LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Roman Hornung, Anne-Laure Boulesteix

Interaction Forests: Identifying and exploiting interpretable quantitative and qualitative interaction effects

Technical Report Number 237, 2021
Department of Statistics
University of Munich

http://www.statistik.uni-muenchen.de

Interaction Forests: Identifying and exploiting interpretable quantitative and qualitative interaction effects

Roman Hornung[1*] Anne-Laure Boulesteix[1]

March 23, 2021

[1] Institute for Medical Information Processing, Biometry and Epidemiology, University of Munich, Munich, 81377, Germany

Abstract

Although interaction effects can be exploited to improve predictions and allow for valuable insights into covariate interplay, they are given little attention in analysis. We introduce *interaction forests*, which are a variant of random forests for categorical, continuous, and survival outcomes, explicitly considering quantitative and qualitative interaction effects in bivariable splits performed by the trees constituting the forests. The new *effect importance measure* (EIM) associated with interaction forests allows ranking of the covariate pairs with respect to their interaction effects' importance for prediction. Using EIM, separate importance value lists for univariable effects, quantitative interaction effects, and qualitative interaction effects are obtained. In the spirit of interpretable machine learning, the bivariable split types of interaction forests target well interpretable interaction effects that are easy to communicate. To learn about the nature of the interplay between identified interacting covariate pairs it is convenient to visualise their estimated bivariable influence. We provide functions that perform this task in the R package *diversityForest* that implements interaction forests. In a large-scale empirical study using 220 data sets, interaction forests tended to deliver better predictions than conventional random forests and competing random forest variants that use multivariable splitting. In a simulation study, EIM delivered considerably better rankings for the relevant quantitative and qualitative interaction effects than competing approaches. These results indicate that interaction forests are suitable tools for the challenging task of identifying and making use of well interpretable interaction effects in predictive modelling.

1 Introduction

In predictive modelling, two covariate variables interact if the effect of one variable on the outcome depends on the value of the other variable. Identifying interaction effects allows for important new insights into the interplay

*Corresponding author. Email: hornung@ibe.med.uni-muenchen.de.

between variables. For example, a variable (e.g., a medical treatment), may have a strong effect, but only for a subgroup of the observations that have certain values of a different variable. Beyond providing valuable insights, considering these effects can also improve the performance of automatic prediction rules.

Interaction effects between two variables A and B are often categorised into two types: quantitative and qualitative interaction effects (Peto, 1982). With quantitative interactions, the strength of the effect of variable A depends on the value of variable B, but the direction of that effect does not change in dependency of B. That is, independent of the value of variable B, variable A either has a positive or negative effect on the outcome. In contrast, in the case of qualitative interactions, the direction of the effect of variable A changes depending on the value of variable B. For example, variable A may have a positive effect on the outcome if the value of variable B is low, but a negative effect if the value is high.

Random forests (Breiman, 2001) are one of the most popular machine learning algorithms, known for their ability to capture complex non-linear relationships between the (covariate) variables and the outcome. They are ensembles of tree predictors, where the predictions of the forests are obtained by summarising the predictions of the individual trees. It is often assumed that random forests would exploit interaction effects between the variables very effectively (see, e.g., the literature references given in Wright *et al.* (2016)). The recursive splitting performed by the trees in random forests indeed models interaction effects between variables. However, an interaction effect between two variables is only modeled by a tree if there are branches in the tree that use each of the two variables at least once for splitting. Given the fact that the splits are selected by evaluating the predictive performances of the variables individually, interaction effects between variables without particularly strong marginal effects are not sufficiently accounted for. This is because a pair of interacting variables is only selected for splitting if at least one of them has a strong enough marginal effect to be selected for the first of the splits in these interacting variables (Wright *et al.*, 2016). Interaction effects between variables that have a strong effect only when used together may thus not be modelled sufficiently by the trees in random forests. Kim and Loh (2001) (for classification tress) and Loh (2002) (for regression trees) devised simple methods to allow selection of variables involved in interactions that are free of (strong) marginal effects. Given their simplicity, these methods can be expected to miss many pairs of interacting variables. In particular, the splitting is performed in a univariable fashion with these approaches, which hampers the methods' ability to model the interaction effects effectively. Interactions can be considered directly when performing multivariable splitting, that is, when using several variables for the same split. In the following, trees that use several variables for the same splits will be denoted multivariate trees, while trees that use classical univariable,

or "axis-aligned" splitting, will be denoted univariate trees.

A key feature of random forests, and an important milestone in the field of interpretable machine learning in general (Molnar *et al.*, 2020), is their ability to rank the variables in descending order with respect to their importance in prediction using Variable Importance Measures (VIMs). This allows identification of variables most valuable for prediction. However, the rankings obtained via the VIMs only allow conclusions on the importance of the individual variables. They are not suitable for identifying interaction effects valuable in prediction.

In this paper, the random forest variant "interaction forest" is introduced. It employs multivariable splitting and allows ranking of quantitative and qualitative interaction effects between variable pairs in descending order with respect to their importance for prediction. This identifies relevant interaction effects that potentially improve prediction and give insights into the dependency pattern of the outcome on the covariate variables that would be unobservable when focusing only on univariable effects. We introduce a variable importance measure, Effect Importance Measure (EIM), with interaction forests which can be used to rank univariable effects and quantitative and qualitative interaction effects separately. We use the term "effect importance" here to stress that the goal of EIM is to rank the strengths of univariable and interaction effects separately. In contrast, VIMs associated with other random forest-type prediction methods rank the impacts of the different variables without differentiating between main and interaction effects.

Quantitative and qualitative interaction effects are represented by different split types in the trees of interaction forests. After forest construction, the importance of these split types is measured separately for each considered variable pair. The importance scores obtained for these different split types for each variable pair can then be interpreted in terms of the degrees of association between the members of the variable pair with respect to the two different interaction effect types considered. The split types are designed to model well interpretable interaction effects. Therefore, the best-ranking variable pairs tend to feature interaction effects that can be well interpreted and consequently, well communicated. Our approach is strongly connected to the rapidly emerging field of interpretable machine learning; see Molnar *et al.* (2020) for an overview. This field is generally concerned with extracting interpretable knowledge from machine learning models. Machine learning tends to focus primarily on achieving strong predictive performance and less on obtaining insights into the dependency pattern between the outcome and the variables. The latter is, however, important for drawing conclusions on the subject matters studied using machine learning methods.

Apart from allowing us to identify and rank important interpretable interaction effects, interaction forests tend to also feature a higher predictive performance than conventional random forests and competing ran-

dom forest-based approaches that utilise multivariable splitting, as is revealed in the analyses shown in this paper. Interaction forests are implemented for categorical, continuous, and survival outcomes in the R package `diversityForest` (currently version 0.3.0). The package is available online from the CRAN repository.

The rest of the paper is structured as follows: Section 2 gives an overview of existing approaches to construct multivariate trees, random forest-based approaches that use multivariate trees, and random forest-based approaches to detect interaction effects. In Section 3, the interaction forest algorithm is described. Two comparison studies with other approaches are presented in Section 4, where one of these uses real data sets and focuses on predictive performance, while the other uses simulated data and focuses on interaction detection. In Section 5 we summarise the main conclusions from the paper and discuss further topics.

2 Existing work on multivariate trees, multivariate tree ensembles, and approaches to identifying interactions from tree ensembles

2.1 Multivariate tree approaches

The literature on multivariate trees is rich. An early example are multivariate CARTs (Breiman *et al.*, 1984) and an important recent contribution are so-called "optimal" multivariate classification trees (Bertsimas and Dunn, 2017). The latter trees differ from conventional decision trees in that the splits are not found individually, but rather entire trees are constructed at once in such a way that the training error is minimised. We provide an in-depth discussion on multivariate tree approaches in Section A.1 of Supplementary Material 1.

2.2 Random forest-based approaches that use multivariate trees

The following random forest-based approaches use multivariable splitting: rotation forests (Rodríguez *et al.*, 2006), mixed ensembles of univariate trees and mean margins decision trees (Gashler *et al.*, 2008), oblique random forests (Menze *et al.*, 2011), and canonical correlation forests (Rainforth and Wood, 2015). We describe these approaches in Section A.2 of Supplementary Material 1.

2.3 Approaches to identifying interactions from tree ensembles

There exists a variety of approaches that aim to identify interactions using tree ensembles. The majority of these approaches use classical univariable splitting (Ishwaran, 2007; Kelly and Okada, 2012; Bureau *et al.*, 2005; Dazard *et al.*, 2018; Chen and Zhang, 2013; Li *et al.*, 2016; Basu *et al.*, 2018; Jiang *et al.*, 2009). Ng and Breiman (2005) and Yoshida and Koike (2011) perform multivariable splitting implicitly by using univariable splits in synthetic variables formed from pairs of variables. Sorokina *et al.* (2008) use so-called additive groves of trees (Sorokina *et al.*, 2007) and Dirichlet process forests (Du and Linero, 2019) take a Bayesian perspective. For more detailed descriptions of these approaches, refer to Section A.3 of Supplementary Material 1. The main conceptional difference between interaction forests and these approaches, is that interaction forests target well interpretable interaction effects. Identifying different types of interaction effects is arguably less important because it is difficult to utilise the knowledge on existing interaction effects, if these are overly complex.

In general, there seems to be much debate in the scientific community to what degree it is possible to identify interaction effects using random forests (Boulesteix *et al.*, 2015b). For an overview on the meaning of the term "interaction" with a focus on random forest methodology and the differentiation of interactions from related, but different concepts, see Boulesteix *et al.* (2015a).

3 The interaction forest algorithm

3.1 Data format

Let $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, $i = 1, ..., n$, denote the available data, where $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are the values of the covariate variables and the outcome values, respectively. The variable vector $\boldsymbol{x}_i$ is of length p, where each entry x_{ij}, $j = 1, \ldots, p$, contains the value of a particular variable for the ith observation, and $\boldsymbol{y}_i$ is a scalar in most cases, for example $\boldsymbol{y}_i \in \{0, 1\}$ for binary outcomes. The outcome may also take the form of a vector, for example $\boldsymbol{y}_i = \{y_{i,1}, y_{i,2}\}$ with $y_{i,1} \in \mathbb{R}_{>0}$ and $y_{i,2} \in \{0, 1\}$ for survival outcomes, where $y_{i,1}$ gives the survival/censoring time and $y_{i,2}$ the censoring indicator. While in the case of training data, both, $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are known, only $\boldsymbol{x}_i$ is known in the case of test data.

3.2 Training and prediction

3.2.1 Interaction forests as specific diversity forests

On a technical note, interaction forests are specific diversity forests (Hornung, 2020), which are themselves specific random forests according to the definition of the term "random forest" in the original random forest article by Breiman (2001). Note that this original definition was much more general than the specific procedure commonly referred to by this term today. The diversity forest algorithm is not a specific algorithm, but an alternative candidate split sampling scheme that makes complex split procedures in random forests possible computationally by drastically reducing the numbers of candidate splits that need to be evaluated for each split. It also avoids the well-known variable selection bias in conventional random forests that has the effect that variables with many possible splits are selected too frequently for splitting (Strobl *et al.*, 2007). The candidate split sets differ for each split in diversity forests and each of them is obtained in the following way: For $l = 1, \ldots, nsplits$ (denoted $npairs$ in the case of interaction forests): 1) Sample one so-called split problem; 2) Sample a single or few splits from the split problem sampled in the first step and add this (or these) split(s) to the candidate split set. For example, in the case of conventional univariable splitting, a split problem consists of all possible splits in one of the variables. Here, to obtain a candidate set for a split, we repeatedly draw variables and, instead of trying out all possible splits in these variables (as would be done in conventional random forests), consider only one randomly drawn split in each variable. Note that this procedure for univariable splitting is very similar to extremely randomized trees (Geurts *et al.*, 2006); the only difference is that with diversity forests the split problems (i.e., variables in the case of conventional univariable splitting) are drawn with instead of without replacement.

In the case of interaction forests, a split problem consists of the collection of all possible splits of all considered split types (cf. Section 3.2.4) in a specific pair of variables. Here, as will be described in Section 3.2.5, we merely sample one split for each of the considered split types from each split problem into the candidate split set. Given the fact that the split problems contain very large numbers of splits for interaction forests, it would be too demanding computationally to try out all splits in the split problems.

3.2.2 General remarks on interaction forests

As conventional random forests, interaction forest prediction rules consist of large numbers of decision trees, where each of these is learned using recursive binary splitting of a bootstrap sample or a random subset of the training data. Therefore, for each tree, there are several observations that are not used for learning. These observations are commonly referred to as

the out-of-bag (OOB) observations associated with a tree, where the set of all OOB observations of a tree is denoted as its OOB sample. The OOB observations can be used in prediction error estimation and tuning parameter optimisation, and they are required for the calculation of the EIM values (see Section 3.3 for details).

Apart from the pre-processing described in Section 3.2.3, interaction forests differ from conventional random forests only in the considered split types (see Section 3.2.4) and in the way these splits are chosen during the construction of the forest (see Section 3.2.5). Prediction is performed analogously as with conventional random forest; for details see Section B.1 of Supplementary Material 1. In the rest of this subsection, we will describe the differences of interaction forests to conventional random forests (Sections 3.2.3 to 3.2.5) and comment on the default values of the involved hyperparameters (Section 3.2.6).

3.2.3 Pre-selection of variable pairs that show indications of interaction effects and handling of unordered categorical covariate variables

For data sets with low or moderate numbers of variables, we consider all possible pairs of variables for splitting in interaction forests. For higher dimensional data, the number of possible variable pairs becomes very large, making it impossible to consider all the pairs. As a solution, we use an automatised procedure in the interaction forest algorithm to pre-select 5000 variable pairs that show indications of interaction effects if the number of variables is larger than 100; that is, if the number of possible variable pairs is larger than 5000. We detail this automatised procedure in Section B.2 of Supplementary Material 1. When constructing the trees in interaction forests, we only consider the pre-selected promising variable pairs for splitting. If the number of pre-selected variable pairs would have been set to a larger value than 5000, each pair would not have been considered frequently enough in splitting to obtain stable EIM values using computationally well feasible numbers of trees. It cannot be excluded that some relevant interaction effects are missed by this proceeding. However, 5000 pre-selected variable pairs seems large enough for catching most pairs that show reasonably strong indications of interaction effects.

The split types used in interaction forests that model well interpretable interaction effects (cf. Section 3.2.4) do not apply directly to unordered categorical variables. This issue is dealt with by converting these variables into ordered variables in the same way as when using the option `respect.unordered.factors="order"` implemented in the R package `ranger` (version 0.12.1) (Wright and Ziegler, 2017). Wright and König (2019) describe and empirically evaluate this option for conventional random forest. The idea of this option is to take the outcome into account for

ordering the categories in such a way that when moving along the ordering of the categories, the outcome tends to change in a consistent direction. See Section B.3 of Supplementary Material 1 or Wright and König (2019) for more details. After ordering the categories of the unordered categorical variables, each categorical variable is coded as $1, \ldots, J$, where J denotes the number of categories of the variable.

3.2.4 Split types

We consider six different types of splits, which are visualised in Figure 1 and described here. The univariable splits take the same form as in conventional random forests, that is, $x_j < p_u^{(j)}$ vs. $x_j \geq p_u^{(j)}$, where x_j denotes a variable and $p_u^{(j)}$ a split point in that variable.

The bivariable splits are divided into two different basic types: those associated with quantitative and those associated with qualitative interactions. These two types of splits will be referred to as quantitative and qualitative splits, respectively, in the following. As discussed in the introduction, in the case of a quantitative interaction, the strength of the effect of variable x_{j_2} depends on the value of variable x_{j_1}. If the quantitative interaction is sufficiently strong, and if the influences of x_{j_1} given x_{j_2} and x_{j_2} given x_{j_1} are monotonically increasing or decreasing, the outcome of the observations in one of the following four groups will differ systematically from that of all other observations: 1) x_{j_1} small and x_{j_2} small, 2) x_{j_1} small and x_{j_2} large, 3) x_{j_1} large and x_{j_2} small, 4) x_{j_1} large and x_{j_2} large. Therefore, we distinguish four different types of quantitative splits, each associated with one of these four groups (Figure 1). These splits divide the current node into two child nodes based on whether each observation falls into the respective group associated with the split. For example, a split associated with the first of the above four groups "x_{j_1} small and x_{j_2} small" takes the following form: $x_{j_1} < p_b^{(j_1)} \cap x_{j_2} < p_b^{(j_2)}$ vs. $(x_{j_1} < p_b^{(j_1)} \cap x_{j_2} < p_b^{(j_2)})^{\complement}$, where $^{\complement}$ denotes the complementary set. The other three types of quantitative splits are defined analogously.

In the case of a qualitative interaction, the direction of the effect of x_{j_2} is different for small values of x_{j_1} than for large values of x_{j_1}. As in the case of the quantitative interactions, we focus on qualitative interactions for which x_{j_1} given x_{j_2} and x_{j_2} given x_{j_1} are monotonically increasing or decreasing. For such interactions, the direction of the influence of x_{j_2} will be opposite between observations with small and large x_{j_1} values. If a qualitative interaction of this type is sufficiently strong, the outcome of observations with small x_{j_1} values, and simultaneously small x_{j_2} values, will be similar to that of observations with large x_{j_1} values and at the same time large x_{j_2} values. Moreover, the outcome of observations with small x_{j_1} values and at the same time large x_{j_2} values will be similar to the outcome

of observations with large x_{j_1} values and at the same time small x_{j_2} values. Based on these considerations, qualitative splits take the following form in interaction forests: $(x_{j_1} < p_b^{(j_1)} \cap x_{j_2} < p_b^{(j_2)}) \cup (x_{j_1} > p_b^{(j_1)} \cap x_{j_2} > p_b^{(j_2)})$ vs. $(x_{j_1} \leq p_b^{(j_1)} \cap x_{j_2} \geq p_b^{(j_2)}) \cup (x_{j_1} \geq p_b^{(j_1)} \cap x_{j_2} \leq p_b^{(j_2)})$.

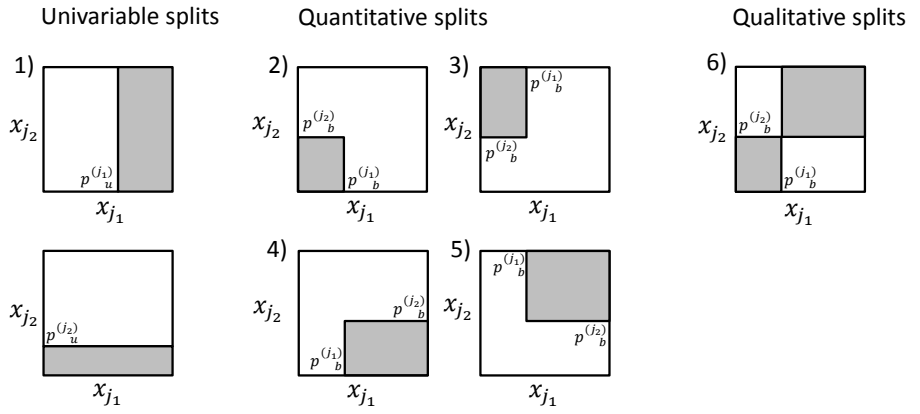

Figure 1: Split types considered in the interaction forest algorithm. Each square visualises the variable space spanned by two variables x_{j_1} and x_{j_2}. The points $p_u^{(j_1)}$, $p_u^{(j_2)}$, and $(p_b^{(j_1)}, p_b^{(j_2)})$ denote the split points for univariable and bivariable splits, respectively. The white and gray areas depict the regions associated with the two child nodes of the splits. For each drawn variable pair in the candidate split sampling one candidate split of each of these split types is drawn, see Section 3.2.5 for details.

The forms of the quantitative and qualitative splits given above represent simplifications of actual quantitative and qualitative interaction effects and do not cover all possible types of interaction effects. For example, quantitative interaction effects for which x_{j_2} is only influential for small and large x_{j_1} values, but not moderate x_{j_1} values are not covered. These split forms were chosen for three reasons. First, in the spirit of interpretable machine learning, these split forms model interaction effect types that are well interpretable and communicable. Second, if the split forms would be more sophisticated, each interacting pair would have a larger proportion of its possible splits be nonsensical with its true interaction effect. By contrast, while the simple split forms of interaction forests are not suitable for reproducing the true interactions exactly, many of the possible splits will represent them sufficiently well. Sampling adequate candidate splits with high frequency is especially important in the case of larger numbers of variables, since more variables means the number of possible pairs is large. Third, the forms of the quantitative and qualitative splits were chosen to be particularly different from each other. This makes the algorithm better able to

differentiate between quantitative and qualitative interactions. If the forms of the splits were more sophisticated, many of the sampled splits would not be very specific for quantitative and qualitative interactions, respectively.

3.2.5 Split selection in the tree construction

As noted above, the trees in interaction forests are grown using recursive binary splitting. The fact that the candidate split sets in interaction forests are sampled randomly from iteratively re-sampled split problems, makes them specific diversity forests. A split problem in interaction forests is the collection of all possible splits in a single pair of variables that are of the six split types shown in Figure 1.

Each node split is selected in the following way in the construction of the trees:

1. *Drawing of the candidate split set:*
   The candidate split set, consisting of univariable and bivariable candidate splits in a number $npairs$ of randomly sampled variable pairs, is drawn in the following way ($npairs$ is similar to the parameter $mtry$ in random forest):
   For $pair = 1, \ldots, npairs$:

   (a) *Drawing of the variable pair to consider.*
   Draw randomly a variable pair x_{j_1} and x_{j_2} from the set of considered variable pairs, which, for $p \leq 100$, represents all possible variable pairs or, for $p > 100$, a subset selected by the procedure described in Section B.2 of Supplementary Material 1.

   (b) *Drawing of univariable splits.*
   Add two univariable split points $p_u^{(j_1)}$ and $p_u^{(j_2)}$, one for each variable x_{j_1} and x_{j_2} to the candidate split set. These split points are drawn in the following way: Sort the unique values of the variables, calculate the midpoints between the adjacent sorted unique values, and randomly draw one of these midpoints.

   (c) *Drawing of bivariable splits.*
   In this step, five bivariable splits in the variable pair x_{j_1} and x_{j_2} are added to the candidate split set: one of each of the four quantitative split types and one qualitative split (cf. Section 3.2.4).
   Each of these five splits uses the same split point $(p_b^{(j_1)}, p_b^{(j_2)})$, which is drawn in the following way:

      i. Draw $p_b^{(j_1)}$ by taking the average of two randomly drawn unique x_{j_1} values.

      ii. Draw $p_b^{(j_2)}$ under the constraint that each quadrant in the two-dimensional coordinate system with origin $(p_b^{(j_1)}, p_b^{(j_2)})$

will contain at least one observation. The latter constraint ensures that all five bivariable splits will be valid. The procedure used for drawing $p_b^{(j_2)}$ is described in detail in Section B.4 of Supplementary Material 1.

2. *Calculation of the split criterion values associated with the different candidate splits.*
   For each candidate split in the candidate split set obtained in the first step, perform the following:

   (a) Assign each of the observations in the current node to one of the two child nodes using the respective candidate split.

   (b) Calculate the value of the considered split criterion (e.g., the Gini impurity for categorical outcomes) for the division of the current node obtained in (a).

3. *Selection of the best candidate split.*
   Choose the split out of all candidates considered in the first step that is associated with the best split criterion value calculated in the second step.

3.2.6 Hyperparameter values

Most hyperparameters of interaction forests, such as the minimum node size or the number of trees, are the same as those in conventional random forests. In general, the performance of random forests has been seen to be quite insensitive to changes in their hyperparameter values (Probst *et al.*, 2019). As noted before, interaction forests are specific diversity forests. In Hornung (2020), it was seen that the performance of diversity forests is also quite insensitive to changes in the size of the candidate split sets, that is, the numbers of candidate splits repeatedly sampled for each split, which is $npairs \times 7$ in interaction forests (cf. Section 3.2.5). For these reasons it should generally not be necessary to optimise the default hyperparameter values used in the implementation of interaction forests in the R package `diversityForest`, version 0.3.0. In Section B.5 of Supplementary Material 1 the default values we use for the main hyperparameters are presented, where we also provide reasonings for choosing each value.

3.3 Effect Importance Measure (EIM): Ranking univariable effects, quantitative, and qualitative interaction effects separately

The Effect Importance Measure (EIM), associated with interaction forests, delivers separate ranking lists for the univariable effects and the quantitative and qualitative interaction effects. The EIM values obtained for these

three different types of effect will be denoted univariable, quantitative, and qualitative EIM values, respectively.

EIM uses a variant of importance measure suggested by Hapfelmeier *et al.* (2014) that is similar to the classical permutation VIM (Breiman, 2001) of random forests. In Hapfelmeier *et al.* (2014), the main motivation for introducing PropRandom was that it can deal with missing values in the data. We use it for calculating the EIM values, because it is more efficient than the permutation VIM with respect to computing time. As with the classical permutation VIM, with the approach by Hapfelmeier *et al.* (2014), the importance value of each variable is based on comparing the prediction accuracies of the trees on their OOB samples with estimates of the trees' prediction accuracies we would expect if the variable was not available during prediction. The single but important difference to the classical permutation VIM is that the trees' prediction accuracies expected if the variable was not available are obtained in the following way in Hapfelmeier *et al.* (2014): The OOB observations are dropped down the trees and in each instance in which a split uses the variable of interest, the OOB observations are assigned randomly to one of the two child nodes. The probabilities used for the two child nodes are set proportional to their sizes.

The latter procedure does not transfer directly to interaction forests. In contrast to conventional random forests, we need a vector for each of the considered six split types, rather than a single vector of variable importance values (cf. Section 3.2.4). When calculating the importance value of a variable or variable pair with respect to a specific split type, we assign the OOB observations randomly to the child nodes in each instance that a split uses the variable or variable pair of interest, and simultaneously is of the split type of interest (from the six considered split types). For example, the prediction accuracy of a tree we would expect if a specific variable pair x_{j_1} and x_{j_2} would have no effect in prediction with respect to the qualitative split type (number 6 in Figure 1) would be estimated as follows: When dropping the OOB observations down the tree, assign them randomly to one of the two child nodes only for splits that use both of these variables x_{j_1} and x_{j_2} and are qualitative splits at the same time. In the following, we will refer to the procedure by Hapfelmeier *et al.* (2014) as *PropRandom* for "proportional randomisation".

The EIM values for univariable effects, quantitative, and qualitative interaction effects are calculated as follows:

1. *Calculate the OOB prediction accuracies of the trees.*

2. *Calculate univariable EIM values.*
   First, calculate the prediction accuracies of the trees with the influences of the respective univariable effects eliminated using the trees' OOB observations, where the predictions for the OOB observations

are obtained by applying PropRandom with respect to the respective univariable splits. Second, subtract these prediction accuracies from those calculated in the first step and average these differences across all trees to obtain the univariable EIM values.

3. *Calculate qualitative EIM values.*
   Using PropRandom, the qualitative EIM values are calculated for each considered variable pair analogously to the univariable EIM values.

4. *Calculate quantitative EIM values.*
   As a first step, for each of the four different quantitative split types (cf. again Section 3.2.4), we apply the same procedure as in the cases of the univariable and qualitative splits.

   This results in four vectors of quantitative EIM values. However, the four different quantitative interaction effects targeted by the four quantitative split types are mutally exclusive, which is why each variable pair can feature only one of these four interaction effect types. Therefore, we need to classify the quantitative interaction effect associated with each considered variable pair as one of the four quantitative interaction effect types.

   Prior to classifying the effects, however, we need to take care of a different issue: The raw quantitative EIM values are not yet specific for the targeted interaction effect types. More precisely, two variables that both feature strong univariable effects, but no interaction effect, will have large raw quantitative EIM values for two of the four quantitative split types. For example, consider the scatter plot of the variable pair with only main effects in Figure 4. Here, the quantitative EIM value associated with split type two in Figure 1 will be large, but also that associated with split type five. The corners of these split types in Figure 1 are opposing. We make use of the latter when adjusting the raw quantitative EIM values to make them specific for quantitative interaction effects. For reasons of brevity, we describe the procedure used for adjusting the raw EIM values in Section B.6 of Supplementary Material 1.

   The quantitative interaction effect types assigned to the respective quantitative interaction effects are the ones associated with the largest of the four adjusted quantitative EIM values. These largest adjusted quantitative EIM values are also the final quantitative EIM values.

Note that, for larger numbers of variables ($p > 100$), a subset of 5000 promising variable pairs is pre-selected prior to the construction of the interaction forest (cf. Section 3.2.3). Therefore, it is possible that not all variables were available for splitting in the construction of the interaction forest, but only those occurring in at least one of the pre-selected promising variable pairs.

Univariable EIM values associated with variables unavailable for splitting are set to zero. For higher dimensional data this can have the effect that some important variables receive univariable EIM values of zero if these variables are not featured in the pre-selected variable pairs. Another issue associated with univariable EIM values for higher dimensional data is that variables that are featured in a larger number of pre-selected variable pairs can receive too large univariable EIM values (cf. also the exemplary real data analysis in Section C.4 of Supplementary Material 1). These variables tend to be used more often for splitting because they are not only contained in more pre-selected variable pairs than others, but also in more candidate split sets since we sample from the (pre-selected) variable pairs in the candidate split sampling and not from the individual variables (cf. Section 3.2.5). Therefore, variables that are in more pre-selected variable pairs are considered more often for splitting, which is why they tend to receive larger univariable EIM values. Given these issues, we strongly recommend interpreting the univariable EIM values with caution, if the data are high-dimensional. If it is of interest to measure the univariable importance of the variables for high-dimensional data, an additional conventional random forest should be constructed and the VIM values of this random forest be used for ranking the univariable effects.

3.4 Visual exploration and classification of the interaction effects

Using the EIM values, the quantitative and qualitative interaction effects can be ranked with respect to their importance in prediction. However, the quantitative and qualitative EIM values do not yet allow to draw conclusions on the exact forms of the interaction effects. The latter is, however, crucial for learning about the nature of the interplay between the variables in interacting pairs. This makes it also possible to identify false positive results, meaning pairs of variables that received high quantitative or qualitative EIM values, but do not interact. The latter check for false positive results is important because to date there is no option to test whether the EIM values are significantly different from zero. Even if there are no interacting variable pairs in the data set, the quantitative and qualitative EIM values of some variable pairs will take the first place.

The easiest interpretable way of studying the natures of interaction effects between pairs of variables is visual exploration. In the `diversityForest` R package we provide functions that apply flexible regression techniques to variable pairs and plot their fits. An example of such a plot is shown in Figure 2. This plot was taken from the exemplary real data analysis in Section C.1 of Supplementary Material 1. Note that it is always important to not overinterpret details of the fits of flexible regression techniques due to their tendency to overfit the observed data.

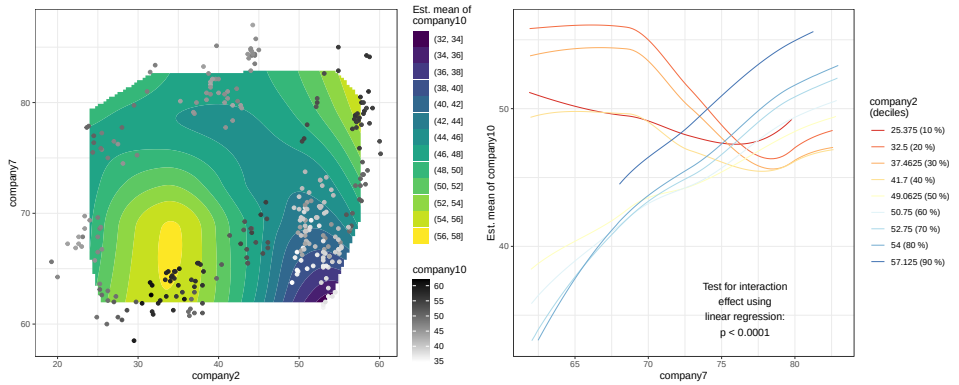

Figure 2: Example of plot produced by the `diversityForest` R package to visualise the estimated bivariable influence of a variable pair. The estimated influence of the stock prices of two aerospace companies (`company2` and `company7`) on the stock price of a third aerospace company (`company10`) is shown. The contour plot in the left panel shows a two-dimensional LOESS fit. For reasons of clarity, the points in the left panels do not show all 950 observations in the data set, but a random subset of 300 observations. The lines in the right panel show cross sections of the two-dimensional LOESS fits in the left panel. This data set was obtained from the open science online platform OpenML (Vanschoren *et al.*, 2013) under the data set ID 223.

In many applications, a specific variable is of main interest. For example, in a medical study the interest may lie in investigating how a specific treatment influences the outcome versus a placebo. We offer an option in `diversityForest` that allows the user to focus on a specific variable when visually examining the bivariable influences of variable pairs with large quantitative or qualitative EIM values.

We strongly encourage readers interested in applying interaction forests to their own data to consult Section C of Supplementary Material 1. Here, we show illustrative interaction forest analyses that use real data sets, where we demonstrate important functionalities from the `diversityForest` package together with the corresponding R commands.

4 Comparison study of interaction forests with existing alternatives

We compared the performance of interaction forests with that of competing approaches. The performance was evaluated both in terms of prediction and in terms of the ability to identify interaction effects. Predictive performance was evaluated using a collection of 220 real data sets. We used simulated

data for evaluating the performance of the algorithms with respect to identifying interaction effects, because for real data the interaction structure between the variables is not known. All R code and data used to perform and evaluate the analyses are made available in Supplementary Material 2.

4.1 Comparison with respect to predictive performance using real data sets

4.1.1 Data

The data material consists of 220 publicly available data sets with binary outcome and various numbers of covariate variables. Details on these data sets and their pre-processing can be found in Hornung (2020) where these data sets were used in identical form as in this paper. Information on their acquisition can be found in Couronné *et al.* (2018), where this collection of data sets was used for the first time.

4.1.2 Study design

The following random forest-based methods were included in this comparison: interaction forests (IF), random forests (RF) (Breiman, 2001), canonical correlation forests (Rainforth and Wood, 2015) (CaF), oblique random forests (Menze *et al.*, 2011) (ObF), and rotation forests (Rodríguez *et al.*, 2006) (RoF). Note that except for RF, all of these approaches use multivariable splitting. Because RF can be seen as a baseline method in this study, the configuration of RF was adjusted to the default configuration of IF (cf. last paragraph of Section 3.2.3 and Section 3.2.6), in order to put neither method at an advantage or disadvantage. For CaF, ObF, and RoF we used the default configurations provided in the respective R implementations. As a validation scheme, we used five times repeated 5-fold stratified cross-validation. As performance measures, we used the area under the receiver operating characteristic curve (AUC), the accuracy (ACC), and the Brier score (Brier). Further details on the design of this comparison study are given in Section D.1 of Supplementary Material 1.

4.1.3 Results

Table 1 shows the performances of the methods summarised across all 220 data sets. IF performed best with respect to the median for all three performance metrics. Treating all data sets as independent, each metric was tested for significant differences in terms of the median between IF, and each of the four competitors using paired Wilcoxon tests. Subsequently, we adjusted the p-values for multiple testing separately for each metric using the Holm-Bonferroni method. Here, IF was significantly better than the competitors with respect to all metrics except for RF with respect to the

	AUC	ACC	Brier
IF	0.9182 [0.7820, 0.9862]	0.8822 [0.7664, 0.9499]	0.0890 [0.0425, 0.1641]
RF	0.9140 [0.7815, 0.9829]	0.8787 [0.7670, 0.9512]	0.0923 [0.0401, 0.1645]
CaF	0.8842 [0.7660, 0.9781]	0.8761 [0.7555, 0.9468]	0.0962 [0.0391, 0.1748]
ObF	0.9051 [0.7721, 0.9824]	0.8644 [0.7356, 0.9465]	0.0985 [0.0461, 0.1818]
RoF	0.8632 [0.7652, 0.9685]	0.8676 [0.7544, 0.9421]	0.1016 [0.0437, 0.1686]

Table 1: Performances of the methods summarised across the 220 data sets. The numbers show the medians of the cross-validated metrics across the data sets. The numbers in square brackets show the 25% quantiles and 75% quantiles (i.e., the first and third quartiles) of the cross-validated metrics obtained for each data set. Larger AUC values, larger ACC values, and smaller Brier values indicate a better performance.

ACC (adjusted p-values: AUC: IF vs. RF: $p < 0.001$, IF vs. CaF: $p < 0.001$, IF vs. ObF: $p = 0.001$, IF vs. RoF: $p < 0.001$; ACC: IF vs. RF: $p = 0.247$, IF vs. CaF: $p = 0.002$, IF vs. ObF: $p = 0.002$, IF vs. RoF: $p < 0.001$; Brier: IF vs. RF: $p = 0.009$, IF vs. CaF: $p < 0.001$, IF vs. ObF: $p < 0.001$, IF vs. RoF: $p < 0.001$). It must be noted that these p-values should be interpreted with caution because the data sets are not all independent as several of them form groups in the sense that they constitute versions of the same data set (details on the data sets can be found in the supplementary files of Hornung (2020)). However, apart from the test "IF vs. RF" for the ACC all p-values are considerably smaller than the significance threshold $\alpha = 0.05$. The effect sizes of the tests were largely in the small to moderate range (AUC: IF vs. RF: $r = 0.33$, IF vs. CaF: $r = 0.37$, IF vs. ObF: $r = 0.22$, IF vs. RoF: $r = 0.52$; ACC: IF vs. RF: $r = 0.08$, IF vs. CaF: $r = 0.22$, IF vs. ObF: $r = 0.23$, IF vs. RoF: $r = 0.26$; Brier: IF vs. RF: $r = 0.18$, IF vs. CaF: $r = 0.24$, IF vs. ObF: $r = 0.33$, IF vs. RoF: $r = 0.34$). RF featured the second best median performance for all three metrics. The rankings between the remaining methods are not consistent across the different metrics in Table 1.

Figure 3 shows the ranks each method achieved among the other methods for each data set. Smaller values of the ranks correspond to a better performance. Small and large values of the ranks will often be referred to as "good ranks" and "bad ranks", respectively. IF tended to achieve considerably better ranks than the other methods with respect to the AUC and the Brier. While there are no notable visible indications of an improvement of IF over RF in terms of the ACC in Figure 3, IF featured the best mean rank among the methods (results not shown). RoF achieved the worst ranks among all methods for all three metrics. CaF achieved the second worst ranks with respect to the AUC and strongly varying ranks with respect to the Brier. In the case of the latter, CaF performed best and worst over proportionally often. In fact, CaF took the best place the most often among the five methods with respect to the Brier score. Further analysis revealed that few of the data sets for which CaF performed best with respect to the Brier

involved many variables. This suggests that CaF is particularly effective for data sets with few variables. It must be noted that the comparison to CaF is not completely fair in the cases of the AUC and the Brier. This is because the latter two metrics evaluate the performance with respect to class probability estimation. CaF did not originally offer the possibility to obtain class probability estimates, which is why we simply used the proportions of the trees predicting either class as class probability estimates. It might be possible to conceive more efficient procedures for obtaining class probability estimates using CaF.

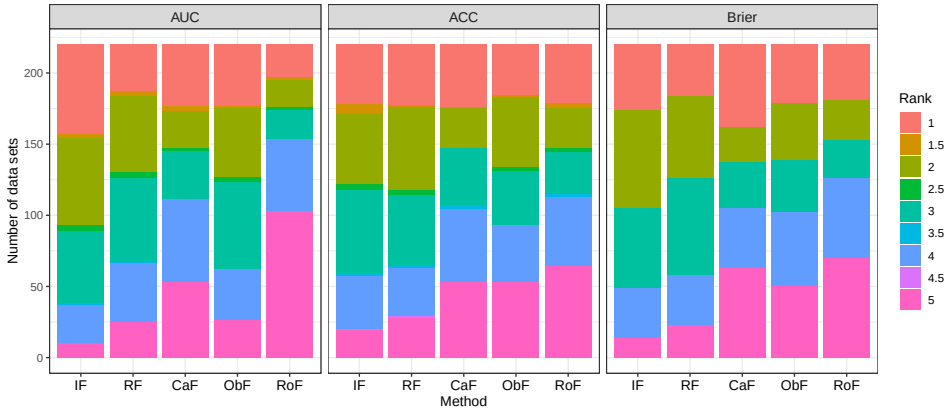

Figure 3: Ranks of the methods with respect to the different performance metrics. Each stacked bar shows the number of data sets for which the respective method achieved the indicated rank among all other methods.

Using LOESS regression, we also investigated the dependencies of the mean ranks of the methods with respect to the three metrics on the number of variables in the data sets and on the sample sizes. Here, we saw no notable dependency of the mean rank of IF on the number of variables. For the ACC and the Brier, the mean rank of IF also did not depend notably on the sample size. In the case of the AUC, IF achieved slightly better mean ranks for large sample sizes. We describe the results of this LOESS analysis in detail in Section D.2 of Supplementary Material 1.

4.2 Comparison with respect to performance in interaction effect importance ranking using simulated data

4.2.1 Study design

Using simulated data, we compared the ranks of interacting variables obtained using the quantitative EIM (IF-EIM-quant) and qualitative EIM (IF-EIM-qual) of IF with the corresponding rankings obtained using alternative

approaches. We also compared the univariable rankings obtained using the univariable EIM (IF-EIM-univ) with those obtained using the classical permutation VIM associated with RF (RF-V). When calculating the ranks, ties in the importance scores were broken randomly.

The following alternative approaches for ranking the predictive importance of interaction effects were included in the comparison study: the paired association measure (PA) (Ishwaran, 2007), the Interaction Minimal Depth Maximal Subtree measure (IMDMS) (Dazard *et al.*, 2018), and the stability score of iterative random forests (iRF) (Basu *et al.*, 2018). There do not seem to be publicly available R implementations for any of the other approaches mentioned in Section 2.3 (for descriptions of these approaches, see Section A.3 of Supplementary Material 1). As a baseline method we also included a naive metric calculated using the permutation VIM of RF (RF-V-pairs): For each pair of variables, we measured the importance of the interaction effect as the mean of the permutation VIM values obtained for the two variables. This metric does not, similar to Bureau *et al.* (2005), measure the importance of the interaction between the features, but focuses only on the importance of the individual features. Any measure that focuses on interaction detection should be superior to this naive metric.

As in the first comparison study (cf. Section 4.1), we adjusted the configuration of RF to the default configuration of IF. For the remaining methods, except for in the case of the number of trees, we used the default configurations of these methods provided in the respective R implementations. Because IF uses 20000 trees per default when calculating EIM values (cf. Section 3.2.6), we used 20000 trees for all methods to avoid potentially putting the other methods at a disadvantage. An exception was the largest sample size considered in the simulation ($n = 1000$): Here, in the cases of iRF and PA using 20000 trees was not feasible computationally, which is why we used these two methods' default choices for the numbers of trees in the case of the largest sample size. For more details on this, see Section E.1 of Supplementary Material 1.

While EIM, PA, and IMDMS return interaction importance scores for each pair of variables, iRF returns comparably short lists of tuples that are candidates for tuples featuring high-order interactions. Therefore, the results obtained using iRF are not directly comparable to those obtained using EIM, PA, and IMDMS. We describe the procedure we used to obtain ranks for all variable pairs with iRF in Section E.1 of Supplementary Material 1. In this procedure, we decided to err on the side of giving iRF an advantage rather than risking a disadvantage. For example, we did not calculate ranks of the interacting (and non-interacting) pairs attributed by iRF for simulation iterations for which iRF did not select the respective pairs; treating these ranks as missing. This disregards the fact that iRF was clearly not successful in identifying the interacting pair for these simulation iterations. Another possibility would have been to impute the ranks attributed by iRF

for these iterations. However, it would have been necessary to make assumptions that potentially put iRF at a disadvantage. We wanted to avoid a disadvantage, because our goal was to investigate whether IF tends to outperform the competing approaches or not. If IF would still perform better than iRF, even if the latter is put at an advantage, we would have more certainty that IF truly performs better than iRF in the investigated context.

The simulation involved a binary, balanced outcome and 68 continuous predictor variables. Fifty variables had no effect and the remaining involved both univariable effects and pairs with quantitative or qualitative interaction effects. For each effect type, three different levels of strength were considered: Strong, moderate, and weak. Three different sample sizes: 100, 500, and 1000 were considered and for each sample size 200 data sets were generated. The effects simulated for the different variables are listed in Table 2. The simulation design is detailed in Section E.3 of Supplementary Material 1. Exemplary simulated data is shown in Figure 4. The first panel shows a scatter plot between two variables with only main effects, but no interaction effects. For both variables, observations in the second outcome class tend to feature larger values than those in the first outcome class. Therefore, the observations from the second class concentrate in the upper-right corner of the plot and the observations from the first class in the lower-left corner. The middle panel of Figure 4 shows a scatter plot between two variables with a quantitative interaction effect. Here, the observations from the second class tend to concentrate in the lower-right corner of the plot. For values of X_7 smaller than about zero, there is no influence of X_8. In this region, the green points (second class) are scattered evenly among the red points (first class); the density of green points does not increase or decrease with larger values of X_8. In contrast, in the case of larger values of X_7, there are much more observations from the second class for smaller values of X_8 than for larger values. This means that the probability for observing observations from the second class is increasingly strongly and negatively influenced by X_8 for larger values of X_7. A scatter plot between two variables with a qualitative interaction effect is shown in the right panel of Figure 4. Here, in the case of smaller values of X_{13} up until about one, there are much more values from the second class for smaller X_{14} values. Conversely, in the case of larger values of X_{13}, there are increasingly more values from the second class for larger X_{14} values. Thus, for small X_{13} values X_{14} has a negative influence on the probability for observing observations from the second class, but a positive influence for larger X_{13} values.

4.2.2 Results

Table 3 shows the median ranks the variables with main effects only received with respect to the univariable EIM values calculated using IF and with respect to the values of the classical permutation VIM calculated using RF.

Predictor variables	Effect type
$X_1, \ldots, X_6$	only main effect, no interaction effect
X_1, X_2	strong effect
X_3, X_4	moderate effect
X_5, X_6	weak effect
$X_7, \ldots, X_{12}$	quantitative interaction effect
$\{X_7, X_8\}$	strong effect
$\{X_9, X_{10}\}$	moderate effect
$\{X_{11}, X_{12}\}$	weak effect
$X_{13}, \ldots, X_{18}$	qualitative interaction effect
$\{X_{13}, X_{14}\}$	strong effect
$\{X_{15}, X_{16}\}$	moderate effect
$\{X_{17}, X_{18}\}$	weak effect
$X_{19}, \ldots, X_{68}$	no effect

Table 2: Simulation design. Interacting variables are enclosed in curly brackets.

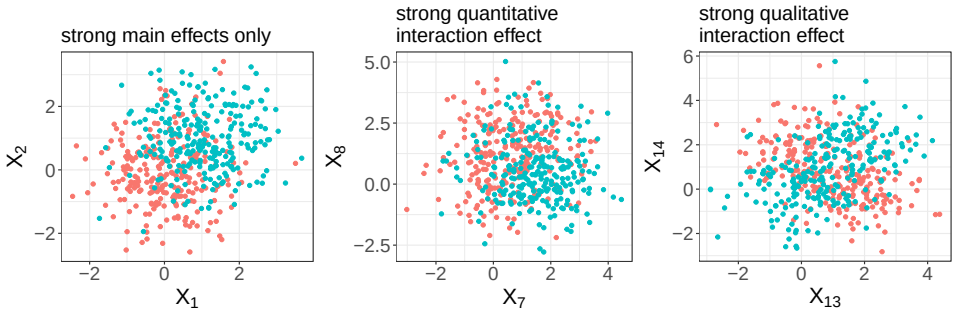

Figure 4: Exemplary pairs of variables with strong effects in a simulated data set (sample size: 500). Each point corresponds to an observation in the data set. The two colors distinguish the two outcome classes, where red and green points show observations from the first and second class, respectively. Corresponding pairs of variables for moderate and weak effects are shown in addition in Supplementary Figure S43 .

The ranks are small and very similar between the two methods. The ranks obtained for the individual simulated data sets are shown in Supplementary Figure S44. Note that, while the variables with the strongest main effect received a median rank of two for both methods, these variables received the best rank more often than the second-best rank for all sample size scenarios (results not shown).

The median ranks obtained for the variable pairs with quantitative interaction effects are shown in Table 4, and the corresponding ranks for all data sets in Supplementary Figure S45.

For the smallest considered sample size, none of the methods were able to consistently identify the variables with quantitative interaction effects.

Effect:	Strong	Moderate	Weak
	n = 100		
IF-EIM-univ	2.0 [1.0, 3.0]	4.0 [3.0, 7.0]	9.0 [6.0, 16.0]
RF-V	2.0 [1.0, 3.0]	4.0 [3.0, 7.0]	10.0 [6.0, 17.0]
	n = 500		
IF-EIM-univ	2.0 [1.0, 2.0]	4.0 [3.0, 5.0]	9.0 [8.0, 10.0]
RF-V	2.0 [1.0, 2.0]	4.0 [3.0, 5.0]	9.0 [7.0, 10.0]
	n = 1000		
IF-EIM-univ	2.0 [1.0, 2.0]	4.0 [3.0, 5.0]	9.0 [8.0, 10.0]
RF-V	2.0 [1.0, 2.0]	4.0 [3.0, 5.0]	9.0 [8.0, 10.0]

Table 3: Simulation results – univariable effects. The numbers show the median ranks the respective variables obtained across the simulated data sets. The numbers in square brackets show the 25% quantiles and 75% quantiles of the ranks obtained for the simulated data sets. Note that for the variables with main effects only, each effect strength was represented by two variables in the simulation design. When calculating the quantities shown in the table, we simply pooled the ranks obtained for both variables of each effect strength.

The median ranks obtained with iRF are the smallest for this sample size. However, this result is misleading as for a large proportion of the simulation iterations the respective variable pairs were not selected at all by iRF. In these cases, iRF was clearly not successful in identifying these quantitative interaction effects. Beyond the results obtained for iRF, the best median rank is 19 in the small sample size scenario and was achieved using IF-EIM-quant for the strongest quantitative interaction effect. However, in practice a variable pair receiving a rank of 19 would most likely not be detected, as there would be too many other variable pairs with better ranks in that case. Nevertheless, for 25% of the simulation iterations the strongest quantitative interaction effect achieved a rank better than 5 in the small sample size scenario for IF-EIM-quant (Table 4).

For moderate and large sample sizes, IF-EIM-quant delivered a median rank of one for the variable pair with strong quantitative interaction effect. The variable pair with moderate quantitative interaction effect also received good median ranks using IF-EIM-quant for these sample sizes: a median rank of seven for sample size 500 and a median rank of three for sample size 1000. The weak quantitative interaction effect was not detectable using IF-EIM-quant. Apart from PA for the strong quantitative interaction effect, the other methods did not deliver median ranks in ranges that would be of use in practice. Generally, PA performed better than IMDMS and RF-V-pairs performed worst. Apart from the variable pair with strong quantitative interaction pair, iRF did not select the pairs with quantitative interaction effects for the (great) majority of simulation iterations for moderate and large sample sizes.

Effect:	Strong	Moderate	Weak
	n = 100		
IF-EIM-quant	19.0 [5.0, 75.8]	141.0 [33.0, 452.0]	675.0 [237.0, 1361.5]
RF-V-pairs	199.0 [79.5, 285.2]	329.5 [208.8, 491.2]	704.0 [493.5, 1066.5]
PA	107.5 [28.8, 579.8]	324.5 [91.0, 756.5]	729.0 [288.5, 1411.2]
IMDMS	77.5 [20.0, 189.2]	259.5 [111.8, 442.5]	499.5 [300.2, 872.5]
iRF	16.0 [5.5, 25.5] (46%)	29.5 [18.2, 36.8] (17%)	37.0 [30.0, 50.0] (2%)
	n = 500		
IF-EIM-quant	1.0 [1.0, 2.0]	7.0 [4.0, 20.0]	100.5 [35.0, 251.5]
RF-V-pairs	138.0 [79.8, 156.2]	331.0 [268.0, 392.2]	532.5 [457.0, 593.0]
PA	11.0 [5.0, 21.2]	34.0 [17.0, 149.8]	294.0 [100.0, 946.2]
IMDMS	22.0 [16.8, 30.0]	147.5 [71.2, 257.2]	510.5 [382.2, 589.5]
iRF	26.0 [18.0, 38.0] (85%)	59.0 [43.5, 73.0] (18%)	– [–, –] (0%)
	n = 1000		
IF-EIM-quant	1.0 [1.0, 1.0]	3.0 [2.0, 5.0]	43.0 [20.0, 108.0]
RF-V-pairs	138.5 [86.8, 142.0]	332.0 [271.0, 389.0]	570.0 [513.0, 626.0]
PA	11.0 [5.8, 46.2]	35.0 [14.0, 186.2]	360.0 [117.5, 955.8]
IMDMS	24.0 [19.0, 29.0]	160.0 [86.5, 211.8]	515.0 [442.8, 592.5]
iRF	28.5 [17.0, 44.0] (99%)	77.0 [65.0, 91.0] (22%)	87.0 [87.0, 87.0] (0%)

Table 4: Simulation results – quantitative interaction effects. The numbers show the median ranks that the respective variable pairs obtained across the simulated data sets. The numbers in square brackets show the 25% quantiles and 75% quantiles of the ranks obtained for the simulated data sets. In the case of iRF, the percentages of the simulated data sets for which the respective pairs were selected using iRF are shown in addition. For the moderate sample size scenario ($n = 500$) iRF did not select the variable pair with the weakest quantitative interaction effect for any of the simulated data sets, which is why the corresponding entry in the table is empty.

The summarised results obtained for the qualitative interaction effects are shown in Table 5 and the corresponding results for all data sets in Supplementary Figure S46. Here, IF-EIM-qual delivered better median ranks than IF-EIM-quant for the quantitative interaction effects (Table 4). While for the smallest sample size, only the variable pair with strongest effect received very low median ranks using IF-EIM-qual, in the case of moderate and large sample sizes, this method delivered very low median ranks for all three variable pairs. The other methods delivered worse results compared to the case of the quantitative interaction effects. PA again performed second-best here. However, the smallest median rank obtained using PA is 20.5, which is not small enough for practical purposes. The median ranks obtained using PA are considerably worse for sample size 1000 than for 500. This result is most likely attributable to the fact that, as described in Section 4.2.1, for the largest sample size, the computational burden associated with PA made it impossible to use 20000 trees per forest with this method. As seen in Table 5, iRF rarely selected variable pairs with qualitative interaction effects. In fact, only for one simulated dataset did iRF select a variable pair

Effect:	Strong	Moderate	Weak
	n = 100		
IF-EIM-qual	1.0 [1.0, 3.0]	10.0 [2.0, 217.5]	263.0 [21.8, 1058.8]
RF-V-pairs	1265.5 [813.2, 1721.2]	1323.5 [957.8, 1786.5]	1439.5 [1028.2, 1853.8]
PA	145.5 [48.5, 428.2]	403.5 [111.8, 1076.8]	932.5 [437.0, 1676.0]
IMDMS	800.5 [582.0, 1170.2]	906.5 [629.8, 1394.5]	1129.5 [804.0, 1495.2]
iRF	35.0 [35.0, 35.0] (0%)	– [–, –] (0%)	– [–, –] (0%)
	n = 500		
IF-EIM-qual	1.0 [1.0, 1.0]	2.0 [2.0, 2.0]	3.0 [3.0, 5.0]
RF-V-pairs	837.5 [748.5, 1032.8]	1022.5 [802.2, 1311.0]	1256.0 [973.5, 1577.0]
PA	20.5 [15.8, 26.0]	37.0 [26.0, 72.0]	152.0 [79.0, 321.5]
IMDMS	740.0 [692.8, 796.5]	801.5 [739.0, 926.0]	919.0 [772.2, 1116.8]
iRF	– [–, –] (0%)	– [–, –] (0%)	– [–, –] (0%)
	n = 1000		
IF-EIM-qual	1.0 [1.0, 1.0]	2.0 [2.0, 2.0]	3.0 [3.0, 3.0]
RF-V-pairs	745.0 [739.0, 782.5]	825.0 [759.0, 924.2]	1149.5 [947.0, 1470.0]
PA	52.0 [26.0, 134.0]	188.5 [106.5, 374.5]	796.0 [297.5, 1538.5]
IMDMS	739.0 [739.0, 740.0]	740.0 [739.0, 779.2]	849.5 [770.0, 955.2]
iRF	– [–, –] (0%)	– [–, –] (0%)	– [–, –] (0%)

Table 5: Simulation results – qualitative interaction effects. The numbers show the median ranks the respective variable pairs obtained across the simulated data sets. The numbers in square brackets show the 25% quantiles and 75% quantiles of the ranks obtained for the simulated data sets. In the case of iRF, the percentages of the simulated data sets for which the respective pairs were selected using iRF are shown in addition. The empty entries for iRF show cases for which iRF did not select the corresponding variable pair for any of the simulated data sets.

of this kind, which was in the case of the variable pair with strongest qualitative interaction effect in the small sample size scenario. RF-V-pairs again delivered the worst median ranks. It is a reassuring result that the baseline method RF-V-pairs performed worst as it suggests that all the other studied approaches are able to differentiate between variable pairs with and without interaction effect at least to some extent.

As IF-EIM-quant is a measure targeting quantitative interaction effects, it should not only rank true quantitative interaction effects good, but also true qualitative interaction effects bad. If IF-EIM-quant would rank both quantitative and qualitative interaction effects good, it would no longer be possible to interpret variable pairs with large IF-EIM-quant values as candidates for variable pairs with strong quantitative interactions. Analogous considerations can be made with respect to IF-EIM-qual. Moreover, variable pairs that feature main effects only, but no interaction effects, should be ranked bad by IF-EIM-quant and IF-EIM-qual in order to avoid incorrectly classifying such variable pairs as featuring interactions. To study how specific the values of IF-EIM-quant and IF-EIM-qual are with respect to the interaction effect types they target, we calculated the median ranks variable

pairs with qualitative interaction effects obtained for IF-EIM-quant, the median ranks variable pairs with quantitative interaction effects obtained for IF-EIM-qual, and the median ranks variable pairs with only main effects, but no interaction effects, obtained for both metrics, IF-EIM-quant and IF-EIM-qual.

These results are shown in Table 6. The median ranks obtained using IF-EIM-qual and IF-EIM-quant for the variable pairs with quantitative and qualitative interaction effects, respectively, are bad, which suggests that the values of these metrics are sufficiently specific for the types of interaction effects they target. For the variable pairs with main effects only, the results differ between IF-EIM-qual and IF-EIM-quant: For IF-EIM-qual, the median ranks obtained for these variable pairs are very bad. This result is not surprising because bivariable influences of variable pairs with qualitative interaction effects differ strongly from bivariable influences of variables with main effects only. For IF-EIM-quant, the median ranks obtained for the variable pairs with main effects only are better than in the case of IF-EIM-qual. However, the reasonably strong quantitative interaction effects are still well identifiable using IF-EIM-quant, because the median ranks obtained for the variable pairs with only main effects are generally much worse than those obtained for the variable pairs with quantitative interaction effects (cf. Table 4). The only exception is that the variable pair, for which both variables had the strongest main effects, received better median ranks than the variable pair with weak quantitative interaction effect. The median ranks obtained for the pair with weak quantitative interaction effect were, however, not in ranges that would be of use in practice anyway. We also computed the median ranks obtained for these variable pairs with main effects only using the competing methods (Supplementary Table S1). Here, except for in the case of the variable pair with the weakest effects, the median ranks obtained using IF-EIM-quant (and IF-EIM-qual) were worse than those obtained using the other methods with the exception of RF-V-pairs. The latter interestingly delivered slightly worse ranks in some settings with weak and moderate effects. The variable pair with strongest main effects, but no interaction effects, had a median rank of 1 for all methods except for IF-EIM-qual and IF-EIM-quant. These results strongly suggest that IF-EIM-qual and IF-EIM-quant are better able to differentiate between interacting variable pairs and variable pairs that feature main effects only than the other compared approaches.

5 Discussion

In this paper, we have proposed and evaluated a random forest type method, interaction forests, that allows ranking of the importance of quantitative and qualitative interaction effects between variable pairs using a novel measure,

	n = 100	n = 500	n = 1000
IF-EIM-qual: Quantitative interaction effects			
Strong	533.0 [177.2, 1190.0]	235.0 [40.0, 744.0]	86.5 [14.0, 397.0]
Moderate	705.5 [249.8, 1269.5]	403.5 [92.0, 1020.2]	299.0 [34.0, 788.2]
Weak	1020.5 [470.0, 1640.0]	784.5 [339.8, 1360.2]	664.5 [233.8, 1266.2]
IF-EIM-quant: Qualitative interaction effects			
Strong	252.0 [110.0, 515.5]	177.0 [81.5, 307.2]	169.0 [72.5, 293.5]
Moderate	429.5 [170.5, 838.0]	254.5 [136.2, 517.0]	279.5 [165.8, 478.8]
Weak	938.5 [462.2, 1450.2]	466.0 [275.5, 800.5]	500.0 [334.8, 891.2]
IF-EIM-qual: Pairs with only main effects			
Strong	992.5 [591.2, 1414.2]	741.0 [382.0, 1117.5]	665.0 [372.8, 1031.0]
Moderate	1033.0 [606.2, 1460.5]	816.0 [411.5, 1350.8]	739.5 [345.0, 1134.2]
Weak	1004.5 [571.2, 1589.0]	951.0 [487.5, 1443.2]	788.5 [328.2, 1328.0]
IF-EIM-quant: Pairs with only main effects			
Strong	28.0 [6.0, 114.8]	23.5 [6.8, 124.0]	17.0 [5.0, 72.0]
Moderate	93.5 [20.0, 329.8]	52.0 [11.0, 211.2]	31.0 [12.0, 140.0]
Weak	338.0 [154.8, 1028.2]	192.0 [65.0, 465.5]	111.0 [29.8, 346.8]

Table 6: Simulation results – specificity of IF-EIM-qual and IF-EIM-quant. The first (upper) part of the table shows the median ranks the variable pairs with quantitative interaction effects obtained using IF-EIM-qual. The second part of the table shows the median ranks the variable pairs with qualitative interaction effects obtained using IF-EIM-quant. The third and fourth part of the table show the median ranks variable pairs with only main effects, but no interaction effects obtained using IF-EIM-qual and IF-EIM-quant, respectively. For the variable pairs with only main effects, but no interaction effects, we considered the pair with the two variables that both have strong effects ("Strong"), the pair with the two variables that both have moderate effects ("Moderate"), and the pair with the two variables that both have weak effects ("Weak", cf. Table 2). The median ranks were obtained across the simulated data sets. The numbers in square brackets show the 25% quantiles and 75% quantiles of the ranks obtained for the simulated data sets.

the Effect Importance Measure (EIM). The concept of interaction forests is quite similar to that of the well-established conventional random forests. However, interaction forests do not suffer from the disadvantage of conventional random forests that interaction effects between variable pairs without particularly strong marginal effects are not taken into account sufficiently. Using a large real-data study we showed that interaction forests tend to feature a better predictive performance than random forests and competing random forest-based approaches that use multivariable splitting. A simulation study suggested that EIM is better able to detect variable pairs with quantitative and qualitative interaction effects than competing approaches. Here, the rankings obtained from the EIM value lists for quantitative interaction effects on the one hand and qualitative interaction effects on the other were confirmed to be reasonably specific for each of these two types

of interaction effects.

In the spirit of interpretable machine learning, interaction forests focus on well interpretable types of interaction effects that are easy to communicate. We propose estimating the bivariable influences of variable pairs with large quantitative and qualitative EIM values using flexible regression techniques and then visualising them. In our R package `diversityForest` we offer functions that perform this task in an automated way. This makes it possible to learn what forms the interaction effects between the individual interacting variable pairs take, which is crucial for interpretive purposes.

Interaction forests are specific diversity forests where the distinguishing feature of the latter is that they make using complex split procedures possible by strongly limiting the number of tried candidate splits in tree construction. While Hornung (2020) evaluated diversity forests exclusively for classical univariable, binary splitting, interaction forests are the first diversity forest method that uses a complex split procedure. The latter procedure would not have been feasible computationally using conventional split selection, which illustrates the practicability of diversity forests. Note that, beyond the problem of the computational burden, it might be anyway counterproductive regarding predictive performance to optimise multivariable splits over all possible splits in the context of a random forest type method. Given the flexibility of multivariable splits, the relevant predictive information in the training data would likely be modelled by only a few consecutive locally optimised multivariable splits. This would result in small trees that make similar predictions. However, as shown by Breiman (2001), in order to achieve a good predictive performance, the predictions of the trees in a random forest should be diverse.

While the splitting procedure used in interaction forests is complex, it is only as complex as necessary for modelling the targeted types of easily interpretable interaction effects. These types of interaction effects can be communicated clearly using statements of the following kind: "The strength of the positive (negative) effect of variable A on the outcome depends on the level of variable B." for quantitative interactions, and "For observations with small values of variable B, the effect of variable A is positive (negative), but for observations with large values of B, the effect of A is negative (positive)." for qualitative interactions. Obtaining separate EIM value lists for quantitative and qualitative interactions is necessary from a technical point of view, because the (adjusted) quantitative and qualitative EIM values do not live on the same scale, making it difficult to pool these values. However, differentiating between quantitative and qualitative interaction effects is also meaningful because these two types are interpreted in different ways and qualitative interactions can be expected to be rarer than quantitative interactions.

The simulation study suggested that interaction forests perform better with respect to detecting qualitative than quantitative interaction effects.

This result is not surprising, because qualitative interaction effects are in general stronger than quantitative interaction effects. Therefore, variable pairs featuring qualitative interaction effects are more distinguishable from non-interacting variable pairs than are variable pairs featuring quantitative interaction effects. In contrast to the result obtained using interaction forests, the competing methods performed better with respect to detecting quantitative than qualitative interaction effects. This is likely related to our observation that these methods seem less effective than interaction forests with respect to differentiating interacting variable pairs from variable pairs with strong main effects only. While EIM performed well in detecting quantitative and qualitative interaction effects for moderate and large sample sizes, for small sample sizes only strong qualitative interaction effects were detectable consistently using EIM.

The implementation of interaction forests in the R package `diversityForest` is closely based on the popular random forest implementation in the R package `ranger` (Wright and Ziegler, 2017). Tree construction and the calculation of the EIM values is performed in C++ using all cores available on the system by default. Using a desktop computer (2.5 GHz, 32 GB RAM, 12 cores), we performed a quick runtime check with data from the simulation study: Constructing 20000 trees and calculating univariable, quantitative, and qualitative EIM values took 0.20 minutes, 1.39 minutes, and 3.30 minutes for samples sizes 100, 500, and 1000, respectively. In the simulation study, the number of variables was 68 with 2278 possible variable pairs. For smaller numbers of variables, the numbers of possible variable pairs become much smaller, which is why the calculations are much faster for lower dimensional data. Moreover, the number of considered (i.e., pre-selected) variable pairs is at most 5000 in interaction forests. Therefore, even for high-dimensional data (e.g., genomics data) the computing times do not become larger than roughly two times those for the simulated data. While these computation times are reasonable, the calculation of the permutation VIM values of classical random forest (using `ranger`) is nevertheless considerably faster than the calculation of the EIM values, since for the former is not necessary to consider the variable pairs.

While interaction forests are implemented for categorical, continuous, and survival outcomes in `diversityForest`, we have evaluated its performance exclusively for categorical outcomes. However, with the exception of the split criterion used, the components of the interaction forests algorithm (e.g., the split types and the EIM calculation procedure) are the same for continuous and survival outcomes. Therefore, the results obtained in this paper should be largely transferable to continuous and survival outcomes.

An interesting venue for future research would be the development of a procedure for statistically testing whether the values of EIM differ significantly from zero. For the classical permutation VIM, a number of such approaches have already been developed. Many of these methods are based on

permutation strategies (Tang *et al.*, 2009; Altmann *et al.*, 2010; Hapfelmeier and Ulm, 2013) and require large numbers of re-computations of the original forest under permutations of the data. Therefore, they would not be suitable in the case of interaction forests as the latter are computationally more demanding than conventional random forests. The Vita (Janitza *et al.*, 2018) approach could, however, be a promising candidate for implementing such a test for the EIM values. Because Vita is based on the empirical distribution of the VIM values, it is suitable specifically for data with many variables, where a large proportion of these are uninformative. However, this does not pose an issue in the cases of the quantitative and qualitative EIM values: First, the number of possible variable pairs soon becomes large beyond data sets with only few variables. Thus, there are generally many quantitative and qualitative EIM values which makes it possible to work with the empirical distributions of these values. Second, interaction effects are usually sparse. Independent of the issue of testing, interaction forests can be used as exploratory tools to obtain interesting interpretable insights into the interplay between variables in non-parametric predictive modelling.

Acknowledgements The authors thank Maximilian Mandl, Christina Nießl, and Raphael Rehms for helpful comments and Anna Jacob for valuable language corrections. This work was supported by the German Science Foundation (DFG-Einzelförderung HO6422/1-2 to Roman Hornung and BO3139/6-2 to Anne-Laure Boulesteix).

Supplementary Material

Supplementary Material 1: PDF file with further contents referred to in the paper

URL (**NOTE**: the link does not work when clicking on it; it has to be copied into the browser window): `http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/070_drittmittel/hornung/interactionforest_suppfiles/suppmat1_hornungboulesteix2021.pdf`

Supplementary Material 2: Electronic Appendix

URL (**NOTE**: the link does not work when clicking on it; it has to be copied into the browser window): `http://www.ibe.med.uni-muenchen.de/organisation/mitarbeiter/070_drittmittel/hornung/interactionforest_suppfiles/suppmat2_hornungboulesteix2021.zip`

The above folder contains all R Code written to perform the analyses presented in this paper and in Supplementary Material 1 as well as the pre-processed versions of the data sets as used in the analyses.

References

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, **26**, 1340–1347.

Basu, S., Kumbier, K., Brown, J. B., and Yu, B. (2018). Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences (PNAS)*, **115**(8), 1943–1948.

Bertsimas, D. and Dunn, J. (2017). Optimal classification trees. *Machine Learning*, **106**, 1039–1082.

Boulesteix, A.-L., Janitza, S., Hapfelmeier, A., Van Steen, K., and Strobl, C. (2015a). Letter to the editor: on the term 'interaction' and related phrases in the literature on random forests. *Briefings in Bioinformatics*, **16**(2), 338–345.

Boulesteix, A.-L., Stierle, V., and Hapfelmeier, A. (2015b). Publication bias in methodological computational research. *Cancer Informatics*, **14**, 11–19.

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

Breiman, L., Friedman, J. H., Olshen, R. A., and Ston, C. J. (1984). *Classification and Regression Trees*. Wadsworth International Group, Monterey, CA.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P., and Eerdewegh, P. V. (2005). Identifying SNPs predictive of phenotype using random forests. *Genetic Epidemiology*, **28**, 171–182.

Chen, Z. and Zhang, W. (2013). Integrative analysis using module-guided random forests reveals correlated genetic factors related to mouse weight. *PLoS Computational Biology*, **9**(3), e1002956.

Couronné, R., Probst, P., and Boulesteix, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, **19**, 270.

Dazard, J.-E., Ishwaran, H., Mehlotra, R., Weinberg, A., and Zimmerman, P. (2018). Ensemble survival tree models to reveal pairwise interactions of variables with time-to-events outcomes in low-dimensional setting. *Statistical Applications in Genetics and Molecular Biology*, **17**(1), 20170038.

Du, J. and Linero, A. (2019). Interaction Detection with Bayesian Decision Tree Ensembles. In K. Chaudhuri and M. Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 108–117.

Gashler, M., Giraud-Carrier, C., and Martinez, T. (2008). Decision tree ensemble: Small heterogeneous is better than large homogeneous. In M. A. Wani, X.-W. Chen, D. Casasent, L. A. Kurgan, T. Hu, and K. Hafeez, editors, *Seventh International Conference on Machine Learning and Applications*, pages 900–905.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, **63**(1), 3–42.

Hapfelmeier, A. and Ulm, K. (2013). A new variable selection approach using random forests. *Computational Statistics & Data Analysis*, **60**, 50–69.

Hapfelmeier, A., Hothorn, T., Ulm, K., and Strobl, C. (2014). A new variable importance measure for random forests with missing data. *Statistics and Computing*, **24**, 21–34.

Hornung, R. (2020). Diversity forests: Using split sampling to allow for complex split procedures in random forest. Technical report 234, Department of Statistics, University of Munich.

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, **1**, 519–537.

Janitza, S., Celik, E., and Boulesteix, A.-L. (2018). A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, **12**, 885–915.

Jiang, R., Tang, W., Wu, X., and Fu, W. (2009). A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics*, **10**(Suppl. 1), S65.

Kelly, C. and Okada, K. (2012). Variable interaction measures with random forest classifiers. In *Proceedings of the 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 154–157.

Kim, H. and Loh, W.-Y. (2001). Classification trees with unbiased multiway splits. *Journal of the American Statistical Association*, **96**(454), 589–604.

Li, J., Malley, J. D., Andrew, A. S., Karagas, M. R., and Moore, J. H. (2016). Detecting gene-gene interactions using a permutation-based random forest method. *BioData Mining*, **9**, 14.

Loh, W.-Y. (2002). Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica*, **12**, 361–386.

Menze, B. H., Kelm, B. M., Splitthoff, D. N., Koethe, U., and Hamprecht, F. A. (2011). On oblique random forests. In D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, editors, *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 453–469.

Molnar, C., Casalicchio, G., and Bischl, B. (2020). Interpretable machine learning - a brief history, state-of-the-art and challenges. arXiv:2010.09337.

Ng, V. W. and Breiman, L. (2005). Bivariate variable selection for classification problem. Technical report 692, Department of Statistics, University of California, Berkeley, CA.

Peto, R. (1982). Statistical aspects of cancer trials. In K. E. Halnam, editor, *Treatment of Cancer*. Chapman & Hall: London.

Probst, P., Boulesteix, A.-L., and Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, **20**(53), 1–32.

Rainforth, T. and Wood, F. (2015). Canonical correlation forests. arXiv:1507.05444.

Rodríguez, J. J., Kuncheva, L. I., and Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**(10), 1619–1630.

Sorokina, D., Caruana, R., and Riedewald, M. (2007). Additive groves of regression trees. In J. N. Kok, J. Koronacki, R. L. Mantaras, S. M. S, D. Mladenič, and A. Skowron, editors, *Proceedings of the 18th European conference on Machine Learning*, pages 323–334.

Sorokina, D., Caruana, R., Riedewald, M., and Fink, D. (2008). Detecting statistical interactions with additive groves of trees. In W. Cohen, A. K. McCallum, and S. T. Roweis, editors, *Proceedings of the 25th international conference on Machine learning*, pages 1000–1007.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, **8**, 25.

Tang, R., Sinnwell, J. P., Li, J., Rider, D. N., de Andrade, M., and Biernacka, J. M. (2009). Identification of genes and haplotypes that predict rheumatoid arthritis using random forest. *BMC Proceedings*, **3**, S68.

Vanschoren, J., van Rijn, J. N., Bischl, B., and Torgo, L. (2013). OpenML: Networked Science in Machine Learning. *SIGKDD Explorations*, **15**(2), 49–60.

Wright, M. N. and König, I. R. (2019). Splitting on categorical predictors in random forests. *PeerJ*, **7**, e6339.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, **77**, 1–17.

Wright, M. N., Ziegler, A., and König, I. R. (2016). Do little interactions get lost in dark random forests? *BMC Bioinformatics*, **17**, 145.

Yoshida, M. and Koike, A. (2011). SNPInterForest: A new method for detecting epistatic interactions. *BMC Bioinformatics*, **12**, 469.