



Scrutinizing XAI using linear ground-truth data with suppressor variables

Rick Wilming¹ · Céline Budding² · Klaus-Robert Müller^{1,3,4,5} · Stefan Haufe^{1,6,7}

Received: 2 May 2021 / Revised: 30 November 2021 / Accepted: 13 March 2022 /
Published online: 13 April 2022
© The Author(s) 2022

Abstract

Machine learning (ML) is increasingly often used to inform high-stakes decisions. As complex ML models (e.g., deep neural networks) are often considered black boxes, a wealth of procedures has been developed to shed light on their inner workings and the ways in which their predictions come about, defining the field of ‘explainable AI’ (XAI). Saliency methods rank input features according to some measure of ‘importance’. Such methods are difficult to validate since a formal definition of feature importance is, thus far, lacking. It has been demonstrated that some saliency methods can highlight features that have no statistical association with the prediction target (suppressor variables). To avoid misinterpretations due to such behavior, we propose the actual presence of such an association as a necessary condition and objective preliminary definition for feature importance. We carefully crafted a ground-truth dataset in which all statistical dependencies are well-defined and linear, serving as a benchmark to study the problem of suppressor variables. We evaluate common explanation methods including LRP, DTD, PatternNet, PatternAttribution, LIME, Anchors, SHAP, and permutation-based methods with respect to our objective definition. We show that most of these methods are unable to distinguish important features from suppressors in this setting.

Keywords Explainable AI · Saliency methods · Ground truth · Benchmark · Linear classification · Suppressor variables

1 Introduction

With AlexNet (Krizhevsky et al. 2012) winning the ImageNet competition, the machine learning (ML) community started into a new era. Within few years, novel models achieved massive leaps in performance for challenging problems in computer vision, natural language processing, and reinforcement learning (e.g., Jaderberg et al. 2015; LeCun et al. 2015; Silver et al. 2017). In several real-world tasks, ML models became on par with

Editors: Annalisa Appice, Sergio Escalera, Jose A. Gamez, Heike Trautmann.

✉ Stefan Haufe
haufe@tu-berlin.de

Extended author information available on the last page of the article

human experts or achieved even super-human performance (Silver et al. 2017). Nowadays, there are increasing efforts to also leverage their predictive power in fields such as health-care and criminal justice, where they may support high-stake decisions that have a profound impact on human lives (Lapuschkin et al. 2019; Rudin 2019).

The complexity of current ML models makes it hard for humans to understand the ways in which their predictions come about. Especially in highly safety- or otherwise critical fields such as medicine, finance, or automatic driving, ethical and legal considerations have led to the demand that predictions of ML models should be ‘transparent’, establishing the field of ‘interpretable’ or ‘explainable’ AI (XAI, e.g., Dombrowski et al. 2022; Samek et al. 2019). Current XAI approaches can be categorized along various dimensions (Arrieta et al. 2020). Some methods provide ‘explanations’ for single input examples (*instance-based*), while others can be applied to entire models only (*global*). A common paradigm is to provide ‘importance’ or ‘relevance’ scores for single input features. Respective methods are called ‘saliency’ or ‘heat’ mapping approaches. Another distinction is made between model-agnostic methods (e.g. Štrumbelj and Kononenko 2014; Lundberg and Lee 2017; Ribeiro et al. 2016), which are based on a model’s output only, and methods that are tailored to a specific class of models (e.g., neural networks, Bach et al. 2015; Binder et al. 2016; Kim et al. 2018; Montavon et al. 2017; Montavon et al. 2018; Samek et al. 2021; Springenberg et al. 2015; Zeiler and Fergus 2014). Finally, linear ML models with sufficiently few features as well as shallow decision trees have been considered intrinsically ‘interpretable’ (Rudin 2019), a notion that has also been challenged (Haufe et al. 2014; Lipton 2018; Poursabzi-Sangdeh et al. 2021), and that is further scrutinized here.

It is understood that XAI methods can serve quality control purposes only under the provision of being trustworthy themselves. However, it is still under scientific debate what specific formal problems XAI is supposed to solve and what requirements respective methods should fulfill (Doshi-Velez and Kim 2017; Lipton 2018; Murdoch et al. 2019). Existing formulations of such requirements are often relatively vague and lack precise mathematical language. Terms like ‘explainable’ or ‘interpretable’ are used by many XAI authors without specifying *how* results of a given method should be interpreted, i.e., what exact formal conclusions can be deduced. Authors of XAI papers frequently suggest interpretations that are either not formally justified or not precise enough to be formally verified. LIME (Ribeiro et al. 2016), for example, includes the following example: “A model predicts that a patient has the flu, and LIME highlights the symptoms in the patient’s history that led to the prediction. Sneezing and headache are portrayed as contributing to the ‘flu’ prediction, while ‘no fatigue’ is evidence against it. With these, a doctor can make an informed decision about whether to trust the model’s prediction.” As we will discuss, nescience about the capabilities of XAI methods can lead to misinterpretations in practice.

Importantly, the lack of quantifiable formal criteria also currently prohibits the objective validation of XAI. Rather than using ground-truth data, current validation schemes are often either restricted to subjective qualitative assessments or use surrogate performance metrics such as the change in model output or performance when manipulating or omitting single features (e.g., Alvarez-Melis and Jaakkola 2018; Fong and Vedaldi 2017; Hooker et al. 2019; Samek et al. 2016). In this paper, we aim to make a first step towards an *objective* validation of saliency methods. To this end, we devise a purely data-driven criterion of feature importance, which defines the superset of features that any XAI method may reasonably identify. Based on this definition, we generate simple synthetic ground-truth data with linear structure, which we use to quantitatively benchmark a multitude of existing XAI approaches including LIME (Ribeiro et al. 2016), SHAP (Lundberg and Lee 2017), and LRP (Bach et al. 2015) with respect to their *explanation performance*.

2 Formalization of feature importance

Let us consider a supervised prediction task, where a model $f^\theta : \mathbb{R}^D \rightarrow \mathcal{Y}$ learns a mapping from a D -dimensional feature space to a label space \mathcal{Y} from a set of N *i.i.d.* training examples $\mathcal{D} = \{(\mathbf{x}^n, y^n)\}_{n=1}^N$, $\mathbf{x}^n \in \mathcal{F} \subseteq \mathbb{R}^D$, $y^n \in \mathcal{Y}$, $n \in \{1, \dots, N\}$. The \mathbf{x}^n and y^n are realizations of random variables \mathbf{X} and Y with joint probability density function $p_{\mathbf{X},Y}(\mathbf{x}, y)$. Saliency maps may either be obtained for entire ML models or on a single-instance basis. It is, thus, a function $\mathbf{s}(f^\theta, \mathbf{x}^*, \mathcal{D}) \in \mathbb{R}^D$ that depends on the model f^θ as well (optionally) the training data \mathcal{D} and/or an input example \mathbf{x}^* . The map \mathbf{s} is supposed to quantify the ‘importance’ of each feature $d \in \{1, \dots, D\}$ either for the prediction of the sample \mathbf{x}^* or for the predictions of the model f^θ in general, according to some criterion. Ideally, one would like to have a way of defining the ‘correct’ saliency map for a certain combination of model and data. Coming up with such a definition is, however, difficult. We, therefore, constrain ourselves to the simpler problem of partitioning the set of features into ‘important’ and ‘unimportant’ ones. Thus, we are looking for functions \mathbf{h} , where $\mathbf{h}(f^\theta, \mathbf{x}^*, \mathcal{D}) \in \{0, 1\}^D$. Here, $\mathcal{F}^+ := \{d : h_d(f^\theta, \mathbf{x}^*, \mathcal{D}) = 1\}$ is the set of ‘important’ features and $\mathcal{F}^- := \{d : h_d(f^\theta, \mathbf{x}^*, \mathcal{D}) = 0\}$ is the set of ‘unimportant’ features. For a given saliency map \mathbf{s} , a corresponding dichotomization function \mathbf{h} can be obtained by thresholding its output, for example based on a statistical hypothesis test.

2.1 Importance as influence on the model decision

The indicator function \mathbf{h} facilitates possible formalizations of ‘importance’. Most current saliency methods – implicitly or explicitly – usually seek to identify those features that significantly influence the decision of a model. For some models, the corresponding sets, $\mathcal{F}_{\text{model}}^+$ and $\mathcal{F}_{\text{model}}^-$, can indeed be defined in a straightforward manner. Examples are linear models, for which $\mathcal{F}_{\text{model}}^+$ can be defined as the set of features with non-zero (not significantly different from zero) model coefficients. For more complex models, such a direct definition is, however, more difficult, and we refrain from attempting a more precise formalization here.

2.2 Importance as statistical relation to the target

It is often implicitly assumed that XAI methods provide qualitative or even quantitative insight about statistical or mechanistic relationships between the input and output variables of a model (Binder et al. 2016; Ribeiro et al. 2016). In other words, it is asserted that \mathcal{F}^+ must contain only features that are at least in some way structurally or statistically related to the prediction target. As an example, a brain region that is highlighted by an XAI method as ‘important’ for predicting a neurological disease will typically be interpreted as a correlate or even causal drive of that disease and be discussed as such. Such interpretations are, however, invalid, as it is possible that features lacking any structural or statistical relationship to the prediction target do significantly reduce the model’s prediction error (thus, are in $\mathcal{F}_{\text{model}}^+$) (Haufe et al. 2014). Such features have been termed suppressor variables (Conger 1974; Friedman and Wall 2005).

Suppressor variables may contain side information, for example, on the correlation structure of the noise, that can be used by a model to predict better. But they themselves do not provide any satisfactory ‘explanation’ about the actual relationship between input and output variables (Haufe et al. 2014). Such features are prone to be misinterpreted, which could have

severe consequences in high-stakes domains. We, therefore, argue that a genuine statistical dependency between a feature and the response variable should be a prerequisite for that feature to be considered important. In other words, the set of important features identified by any saliency method should be a subset of a set $\mathcal{F}_{\text{dep}}^+$ that can be defined based on the data alone as

$$\mathcal{F}_{\text{dep}}^+ := \{d \mid X_d \not\perp Y\}, \quad (1)$$

where

$$X_d \not\perp Y \Leftrightarrow p(x_d, y) \neq p(x_d) \cdot p(y)$$

for some choice of x_d, y . The set of unimportant features is defined as the complement

$$\mathcal{F}_{\text{dep}}^- = \{d \mid X_d \perp Y\} = \{1, \dots, D\} - \mathcal{F}_{\text{dep}}^+.$$

Notably, a data-driven mathematical definition of feature importance such as ours also provides a recipe to generate ground-truth reference data with known sets of important features. This paves the way for an objective evaluation of XAI methods, which is the purpose of this paper.

3 Suppressor variables

To understand how suppressor variables can cause misinterpretations for existing saliency methods, consider the following linear generative model (c.f., Haufe et al. 2014) with two features $x_1, x_2 \in \mathbb{R}$ and a response (target) variable $y \in \mathbb{R}$:

$$\begin{aligned} x_1 &= a_1 \zeta + d_1 \rho \\ x_2 &= d_2 \rho \\ y &= \zeta. \end{aligned} \quad (2)$$

Here, ζ and $\rho \in \mathbb{R}$ are random variables called the *signal* and *distractor*, respectively, and a_1, b_1 and $b_2 \in \mathbb{R}$ are non-zero coefficients. The mixing weight vectors $\mathbf{a} = [a_1, 0]^\top$ and $\mathbf{b} = [b_1, b_2]^\top$ are called *signal* and *distractor patterns*, respectively (see Haufe et al. 2014; Kindermans et al. 2018). The learning task is to predict labels y from features $\mathbf{x} = [x_1, x_2]^\top$. This task is solvable using x_1 alone, since x_1 and y share the common signal ζ . However, the presence of the distractor in x_1 limits the achievable prediction accuracy. Since x_2 and y do not share a common term, no prediction of y above chance-level is possible using x_2 . However, a bivariate model using both features can eliminate the distractor so that the label can be perfectly recovered. Specifically, the linear model $f^{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ with demixing weight vector (also called *extraction filter*) $\mathbf{w} = [1/a_1, -d_1/(a_1 d_2)]^\top$ achieves that:

$$\mathbf{w}^\top \mathbf{x} = \zeta + \frac{d_1}{a_1} \rho - \frac{d_1}{a_1 d_2} d_2 \rho = \zeta = y. \quad (3)$$

According to the terminology introduced above, the set of features statistically related to the target is $\mathcal{F}_{\text{dep}}^+ = \{1\}$, while the set of ‘influential’ features is $\mathcal{F}_{\text{model}}^+ = \{1, 2\}$. The influence of x_2 on the prediction is certified by the non-zero coefficient $w_2 = -d_1/(a_1 d_2)$ of the optimal prediction model. Depending on the coefficients a_1, d_1 and d_2 , this influence can

have positive or negative polarity, and its strength can be smaller or bigger than $w_1 = 1/d_1$, provoking diverse conjectures about the nature of its influence. However, x_2 itself has no statistical relationship to y by construction. It is, therefore, a suppressor variable (Conger 1974; Friedman and Wall 2005; Haufe et al. 2014; Horst et al. 1941).

In a real problem setting, y could be a disease to be diagnosed and ζ could be a (perfect) physiological marker. The baseline level of that marker could, however, be different for the two sexes, encoded in the distractor ρ , even if the prevalence of the disease does not depend on sex. A bivariate ML model can subtract the sex-specific baseline and, thereby, diagnose the disease more accurately than a univariate model based on the measured marker alone. A clinician confronted with the influence of sex on the model decision may thus erroneously conclude that sex is a factor that also correlates with or even causally influences the presence of the disease. Such examples refute the widely accepted notion (Ribeiro et al. 2016; Rudin 2019) that linear models are easy to interpret (Haufe et al. 2014).

4 Methods

We generate synthetic images of two classes in the spirit of the linear example introduced in Sect. 3; thus with known sets $\mathcal{F}_{\text{dep}}^+$ of class-specific pixels. These are then used to train linear classifiers to discriminate the two classes. The resulting models are analyzed by a multitude of XAI approaches in order to obtain ‘saliency’ maps. The performance of these methods w.r.t. recovering (only) pixels in $\mathcal{F}_{\text{dep}}^+$ is then quantitatively assessed using appropriate metrics. The techniques used in these steps are described in the following.

4.1 Data generation

Following Haufe et al. (2014), we extend the two-dimensional example of a suppressor variable (2) and create synthetic data sets $\mathcal{D} = \{(\mathbf{x}^n, y^n)\}_{n=1}^N$ of *i.i.d.* observations ($\mathbf{x}^n \in \mathbb{R}^D, y^n \in \{-1, 1\}$) according to the generative model

$$\mathbf{x} = \lambda_1 \mathbf{a}y + \lambda_2 \mathbf{d}\rho + \lambda_3 \boldsymbol{\eta}, \quad (4)$$

with activation pattern $\mathbf{a} \in \mathbb{R}^D$ and distractor pattern $\mathbf{d} \in \mathbb{R}^D$. The signal is directly encoded using binary class labels $y \in \{-1, 1\} \sim \text{Bernoulli}(1/2)$. The distractor $\rho \in \mathbb{R} \sim \mathcal{N}(0, 1)$ is sampled from a standard normal distribution. The multivariate noise $\boldsymbol{\eta} \in \mathbb{R}^D \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ is modelled according to a D -dimensional multivariate Gaussian distribution with zero mean and random covariance matrix $\boldsymbol{\Sigma} = \mathbf{V}\mathbf{E}\mathbf{V}^\top$, where \mathbf{V} is uniformly sampled from the space of orthogonal matrices and $\mathbf{E} = \text{diag}(\mathbf{e})$ is a diagonal matrix of eigenvalues sampled as $\mathbf{e} = [e_1 + c, \dots, e_D + c]$, where $c := \max(\mathbf{e})/100$ and $e_d \sim \mathcal{U}(0, 1), d \in \{1, \dots, D\}$. The signal, distractor, and noise components $\mathbf{a}[y^1, \dots, y^N]$, $\mathbf{d}[\rho^1, \dots, \rho^N]$, and $[\boldsymbol{\eta}^1, \dots, \boldsymbol{\eta}^N] \in \mathbb{R}^{D \times N}$ are normalized by their respective Frobenius norms, e.g., $[\boldsymbol{\eta}^1, \dots, \boldsymbol{\eta}^N] \leftarrow [\boldsymbol{\eta}^1, \dots, \boldsymbol{\eta}^N] / \sqrt{\sum_{n=1}^N \sum_{d=1}^D |\eta_d^n|^2}$. The observed features are then obtained as a weighted sum of the three components, where the factors λ_i with $\sum_i^3 \lambda_i = 1$ adjust the influence of each component in the sum, defining the signal-to-noise ratio (SNR). For a given SNR factor λ_1 , we here set $\lambda_2 = \lambda_3 = (1 - \lambda_1)/2$.

Note that the activation pattern \mathbf{a} represents the ground truth feature importance map according to our definition (1). To be comprehensible by a human, it would be desirable for \mathbf{a} to have a simple structure (e.g. sparse, compact). As it is an intrinsic property of the data,

this cannot be ensured in practice, though. However, estimates of \mathbf{a} can be biased towards having simple structure.

4.2 Classifiers

The generative model (4) leads to Gaussian class-conditional distributions with equal covariance matrices for both classes. Thus, Bayes-optimal classification can be achieved using a linear discriminant function $f^{\mathbf{w}} : \mathbb{R}^D \rightarrow \mathbb{R}$ parametrized by a weight vector \mathbf{w} . We here use two different implementations of linear logistic regression (LLR). The first one is part of scikit-learn (Pedregosa et al. 2011), where we use the default parameters, no regularization, and no intercept. This implementation is employed in combination with model-agnostic XAI methods (Pattern, PFI, EMR, FIRM, SHAP, Pearson Correlation, Anchors and LIME, see below). The second implementation is a single-layer neural network (NN) with two output neurons and softmax activation function, which was built in Keras with a Tensorflow backend. The network is trained without regularization using the Adam optimizer. This implementation is employed for model-based XAI methods defined on neural networks only (Deep Taylor, LRP, PatternNet and PatternAttribution, see description below). Note that, while both implementations should in theory lead to the same discriminant function, slight discrepancies in their weight vectors are observed in practice (see supplementary Figure S8).

4.3 XAI methods

We assess the following model-agnostic and NN-based saliency methods. All methods can be used to generate global maps, while only some are capable of generating sample-based heat maps. All methods provide continuous-valued saliency maps \mathbf{s} , which are compared to the binary ground truth (encoded in the sets $\mathcal{F}_{\text{dep}}^+$ and $\mathcal{F}_{\text{dep}}^-$) using metrics from signal detection theory (see Sect. 4.4). Thus, we do not require any method to provide a dichotomization function \mathbf{h} .

Linear model weights (extraction filters)

For linear models $f^{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b$, the model weights \mathbf{w} are most commonly used for interpretation. Thus, the function

$$\mathbf{s}^{\text{filter}}(D) = \mathbf{w} \quad (5)$$

provides a global saliency map, which we call the *linear extraction filter*. Notably, saliency methods based on the gradient of the model output with respect to the input features (Bae-hrens et al. 2010; Simonyan et al. 2013) reduce to the extraction filter for linear prediction models (Kindermans et al. 2018).

As has been noted in Sect. 3 and, in more depth, in Haufe et al. (2014), extraction filters are prone to highlight suppressor variables. This also holds for sparse weight vectors, as the inclusion of suppressor variables in the model may be necessary to achieve optimal performance (Haufe et al. 2014).

Linear activation pattern

The set $\mathcal{F}_{\text{dep}}^+$ can be estimated from empirical data by testing the dependency between the target y and each feature x_d using a statistical test for general non-linear associations (e.g., Gretton et al. 2007). For the linear generative model studied here, it is sufficient to evaluate the sample covariance $\text{Cov}[x_d, y]$ between each input feature $x_d, d \in \{1, \dots, D\}$

and the target variable y . To obtain a model-specific output, we can further replace the target variable y by its model approximation $\mathbf{w}^\top \mathbf{x}$, leading to $s_d(\mathcal{D}) := \text{Cov}[x_d, \mathbf{w}^\top \mathbf{x}]$, for $d \in \{1, \dots, D\}$. The resulting global saliency map

$$\mathbf{s}^{\text{pattern}}(\mathcal{D}) = \mathbf{S}_x \mathbf{w}, \quad (6)$$

where $\mathbf{S}_x = \text{Cov}[\mathbf{x}, \mathbf{x}]$ is the sample data covariance matrix, called the *linear activation pattern* (Haufe et al. 2014). The linear activation pattern is a global saliency map for linear models that does not highlight spurious suppressor variables (Haufe et al. 2014). Note that $\mathbf{s}^{\text{pattern}}$ is a consistent estimator for the coefficients \mathbf{a} of the generative model in our suppressor variable example (2). In particular, $s_2^{\text{pattern}} \rightarrow 0$ for $N \rightarrow \infty$.

Pearson correlation Since the linear activation pattern corresponds to the covariance of each feature with either the target variable or model output, a natural idea is to replace it with Pearson's correlation

$$s_d^{\text{corr}}(\mathcal{D}) := \text{Corr}[x_d, \mathbf{w}^\top \mathbf{x}] = \frac{\text{Cov}[x_d, \mathbf{w}^\top \mathbf{x}]}{\sqrt{\text{Var}[x_d] \text{Var}[\mathbf{w}^\top \mathbf{x}]}} \quad (7)$$

in order to obtain a normalized measure of feature importance. However, due to the normalization terms in the denominator, this measure is more strongly affected by noise than the covariance-based activation pattern.

Permutation feature importance (PFI) and empirical model reliance (EMR)

The PFI approach was introduced by Breiman (2001) to assess the influence of features on the performance of random forests. The idea is to shuffle the values of one feature of interest, keep the remaining features fix, and to observe the effect on the misclassification rate. In a same style, Fisher et al. (2019) introduced the notion of model reliance, a framework for permutation feature importance approaches. In our work, we utilize the empirical model reliance (EMR), which measures the change of the loss function after shuffling the values of the feature of interest. As such, PFI and EMR provide global saliency maps.

Feature importance ranking measure (FIRM) The feature importance ranking measure (FIRM, Zien et al. 2009) takes the underlying correlations of the features into account, by leveraging the conditional expectation of the model's output function, given the feature of interest, and measuring its deviation. As such, FIRM provides a global saliency map. While intractable for arbitrary models and data distributions, FIRM admits a closed-form solution for linear models and Gaussian distributed data, which we implemented. Notably, under these assumptions, FIRM is equivalent to the linear activation pattern (see above) up to a re-scaling of each feature by its standard deviation (Haufe et al. 2014).

Local interpretable model-agnostic explanations (LIME) To generate a saliency map for a model's prediction on a single example, LIME (Ribeiro et al. 2016) samples instances around that instance, and weights the samples according to their proximity to it. LIME then learns a linear surrogate model in the vicinity of the instance of interest, trying to linearly approximate the local behavior of the model, which is then interpreted by examining the weight vector (extraction filter) of that linear model. As such, LIME inherits the conceptual drawbacks of methods directly interpreting gradients or model weights in the presence of suppressor variables.

Shapley additive explanations (SHAP) The Shapley value (Shapley 1953) is a game theoretic approach to measure the influence of a feature on the decisions of a model on

a single example. Since its computation is intractable for most real-world settings, an approximation called SHAP (Lundberg and Lee 2017) has become widely popular. We use the linear SHAP method, including the option to account for correlations between features.

Anchors Anchors (Ribeiro et al. 2018) seeks to identify a sparse set of important features for single instances, which lead to consistent predictions in the vicinity of the instance. Features for which changes in value have almost no effect on the model's performance are considered unimportant.

Neural-network-specific methods In addition to the model-agnostic XAI methods introduced above, a number of model-specific methods tailored to neural network architectures are considered. All these methods are based on modified backpropagation, but deal with nonlinearities in the network in a different way. For all methods, the implementation in the *innvestigate*¹ (Alber et al. 2019) package is used.

Simonyan and Zisserman (2015) proposed a sensitivity analysis, where pixels for which the model output is more affected by a shift in the input signal are considered more important. To this end, the gradient of the output with respect to the input signal is calculated. However, as discussed above, the gradient of a linear model reduces to its model weights (extraction filters): $Grad_{NN} = w_{NN}$. DeConvNet (Zeiler and Fergus 2014) and Guided Backpropagation (Springenberg et al. 2015) are two additional methods that again reduce to the gradient/extraction filter for linear models.

PatternNet (Kindermans et al. 2018) is conceptually similar to gradient analysis. However, rather than model weights, activation patterns are estimated per node and backpropagated through the network. For linear networks, PatternNet coincides with the linear activation pattern approach, although we observe slight deviations between the methods in practice.

Lastly, layer-wise relevance propagation (LRP, Bach et al. 2015), Deep Taylor Decomposition (DTD, Montavon et al. 2017), and PatternAttribution (Kindermans et al. 2018) aim to visualize how much the different dimensions of the input contribute to the output through the layers. As such, each node in the network is assigned a certain amount of 'relevance', while keeping the total 'relevance' per layer constant. For LRP, two different variants ('rules') are included: the z -rule and the $\alpha\beta$ -rule (Bach et al. 2015). Deep Taylor Decomposition (DTD) approximates the subfunctions learned by the different nodes by applying a Taylor decomposition around a root point and pooling the relevance over all neurons. Lastly, PatternAttribution (Kindermans et al. 2018) estimates the root point from the data based on the PatternNet approach.

4.4 Measures of explanation performance

While numerous subjective criteria for evaluating the success of XAI methods have been proposed (e.g., Nguyen and Martínez 2020; Schmidt and Biessmann 2019), we here aim to provide objective, data-dependent, criteria using definition (1). Since, we know that statistical differences between classes are only present in features belonging to the set \mathcal{F}_{dep}^+ , while features \mathcal{F}_{dep}^- are entirely driven by non-class-specific, fluctuations, the dichotomization

¹ <https://github.com/albermax/innvestigate>.

$$h_d^{\text{true}} = \begin{cases} 1, & d \in \mathcal{F}_{\text{dep}}^+ \\ 0, & d \in \mathcal{F}_{\text{dep}}^- \end{cases} \quad (8)$$

is used as a ground truth both for global and instance-based ‘explanations’. This binary ground truth is compared to the continuous-valued saliency map $\mathbf{s}(f^\theta, \mathbf{x}^*, \mathcal{D}) \in \mathbb{R}^D$ of each XAI method.

Explanation performance is measured by comparing \mathbf{h}^{true} with \mathbf{s} . To this end, saliency maps are rectified by taking the absolute value $|\mathbf{s}|$. As performance metrics, we use the area under receiver operating curve (AUROC) and the precision for a fixed specificity of 90% (PREC90), which is obtained using the lowest threshold on \mathbf{s} for which the specificity is greater or equal to 90%. Results were similar when AUROC was replaced by average precision (see supplementary material). The PREC90 metric is based on the following consideration: while $\mathcal{F}_{\text{dep}}^+$ defines the set of features that any XAI method *may* highlight, a particular machine learning model may actually use only a subset of them. Thus, we would like to penalize false negatives (features that are in $\mathcal{F}_{\text{dep}}^+$ but receive a low score according to \mathbf{s}) much less than false positives (features in $\mathcal{F}_{\text{dep}}^-$ that receive a high score). To this end, we evaluate the precision (fraction of truly important features among those estimated to be important) at a high decision threshold based on the consideration that good XAI methods should assign very high saliency scores only to truly important features. Truly important features receiving low scores thus do not influence this metric.

Performance metrics are evaluated per model for global XAI methods and per sample for instance-based XAI methods. In addition, instance-based rectified saliency maps are averaged to also yield global maps

$$\mathbf{s}^{\text{global}}(f^\theta, \mathcal{D}) = \frac{1}{N} \sum_{n=1}^N |\mathbf{s}^{\text{instance}}(f^\theta, \mathbf{x}^n, \mathcal{D})|, \quad (9)$$

the performance of which is also evaluated. Note that, in our setting, the input-output relationships between features and target are static. Thus, the same, global, ground-truth saliency map is expected to be reconstructed by each local explanation on average. While individual explanations may be heavily corrupted by noise, this effect should be suppressed when averaging rectified heat maps across all samples, which is, therefore, considered a meaningful way to derive global explanations for instance-based XAI methods. Thus, instance-based XAI methods are evaluated both in terms of global and single-instance performance, while global XAI methods are only evaluated with respect to the former.

5 Experiments

We conduct a set of experiments aimed to address the following questions: (i) which XAI methods are best able to differentiate between important (that is, class-specific) features and non-important features, (ii) how does the signal-to-noise ratio of the data (through the accuracy of the classifier) affect the explanation performance of each method. Python code to reproduce our experiments is provided on [github](https://github.com)². We generate $K = 100$ datasets with $N = 1000$ samples each according to the model specified in Sect. 4.1. The prediction task

² <https://github.com/braindatalab/scrutinizing-xai>.

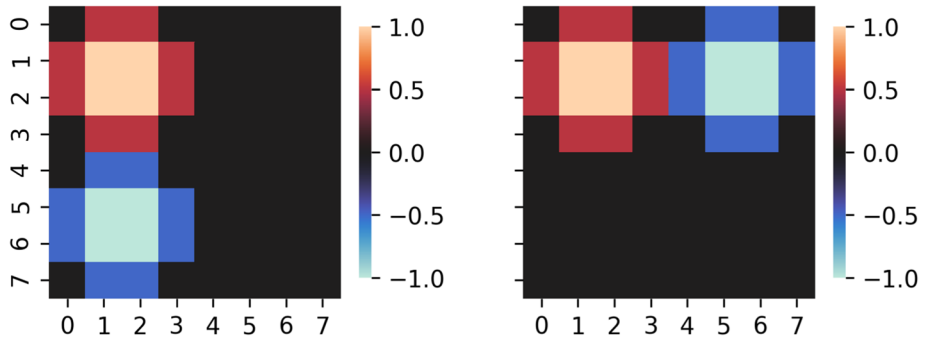


Fig. 1 The signal activation pattern \mathbf{a} (left) and the distractor activation pattern \mathbf{d} (right) used in our experiments can be visualized as images of 8×8 pixels size. The signal pattern consists of two blobs with opposite signs: one in the upper left and one in the lower left corner, while the distractor pattern consists of blobs in the upper left and upper right corners. Thus, the two components spatially overlap in the upper left corner

is to discriminate between two categories encoded in the binary variable $y^n, n \in \{1, \dots, N\}$. Our feature space are images of size 8×8 , thus $D = 8^2 = 64$. Figure 1 depicts the static signal and distractor patterns $\mathbf{a} \in \mathbb{R}^{64}$ and $\mathbf{d} \in \mathbb{R}^{64}$, which are identical across all experiments. As can be seen, signal and distractor overlap in the upper left corner of the image, while the lower left corner is occupied by the signal only and the upper right corner is occupied by the distractor only. All pixels are moreover affected by multivariate correlated Gaussian noise $\boldsymbol{\eta}$.

With the signal pattern \mathbf{a} , we control the statistical dependencies between features and classification target in our synthetic data. Therefore, the *ground truth* set of important features in our experiments is given by

$$\mathcal{F}_{\text{dep}}^+ = \{d \mid a_d \neq 0, 1 \leq d \leq 64\} . \tag{10}$$

Note that noise and distractor components both do not contain any class-specific information. The distractor, thus, merely serves as a strong one-dimensional noise component with predefined characteristic spatial pattern. Its main purpose in our experiments is to facilitate the visual assessment of saliency maps, where any importance assigned to the right half of the image represents a false positive.

For each dataset, class labels y^n , distractor values ρ^n , and noise vectors $\boldsymbol{\eta}^n$ are sampled independently from their respective distributions described in Sect. 4.1. Five different SNRs are analyzed, corresponding to five different choices of the parameter $\lambda_1 \in \{0.0, 0.02, 0.04, 0.06, 0.08\}$. Each resulting dataset $\mathcal{D}_{k,\lambda_1}, k \in \{1, \dots, 100\}$ is divided into a train set $\mathcal{D}_{k,\lambda_1}^{\text{train}}$ and a validation set $\mathcal{D}_{k,\lambda_1}^{\text{val}}$, with samples sizes $N^{\text{train}} = 800$ and $N^{\text{val}} = 200$, respectively.

Linear logistic regression classifiers f^w are fitted on $\mathcal{D}_{k,\lambda_1}^{\text{train}}$ and applied to $\mathcal{D}_{k,\lambda_1}^{\text{val}}$. The logistic regression implemented in scikit-learn is trained with a maximum number of 1000 iterations. The neural network based implementation is trained for 200 epochs with a learning rate of 0.1. Since the neural network has two output neurons, its effective extraction filter \mathbf{w}^{NN} was calculated as the difference $\mathbf{w}^{\text{NN}} = \mathbf{w}_1^{\text{NN}} - \mathbf{w}_2^{\text{NN}}$.

Saliency maps $\mathbf{s}(f^w, \mathbf{x}^n, \mathcal{D}_{k,\lambda_1}^{\text{train}})$ are obtained for each of the methods described in Sect. 4.3 and for all datasets $\mathcal{D}_{k,\lambda_1}^{\text{train}}$, where instance-based maps are evaluated on all input

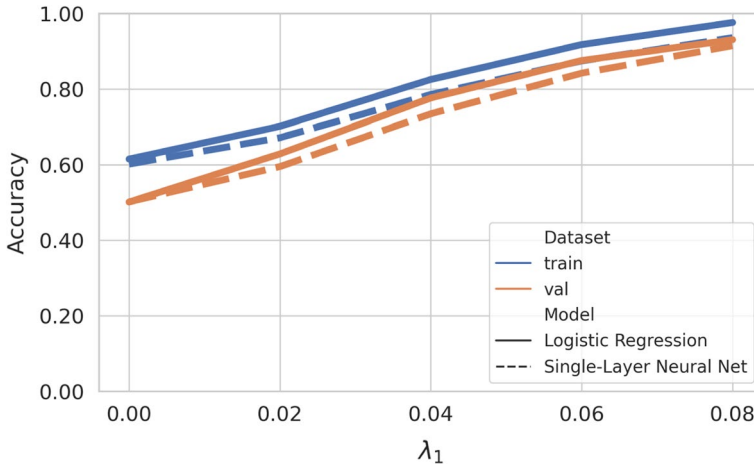


Fig. 2 With increasing signal-to-noise ratio (determined through the parameter λ_1 of our generative model (4)), the classification accuracy of the logistic regression and the single-layer neural network increases, reaching near-perfect accuracy for $\lambda_1 = 0.08$

examples \mathbf{x}^n from the corresponding validation sets $\mathcal{D}_{k,\lambda_1}^{\text{val}}$. All XAI methods are applied using the default parameters, with the exception of $\text{LRP}_{\alpha\beta}$, for which we set $\alpha = 2$ and $\beta = 1$. That is, for SHAP, we use the LinearExplainer with Impute Masker, set `feature_perturbation=correlation_dependent` and `perform the sampling with samples = 1000`. For LIME, we use `kernel_width = $\sqrt{64 * 0.75}$` , `kernel = $\exp(-x^2/\text{kernel_width}^2)^{1/2}$` , `discretize_continuous = False`, and `feature_selection = highest_weights`. For ANCHORS, we use `threshold = 0.95`, `delta = 0.1`, `discretizer = 'quartile'`, `tau = 0.15`, `batch_size = 100`, `coverage_samples = 10000`, `beam_size = 1`, `stop_on_first = False`, `max_anchor_size = 64`, `min_samples_start = 100`, `n_covered_ex = 10`, `binary_cache_size = 10000`, `cache_margin = 1000`.

6 Results

Figure 2 shows the classification accuracy achieved by the LLR classifiers as a function of the SNR parameter λ_1 . Both implementations reach near-perfect training and validation accuracy at an SNR of $\lambda_1 = 0.08$. For $\lambda_1 = 0$ (no class-specific information present), the validation accuracy attains chance level (0.5), as expected, while the training accuracy of 0.6 indicates a small degree of overfitting.

6.1 Qualitative assessment of saliency maps

Figure 3 depicts examples of global heat maps obtained for a randomly drawn dataset \mathcal{D}_k for three different SNRs, $\lambda_1 \in \{0.0, 0.04, 0.08\}$. Shown are rectified quantities obtained by taking the absolute value. Instance-based maps were averaged over all instances of the validation dataset to obtain global maps. As expected, at $\lambda_1 = 0$ (lack of class-specific information; therefore, chance-level classification), none of the

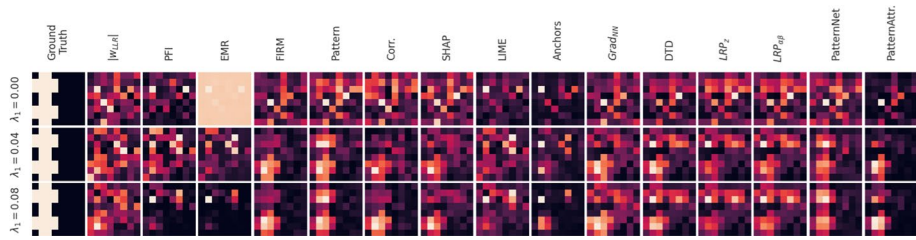


Fig. 3 Global saliency maps obtained from various XAI methods on a single dataset. Rows represent three different choices of the SNR parameter λ_1 . In the top row, no class-related information is present, yielding chance-level classification, while for the bottom row near-perfect classification accuracy is obtained. The ‘ground truth’ set of important features is defined as the set of pixels with for which a statistical relationship to the class label is modeled, i.e. the set of pixels with nonzero signal patterns defined in (10). Notably, a number of XAI methods assign significant importance to pixels in the right half of the image, which are statistically unrelated to the class label (suppressor variables) by construction

saliency maps resembles to the ground-truth importance map given by the pattern of the simulated signal. For $\lambda_1 = 0.04$ and $\lambda_1 = 0.08$, vast differences between different methods appear, though. The saliency maps of the linear Pattern as well as PatternNet deliver the best results on this dataset, recovering the two blobs of the ground-truth signal pattern most closely while correctly ignoring the right half of the image. FIRM, Correlation and PatternAttribution do recover the lower left blob of the signal pattern but assign much less importance to the upper left blob, where signal and distractor patterns overlap. All other methods including extraction filters w^{LLR} and $Grad_{NN}$, PFI, EMR, SHAP, LIME, Anchors, DTD, and the two LRP variants assign significant importance to the upper right corner, in which no class-related signal is present (thus, to suppressor variables), even for high SNR ($\lambda_1 = 0.08$). For some methods, the importance assigned to the distractor-only upper right blob is of the same order as the importance assigned to the upper left corner, in which signal and distractor overlap (PFI, EMR, LIME, Anchors, DTD and LRP). Interestingly, the signal-only lower left corner is assigned much less importance than the distractor-only upper right corner by some methods (PFI, EMR, LRP). This indicates that these methods mainly focus on the process of optimally *extracting* the signal component with the distractor component rather than localizing the signal itself. Saliency maps provided by PFI, EMR, and Anchors are sparsest, focus on a small number of important as well as unimportant features, while gradient and extraction filter maps are the least sparse.

From the randomly chosen dataset \mathcal{D}_k used to create Fig. 3, we further picked a single random instance that was correctly classified by both LLR implementations. Saliency maps for this instance are shown in Fig. 4. At high SNR, LIME, Anchors, DTD, and LRP still assign importance to the right half of the image, where no statistical relation to the class label is present by construction, while PatternNet, PatternAttribution, and SHAP do not. Interestingly, the instance-based saliency maps of the best performing methods, PatternNet and PatternAttribution, closely match the global maps obtained from these methods, even though they barely resemble features of the instance they were computed for. This suggests that these methods are strongly dominated by the global statistics of the training data rather than the properties of the individual input sample.

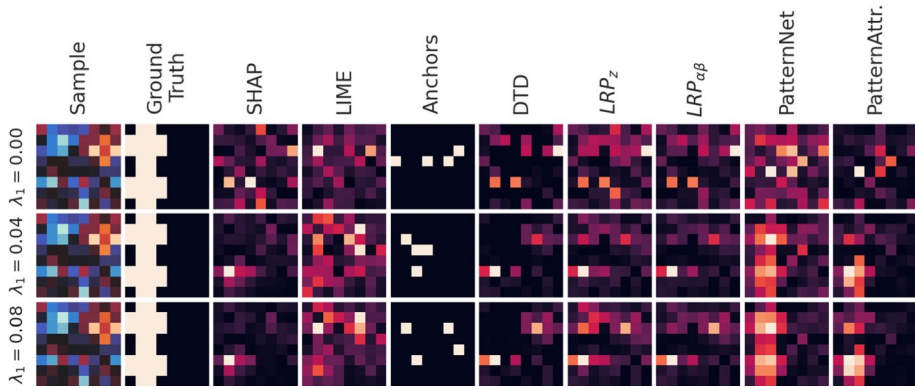


Fig. 4 Saliency maps obtained for a randomly chosen single instance. At high SNR, PatternNet and PatternAttribution best reconstruct the ground truth signal pattern, while SHAP, LIME, DTD, and LRP assign importance to the right half of the image, where no statistical relation to the class label is present by construction

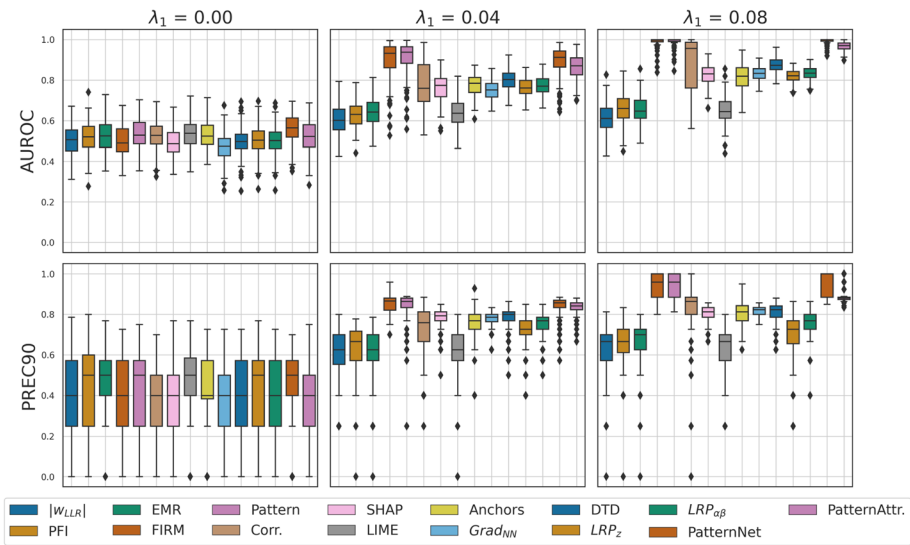


Fig. 5 Quantitative explanation performance of global saliency maps attained by various XAI approaches. Performance was measured by the area under the receiver-operating curve (AUROC) and the precision at $\approx 90\%$ specificity. While chance-level performance is uniformly observed in absence of any class-related signal, stark differences between methods emerge for medium and high SNR ($\lambda_1 = 0.04$, and $\lambda_1 = 0.08$). Among the global XAI methods, the linear Pattern and FIRM consistently provide the best ‘explanations’ according to both performance metrics. Among the instance-based methods, the saliency maps obtained by PatternNet and PatternAttribution (averaged across all instances of the validation set) show the strongest explanation performance

6.2 Quantification of explanation performance

Figure 5 depicts the explanation performance of the global saliency maps provided by the considered XAI methods across 100 experiments. Shown are the median performance as

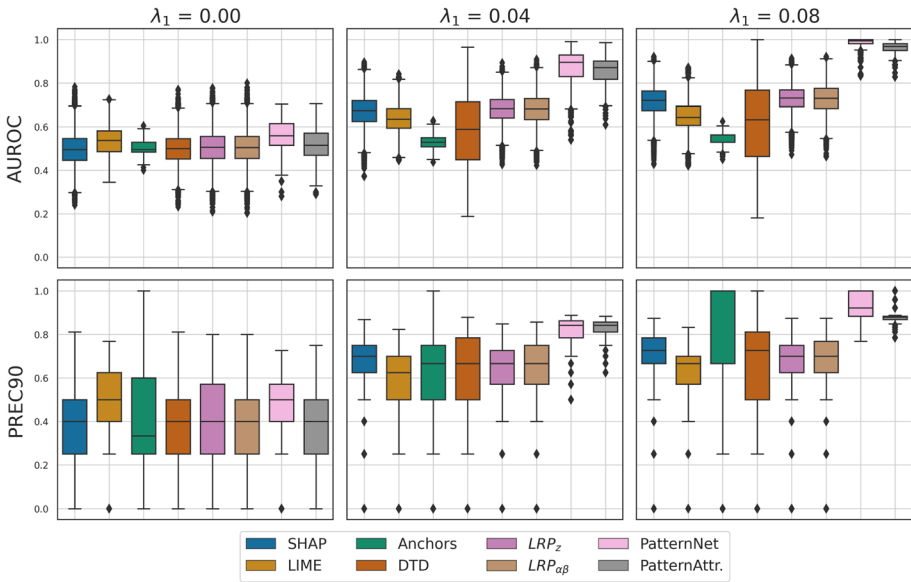


Fig. 6 Explanation performance of instance-based saliency maps of neural-network-based XAI methods. All methods perform at chance level in absence of class-related information. For medium and high SNR, PatternNet and PatternAttribution show the strongest capability to assign importance to those feature that are associated with the classification target by construction, while ignoring suppressor features

well as the lower and upper quartiles, as well as outliers. As expected, the explanation performance of all methods (that is, the ability to distinguish truly important from unimportant pixels in the image) is not significantly different from chance level (AUROC = 0.5) when indeed no class-specific information is present ($\lambda_1 = 0$). At higher SNR ($\lambda_1 = 0.04$, and $\lambda_1 = 0.08$), most methods deviate from chance-level; however, significant differences are observed between methods. The median performances of LIME, PFI, and EMR quickly saturate at a low level around AUROC = 0.6. FIRM, Pattern, PatternNet and PatternAttribution have consistently higher AUROC and PREC90 scores than other methods, approaching perfect performance for $\lambda_1 = 0.08$. The Pearson correlation between feature and class label can be considered as a runner-up, but is characterized by higher variance compared to the (covariance-based) Pattern. This can be explained by the fact that the presence of noise in the upper left corner (where both the signal and distractor are present) diminishes the correlation but not the covariance between the class label and the features in that corner. SHAP, DTD and LRP achieve moderate performance, while PFI, EMR, and Anchors do not perform well at all. This can only partially be explained by the sparsity of their saliency maps, which is penalized by the AUROC metric but not the PREC90 metric. Indeed, the difference between PFI, EMR, and Anchors on one hand and the rest of the methods on the other hand is smaller for the PREC90 than for the AUROC metric. However, the ranking of methods is similar for both metrics.

In Fig. 6, quantitative results attained – obtained per instance without averaging – are shown. As observed for the global saliency maps, PatternNet and PatternAttribution the highest scores for both medium and high SNR, followed by the gradient of the neural network. Variants of LRP have achieve moderate performance in all settings. A high variance is, however, observed for DTD.

7 Discussion & related work

‘Explainable’ artificial intelligence (XAI) is a highly relevant field that has already produced a vast body of literature. But many existing XAI approaches do not come with a theory on *how* their results should be interpreted, i.e., what formal statements can be reasonably derived from them. We here formalize a minimal assumption that (as we believe) humans typically make when being offered ‘explanations’. Namely, that the input features highlighted by an XAI method must have an actual statistical relationship to the prediction target. Using empirical experiments and well-controlled synthetic data we demonstrate, however, that this is not guaranteed for a substantial number of state-of-the-art XAI approaches, inviting misinterpretations. In this light, interpretations such as those suggested for LIME in Ribeiro et al. (2016) (see introduction) seem to be unjustified, because LIME cannot rule out the influence of suppressor variables. False-positive associations between features and a disease thereby do not seem to be the only possible misinterpretations. A doctor confronted with the high importance of a variable known to be unrelated to a disease (a suppressor) may not be able to recognize that but may rather erroneously come to the conclusion that the model is not trustworthy.

Our synthetic data were specifically designed to include suppressor variables, which are statistically independent of the prediction target but improve the prediction in combination with other variables. More specifically, suppressor variables display a *conditional dependency* on the target given other features. In example (2), for example, the suppressor x_2 is independent of y but becomes dependent on y given x_1 . A multivariate model can leverage this conditional dependency to improve its prediction – here, by removing shared noise from feature x_1 . However, ‘influential’ features showing only such conditional dependencies can be of little interest in practice and need to be interpreted differently than features exhibiting a direct statistical relationship.

In our simulation, various XAI methods were found to be unable to reject suppressor variables as being unimportant. This failure was found to be aggravated in a setting where all signal-containing features were contaminated with the distractor (thus, the lower left blob in the signal pattern was absent), see supplementary material. While – based on the consideration made above – this behavior is expected for methods based on interpreting model weights, such as LIME or gradient-based approaches, it was also observed for PFI, EMR, SHAP, Anchors, DTD and LRP. While we suggest an explanation for that behavior in the following paragraph, future work will be required to theoretically study the behavior of each method in the presence of suppressors.

The degree to which suppressor variables affect model explanations in practice is hard to estimate and may differ considerably between domains and applications. Thus, the quantitative results presented here are not claimed to universally hold. To rule out adverse effects to due suppressor variables or other detrimental data properties in a particular application, it should become common practice to conduct simulations with realistic domain-specific ground-truth data.

7.1 Insufficiency of model-driven XAI

In principle, one may argue that different types of interpretations can be useful in different contexts. For example, the identification of input dimensions that have a strong ‘influence’ on a model’s output may be useful to study the general behavior of that model (e.g.

for debugging purposes). However, we argue that it is insufficient to analyze any model without taking into account the distribution of the data it was trained on. The difficulty of several XAI methods to reject suppressor variables can be explained by their inability to recognize that suppressor variables and truly target-related variables (e.g., x_1 and x_2 in example (2)) are *correlated* and thus cannot be manipulated independently, limiting the degrees of freedom in which individual features can influence the model output. This limitation is not only inherent to several XAI methods but also to empirical ‘validation’ schemes based on the manipulation of single input features. As the resulting surrogates do not follow the true distribution of the training data, limited insight about the actual behavior of the model when used for its intended purpose can be gained.

In contrast to existing predominantly model-driven and data-agnostic XAI approaches, we here provide a definition of feature importance that is purely data-driven, namely the presence of a univariate statistical interaction to the prediction target. Importantly, this definition can also be tested on empirical data using statistical tests for non-linear interactions (Gretton et al. 2007). In the linear case studied here, it is sufficient to analyze the covariance between features and prediction target, as described in (Haufe et al. 2014), to obtain a saliency map with optimal explanation performance according to our metrics. The results of instance-based extensions of the linear covariance pattern, such as PatternNet and Pattern-Attribution (Kindermans et al. 2018), however, suggest that the global covariance structure of the training may strongly dominate saliency maps obtained for single instances, which should be a subject of further investigation.

In general, features and target variables can be considered to be part of a system of random variables whose relationships are governed by structural equations (such as Eqs. (2) and (4)). These structural relationships determine the possible statistical relationships of the involved random variables, and thus give rise to an even more fundamental definition of feature importance compared to our current definition based on actual statistical dependencies (Eq. (1)). In fact, one can construct artificial settings, where structural relationships between features and target exist but do not manifest in statistical dependencies due to cancellation effects. However, we consider such situations rather irrelevant in practice. Typically, definitions based on structural and statistical relationships will coincide, which is also the case in our experimental setting. The advantage of definition (1) is that, while structural relationships are hard to assess in practice, the mere presence of statistical relationships may be assessed empirically, offering a general way to estimate feature importance in practice.

Our definition encompasses the superset of features that *may* be found important by any combination of machine learning model and saliency method. In fact, a model may not necessarily use all features contained in the set $\mathcal{F}_{\text{dep}}^+$ to achieve its prediction task. This is accounted for by our performance metric PREC90, which is designed to ignore most false negative omissions of important features. Practically, it may be desirable to *fuse* data- and model-driven saliency maps, e.g. by taking the intersection between the estimated set $\mathcal{F}_{\text{dep}}^+$ and the set identified by a conventional XAI method.

7.2 Existing validation approaches

One can distinguish three categories of existing evaluation techniques for XAI methods. (i) evaluating the sensitivity or robustness of explanations to model modifications and input perturbations, (ii) using interdisciplinary and human-centered techniques to evaluate

explanations, and (iii) establishing a controlled setting by leveraging a-priori knowledge about relevant features.

Sensitivity-and robustness-centered evaluations Assessing the robustness and sensitivity of saliency maps in response to input perturbations and model changes is a common strategy underlying XAI approaches and their validation. However, such approaches do not establish a notion of correctness of explanations but merely formulate additional criteria (sanity checks) (see, e.g., Doshi-Velez and Kim 2017). For example, Alvarez-Melis and Jaakkola (2018) assessed model ‘explanations’ regarding their robustness – asserting that similar inputs should lead to similar explanations – and showed that LIME and SHAP do not fulfill this requirement. Several studies developed tests to detect inadequate explanations (Adebayo et al. 2018; Ancona et al. 2018). Adebayo et al. demonstrated that, for some XAI methods, the identified features of trained models are akin to the ones identified by randomized models. Hooker et al. (2019) came to similar conclusions. These features often represent low-level properties of the inputs, such as edges in images, which do not necessarily carry information about the prediction target (Adebayo et al. 2018; Sixt et al. 2020).

Human-centered evaluations Human judgement is also often used to evaluate XAI methods (e.g., Baehrens et al. 2010; Poursabzi-Sangdeh et al. 2021; Lage et al. 2018; Schmidt and Biessmann 2019). To this end, the extent to which the use of model ‘explanations’ can help a human to accomplish a task or to predict a model’s behavior is typically measured. Another possibility is to define ground-truth explanations directly through human expert judgement (Park et al. 2018). As such, human-centered approaches also do not establish a mathematically sound ground-truth, as human evaluations can be highly biased. In contrast, we here exclusively focus on formally-grounded evaluation techniques, (c.f., Doshi-Velez and Kim 2017).

Ground-truth-centered evaluations Few works have attempted to use objective criteria and/or ground truth data to assess XAI methods. Kim et al. (2018) used synthetic data to obtain qualitative ‘explanations’, which were then evaluated by humans, while Yang and Kim (2019) derived quantitative statements from synthetic data. However, in both cases, the ground-truth was not defined as a verifiable property of the data but as a ‘relative feature importance’ representing how ‘important’ a feature is to a model relatively to another model. In other works, the importance of features was defined through a generative process similar to ours (Ismail et al. 2019; Tjoa and Guan 2020). Yet, these works have not provided a formal, data-driven, definition of feature importance that would provide theoretical basis for their ground truth. Moreover, correlated noise settings leading to emergence of suppressor variables, have not been systematically studied in these works but have been shown to have a profound impact on the conclusions that can be drawn from XAI methods here.

7.3 Limitations and outlook

The present paper focuses on data with Gaussian class-conditional distributions with equal covariance, where linear machine learning models are Bayes-optimal. While this represents a well-controlled baseline setting, it is unlikely that solutions for the linear case transfer to general non-linear settings. Non-linear extensions of the activation pattern approach, such as PatternNet and PatternAttribution (Kindermans et al. 2018), exist but have not been validated on ground-truth data. Our future work will address this gap by simulating non-linear suppressor variables, emerging through non-linear interactions between features.

In non-linear settings, a clear distinction between features statistically related to the target or not may not always be possible. In tasks like image categorization, class-specific information might be contained in different features for each sample, depending on where in the image the target object is located. In extreme cases, objects of all classes may allocate the same locations leading to similar univariate marginal feature distributions whereas the discriminative information is contained in dependencies between features. Future work will be concerned with providing ground-truth definitions better reflecting this case.

8 Conclusion

We have formalized feature importance in an objective, purely data-driven, way as the presence of a statistical dependency between feature and prediction target. We have further described suppressor variables as variables with no such statistical dependency that are, nevertheless, typically identified as important according to criteria that are prevalent in the XAI community. Based on linear ground-truth data, generated to reflect our definition of feature importance, we designed a quantitative benchmark including metrics of ‘explanation performance’, using which we empirically demonstrated that many currently popular XAI methods perform poorly in the presence of so-called suppressor variables. Future work needs to further investigate non-linear cases and conceive well-defined notions of feature importance for specific non-linear settings. These should ultimately inform the development of novel XAI methods.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10994-022-06167-y>.

Funding Open Access funding enabled and organized by Projekt DEAL. This result is part of a project that has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (Grant agreement No. 758985). KRM also acknowledges support by the German Ministry for Education and Research as BIFOLD – Berlin Institute for the Foundations of Learning and Data (ref. 01IS18025A and ref. 01IS18037A), and the German Research Foundation (DFG) as Math+: Berlin Mathematics Research Center (EXC 2046/1, project-ID: 390685689), Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea Government (No. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University).

Availability of data and material All data used here can be generated using the provided code.

Code availability <https://github.com/braindatalab/scrutinizing-xai>.

Declarations

Conflicts of interest/Competing interests The authors declare no conflicts of interest/competing interests.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Curran Associates Inc., Montréal, Canada, NIPS'18, pp 9525–9536.
- Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K. T., Montavon, G., et al. (2019). Investigate eural networks! *J Mach Learn Res*, 20(93), 1–8.
- Alvarez-Melis, D., & Jaakkola, T.S. (2018). On the Robustness of Interpretability Methods. [arXiv:180608049](https://arxiv.org/abs/1806.08049) [cs, stat] [ArXiv: 1806.08049](https://arxiv.org/abs/1806.08049).
- Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2018). Towards better understanding of gradient-based attribution methods for deep neural networks. In: ICLR.
- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58, 82–115.
- Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS One*, 10(7), e0130140.
- Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., & Müller, K. R. (2010). How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11, 1803–1831.
- Binder, A., Bach, S., Montavon, G., Müller, K. R., & Samek, W. (2016). Layer-Wise Relevance Propagation for Deep Neural Network Architectures. In K. J. Kim & N. Joukov (Eds.), *Information Science and Applications (ICISA) 2016* (pp. 913–922). Lecture Notes in Electrical Engineering: Springer, Singapore.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Conger, A. J. (1974). A revised definition for suppressor variables: a guide to their identification and interpretation, a revised definition for suppressor variables: A guide to their identification and interpretation. *Educational and Psychological Measurement*, 34(1), 35–46.
- Dombrowski, A. K., Anders, C. J., Müller, K. R., & Kessel, P. (2022). Towards robust explanations for deep neural networks. *Pattern Recognition*, 121, 108194.
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. [arXiv: 1702.08608](https://arxiv.org/abs/1702.08608) [cs, stat] [ArXiv: 1702.08608](https://arxiv.org/abs/1702.08608).
- Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177), 1–81.
- Fong, R.C., & Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp 3429–3437.
- Friedman, L., & Wall, M. (2005). Graphical views of suppression and multicollinearity in multiple linear regression. *The American Statistician*, 59(2), 127–136.
- Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., Smola, A. J., et al. (2007). A kernel statistical test of independence. *Nips Citeseer*, 20, 585–592.
- Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87, 96–110.
- Hooker, S., Erhan, D., Kindermans, P.J., & Kim, B. (2019). A benchmark for interpretability methods in deep neural networks. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R (eds) *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol 32, pp 9737–9748.
- Horst, P., Col Wallin, P., Col Guttman, L., Brim Col Wallin, F., Clausen, J.A., Col Reed, R., & Col Rosenthal, E. (1941). The prediction of personal adjustment: A survey of logical problems and research techniques, with illustrative application to problems of vocational selection, school success, marriage, and crime. Social science research council.
- Ismail, A.A., Gunady, M., Pessoa, L., Corrada Bravo, H., & Feizi, S. (2019). Input-cell attention reduces vanishing saliency of recurrent neural networks. In: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (eds) *Advances in Neural Information Processing Systems*, Curran Associates, Inc., vol 32, pp 10814–10824.
- Jaderberg, M., Simonyan, K., Zisserman, A., & Kavukcuoglu, K. (2015). Spatial transformer networks. In: *Proceedings of the 28th International Conference on Neural Information Processing Systems*-(Vol. 2, pp 2017–2025).

- Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: *International Conference on Machine Learning*, PMLR, pp 2668–2677.
- Kindermans, P., Schütt, K.T., Alber, M., Müller, K., Erhan, D., Kim, B., & Dähne, S. (2018). Learning how to explain neural networks: Patternnet and patternattribution. In: ICLR.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105.
- Lage, I., Ross, A., Gershman, S.J., Kim, B., & Doshi-Velez, F. (2018). Human-in-the-Loop Interpretability Prior. In: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (eds) *Advances in Neural Information Processing Systems 31*, Curran Associates, Inc., pp 10159–10168.
- Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K. R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10(1), 1096.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31–57.
- Lundberg, S.M., & Lee, S.I. (2017). A Unified Approach to Interpreting Model Predictions. In: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds) *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., pp 4765–4774.
- Montavon, G., Bach, S., Binder, A., Samek, W., & Müller, K. R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65, 211–222.
- Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73, 1–15.
- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44), 22071–22080.
- Nguyen, A.p., & Martínez, M.R. (2020). On quantitative aspects of model interpretability. [arXiv:2007.07584](https://arxiv.org/abs/2007.07584) [cs, stat] [ArXiv: 2007.07584](https://arxiv.org/abs/2007.07584).
- Park, D.H., Hendricks, L.A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., & Rohrbach, M. (2018). Multimodal explanations: Justifying decisions and pointing to the evidence. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., & others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830
- Poursabzi-Sangdeh, F., Goldstein, D.G., Hofman, J.M., Wortman Vaughan, J.W., & Wallach, H. (2021). Manipulating and measuring model interpretability. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp 1–52.
- Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). “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*, pp 1135–1144.
- Ribeiro, M.T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. In: AAAI.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
- Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K. R. (2016). Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11), 2660–2673.
- Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning* (Vol. 11700). New York: Springer.
- Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K. R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3), 247–278.
- Schmidt, P., & Biessmann, F. (2019). Quantifying interpretability and trust in machine learning systems. [arXiv:1901.08558](https://arxiv.org/abs/1901.08558) [cs, stat] [ArXiv: 1901.08558](https://arxiv.org/abs/1901.08558).
- Shapley, L. S. (1953). A value for n-person games. *Contributions to the theory of games*, 2(28), 307–317.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676), 354–359.
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) [cs] [ArXiv: 1409.1556](https://arxiv.org/abs/1409.1556).

- Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint [arXiv:13126034](https://arxiv.org/abs/1312.6034).
- Sixt, L., Granz, M., & Landgraf, T. (2020). When explanations lie: Why many modified bp attributions fail. In: *International Conference on Machine Learning*, PMLR, pp 9046–9057.
- Springenberg, J.T., Dosovitskiy, A., Brox, T., & Riedmiller, M.A. (2015). Striving for simplicity: The all convolutional net. CoRR [arXiv:1412.6806](https://arxiv.org/abs/1412.6806).
- Tjoa, E., & Guan, C. (2020). Quantifying explainability of saliency methods in deep neural networks. [arXiv:200902899](https://arxiv.org/abs/200902899) [cs] [ArXiv: 2009.02899](https://arxiv.org/abs/2009.02899).
- Yang, M., & Kim, B. (2019). Benchmarking attribution methods with relative feature importance. [arXiv:190709701](https://arxiv.org/abs/190709701) [cs, stat] [ArXiv: 1907.09701](https://arxiv.org/abs/1907.09701).
- Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer Vision - ECCV 2014* (pp. 818–833). Lecture Notes in Computer Science: Springer International Publishing, Cham.
- Zien, A., Krämer, N., Sonnenburg, S., & Rätsch, G. (2009). The Feature Importance Ranking Measure. In W. Buntine, M. Grobelnik, D. Mladenić, & J. Shawe-Taylor (Eds.), *Machine Learning and Knowledge Discovery in Databases* (pp. 694–709). Lecture Notes in Computer Science: Springer, Berlin, Heidelberg.
- Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3), 647–665.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Rick Wilming¹  · Céline Budding² · Klaus-Robert Müller^{1,3,4,5}  · Stefan Haufe^{1,6,7} 

Rick Wilming
rick.wilming@tu-berlin.de

Céline Budding
c.e.budding@tue.nl

Klaus-Robert Müller
klaus-robert.mueller@tu-berlin.de

¹ Technische Universität, Berlin, Germany

² Eindhoven University of Technology, Eindhoven, The Netherlands

³ BIFOLD – Berlin Institute for the Foundations of Learning and Data, Berlin, Germany

⁴ Korea University, Seoul, South Korea

⁵ Max Planck Institute for Informatics, Saarbrücken, Germany

⁶ Physikalisch-Technische Bundesanstalt, Berlin, Germany

⁷ Charité – Universitätsmedizin, Berlin, Germany