Conditional Feature Importance with Generative Modeling Using Adversarial Random Forests

Kristin Blesch[*1,2], **Niklas Koenen**[*1,2], **Jan Kapar**[1,2], **Pegah Golchian**[1,2], **Lukas Burk**[1,2], **Markus Loecher**[3], **Marvin N. Wright**[†1,2,4]

[1]Leibniz Institute for Prevention Research & Epidemiology – BIPS, Germany
[2]Faculty of Mathematics and Computer Science, University of Bremen, Germany
[3]Department of Business and Economics, Berlin School of Economics and Law, Germany
[4]Department of Public Health, University of Copenhagen, Denmark
{blesch, koenen, kapar, golchian, burk, wright}@leibniz-bips.de, markus.loecher@hwr-berlin.de

Abstract

This paper proposes a method for measuring conditional feature importance via generative modeling. In explainable artificial intelligence (XAI), conditional feature importance assesses the impact of a feature on a prediction model's performance given the information of other features. Model-agnostic post hoc methods to do so typically evaluate changes in the predictive performance under on-manifold feature value manipulations. Such procedures require creating feature values that respect conditional feature distributions, which can be challenging in practice. Recent advancements in generative modeling can facilitate this. For tabular data, which may consist of both categorical and continuous features, the adversarial random forest (ARF) stands out as a generative model that can generate on-manifold data points without requiring intensive tuning efforts or computational resources, making it a promising candidate model for subroutines in XAI methods. This paper proposes cARFi (conditional ARF feature importance), a method for measuring conditional feature importance through feature values sampled from ARF-estimated conditional distributions. cARFi requires only little tuning to yield robust importance scores that can flexibly adapt for conditional or marginal notions of feature importance, including straightforward extensions to condition on feature subsets and allows for inferring the significance of feature importances through statistical tests.

Code — https://github.com/bips-hb/cARFi_paper

1 Introduction

Explainable artificial intelligence (XAI) aims to shed light on the opaque behavior of machine learning algorithms, which includes assessing the importance of features for a predictive algorithm. Model-agnostic post hoc methods attribute scores to input features according to their relevance for the prediction in an arbitrary, already fitted supervised machine learning model (Molnar 2020; Murdoch et al. 2019). Refined conceptualizations include, for example, methods aiming for insights on the prediction of individual observations, like Shapley additive explanations (Lundberg and Lee 2017), or a feature importance focus on the model's overall behavior, yielding global-level explanations.

A crucial distinction in feature importance concepts is between conditional and marginal viewpoints (Strobl et al. 2008; Watson and Wright 2021): Marginal feature importance evaluates a feature's impact irrespective of other features included in the model, whereas conditional feature importance takes the predictive information of other features into account. The presence of dependency structures, which real-world datasets frequently exhibit, plays a pivotal role in this distinction because a feature's impact on the prediction *given*, i.e., on top of the predictive information provided by correlated features, alters the importance score attributed (Watson and Wright 2021). To that end, it is worth highlighting that also in-between measures exist, e.g., conditioning on only few features as in relative feature importance (König et al. 2021) or having a parameter that alters the "strictness" of the conditional distribution regarding the degree of extrapolation allowed (see further our discussion on tree-depth within cARFi in Secs. 3 and 4.1). Overlooking such nuances poses a pitfall that can result in misinterpretations (Molnar et al. 2022). However, determining marginal, conditional, or in-between notions of feature importance in practice necessitates suitable and ready-to-use methods.

Especially methods for model-agnostic post hoc conditional feature importance measurement on a global level of explanation urge for improvement. Recap that such methods aim to evaluate the change in a model's performance when erasing the feature of interest's predictive information from the dataset, while accounting for the information provided by other features (Fisher, Rudin, and Dominici 2019; Watson and Wright 2021). Approaches may suggest removing the feature of interest from the model altogether, but this commonly involves computationally expensive model refits (Lei et al. 2018). More commonly, approaches rely on strategies that remove the dependency on the target while maintaining the features' joint distributional behavior. These methods aim to evaluate the change in predictive performance when replacing the feature of interest's values x_j with $\tilde{x}_j$ that respect the conditional distribution $p(x_j|\mathbf{X}_C = \mathbf{x}_C)$, where $\mathbf{X}_C$ indicates the features to con-

[*]These authors contributed equally.
[†]M. Loecher and M.N.Wright share last authorship.

dition on (typically all features in the model except for X_j). To that end, conditional feature importance approaches may assume that the data is separable into subgroups (Molnar et al. 2023) or draw on related frameworks such as knockoffs (Watson and Wright 2021) to circumvent the direct modeling of $p(x_j|\mathbf{X}_C = \mathbf{x}_C)$, which is challenging in practice.

Generative modeling involves generating new data samples by learning the joint distribution of $\mathbf{X}$. Beyond that, some generative models have the ability to also derive (arbitrary) conditional densities $p(x_j|\mathbf{X}_C = \mathbf{x}_C)$ from the joint density without model retraining, which is particularly useful when customizing conditioning sets, e.g., as in relative feature importance. Further, generative models enable the sampling of large amounts of data, which can increase the robustness of feature importance scores. However, generative models are typically computationally intensive and require high effort in parameter tuning (Goodfellow et al. 2014; Xu et al. 2019). To leverage the potential of generative models as a subroutine in feature importance measurement for a broader application range, it is crucial to focus on generative models suitable for continuous and categorical features (mixed data), which require minimal tuning and computational resources.

This paper combines recent advancements in generative modeling with XAI and introduces conditional ARF feature importance (cARFi), a method for conditional feature importance measurement using adversarial random forests (ARF, Watson et al. 2023). An ARF is an off-the-shelf generative model that is particularly suitable for mixed tabular data, executing in due runtime with comparably little tuning efforts, and allows to efficiently estimate and sample from conditional densities. cARFi is robust, flexibly adjustable for various conditioning sets and even allows for applying statistically valid inference to test for nonzero feature importance.

The remainder of this paper is structured as follows: Sec. 2 introduces relevant background and related work. Then, in Sec. 3, the proposed method is described in more depth and theoretical properties are discussed. Sec. 4 contains a proof of concept, followed by a simulation study and real data example to evaluate the performance of cARFi under different conditioning sets. Finally, we conclude our findings (Sec. 5) and discuss results in Sec. 6.

2 Background and Related Work

Conditional Feature Importance The measurement of conditional feature importance can be approached from various standpoints. Frameworks may focus on leveraging model-specific traits (Strobl et al. 2008), study the topic across entire model classes (Fisher, Rudin, and Dominici 2019), or compare model refits that omit the feature of interest (leave-one-covatiate out; LOCO, Lei et al. 2018). This paper focuses on model-agnostic methods for a single given model, for which several approaches exist.

SAGE values (Covert, Lundberg, and Lee 2020) draw on the game-theoretic concept of Shapley values and, in theory, can measure global conditional feature importance by analyzing a feature's additional contribution across various feature subgroups. In practice, however, approximations that rely on marginal sampling, as in KernelSHAP (Lundberg and Lee 2017), are frequently used instead of the challenging sampling of values from conditional distributions. As a consequence, this turns the method into a marginal feature importance measure.

Another approach is analyzing the effect of perturbing feature values within conditional subgroups (CS, Molnar et al. 2023). CS works analogous to permutation feature importance (PFI, Breiman 2001), but respects the data manifold. The method requires the data to be separable in suitable subgroups that inherent sufficient amounts of data, for which Molnar et al. (2023) suggest a tree-based search procedure. While CS, in principle, can also work with conditioning on feature subsets, it does require rerunning the entire procedure for each conditioning set of features. Further, CS evaluates permutations of feature values within subgroups but does not allow for estimating and sampling from the respective conditional distributions.

A method that, in contrast to SAGE and CS, accommodates a procedure for testing nonzero feature importance is the conditional predictive impact (CPI, Watson and Wright 2021). This method relies on model-X knockoffs (Candès et al. 2018), which are synthetic data samples that imitate the original data and satisfy desirable properties. While knockoff samplers are also generative models, they impose additional properties on the synthesized data, such as equality in the joint distribution of original and generated data under swapping operations (Candès et al. 2018), which may be overly strict, and hence disadvantageous for certain tasks (Blesch, Wright, and Watson 2023). Further, knockoff-based procedures may yield unstable results and multiple knockoff imputation might be favorable (Gimenez and Zou 2019), yet computationally intensive (see further Sec. 3).

Apparent stumbling blocks of the conditional feature importance methods discussed above are (1) access to conditional distributions and (2) ensuring that sufficient amounts of data are available either in the subgroups or generated by knockoffs, to calculate meaningful and robust feature attributions. cARFi tackles both aspects simultaneously, allowing for the generation of multiple values from conditional distributions in due time. Further, cARFi allows for selecting a hyperparameter that balances the strictness of the conditioning set, making it a flexible measure for both conditional, marginal, and in-between feature importance.

Relative Feature Importance In relative feature importance, the evaluation is conditional on *few* other features in the model (König et al. 2021). In a certain sense, this concept tones down conditional feature importance which accounts for *all* other features in the model.

Even though relative feature importance is barely discussed in XAI, this concept is highly relevant in practice. For example, if the underlying causal structure includes a collider, e.g., both X and Y share a common effect on Z in the causal graph $X \rightarrow Z \leftarrow Y$. Conditioning on the collider Z creates a non-causal association between X and Y, which can mislead interpretations. Similarly, features may be considered "good" or "bad" controls, which should or should not be included as control variables in a model (Cinelli, Forney, and Pearl 2024). Translating this to XAI, users may

want to control for only few features, for example, if they are known to be confounding features, but ignore the effects of other features on the importance.

Conditional feature importance measures can adapt to the concept of relative importance, yet current methods are impractical for this. For example, CS needs to find updated subgroups, and knockoff samplers within CPI must be refit for altered conditioning sets due to the knockoff's properties. As noted in König et al. (2021), any procedure yielding feature values from the respective conditional distribution works as a subroutine in relative feature importance measurement and may even respect features absent during model training. The challenge, however, is finding a synthesizer that does not rely on strict parametric assumptions, requires refits for changing conditioning sets, or involves vast computational and tuning resources for generating data generally. With recent advancements in generative modeling, this has become feasible and the method proposed in this paper can adapt flexibly to customized conditioning sets without requiring auxiliary calculations.

Generative Modeling The field of generative modeling is concerned with building models that can generate data instances $\tilde{\mathbf{X}}$ that follow the same joint distribution as some given data matrix $\mathbf{X}$. Generative models rapidly advance a variety of machine learning-related tasks, such as text generation with large language models (OpenAI 2023), and offer promising lines of research for XAI.

Tabular data has unique characteristics, such as mixed features, that require careful consideration (Borisov et al. 2024). Typically, generative models are based on deep learning architectures, such as in generative adversarial networks (Goodfellow et al. 2014), normalizing flows (Rezende and Mohamed 2015), variational autoencoders (Kingma and Welling 2014) and diffusion probabilistic models (Ho, Jain, and Abbeel 2020), often combined with transformer-based architectures (Vaswani et al. 2017); for an overview, see Bond-Taylor et al. (2021); Foster (2022). Such architectures can be difficult for reaching convergence, and are typically computationally demanding and tuning intensive. Adaptions to tabular data exist (Xu et al. 2019), yet, tree-based methods may be better suited for tabular data since they require little tuning and naturally handle mixed features (Borisov et al. 2024; Grinsztajn, Oyallon, and Varoquaux 2022). Attempts to generate data with trees in a more convenient, straightforward manner are thus promising and several such methods have been proposed recently (Correia, Peharz, and de Campos 2020; Nock and Guillame-Bert 2023).

From such approaches, ARF (Watson et al. 2023) stands out as a fast and off-the-shelf method for generating high-quality tabular data, requiring only few efforts in tuning and computational resources. Further, it allows for estimating and sampling from the joint as well as conditional distributions, which is essential for the task at hand. The unconditional ARF procedure works as follows:

1. Fit unsupervised random forest (Shi and Horvath 2006): First, permute feature values in the given dataset $\mathbf{X}$ randomly across instances to create naive synthetic dataset $\tilde{\mathbf{X}}$. Then, fit a random forest $\hat{f}^0$ to distinguish instances from $\mathbf{X}$ and $\tilde{\mathbf{X}}$ (labeled accordingly), where splits in the forest's trees pick up the data's dependency structure.
2. If the accuracy of $\hat{f}^0$ is above 50%, new synthetic data is sampled from the leaves of forest $\hat{f}^0$ (generator step) and a new random forest $\hat{f}^1$ is fit to classify real and synthetic data (discriminator step).
3. Data generation and discrimination is continued for k iterations until the accuracy of $\hat{f}^k$ drops down to 50% or below. This indicates that the algorithm has converged, implying that all feature dependencies have been learned and features are mutually independent in the leaves.
4. FORDE step (density estimation): The estimated joint density $\hat{p}_{\text{ARF}}$ can – thanks to the mutual independence assumption of features within the leaves – be formulated as a mixture of products $\hat{p}_l$ of univariate densities $\hat{p}_{lj}$ for leaf l and feature j, which can be estimated with any arbitrary univariate density estimator within the random forest's leaves, weighted by the share of real data π_l that falls into l:
$$\hat{p}_{\text{ARF}}(\mathbf{x}) = \sum_l \pi_l \hat{p}_l(\mathbf{x}) = \sum_l \pi_l \prod_j \hat{p}_{lj}(x_j). \quad (1)$$
5. FORGE step (data generation): Synthetic data is generated by drawing a leaf l from the random forest with probability π_l and then sampling from the estimated univariate densities $\hat{p}_{lj}$ within that leaf.

An ARF model fitted once can generate arbitrary amounts of data from the estimated joint distribution $\hat{p}_{\text{ARF}}(\mathbf{x})$ and derived conditional distributions, as further discussed in Sec. 3. This enables generating multiple values from arbitrary conditional distributions feasible in due time, which positions ARF as particularly suitable to serve as a subroutine in conditional and relative feature importance measurement.

3 Methods

We propose conditional ARF feature importance (cARFi), a robust measure for conditional and relative feature importance that leverages recent advancements in generative modeling to XAI. cARFi relies on the concept of feature importance measurement that links it to changes in the performance of the prediction model $\hat{f}$ (Breiman 2001; Covert, Lundberg, and Lee 2020; Watson and Wright 2021). To attribute importance score FI_j to the feature of interest X_j, we evaluate the difference in expected loss ℓ when removing the effect of X_j on the target Y using the modified dataset $\mathbf{X}^*$. That is, FI_j evaluates $\mathbb{E}[\ell(\hat{f}(\mathbf{X}^*), Y)] - \mathbb{E}[\ell(\hat{f}(\mathbf{X}), Y)]$. This quantity can be approximated empirically by averaging the instance-wise loss differences across the entire dataset $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^N$:

$$\widehat{\text{FI}}_j = \frac{1}{N} \sum_{i=1}^{N} \ell\left(\hat{f}(\mathbf{x}^{*(i)}), y^{(i)}\right) - \ell\left(\hat{f}(\mathbf{x}^{(i)}), y^{(i)}\right). \quad (2)$$

Recap that $\hat{f}$ typically cannot handle missing values directly, hence, we have to ensure that the dimension of $\mathbf{X}^*$ matches that of $\mathbf{X}$ instead of analyzing $\mathbf{X}^* := \mathbf{X} \setminus X_j$ directly. To remove the information contained in X_j on the

target Y, we therefore aim to replace values of X_j with $\tilde{x}_j$ such that $\tilde{X}_j$ is independent of Y. While marginal feature importance focuses on $\tilde{X}_j \perp\!\!\!\perp \{Y, \mathbf{X} \setminus X_j\}$, for conditional and relative feature importance, this independence must be present considering the features in a conditioning set $\mathbf{X}_C$, i.e., $\tilde{X}_j \perp\!\!\!\perp Y|\mathbf{X}_C$. Further, the conditional distributions of $X_j|\mathbf{X}_C$ and $\tilde{X}_j|\mathbf{X}_C$ need to be equal such that the generated data instances are on the data manifold. In sum, we aim for sampling values $\tilde{x}_j^{(i)} \sim p(x_j|\mathbf{X}_C = \mathbf{x}_C)$ that satisfy $\tilde{X}_j \perp\!\!\!\perp Y|\mathbf{X}_C$.

Conditional ARF Feature Importance An ARF can generate such values straightforwardly because of its ability to derive conditional distributions $p(x_j|\mathbf{X}_C = \mathbf{x}_C)$ from the learned joint distribution, with conditional independence to Y satisfied automatically as long as the ARF is fit with only $\mathbf{X}$ (excluding Y). As discussed similarly in Dandl et al. (2024), we can leverage the unique traits of ARFs for conditional and relative feature importance measurement as follows:

- Once $\hat{p}_{\text{ARF}}$ is estimated, ARF allows us to derive estimated conditional densities $\hat{p}_{\text{ARF}}(x_j|\mathbf{X}_C = \mathbf{x}_C)$ for fixed values $\mathbf{x}_C$ with arbitrary conditioning sets C without the need of refitting the ARF:

$$\hat{p}_{\text{ARF}}(x_j|\mathbf{X}_C = \mathbf{x}_C) = \sum_l \pi_l' \hat{p}_{lj}(x_j) \tag{3}$$

  with updated weights $\pi_l' := \pi_l \frac{\hat{p}_l(\mathbf{x}_C)}{\hat{p}_{\text{ARF}}(\mathbf{x}_C)}$. This includes the case of $C = \{1, \ldots, p\} \setminus \{j\}$ for any feature of interest j and for arbitrary subsets of features $\mathbf{X}_S$, $S \subset \{1, \ldots, p\}$ to condition on, as required in relative feature importance measurement.
- With the conditional densities derived, arbitrary amounts of feature values can be sampled in due time as this does not require rerunning the ARF procedure. We can use this computationally cheap sampling to increase the stability (robustness) of the feature importance measure: we propose to sample R values of $\tilde{x}_j^{(i)}$ for each observation i, averaging the change in loss across those replicates before calculating the model-wide feature importance score across all N observations.

In summary, the cARFi method extends the PFI framework by utilizing repeated conditional sampling with ARF. The whole procedure is described in Algorithm 1.

Effect of Hyperparameters Since ARFs solely approximate dependencies across features by their random forests' splits, as reflected in the local independence assumption in Equation (1), the tree depth has decisive implications. Growing deep trees is desirable to learn the correlations present in the real data accurately and thus yield apt estimates for the underlying joint and conditional data distributions. However, the resulting hyperrectangles defined by the random forest's leaves matching the conditions might become very small, leading to only little variation in feature values sampled when calculating cARFi. Consequently, the changes in the loss might become minimal, resulting in lowered power of statistical tests indicating important features. On the contrary, when growing shallow trees, which yield large hyperrectangles, the estimated conditional distributions might not approximate the real distributions well and the statistical test applied might reject too many null hypotheses, inflating the type I error possibly above the predefined significance level.

Algorithm 1: cARFi

Input: $(\mathbf{X}^{\text{train}}, Y^{\text{train}})$, $(\mathbf{X}^{\text{test}}, Y^{\text{test}})$, learner f, feature (set) of interest j, conditioning set C, ARF procedure a, loss function ℓ, number of replicates R

1: learn $\hat{f} \leftarrow f(\mathbf{X}^{\text{train}}, Y^{\text{train}})$
2: fit ARF $\hat{a} \leftarrow a(\mathbf{X}^{\text{train}})$ and estimate density $\hat{p}_{\hat{a}}$
3: sample R feature values for each test instance i:
   for each $i \in [N], r \in [R]$: $\tilde{\mathbf{X}}_j^{\text{test},i(r)} \sim \hat{p}_{\hat{a}}(x_j|\mathbf{X}_C^{\text{test},i})$
4: define $\tilde{\mathbf{X}}^{\text{test},i(r)} := \{\tilde{\mathbf{X}}_j^{\text{test},i(r)}, \mathbf{X}_{-j}^{\text{test},i}\}$ and calculate instance-wise loss difference w.r.t. j:
   $\Delta_j^i \leftarrow \frac{1}{R} \sum_{r=1}^{R} \ell(\hat{f}(\tilde{\mathbf{X}}^{\text{test},i(r)}, Y)) - \ell(\hat{f}(\mathbf{X}^{\text{test},i}, Y))$
5: calculate $\widehat{\text{cARFi}}_j \leftarrow \frac{1}{N} \sum_{i=1}^{N} \Delta_j^i$

Output: $\widehat{\text{cARFi}}_j$

The main parameter controlling for this phenomenon is the minimum node size, i.e., the minimum amount of data points that have to be contained in the terminal nodes of the ARF's trees. In a simulation study, we show how altering this parameter shifts the distribution under the null hypothesis (see Sec. 4.1 and Appendix 2.1). This provides a new perspective on feature importance measurement since it offers the opportunity of in-between feature importance. That is, the notion of marginal and conditional feature importance can be shifted effortlessly by adjusting this parameter, offering new pathways for feature importance measurement.

Another important parameter for this is the finite bounds argument, that, when set to 'local', replaces infinite bound values with the empirical bounds within the leaves. This takes into account the poor extrapolation ability of random forests and helps to keep the generated data on the manifold.

Statistical Testing A crucial property of the cARFi estimator is that it is asymptotically normally distributed, thus providing valid statistical inference for the importance of a set of features S conditioned on a set of other features C. The derivation is based on the detailed explanations for the CPI concept proposed by Watson and Wright (2021) and utilizes the fact that the empirical risk estimator is asymptotically normally distributed. For a loss function ℓ that acts instance-wise, we define the following random variable:

$$\Delta = \frac{1}{R} \sum_{r=1}^{R} \ell\left(\hat{f}(\{\tilde{\mathbf{X}}_S^{(r)}, \mathbf{X}_C\}), Y\right) - \ell(\hat{f}(\mathbf{X}), Y). \tag{4}$$

Assuming that the distribution of $\tilde{\mathbf{X}}_S$ learned by the ARF follows the actual distribution of $\mathbf{X}_S|\mathbf{X}_C$, the samples $\Delta_1, \ldots, \Delta_n$ are also i.i.d. Additionally, drawing R imputations does not affect the i.i.d. condition of the averaged loss

values of an instance. Thus, with a larger number of samples N, our cARFi estimator $\widehat{\text{cARFi}}_S = \frac{1}{N}\sum_{i=1}^{N}\Delta_i$ converges in probability to a Gaussian distribution by the central limit theorem. Consequently, this allows for statistical significance tests as described in Watson and Wright (2021) (e.g., paired t-test and Fisher exact test). However, we must note a limitation on the consistency of this convergence: the convergence is theoretically guaranteed only for models $\hat{f}$ with a finite VC dimension (Vapnik and Chervonenkis 1971). Nonetheless, our simulations with support vector machines, which have an infinite VC dimension, showed consistent results (see Sec. 4.1). Additionally, uncertainties and biases in the tests may arise because the quality of the conditional distributions learned by the ARF and potential uncertainties in the machine learning model $\hat{f}$ are disregarded.

4 Evaluation

We evaluate the performance of cARFi on both simulated and real data. First, we demonstrate that cARFi allows for valid inference procedures and achieves high power in testing for nonzero feature importance. Next, we compare the performance of cARFi to that of CPI, its closest competitor in testing for conditional feature importance, and evaluate cARFi-based relative feature importance, both by drawing on simulation studies from previous literature. Finally, we illustrate cARFi's feature attributions to those of competing methods for a real data example.

4.1 Proof of Concept

To validate that the performance of cARFi is as expected from the theoretical considerations outlined in Sec. 3, we draw on a simulation setup established in prior literature. Using the setup of Watson and Wright (2021), we demonstrate that cARFi enables powerful and statistically valid testing for nonzero conditional feature importance.

In detail, for $M = 10,000$ replicates, we generate $N = 1,000$ instances of features $\mathbf{X} = X_1, \ldots, X_{10}$ from a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$, where $\Sigma_{ij} = 0.5^{|i-j|}$. Using effect sizes $\beta = (0.0, 0.1, \ldots, 0.9)$ and additive noise $\epsilon \sim \mathcal{N}(0, 1)$, we construct target variable Y according to two different settings: (1) linear setting: $\mathbf{Y} = \beta\mathbf{X} + \epsilon$; (2) non-linear setting: $\mathbf{Y} = \beta\mathbf{X}^{'} + \epsilon$, where $x_{ij}^{'} = 1$ if $\mathbf{\Phi}^{-1}(0.25) \leq x_{ij} \leq \mathbf{\Phi}^{-1}(0.75)$, else $x_{ij}^{'} = -1$.

We fit several prediction models $\hat{f}$ to this data, including a (feedforward) neural network, support vector machine, random forest, and linear model. Subsequently, we use the mean squared error to assess ℓ and thus obtain test statistics.

From Fig. 1, we can see that cARFi acts as expected. At effect size 0, the rejection rate is at 5%, effectively controlling type I error at the nominal level of 5%. At positive effect sizes, power increases with effect size and all learners reach 100% power, with the exception of the linear model on nonlinear data. Fig. 1 shows results for a minimum leaf size of 20. In addition, Appendix 2.1 gives results for minimum leaf sizes of 2, 5, 10, 50 and 100, and further details feature importance values for the different effect sizes and empirical distributions of the test statistics. These results show that very deep trees (small leaf size) lead to slightly conservative results, i.e., type I errors below the nominal level and lower power, while shallow trees (large leaf size) show slightly inflated type I errors and higher power.

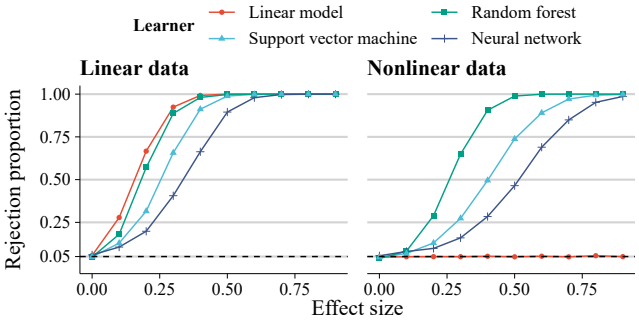

Figure 1: Rejection rates of one-sided paired t-tests at $\alpha = 0.05$ to detect relevant features at different effect sizes, i.e., type I error rates at effect size 0 and power at effect size >0.

4.2 Simulation Study

Next, we compare the performance of cARFi to related methods in terms of time and statistical power using a simulation setting derived in previous literature.[1] In addition, we highlight differences in feature importance measures for various conditioning sets in a simulated and real-world setting.

Mixed Data We evaluate the directed acyclic graph (DAG) introduced in scenario (III) in (Blesch, Watson, and Wright 2024) in the mixed data setting: X_2, X_4 are Gaussian, X_1 and X_3 are categorical with 10 levels. Considering this DAG, the outcome Y is conditionally independent of X_1 (X_2) given X_4 (X_3), or, more formally: $X_1 \perp\!\!\!\perp Y \mid \{X_2, X_3, X_4\}$ and $X_2 \perp\!\!\!\perp Y \mid \{X_1, X_3, X_4\}$, whereas $X_3 \not\perp\!\!\!\perp Y \{X_1, X_2, X_4\}$ and $X_4 \not\perp\!\!\!\perp Y \mid \{X_1, X_2, X_3\}$. Therefore, a conditional feature importance measure should only attribute nonzero importance to variables X_3, X_4, but not to X_1, X_2. The simulation analyzes the rejection rate of the null hypothesis based on the t-test with varying sample sizes for CPI with naive dummy-encoded Gaussian knockoffs, CPI with sequential knockoffs, and cARFi. Additionally, we generate both a single knockoff or ARF-sample ($R = 1$) and multiple ones ($R = 20$), which at this point hasn't been investigated before for the CPI framework. The results using a minimum node size of 20 are shown in Fig. 2.

We observe that cARFi behaves very similarly to CPI with sequential knockoffs in the sense that the type I error is mostly controlled for the conditionally unimportant features (X_1 and X_2) and shows good power for the important features (X_3 and X_4). Consistent with the findings of Blesch, Watson, and Wright (2024), it is notable that the naive dummy-encoded Gaussian knockoffs only correctly identify the categorical variable X_3 as significant at very high sample sizes. Considering the comparison with the

[1] We evaluate the performance of competing methods on simulation setups used in previous literature to promote a direct and fair comparison that is not tailored to advantageously present cARFi.

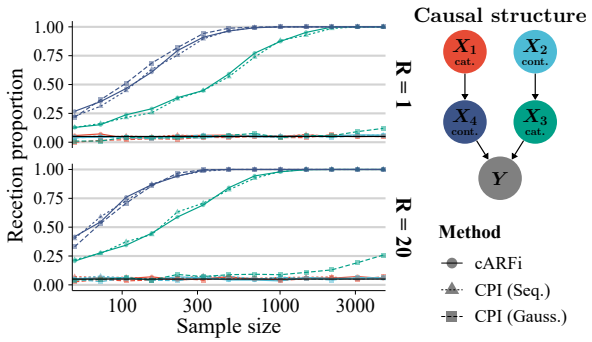

Figure 2: Rejection rates of one-sided paired t-tests at $\alpha = 0.05$ to detect relevant features, i.e. power and type I error rates, across 500 simulation runs. X_1, X_3 are 10-level categoricals, X_2, X_4 are Gaussian. Effect size $\beta = 0.5$ and random forest prediction model.

replicates R shown in the lower plot of Fig. 2 reveals gains in robustness. Using $R = 20$ instead of a single knockoff or ARF-sample, both power and stability of the type I error improve, especially at lower sample sizes. At this point, we want to highlight the advantage of cARFi compared to sequential knockoffs, as generating ARF samples is slightly faster with $R = 1$ and much faster with $R = 20$ than creating the corresponding knockoffs (see Appendix 2.3).

Impact of the Conditioning Set Here, we present results for a modified version of the DAG investigated in section (VI) in (König et al. 2021) as displayed in the leftmost panel of Fig. 3 for two fundamentally different models. For a random forest, the marginal measures PFI and SAGE "leak" importance from X_3 and X_4 to the correlated features X_1, X_2 and X_5 while the linear model inherently conditions on all other features. When we explicitly condition on all other features, the rightmost panel reveals a few interesting insights: cARFi's estimates are close to the "truth" (in a relative sense), whereas CPI assigns zero attribution to X_3 and CS seems unable to properly adjust its marginal measures. Selective conditioning on X_1, X_2, respectively, lowers the score for X_3: $\widehat{\text{PFI}}_3 > \widehat{\text{cARFi}}_3(X_1) > \widehat{\text{cARFi}}_3(X_2) \approx \widehat{\text{cARFi}}_3(X_1, X_2)$. Similarly, conditioning on X_5 (X_1) also lowers the contribution of X_3 (X_4): $\widehat{\text{PFI}}_3 > \widehat{\text{cARFi}}_3(X_5)$ and $\widehat{\text{PFI}}_4 > \widehat{\text{cARFi}}_4(X_1)$. For completeness, we include the various conditional independence relations as well as consistency checks and other details in Appendix 1.

4.3 Bike-Sharing Dataset

Finally, we evaluate the behavior of cARFi under different conditioning sets in a real-world setting using the widely used bike-sharing dataset (Fanaee-T and Gama 2013). The dataset contains hourly records of bike rentals and includes seasonal and meteorological information such as season, weekday, humidity, and temperature. Due to the number of possible combinations of the variables, we limit our analysis to hour and temperature, which are marginally two of the most important features for the trained random forest. Even though we don't know the exact feature importance or the underlying DAG for the causal associations, we can infer some relationships based on natural and logical laws. For example, we know that the season affects the temperature, as it is colder in winter than in summer.

We train a random forest on two-thirds of the 8 645 instances and use the remaining as a holdout for the XAI method. We repeat this setting 50 times and apply PFI (marginal measure, i.e., no conditions), and cARFi with different conditioning sets. For cARFi, we use a minimum node size of 20, $R = 5$, and the root mean squared error (RMSE) as a loss function. The results are presented in Fig. 4.

The left panel of Fig. 4 shows the importance values from PFI as a marginal feature importance measure and cARFi conditioned on all other features. We observe that cARFi often results in a decrease in importance as some effects can be explained through the other features, i.e., they are not conditionally independent. For example, *Hour*, *Temperature*, and *Humidity* show a considerable drop in importance when comparing PFI to the conditional variants. In contrast, the binary variable *Workday* shows only a minor change, which aligns with the intuitive understanding that whether a day is a workday or a holiday is not influenced by seasonal, temporal, or weather conditions. The other two panels illustrate the varying variable importance under different conditions. For instance, we observe that the importance of both *Hour* and *Temperature* increases when conditioned only on *Workday*. This is likely due to the interaction between these variables and *Workday* (Hiabu, Meyer, and Wright 2023). For example, on holidays, bikes are rented at different times compared to weekdays when people commute to work in the morning. Additionally, we observe that the effect of *Hour* on the number of rented bikes decreases when conditioned on *Temperature*, and even more when also conditioned on *Season*. This occurs because the time of day influences temperature, so the effect of *Hour* through *Temperature* on the bike rentals is likely partially absorbed by this condition. The reduction becomes even more apparent when *Season* is added, as it captures broader trends like typical temperature ranges and daylight hours, which are closely tied to both time of day and bike usage, further decreasing *Hour*'s importance for the prediction. We observe a similar effect with *Temperature* when we condition on *Season*: Since these two are strongly correlated, conditioning on *Season* alone drastically reduces *Temperature*'s importance in predicting bike rentals. This effect is further amplified when we also condition on *Hour*, as it correlates with daily temperature changes, capturing much of the variability that *Temperature* would otherwise explain.

5 Conclusion

This paper presents cARFi, a method that leverages a straightforward generative model for the measurement of conditional and relative feature importance. cARFi offers robust and flexible calculation of feature importance in a model-agnostic way, and directly handles mixed data thanks to the random forest based generative procedure. The procedure can adapt to feature subset conditioning without having to rerun the procedure and further introduces an opportunity to smoothly shift between notions of marginal, in-between

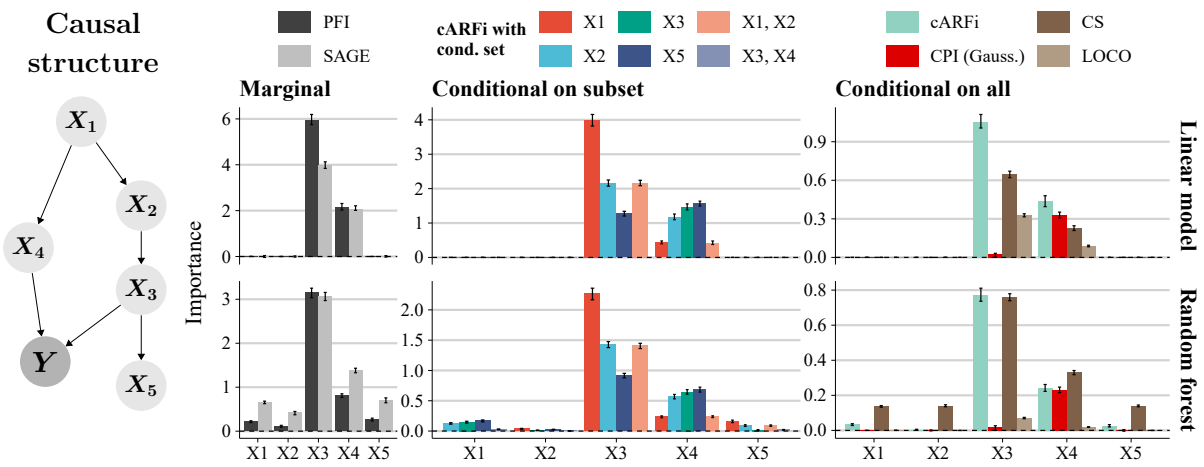

Figure 3: Marginal vs. conditional feature importances for a linear model (LM, upper row) and a random forest (RF, lower row). While the LM coefficients $\beta_1, \beta_2, \beta_5$ are close to zero, the RF assigns marginal importance to X_1, X_2, X_5 due to their strong correlation with X_3, X_4. cARFi resolves these indirect influences by conditioning on the respective feature subsets.

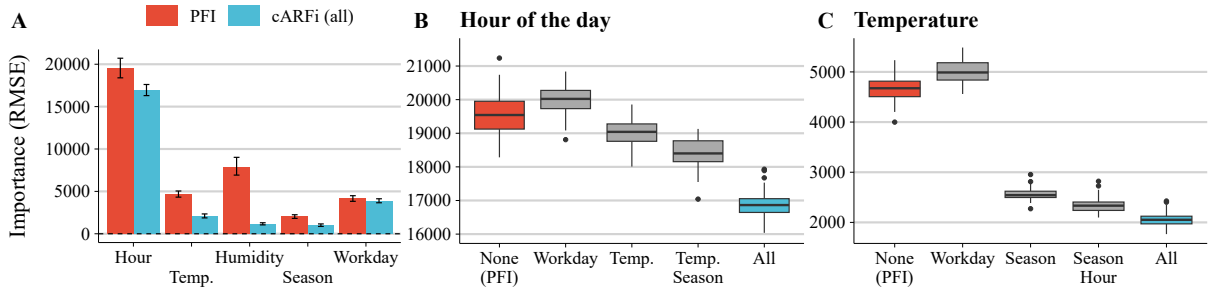

Figure 4: Feature importance values for the variables from the real-world bike rental example. Panel A: PFI and cARFi (conditioned on all other ones) values of all included variables in the random forest model. Panels B and C: cARFi values for *Hour* and *Temperature* for selected conditioning sets, respectively. The RMSE is used as the loss function and 50 repetitions.

and conditional feature importance measurement. In both simulated and empirical examples, the method demonstrates competitive results, and the ease of application is particularly appealing for empirical usage. An implementation of cARFi in the `R` programming language and code for reproducing the results of this paper is available on the corresponding GitHub repository as linked on the first page.

6 Discussion

Even though the robustness of cARFi is analyzed regarding the number of sampled instances, see Fig. 2, it may be further studied. For example, by systematically varying the sample size, dimensionality, underlying distribution of the data and choice of hyperparameters, such as the minimum node size within the ARF subroutine. Delimiting the scope of the paper, we leave such analyses for future research, yet want to highlight the necessity of providing users with comprehensive studies on such considerations. However, the relevance of parameters to focus on will highly depend on the field of application and hence, should be investigated for the specific use case at hand.

That said, studies showcasing the empirical use of cARFi through applied use cases are highly desirable. This paper focuses on the methodological proposition, hence introduces, analyzes and discusses cARFi from an abstract, general standpoint. However, cARFi is designed to facilitate conditional and relative feature importance measurement in real-world applications. Therefore, we encourage applied researchers to challenge the usefulness of cARFi in practice.

In future work, advancements in the rapidly changing fields of generative modeling, XAI and closely related fields could be taken into account and innovative methods developed accordingly. This includes propositions that bridge other XAI methods with generative modeling, e.g., exploiting the fast subset conditioning of ARF for methods such as SHAP (Lundberg and Lee 2017) and SAGE (Covert, Lundberg, and Lee 2020) that heavily rely on such operations. In principle, any well-fitted generative model that can synthesize values in accordance to the requested conditional distributions can work in such subroutines. As another example, cARFi's conditional independence testing procedure may facilitate applying algorithms in causal structure learning. Hence, cARFi can serve as a starting point for future research to propose new algorithms in various academic fields.

Acknowledgements

This work was supported by the German Research Foundation (DFG), grant numbers 437611051 and 459360854 and by the U Bremen Research Alliance/AI Center for Health Care, financially supported by the Federal State of Bremen. We thank Sophie Langbein for valuable discussions. Experiments were run on the Beartooth Computing Environment (University of Wyoming Advanced Research Computing Center 2018).

References

Blesch, K.; Watson, D. S.; and Wright, M. N. 2024. Conditional feature importance for mixed data. *AStA Advances in Statistical Analysis*, 108: 259–278.

Blesch, K.; Wright, M. N.; and Watson, D. 2023. Unfooling SHAP and SAGE: knockoff imputation for Shapley values. In *World Conference on Explainable Artificial Intelligence*, 131–146. Springer.

Bond-Taylor, S.; Leach, A.; Long, Y.; and Willcocks, C. G. 2021. Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 7327–7347.

Borisov, V.; Leemann, T.; Seßler, K.; Haug, J.; Pawelczyk, M.; and Kasneci, G. 2024. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6): 7499–7519.

Breiman, L. 2001. Random forests. *Machine Learning*, 45(1): 5–32.

Candès, E.; Fan, Y.; Janson, L.; and Lv, J. 2018. Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3): 551 – 577.

Cinelli, C.; Forney, A.; and Pearl, J. 2024. A crash course in good and bad controls. *Sociological Methods & Research*, 53(3): 1071–1104.

Correia, A.; Peharz, R.; and de Campos, C. P. 2020. Joints in random forests. *Advances in Neural Information Processing Systems*, 33: 11404–11415.

Covert, I.; Lundberg, S. M.; and Lee, S.-I. 2020. Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33: 17 212–17 223.

Dandl, S.; Blesch, K.; Freiesleben, T.; König, G.; Kapar, J.; Bischl, B.; and Wright, M. N. 2024. CountARFactuals–generating plausible model-agnostic counterfactual explanations with adversarial random forests. In *World Conference on Explainable Artificial Intelligence*, 85–107. Springer.

Fanaee-T, H.; and Gama, J. 2013. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2(2–3): 113–127.

Fisher, A.; Rudin, C.; and 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.

Foster, D. 2022. *Generative deep learning: Teaching Machines To Paint, Write, Compose, and Play*. O'Reilly Media, Inc., 2^{nd} edition.

Gimenez, J. R.; and Zou, J. 2019. Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. In *Proceedings of the* 22^{th} *International Conference on Artificial Intelligence and Statistics*, 2184–2192. PMLR.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Grinsztajn, L.; Oyallon, E.; and Varoquaux, G. 2022. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35: 507–520.

Hiabu, M.; Meyer, J. T.; and Wright, M. N. 2023. Unifying local and global model explanations by functional decomposition of low dimensional structures. In *Proceedings of the* 26^{th} *International Conference on Artificial Intelligence and Statistics*, 7040–7060. PMLR.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational Bayes. In *Proceedings of the* 2^{nd} *International Conference on Learning Representations*.

König, G.; Molnar, C.; Bischl, B.; and Grosse-Wentrup, M. 2021. Relative feature importance. In *25th International Conference on Pattern Recognition (ICPR)*, 9318–9325. IEEE.

Lei, J.; G'Sell, M.; Rinaldo, A.; Tibshirani, R. J.; and Wasserman, L. 2018. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30: 4768–4777.

Molnar, C. 2020. Interpretable Machine Learning. Available at: https://christophm.github.io/interpretable-ml-book/ (Accessed: December 28, 2023).

Molnar, C.; Casalicchio, G.; Grosse-Wentrup, M.; and Bischl, B. 2022. General pitfalls of model-agnostic interpretation methods for machine learning models. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, 39–68. Springer Nature.

Molnar, C.; König, G.; Bischl, B.; and Casalicchio, G. 2023. Model-agnostic feature importance and effects with dependent features: A conditional subgroup approach. *Data Mining and Knowledge Discovery*, 1–39.

Murdoch, W. J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; and Yu, B. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44): 22071–22080.

Nock, R.; and Guillame-Bert, M. 2023. Generative forests. *ArXiv Preprint arXiv:2308.03648*.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv Preprint arXiv:2303.08774*.

Rezende, D.; and Mohamed, S. 2015. Variational inference with normalizing flows. In *Proceedings of the* 32^{th} *International Conference on Machine Learning*, 1530–1538. PMLR.

Shi, T.; and Horvath, S. 2006. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 15(1): 118–138.

Strobl, C.; Boulesteix, A.-L.; Kneib, T.; Augustin, T.; and Zeileis, A. 2008. Conditional variable importance for random forests. *BMC Bioinformatics*, 9: 1–11.

University of Wyoming Advanced Research Computing Center. 2018. UW ARCC Beartooth High Performance Compute Cluster. https://doi.org/10.15786/M2FY47.

Vapnik, V. N.; and Chervonenkis, A. Y. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2): 264–280.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Watson, D. S.; Blesch, K.; Kapar, J.; and Wright, M. N. 2023. Adversarial random forests for density estimation and generative modeling. In *Proceedings of the* 26^{th} *International Conference on Artificial Intelligence and Statistics*, 5357–5375. PMLR.

Watson, D. S.; and Wright, M. N. 2021. Testing conditional independence in supervised learning algorithms. *Machine Learning*, 110(8): 2107–2129.

Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; and Veeramachaneni, K. 2019. Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems*, volume 32.