beyond the variance as they focus more on the entire output distribution.

A famous screening tool, which goes beyond variance-based IM, is the *Hilbert-Schmidt Independence Criterion* (HSIC). Initially introduced by [51], this measure is built upon kernel-based approaches for detecting dependence, and more particularly on cross-covariance operators in reproducing kernel Hilbert spaces (RKHS). HSIC can be seen as a generalized notion of covariance between two random variables and thus makes it possible to capture a very broad spectrum of forms of dependency between variables. For this reason, [69] and then [70] popularized this measure for global SA purposes. Without going too much into details, one considers nonlinear transformations (called "features") of the input $X_j$ and the output $Y$ that lie into two RKHS (here, $\mathcal{X}_j$ and $\mathcal{Y}$) equipped with their characteristic kernels $\kappa_j(\cdot,\cdot)$ and $\kappa(\cdot,\cdot)$ which define inner products, respectively. Then, the HSIC index can be defined as follows:

$$\mathrm{HSIC}(X_j,Y)_{\mathcal{X}_j,\mathcal{Y}} = \mathbb{E}\left[\kappa_j(X_j,X_j^{'})\kappa(Y,Y^{'})\right] \tag{16}$$

$$+\,\mathbb{E}\left[\kappa_j(X_j,X_j^{'})\right]\mathbb{E}\left[\kappa(Y,Y^{'})\right] \tag{17}$$

$$-\,2\mathbb{E}\left[\mathbb{E}[\kappa_j(X_j,X_j^{'})|X_j]\;\mathbb{E}[\kappa(Y,Y^{'})|Y]\right]$$

where $(X_j^{'},Y^{'})$ is an i.i.d. copy of $(X_j,Y)$.

A desirable property of HSIC is that it equals zero if and only if $Y$ and $X_j$ are independent. Moreover, HSIC indices offer the advantage of having a low estimation cost (in practice, a few hundred samples vs. several tens of thousands for the Sobol' indices) and their estimation is independent from the input dimension $d$. In addition, statistical independence tests can be built based on HSIC estimates in order to ensure, despite the finite sample estimation, the significance of the screening. As recently shown, these tests can be very efficient for screening purposes [70]–[72]. Several HSIC-based statistical tests are available: asymptotic versions (i.e., for large sample size), spectral tests [73], permutation-based versions [71] for non-asymptotic case (i.e., case of a small sample size) and even adaptive versions [74]. Finally, one often use a normalized version of the HSIC index, called "R2-HSIC" and defined as follows:

$$\widehat{R^2_{\mathrm{HSIC},j}} = \frac{\widehat{\mathrm{HSIC}}(X_j,Y)}{\sqrt{\widehat{\mathrm{HSIC}}(X_j,X_j)\;\widehat{\mathrm{HSIC}}(Y,Y)}}. \tag{18}$$

This quantity, together with p-values obtained from the tests, enable to perform screening and ranking of the most influential inputs.

## VI. Application example on a public dataset

In this section, various IM described previously are illustrated through the Boston housing dataset (BostonHousing2 contained in the R package mlbench), which comes from a Boston 1970 census. There are $n = 506$ observations, one output ($Y :=$ cmedv which means "median value of owner-occupied home") and $d = 12$ inputs (mostly economical, geographical and social factors).

Firstly, we apply several ML models in order to test the various configurations and estimate the previous IM:

- a linear regression between the output and the inputs gives $\widehat{R}^2_{\text{Lin}} = 0.739$;
- a RF model gives $\widehat{R}^2_{\text{RF}} = 0.893$;
- for conditional-MDA estimation, one computes the RF implementation proposed by [56] where base learners are some conditional inference trees, with a $\widehat{R}^2_{\text{C-MDA}} = 0.883$.

As a result, one can see that the two trees ensemble approaches show a nice improvement in predictivity capabilities.

Numerical results are given in Table II. Ten metrics are estimated, namely: VIF, SRC$^2$, LMG, RWA, PMVD, SHAFF, RF-MDA, CF-MDA, CF-C-MDA and R2-HSIC. Note that, "RF-MDA" corresponds to the MDA estimated from the standard RF algorithm, "CF-MDA" to the MDA for the conditional forest and "CF-C-MDA" to the conditional MDA for the conditional forest. As a remark, one can mention that the first metric (VIF) is provided in order to diagnose whether inputs are multicollinear or not. The next four columns are related to variance-based IM for a linear model, while the four next ones are related to variance-based IM for RF-based algorithms. Finally, the last one is devoted to R2-HSIC indices.

A first primary analysis enables to emphasize three core results:

- firstly, VIF indices indicate that several inputs are highly correlated, as a possible significant collinearity between rad ($\widehat{\mathrm{VIF}} = 7.40$) and tax ($\widehat{\mathrm{VIF}} = 8.88$);
- secondly, the metrics unanimously identify the same couple of most influential inputs: lstat (percentage of lower status of the population) and rm (average number of rooms per dwelling);
- thirdly, one can see how some metrics sum up to the $R^2$ which is a desirable property for an IM.

| Variable | Name | VIF | $SRC^2$ (%) | LMG (%) | RWA (%) | PMVD (%) | SHAFF (%) | RF-MDA | CF-MDA | CF-C-MDA | R2-HSIC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_1$ | crim | 1.79 | 1.09 | 2.79 | 3.29 | 0.72 | 3.89 | 7.5 | 3.2 | 0.24 | 0.28 |
| $X_2$ | zn | 2.30 | 1.51 | 2.50 | 2.81 | 0.67 | 2.34 | 0.4 | 0.4 | 0.0 | 0.16 |
| $X_3$ | indus | 3.95 | 0.10 | 3.74 | 3.66 | 0.06 | 3.17 | 6.7 | 5.1 | 0.1 | 0.29 |
| $X_4$ | nox | 4.39 | 4.79 | 3.31 | 3.68 | 1.54 | 6.16 | 10.1 | 3.5 | 0.3 | 0.29 |
| $X_5$ | rm | 1.93 | **8.59** | **19.01** | **20.59** | **22.71** | **22.18** | **37.1** | **30.2** | **5.1** | **0.45** |
| $X_6$ | age | 3.09 | 0.01 | 2.20 | 2.70 | 0.00 | 4.31 | 3.4 | 1.7 | 0.0 | 0.25 |
| $X_7$ | dis | 3.95 | **12.02** | 3.17 | 1.86 | 2.18 | 0.51 | 7.4 | 1.8 | 0.1 | 0.16 |
| $X_8$ | rad | **7.40** | **9.56** | 2.46 | 2.10 | 0.83 | 3.87 | 1.4 | 0.7 | 0.0 | 0.21 |
| $X_9$ | tax | **8.88** | 6.73 | 3.87 | 3.64 | 1.07 | 6.35 | 3.7 | 3.6 | 0.1 | 0.27 |
| $X_{10}$ | ptratio | 1.78 | 5.15 | 7.93 | 8.70 | 6.48 | 4.64 | 5.0 | 5.9 | 0.2 | 0.23 |
| $X_{11}$ | b | 1.34 | 0.92 | 2.37 | 2.97 | 1.12 | 4.70 | 1.5 | 0.8 | 0.1 | 0.14 |
| $X_{12}$ | lstat | 2.93 | **17.64** | **20.59** | **17.92** | **36.56** | **26.59** | **61.8** | **48.3** | **5.1** | **0.56** |
| Sum | - | - | 68.10 | 73.93 | 73.93 | 73.93 | 88.71 | - | - | - | - |

Table II
RESULTS FROM THE BOSTON HOUSING DATASET STUDY ($Y :=$ CMEDV).

Indeed, one can see that $SRC^2$ estimates do not sum up to $\widehat{R}^2_{\text{Lin}}$ since collinearity affects both rad and tax. Moreover, $SRC^2$ indicates as a second most influential input the dis variable. Thus, one can easily see that $SRC^2$ are not able to provide interpretable IM as soon as correlated inputs are involved. As for LMG, RWA and PMVD, one can see a coherence between the estimates for rm and lstat. However, RWA leads to a switch in the ordering (which can be due to the sample size). While LMG and RWA provide rather close values (as expected), one can notice that PMVD discriminates more that these two IM (e.g., by putting values close to zero for several inputs).

In the context of a RF model, one can see that SHAFF estimates sum up to $\widehat{R}^2_{\text{RF}}$. Again, lstat and rm are well identified. The same conclusion can be drawn from RF-MDA and CF-MDA. However, CF-C-MDA gives a similar importance to these two inputs. One can see that these metrics do not sum up, neither to $\widehat{R}^2_{\text{RF}}$ nor to $\widehat{R}^2_{\text{C-MDA}}$. Moreover, CF-C-MDA estimates others features' importance to be almost zero. This is caused by the correlation between features: standard MDA artificially gives importance to the majority of features.

Finally, the last metric is R2-HSIC which still indicates the same ranking (lstat and rm). However, even if the numerical values of R2-HSIC are normalized in $[0, 1]$, the estimates are not widely spread. For this reason, one should rely on the p-values associated to independence statistical tests (not shown here, for the sake of clarity).

## VII. CONCLUSION

In this paper, several approaches adapted to ML interpretability have been described by distinguishing the linear and nonlinear ML model cases, the independent and dependent features' cases and the low dimensional and high dimensional cases. Some links between SA and ML interpretability have been illustrated and emphasized. Moreover, some of these metrics can be used for various settings: typically, measuring the features' importance (ranking), or feature identification (screening). The illustration shown is based on a public dataset; similar conclusions have been obtained on an EDF dataset related to the concentration of fission products released in the primary circuit's water of nuclear reactors.

Variance-based IM have been particularly scrutinized and Fig. 4 gives a synthesis of the various IM discussed in this paper. In the case of independent inputs, a first approach consists in testing a linear regression model. If this framework is valid, SRC (or $SRC^2$) indices give an immediate interpretation. If the linear model is not valid, more general IM, such as MDA and Sobol' indices, can be computed, either by metamodel-based techniques (e.g., in the small-size sample case), or using RF and permutation-based techniques (e.g., in the large-size sample case). In the case of dependent inputs, as in the previous one, linear-regression based indices exist (namely, LMG and PMVD) but are more costly to compute. If the model is nonlinear, the Shapley effects allocate a share of variance to each input and can also be estimated by various methods. However, it is relevant to note that the PMVD remains particularly interesting as it respects the exclusion principle (i.e., putting a null IM to each input with a non causal relationship with the output but which might be correlated with another causal input). Such an interesting property is not fulfilled neither by LMG nor by Shapley effects. As shown in Fig. 4, a nonlinear IM respecting the exclusion property, as the PME, is needed [65], [66].

To finish with, Fig. 4 provides other well-known screening techniques which have not been discussed in this paper, for the sake of conciseness (e.g., Lasso regression, Elastic net and the recent HSIC-Shapley effects [75]).


## ACKNOWLEDGMENT

We warmly thank Sébastien Da Veiga for his advice on the classification grid and Nicolas Bousquet for valuable discussions about interpretability in machine learning.


## REFERENCES


[1] A. Bibal, M. Lognoul, A. De Streel, and B. Frénay, "Legal requirements on explainability in machine learning," *Artificial Intelligence and Law*, pp. 149–169, 2021.

[2] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

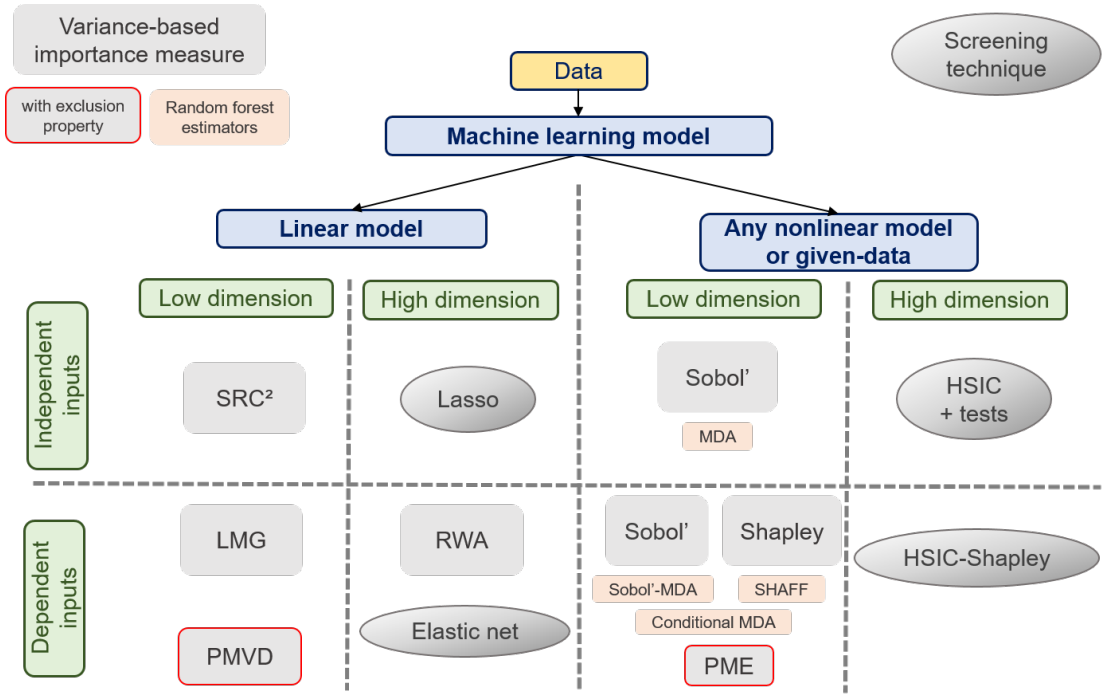


Figure 4. Classification grid of different machine learning importance measures and screening techniques.


[3] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, and A. Holzinger, "Explainable artificial intelligence: Concepts, applications, research challenges and visions," in *Lecture Notes in Computer Science*, A. Holzinger, P. Kieseberg, A. Tjo, and E. Weippl, Eds., vol. 12279. Springer, Cham, 2020.

[4] R. Hawkins, C. Paterson, C. Picardi, Y. Jia, R. Calinescu, and I. Habli, *Guidance on the assurance of machine learning in autonomous systems (AMLAS)*, Assuring Autonomy International Programme (AAIP), University of York, 2021.

[5] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning - A brief history, state-of-the-art and challenges," in *PKDD/ECML Workshops 2020*, 2020, pp. 417–431.

[6] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Salsana, and S. Tarantola, *Global sensitivity analysis - The primer*. Wiley, 2008.

[7] S. Da Veiga, F. Gamboa, B. Iooss, and C. Prieur, *Basics and Trends in Sensitivity Analysis. Theory and Practice in R*. SIAM, 2021.

[8] S. Razavi, A. Jakeman, A. Saltelli, C. Prieur, B. Iooss, E. Borgonovo, E. Plischke, S. Lo Piano, T. Iwanaga, W. Becker, S. Tarantola, J. Guillaume, J. Jakeman, H. Gupta, N. Melillo, G. Rabiti, V. Chabridon, Q. Duan, X. Sun, S. Smith, R. Sheikholeslami, N. Hosseini, M. Asadzadeh, A. Puy, S. Kucherenko, and H. Maier, "The future of sensitivity analysis: An essential discipline for systems modelling and policy making," *Environmental Modelling and Software*, vol. 137, no. 104954, 2020.

[9] B. Iooss, R. Kennet, and P. Secchi, "Different views of interpretability," in *Interpretability for Industry 4.0: Statistical and Machine Learning Approaches*, A. Lepore, B. Palumbo, and J.-M. Poggi, Eds. Springer, 2022, in press.

[10] C. Molnar, *Interpretable machine learning: A guide for making black-box models explainable (2nd ed.)*. github, 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[11] B. Iooss, "Revue sur l'analyse de sensibilité globale de modèles numériques," *Journal de la Société Française de Statistique*, vol. 152, pp. 1–23, 2011.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, 2nd ed. Springer, 2009.

[13] A. Pasanisi and A. Dutfoy, "An industrial viewpoint on uncertainty quantification in simulation: Stakes, methods, tools, examples," in *Uncertainty quantification in scientific computing - 10th IFIP WG 2.5 working conference, WoCoUQ 2011, Boulder, CO, USA, August 1-4, 2011*, ser. IFIP Advances in Information and Communication Technology, A. Dienstfrey and R. Boisvert, Eds., vol. 377. Berlin: Springer, 2012, pp. 27–45.

[14] C. Bénard, "Random forests and interpretability of learning algorithms," Ph.D. dissertation, Sorbonne University, France, 2021.

[15] K. Zhao and T. Hastie, "Causal interpretations of black-box models," *Journal of Business & Economic Statistics*, vol. 39, no. 1, pp. 272–281, 2021.

[16] D. Kurowicka and R. Cooke, *Uncertainty analysis with high dimensional dependence modelling*. Wiley, 2006.

[17] J. Fan and J. Lv, "A selective overview of variable selection in high dimensional feature space," *Statistica Sinica*, vol. 20, no. 1, pp. 101–148, 2010.

[18] C. Yenigün and M. Rizzo, "Variable selection in regression using maximal correlation and distance correlation," *Journal of Statistical Computation and Simulation*, vol. 85, no. 8, pp. 1692–1705, 2015.

[19] D. Tjøstheim, H. Otneim, and B. Støve, "Statistical dependence: Beyond Pearson's $\rho$," *Statistical Science*, vol. 37, no. 1, pp. 90–109, 2022.

[20] U. Grömping, "Variable importance in regression models," *WIREs Comput Stat*, vol. 7, no. 137-152, 2015.

[21] I. Covert, S. Lundberg, and S.-I. Lee, "Understanding global feature contributions with additive importance measures," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[22] F. Bachoc, F. Gamboa, M. Halford, J.-M. Loubes, and L. Risser, "Explaining machine learning models using entropic variable projection," *arXiv:1810.07924*, 2020.

[23] E. Borgonovo, E. Plischke, and G. Rabitti, "Interactions and computer experiments," *Scandinavian Journal of Statistics*, no. 10.1111/sjos.12560, 2022.

[24] C. M. Bishop, *Pattern Recognition and Machine Learning*, ser. Information Science and Statistics. Springer Science + Business Media, 2006.

[25] J. Johnson, "A heuristic method for estimating the relative weight of predictor variables in multiple regression," *Multivariate Behavioral Research*, vol. 35, pp. 1–19, 2000.

[26] J. Johnson and J. LeBreton, "History and use of relative importance indices in organizational research," *Organizational Research Methods*, vol. 7, pp. 238–257, 2004.

[27] U. Grömping, "Estimators of relative importance in linear regression based on variance decomposition," *The American Statistician*, vol. 61, no. 2, 2007.

[28] ——, "Relative importance for linear regression in R: the Package relaimpo," *Journal of Statistical Software*, vol. 17, pp. 1–27, 2006.

[29] A. Antoniadis, S. Lambert-Lacroix, and J.-M. Poggi, "Random forests for global sensitivity analysis: A selective review," *Reliability Engineering & System Safety*, vol. 206, no. 107312, 2021.

[30] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[31] ——, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.

[32] G. Louppe, "Understanding random forests: From theory to practice," *arXiv:1407.7502 [stat]*, Jun. 2015, arXiv: 1407.7502. [Online]. Available: http://arxiv.org/abs/1407.7502

[33] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, p. 25, 2007.

[34] A. Fisher, C. Rudin, and F. Dominici, "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously," *arXiv:1801.01489 [stat]*, Dec. 2019, arXiv: 1801.01489.

[35] B. Gregorutti, "Forêts aléatoires et sélection de variables : analyse des données des enregistreurs de vol pour la sécurité aérienne," Ph.D. dissertation, Université Paris VI, France, 2015.

[36] P. Wei, Z. Lu, and J. Song, "Variable importance analysis: a comprehensive review," *Reliability Engineering & System Safety*, vol. 142, pp. 399–432, 2015.

[37] T. Homma and A. Saltelli, "Importance measures in global sensitivity analysis of non linear models," *Reliability Engineering & System Safety*, vol. 52, pp. 1–17, 1996.

[38] B. Iooss and P. Lemaître, "A review on global sensitivity analysis methods," in *Uncertainty management in Simulation-Optimization of Complex Systems: Algorithms and Applications*, C. Meloni and G. Dellino, Eds. Springer, 2015, pp. 101–122.

[39] I. Sobol', "Sensitivity estimates for non linear mathematical models," *Mathematical Modelling and Computational Experiments*, vol. 1, pp. 407–414, 1993.

[40] B. Li and C. Chen, "First-order sensitivity analysis for hidden neuron selection in layer-wise training of networks," *Neural Processing Letters*, vol. 48, pp. 1105–1121, 2018.

[41] T. Fel, R. Cadene, M. Chalvidal, M. Cord, D. Vigouroux, and T. Serre, "Look at the variance! Efficient black-box explanations with Sobol-based sensitivity analysis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[42] A. Owen and C. Hoyt, "Efficient estimation of the ANOVA mean dimension, with an application to neural net classification," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 9, pp. 708–730, 2021.

[43] D. W. Marquardt, "Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation," *Technometrics*, vol. 12, no. 3, pp. 591–612, 1970.

[44] D. Belsley, E. Kuh, and R. Welsch, *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons, Inc, 1980.

[45] AVSI – AFE 87 Project, "Final Report AFE 87 – Machine Learning," Aerospace Vehicle Systems Institute (AVSI), Tech. Rep., 2020, report No. 87-REP-01.

[46] DEEL Certification Workgoup, "Machine Learning in Certified Systems," DEpendable & Explainable Learning (DEEL), IRT Saint Exupéry, Tech. Rep., 2020, report No. S079L03T00-005.

[47] C. Molnar, G. König, B. Bischl, and G. Casalicchio, "Model-agnostic Feature Importance and Effects with Dependent Features – A Conditional Subgroup Approach," *ArXiv e-prints*, pp. 1–20, 2020.

[48] F. Öztürk and F. Akdeniz, "Ill-conditioning and multicollinearity," *Linear Algebra and its Applications*, vol. 321, pp. 295–305, 2000.

[49] J. Fox and G. Monette, "Generalized Collinearity Diagnostics," *Journal of the American Statistical Association*, vol. 87, no. 417, pp. 178–183, 1992.

[50] W. Hoeffding, "A class of statistics with asymptotically normal distributions," *Annals of Mathematical Statistics*, vol. 19, pp. 293–325, 1948.

[51] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring statistical dependence with Hilbert-Schmidt norms," in *Proceedings Algorithmic Learning Theory*. Springer-Verlag, 2005, pp. 63–77.

[52] C. Genest and B. Rémillard, "Tests of independence and randomness based on the empirical copula process," *TEST*, vol. 13, pp. 335–369, 2004.

[53] B. Feldman, "Relative importance and value," *SSRN Electronic Journal*, 03 2005.

[54] R. H. Lindeman, P. F. Merenda, and R. Z. Gold, *Introduction to bivariate and multivariate analysis*. Glenview, IL: Scott Foresman and Company, 1980.

[55] B. Feldman, "The proportional value of a cooperative game," in *Econometric Society World Congress 2000 Contributed papers*, no. 1140. Econometric Society, 2000.

[56] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC bioinformatics*, vol. 9, p. 307, 08 2008.

[57] C. Bénard, S. D. Veiga, and E. Scornet, "MDA for random forests: inconsistency, and a practical solution via the Sobol-MDA," *Biometrika, in press*, 2022.

[58] T. Mara, S. Tarantola, and P. Annoni, "Non-parametric methods for global sensitivity analysis of model output with dependent inputs," *Environmental Modeling & Software*, vol. 72, pp. 173–183, 2015.

[59] L. Shapley, "A value for n-persons game," in *Contributions to the theory of games II, Annals of mathematic studies*, H. Kuhn and A. Tucker, Eds. Princeton University Press, Princeton, NJ, 1953.

[60] A. Owen, "Sobol' indices and Shapley value," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 2, pp. 245–251, 2014.

[61] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[62] B. Williamson and J. Feng, "Efficient nonparametric statistical inference on population feature importance using shapley values," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 282–10 291.

[63] C. Bénard, G. Biau, S. D. Veiga, and E. Scornet, "SHAFF: Fast and consistent SHApley eFfect estimates via random Forests," in *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, Virtual, March 2022.

[64] B. Iooss and C. Prieur, "Shapley effects for sensitivity analysis with dependent inputs: comparisons with Sobol' indices, numerical estimation and applications," *International Journal for Uncertainty Quantification*, vol. 9, pp. 493–514,, 2019.

[65] M. Hérin, M. Il Idrissi, V. Chabridon, and B. Iooss, "Proportional marginal effects for sensitivity analysis with correlated inputs," in *Proceedings of the 10th International Conference on Sensitivity Analysis of Model Output (SAMO 2022)*, Tallahassee, Florida, USA, March 2022.

[66] ——, "Proportional marginal effects for sensitivity analysis with correlated inputs," *Preprint*, 2022.

[67] S. Kucherenko, B. Feil, N. Shah, and W. Mauntz, "The identification of model effective dimensions using global sensitivity analysis," *Reliability Engineering & System Safety*, vol. 96, pp. 440–449, 2011.

[68] L. Clouvel, P. Mosca, J. Martinez, and G. Delipei, "Shapley and Johnson values for sensitivity analysis of PWR power distribution in fast flux calculation," in *M&C 2019*, Portland, USA, August 2019.

[69] S. Da Veiga, "Global sensitivity analysis with dependence measures," *Journal of Statistical Computation and Simulation*, vol. 85, pp. 1283–1305, 2015.

[70] M. De Lozzo and A. Marrel, "New improvements in the use of dependence measures for sensitivity analysis and screening," *Journal of Statistical Computation and Simulation*, vol. 86, pp. 3038–3058, 2016.

[71] A. Meynaoui, M. Albert, B. Laurent, and A. Marrel, "Adaptive test of independence based on hsic measures," *Preprint*, 2019, URL http://export.arxiv.org/pdf/1902.06441.

[72] A. Marrel and V. Chabridon, "Statistical developments for target and conditional sensitivity analysis: Application on safety studies for nuclear reactor," *Reliability Engineering & System Safety*, vol. 214, p. 107711, 2021.

[73] Q. Zhang, S. Filippi, A. Gretton, and D. Sejdinovic, "Large-scale kernel methods for independence testing," *Statistics and Computing*, vol. 28, pp. 113–130, 2018.

[74] R. E. Amri and A. Marrel, "Optimized hsic-based tests for sensitivity analysis: Application to thermalhydraulic simulation of accidental scenario on nuclear reactor," *Quality and Reliability Engineering International*, vol. 38, no. 3, pp. 1386–1403, 2022.

[75] S. Da Veiga, "Kernel-based anova decomposition and shapley effects–application to global sensitivity analysis," *Preprint*, 2021, arXiv:2101.05487.