

A computationally fast variable importance test for random forests for high-dimensional data

Silke Janitza¹  · Ender Celik¹ ·
Anne-Laure Boulesteix¹

Received: 24 October 2015 / Revised: 13 August 2016 / Accepted: 22 August 2016 /
Published online: 29 November 2016
© Springer-Verlag Berlin Heidelberg 2016

Abstract Random forests are a commonly used tool for classification and for ranking candidate predictors based on the so-called variable importance measures. These measures attribute scores to the variables reflecting their importance. A drawback of variable importance measures is that there is no natural cutoff that can be used to discriminate between important and non-important variables. Several approaches, for example approaches based on hypothesis testing, were developed for addressing this problem. The existing testing approaches require the repeated computation of random forests. While for low-dimensional settings those approaches might be computationally tractable, for high-dimensional settings typically including thousands of candidate predictors, computing time is enormous. In this article a computationally fast heuristic variable importance test is proposed that is appropriate for high-dimensional data where many variables do not carry any information. The testing approach is based on a modified version of the permutation variable importance, which is inspired by cross-validation procedures. The new approach is tested and compared to the approach of Altmann and colleagues using simulation studies, which are based on real data from high-dimensional binary classification settings. The new approach controls the type I error and has at least comparable power at a substantially smaller computation time in the studies. Thus, it might be used as a computationally fast alternative to existing procedures for high-dimensional data settings where many variables do not carry any information. The new approach is implemented in the R package *vita*.

Electronic supplementary material The online version of this article (doi:[10.1007/s11634-016-0276-4](https://doi.org/10.1007/s11634-016-0276-4)) contains supplementary material, which is available to authorized users.

✉ Anne-Laure Boulesteix
boulesteix@ibe.med.uni-muenchen.de

¹ Department of Medical Informatics, Biometry and Epidemiology, University of Munich, Marchioninstr. 15, 81377 Munich, Germany

Keywords Gene selection · Feature selection · Random forests · Variable importance · Variable selection · Variable importance test

Mathematics Subject Classification 62F07 · 65C60 · 62-07

1 Introduction

Since its introduction in 2001 random forests have evolved to a popular classification and regression tool which is applied in many different domains. A random forest consists of several hundreds to thousands of decision trees. The trees are usually built from a bootstrap sample or a subsample of the original data (Breiman 2001). Random forests are fully non-parametric and thus offer a great flexibility. Moreover, they can even be applied in the statistically challenging setting in which the number of variables, p , is higher than the number of observations, n . This makes random forests especially attractive for complex high-dimensional molecular data applications and fast implementations of random forests are already available (e.g., Wright and Ziegler 2016; Schwarz et al. 2010).

A further advantage of random forests is that they also offer so-called variable importance measures, which can be used to rank variables according to their predictive abilities. The random forests method and its implemented importance measures have often been used for the identification of biomarkers (often genes) that can be used to differentiate between diseased and non-diseased subjects (e.g., Reif et al. 2009; Wang-Sattler et al. 2012; Yatsunenko et al. 2012). Identifying relevant genes is of high interest, not only for diagnosis of certain disorders but also to gain valuable insights into the functionality and mechanisms that lead to a specific disorder.

There are two commonly used variable importance measures, the Gini importance and the permutation importance. Several articles have shown that the Gini importance has undesirable properties (Strobl et al. 2007; Nicodemus and Malley 2009; Nicodemus 2011; Boulesteix et al. 2012). For example variables that offer many cutpoints systematically obtain higher Gini importance scores (Strobl et al. 2007). Insofar the permutation variable importance should be preferred. The permutation variable importance measure—also referred to as the mean decrease in accuracy—is computed as the change in prediction accuracy (measured e.g. by the error rate) when the variable of interest is not used for prediction. For non-relevant variables the change in accuracy is solely due to random variations. Thus, the importance scores of non-relevant variables randomly varies around zero. For relevant variables, in contrast, a worsening in prediction performance is expected, which is indicated by a positive importance score.

However, the distributions of non-relevant and relevant variables overlap, and there is no universally applicable threshold that can be used to determine which positive importance scores are large enough so that it is unlikely that these have occurred by chance. Therefore, approaches were proposed to determine the threshold based on the observed data. Often, in applications, a certain percentage of the highest ranked variables are selected (see, e.g., Díaz-Uriarte and De Andres 2006, and references therein). Reif et al. (2009) for example filtered out the 10% of variables with the

highest importance scores and used them for further considerations. However, one should be careful when selecting a pre-specified number of highest ranked variables and considering these as relevant because one would always identify some variables as relevant even in the absence of any associations between the variables and the disorder.

An ad-hoc approach consists in using the absolute value of the smallest observed importance score as a threshold for determining which variables are likely to be relevant, because one can be sure that the smallest observed importance score must have been occurred due purely to chance (Strobl et al. 2009). However, this approach has several disadvantages, two of them being that the threshold depends on one single observed importance score and that it becomes more extreme the more variables there are. It is thus clear that more elaborated approaches are needed. Recently, a variable selection strategy was proposed that uses the smallest observed importance score as a threshold for successively excluding non-relevant variables (Szymczak et al. 2016).

Variable selection strategies like that of Szymczak et al. (2016) might be promising for determining which variables are likely relevant. Other promising methods, which are considered in this paper, are testing procedures (Huynh-Thu et al. 2012). If the null importance distribution is known, p values for observed importance scores can be easily derived. The distribution of the variable importance is, however, unknown and it is difficult—if not impossible—to theoretically derive the null distribution because it depends on several different factors, including factors related to the data, such as correlations between the variables, the signal-to-noise ratio or the total number of variables, and including random forest specific factors, such as the choice of the number of randomly drawn candidate predictor variables for each split. This also explains the fact that there are hardly any theoretical results on the variable importance, which was also noted by others (e.g., Ishwaran 2008; Louppe et al. 2013).

A statistical test based on the supposed normality of a scaled version of the permutation variable importance was proposed by Breiman (2008). However, the procedure of Breiman (2008) was shown to have alarming statistical properties, and should not be used (Strobl and Zeileis 2008). Many testing approaches are based on permutation strategies (Tang et al. 2009; Altmann et al. 2010; Hapfelmeier and Ulm 2013). The advantage of such procedures is that no assumptions are made on the null distribution. However, such procedures are computationally demanding. Recently Hapfelmeier and Ulm (2013) published a comprehensive comparison study of different permutation-based testing approaches. They concluded from their study that their novel approach has higher statistical power than many of the existing approaches and controls the type I error. Their approach works as follows: For each variable that is tested for its association with the response, a large number of random forests (Hapfelmeier and Ulm (2013) used 400 in their studies) has to be computed. Each random forest is constructed based on a different permuted version of the variable and the importance score of the permuted version is computed. The p value for the variable is then computed as the fraction of variable importance scores (obtained for the permuted versions), that are greater than the variable importance of the original (i.e., unpermuted) version of the variable. The computation of p values for all variables thus requires computing as many random forests as predictor variables multiplied by the number of

permutation runs. This approach was developed and investigated for the classical low-dimensional setting which typically includes not more than a dozen covariates. It is obvious that with high-dimensional data such permutation-based approaches become computationally very demanding, and might even become practically infeasible. This is also the reason why the approach of Hapfelmeier and Ulm (2013) is not considered as a competing method to the new testing procedure for high-dimensional data that is proposed in this paper. The approach of Altmann et al. (2010) is considered instead.

Note that, in a statistical test we aim to draw conclusions about the value of a population parameter through the use of the observed sample. In the context of variable importance it is often not clear what this population parameter refers to and if it even exists. Thus, many testing approaches that were proposed for variable importance measures in random forests, should rather be regarded as heuristic methods that enable the selection of variables, instead of real statistical tests in the strict mathematical sense. However, for simplicity and to be consistent with the literature, we refer to such approaches as statistical tests in this paper, although it should be kept in mind that in the strict mathematical sense these are not statistical tests.

In this paper, we present a heuristic variable importance test for high-dimensional data that is computationally very fast because, in contrast to the existing approaches, the new testing procedure is not based on permutations. The testing procedure is applicable to permutation-based importance measures (note that here the term “permutation” refers to the properties of the variable importance measure and not to the test for the variable importance measure). The heuristic approach, which is proposed in this paper, reconstructs the null distribution based on the importance scores of variables that are likely non-relevant. Variables that are likely non-relevant are those for which prediction accuracy does not change at all (indicated by importance scores of zero) or where random variations lead to slightly better prediction accuracies when not using the variable for prediction (indicated by negative importance scores). The non-positive variable importance scores are then used to construct a distribution which is symmetric around zero. This distribution is then used as a null distribution for computing the p values.

The new testing approach is based on an alternative computation of the permutation variable importance measure. This has the drawback that, in contrast to many other testing approaches, the new approach cannot be used to derive p values from importance scores that are obtained as a by-product of the existing random forest implementations. Moreover, the testing approach is not suitable for all types of data—the data must contain a relatively large number of variables without any effect such that a large number of non-positive variables can be retrieved to sufficiently approximate the null distribution. In practice, it is usually unknown whether there is a large number of variables without any effect, but for some applications there might be previous knowledge on this. A high number of variables without any effect is for example typically present with genetic data, such as microarray or SNP data, so that the testing approach is primarily of practical relevance to high-dimensional genomic data settings. However, the method is also applicable to other data and is not restricted to genomic data.

This paper is structured as follows: in Sect. 2 the idea of random forests and their integrated permutation variable importance measure is briefly reviewed. An alter-

native computation of the permutation variable importance measure is subsequently described. After briefly reviewing the approach of [Altmann et al. \(2010\)](#), the new heuristic testing idea is outlined, which makes use of this alternative computation of the variable importance. In Sect. 3 we describe the designs considered in the simulation studies, which are conducted for investigating the new testing approach and for the comparison to the approach of [Altmann et al. \(2010\)](#). Section 4 shows the results of the studies and Sect. 5 gives a brief summary and discussion of the results.

2 Methods

2.1 Random forests

Random forests, developed by [Breiman \(2001\)](#), is an ensemble method that combines several classification trees. It can be used for classification and regression tasks as well as for more special analyses such as for survival analysis ([Ishwaran et al. 2008](#); [Hothorn et al. 2006](#)). In this paper, we focus on the use of random forests for classification tasks. Each tree in random forests is built from a bootstrap sample or from a subsample of the original data. The observations that are not used for the construction of a specific tree are termed *out-of-bag* (OOB) observations. The OOB observations can be used for estimating the error rate of a tree. Besides using only a random sample of the observations in a tree, there is a second random component that increases diversity among trees. This refers to the variables that are considered for a split. At each split in a tree a subset of `mtry` predictor variables is drawn from all candidate predictors and considered for the split. Among those variables, the one that provides the “best” split according to a specific split criterion is selected.

There are different variants of random forests, which basically differ in their splitting criteria. The oldest and most popular variant is that of [Breiman \(2001\)](#), which implements splits based on node impurity measures, such as the Gini index for classification trees. This version preferentially selects certain types of predictor variables for a split; for example, variables that have many possible splits, such as categorical variables with many categories, are preferentially selected, or variables with many missing values (see, e.g., [Kim and Loh 2001](#)). Therefore, one should be careful when using this version, especially if one wants to assess the importance of variables that have different scales. Approaches based on testing procedures performed at each split prevent these issues. A random forest version based on conditional inference tests was proposed by [Hothorn et al. \(2006\)](#), and is the better method of choice if there are, for example, variables of different scales or variables with missing values. However, we chose the version of Breiman for implementing our studies because of its computational speed. To be more precise, we used the R package `randomForest` ([Liaw and Wiener 2002](#)). We chose only settings with continuous predictor variables so that we do not expect that a preferential selection of certain types of variables for a split occurs in our studies. Moreover, we used subsampling (i.e., sampling from the original data without replacement) instead of bootstrapping in order to avoid undesired results induced by the bootstrap ([Strobl et al. 2007](#)).

2.2 Classical permutation variable importance

Let us consider a classification problem, where for an observation with covariates $X = (X_1, \dots, X_p)^\top$ comprising p random variables from a feature space \mathcal{X} , the aim is to derive a prediction $f(X)$ for the observation’s response Y , with both $f(X)$ and Y taking values in $\{1, \dots, k\}$ and k denoting the number of classes. Moreover, we introduce a variable X_j^* , which is an independent replication of X_j and is also independent of the response and of all other predictor variables (Gregorutti et al. 2013). According to previous work by Gregorutti et al. (2013) and Zhu et al. (2015), the variable importance for a variable X_j is defined by

$$VI_j = P(Y \neq f(X_1, \dots, X_j^*, \dots, X_p)) - P(Y \neq f(X_1, \dots, X_j, \dots, X_p)). \tag{1}$$

From now on the variable X_j is termed as relevant if $VI_j > 0$, that is, if the probability for an incorrect classification is larger when X_j^* , the independent replication of X_j , is used for deriving predictions than when X_j had been used. It is important to note that this definition of relevant predictor variables may also include variables that do not have their “own” effect on the response, but are associated with the response due to their correlation with influential predictor variables.

Now let $(x_i, y_i), i = 1, \dots, n$ denote a set of n realizations of independent and identically distributed (i.i.d.) replications of (X, Y) . If the design matrix is denoted by

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

and $y = (y_1, \dots, y_n)^\top$ denotes the response vector, the whole data is represented by

$$Z = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} & y_1 \\ x_{21} & x_{22} & \dots & x_{2p} & y_2 \\ \vdots & \vdots & & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} & y_n \end{bmatrix}.$$

In order to estimate VI_j within the random forest framework one tries to mimic the independent replication X_j^* in Eq. (1) by permuting the values of the j -th column of X (or Z) containing the covariate values for the variable X_j . Then tree predictions and the respective error rates are obtained based on Z and based on the modified data $\tilde{Z}_{(j)}$ with permuted values for variable X_j . The difference between these error rates gives an estimate of the importance of the variable X_j . The empirical counterpart of VI_j is thus obtained as

$$\widehat{VI}_j = \frac{1}{ntree} \sum_{t=1}^{ntree} \frac{1}{|OOB_t|} \sum_{i \in OOB_t} \{I(y_i \neq \hat{y}_{it}^*) - I(y_i \neq \hat{y}_{it})\}, \tag{2}$$

with $I(\cdot)$ denoting the indicator function, OOB_t denoting the set of indices for observations $i \in \{1, \dots, n\}$ that are out-of-bag for tree $t \in \{1, \dots, ntree\}$, with $ntree$ denoting the number of trees in the random forest, and \hat{y}_{it} and \hat{y}_{it}^* denoting the predictions by the t -th tree before and after permuting the values of the variable X_j , respectively. The \widehat{VI}_j defined in Eq. (2) is the importance measure that is used in the available random forests software and is referred to as the classical permutation variable importance from now on. Intuitively, if the variable X_j is not associated with the response, the trees' error rates before and after the permutation will not significantly differ because X_j does not aid in correctly classifying patients, no matter if its original values or the permuted values are used. In this case \widehat{VI}_j should fluctuate around zero. In contrast to that, if X_j is associated with the response, the error rate after permuting the values of the variable X_j is expected to be larger than the error rate before the permutation. In this case \widehat{VI}_j is expected to fluctuate around a positive constant.

2.3 Alternative computation of permutation variable importance

The classical permutation variable importance is not appropriate for the use in the new testing procedure as will be seen later in this paper. This is why an alternative computation of the variable importance is introduced in the following.

2.3.1 Cross-validated variable importance

The alternative computation of the variable importance is not based on the OOB observations, but uses a strategy that is inspired by the cross-validation procedure. The idea is to split the data first into k sets of approximately equal size. Then k random forests are constructed, where the l -th random forest is constructed based on observations that are not part of the l -th set S_l . For the l -th random forest only the observations from S_l are used for computing the variable importance.

For categorical response the *fold-specific variable importance* for predictor variable X_j is defined by

$$\widehat{VI}_j^{CV(l)} = \frac{1}{ntree} \sum_{t=1}^{ntree} \frac{1}{|S_l|} \sum_{i \in S_l} \{I(y_i \neq \hat{y}_{it}^*) - I(y_i \neq \hat{y}_{it})\}, \tag{3}$$

with $ntree$ denoting the number of trees in a random forest, $I(\cdot)$ denoting the indicator function and \hat{y}_{it} and \hat{y}_{it}^* denoting the predictions by the t -th tree before and after permuting the values of X_j , respectively. Note that the predictions \hat{y}_{it} and \hat{y}_{it}^* , $t = 1, \dots, ntree$, are obtained from the random forest, which is constructed based on a subset of the data \mathbf{Z} , in which the lines for the observations $i \in S_l$ are removed. Thus, the observations $i \in S_l$ are not used in tree construction.

The *cross-validated variable importance* for predictor variable X_j is then defined by

$$\widehat{VI}_j^{CV} = \frac{1}{k} \sum_{l=1}^k \widehat{VI}_j^{CV(l)}. \quad (4)$$

2.3.2 Hold-out variable importance

The most simple version of cross-validation results for $k = 2$, so that each of the two sets is once used for creating the random forest and once for deriving importance scores. In general, this method is also known as 2-fold cross-validation or the hold-out method. To differentiate it from cross-validation with $k \geq 3$, from now on we will refer to it as the hold-out method. The corresponding *hold-out variable importance* for predictor variable X_j is given by

$$\widehat{VI}_j^{HO} = \frac{1}{2} \sum_{l=1}^2 \widehat{VI}_j^{CV(l)}, \quad (5)$$

and directly results from setting k to 2 in Eq. (4). Thus, it is a special case of the cross-validated importance measure defined in Eq. (4).

2.3.3 Relatedness to the classical variable importance

In the following, the similarity between the classical variable importance defined in Eq. (2) and the cross-validated variable importance is shown. More precisely, they are approximately equal in the specific case with $k = n$.

If trees are constructed on bootstrap samples then $\lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^n = e^{-1} \approx 0.368$ of the observations are not used for constructing a specific tree, thus $|OOB_t| \approx 0.368n$ (the same holds for subsamples of size $0.632n$). The classical variable importance defined in Eq. (2) can thus be approximated as

$$\frac{1}{ntree} \cdot \frac{1}{0.368n} \sum_{t=1}^{ntree} \sum_{i \in OOB_t} \{I(y_i \neq \hat{y}_{it}^*) - I(y_i \neq \hat{y}_{it})\}, \quad (6)$$

with predictions \hat{y}_{it} and \hat{y}_{it}^* obtained from trees $t \in \{1, \dots, ntree\}$. In this representation the differences in error rates are first summed over the OOB observations and then over all trees. Under specific assumptions, which will be specified hereafter, the sum could be computed first over the trees for which an observation is out-of-bag and then over the observations. For this purpose, let us define the set SF_i (from ‘‘Sub-random Forest’’) that includes the indices of all trees for which the i th observation is out-of-bag and thus was not used to construct the trees. Following the argumentation above and in light of the fact that the number of trees in a random forest is usually very large, the number of trees that do not make use of a given observation can be asymptotically

approximated by $0.368ntree$. Thus, each sub-random forest contains approximately $0.368ntree$ trees. We can now rewrite Eq. (6) as follows:

$$\begin{aligned} & \frac{1}{n} \cdot \frac{1}{0.368ntree} \sum_{i=1}^n \sum_{t \in SF_i} \{I(y_i \neq \hat{y}_{it}^*) - I(y_i \neq \hat{y}_{it})\} \\ & \approx \frac{1}{n} \sum_{i=1}^n \frac{1}{|SF_i|} \sum_{t \in SF_i} \{I(y_i \neq \hat{y}_{it}^*) - I(y_i \neq \hat{y}_{it})\} \end{aligned} \tag{7}$$

with \hat{y}_{it} and \hat{y}_{it}^* , $t \in SF_i$, obtained from the sub-random forest which is constructed based on a subset of the data \mathbf{Z} , in which the line for the i th observation is removed.

The latter representation shows the close relationship to the cross-validated variable importance defined in Eq. (4) (note that with $k = n$, the set S_i includes only the observation i , thus $|S_i| = 1$).

It should be noted that the predictions \hat{y}_{it} and \hat{y}_{it}^* used to compute the classical variable importance and the cross-validated variable importance are not exactly the same. In the classical variable importance, there is only a single random forest and the predictions for the i th observation are obtained from a part of this forest, namely from all trees $t \in \{1, \dots, ntree\}$ that were not constructed based on the data (x_i, y_i) of the i th observation (i.e., from the sub-random forest SF_i). The sub-random forests for two observations i and j are therefore partly constructed on the same trees. This is in contrast to the cross-validated variable importance, in which k separate random forests are constructed. In the specific setting with $k = n$, as many random forests are constructed as observations exist. Accordingly, the predictions for the i th observation are obtained from a random forest that was constructed only for the observation i . Therefore, in contrast to the classical variable importance, the random forests used to derive predictions for two observations i and j are not constructed on the same trees.

2.4 Permutation-based testing approach of Altmann et al. (2010)

The testing approach of Altmann et al. (2010) has originally been proposed as heuristic for correcting importance measures which tend to favour certain types of predictors, such as the Gini importance measure. However, it is applicable to all kinds of importance measures of random forests. Besides its ability to correct importance measures, it outputs p values which are computed from importance scores. This feature enables the user to select relevant variables based on the p values.

In the first step of the method of Altmann, the variable importance scores are obtained for all variables. Any arbitrary importance measure may be used for computing the importance scores. In the second step, importance scores for settings in which the variable is not associated with the response are computed. Altmann et al. (2010) generate these settings by randomly permuting the response variable to break any associations between the response variable and all predictor variables. The data generated in this way is then used to construct a new random forest and to compute the importance scores for the predictor variables. The importance scores can be regarded as realizations drawn from the unknown null distribution. The procedure,

which involves the steps of randomly permuting the response vector, constructing a random forest and computing the importance scores, is repeated S times. For each variable there are S importance scores that can be regarded as realizations from the unknown null distribution. Finally, in the last step of the method of Altmann, the S importance scores are used to compute the p value for the variable. One possibility for deriving the p value consists in computing the fraction of S importance scores that are greater than the original importance score. This approach is referred to as the *non-parametric* approach in this paper since we do not make any assumptions on the distribution of importance scores of unrelated predictor variables. Alternatively, one can assume a parametric distribution such as the Gaussian, Log-normal or Gamma-distribution for the importance scores of unrelated predictor variables. The parameters for the respective distribution are replaced by their maximum likelihood estimates, which are computed based on the S importance scores of the considered variable. Having defined a specific distribution for the variable's null importance, the p value is computed as the probability of observing an importance score that is higher than the original importance score, given this distribution. We refer to this approach as *parametric* approach.

2.5 Naive and new testing approaches

Using the notation introduced in Sect. 2.2, we first define the hypothesis being tested. If X_j is independent of the response, permuting its values will not have any effect so that $f(X_1, \dots, X_j^*, \dots, X_p) = f(X_1, \dots, X_j, \dots, X_p)$ and $VI_j = 0$ (Zhu et al. 2015). If X_j is associated with the response, in contrast, $P(Y \neq f(X_1, \dots, X_j^*, \dots, X_p))$ is larger than $P(Y \neq f(X_1, \dots, X_j, \dots, X_p))$ and $VI_j > 0$. We thus define the null and alternative hypotheses as

$$H_0 : VI_j \leq 0 \quad \text{and} \quad H_1 : VI_j > 0.$$

From the considerations made in Sect. 2.2, the classical permutation variable importance might be expected to randomly vary around the value zero if variables are not associated with the response. In this paper, we investigate a new heuristic test approach which consists in approximating the null distribution based on the observed importance scores of those variables that are likely to be non-relevant. Variables that are likely non-relevant are those for which the error rate does not change at all (indicated by importance scores of zero) or where random variations lead to slightly higher error rates when not using the variable for prediction (indicated by negative importance scores). The non-positive variable importance scores are then used to construct a distribution which is symmetric around zero. This is done by “mirroring” the empirical distribution of the observed negative and zero importance scores on the y -axis. This results in a distribution that is symmetric around zero (see Fig. 1). It will be shown later that the classical permutation variable importance does not fulfil the assumption of a symmetric null distribution, but the hold-out variable importance does.

In the following, a technical description of the construction of the null distribution based on the observed non-positive variable importance scores is given. Let $M_1 =$

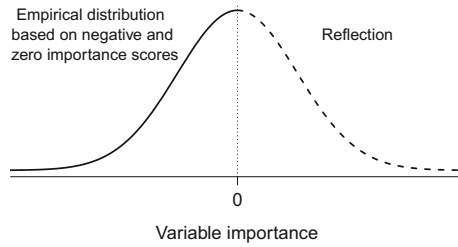


Fig. 1 Reconstruction of the null distribution based on variables that are likely non-relevant (i.e., with negative or zero importance scores). The negative part of the null distribution (solid line) is approximated based on the observed negative and zero importance scores. The positive part (dashed line) is obtained from reflection about the y-axis

$\{\widehat{VI}_j | \widehat{VI}_j < 0; j = 1, \dots, p\}$ denote the negative variable importance scores, and $M_2 = \{\widehat{VI}_j | \widehat{VI}_j = 0; j = 1, \dots, p\}$ is the set of importance scores which are zero. We define the hypothetical importance scores $M_3 = \{-\widehat{VI}_j | \widehat{VI}_j < 0; j = 1, \dots, p\} = -M_1$, which arise from multiplying the negative importance scores by -1 . The null distribution \hat{F}_0 is computed as the empirical cumulative distribution function of $M = M_1 \cup M_2 \cup M_3$. Based on \hat{F}_0 a p value for variable X_j is derived as $1 - \hat{F}_0(\widehat{VI}_j)$. We will term this procedure the “naive testing approach”.

Alternatively, one might use the hold-out variable importance (Eq. (5)) instead of the classical out-of-bag based importance (Eq. (2)). Then the approach is referred to as the “new testing approach”. The basic steps of the new testing approach are sketched in the following.

A new variable importance test for high-dimensional data

Step 1 The data is randomly partitioned into two sets of equal size. Each set is used to create a random forest. The two random forests are used to compute the hold-out variable importance \widehat{VI}_j^{HO} (see Eq. (5)) for variables $X_j, j = 1, \dots, p$.

Step 2 The null distribution for the hold-out variable importance is approximated based on the observed non-positive importance scores. For this purpose we define the sets

$$M_1 = \{\widehat{VI}_j^{HO} | \widehat{VI}_j^{HO} < 0; j = 1, \dots, p\} \text{ (i.e., all negative importance scores),}$$

$$M_2 = \{\widehat{VI}_j^{HO} | \widehat{VI}_j^{HO} = 0; j = 1, \dots, p\} \text{ (i.e., all importance scores of zero)}$$

and

$$M_3 = \{-\widehat{VI}_j^{HO} | \widehat{VI}_j^{HO} < 0; j = 1, \dots, p\} = -M_1 \text{ (i.e., all negative importance scores multiplied by } -1),$$

and consider the empirical cumulative distribution function \hat{F}_0 of $M = M_1 \cup M_2 \cup M_3$.

Step 3 The p value corresponding to the variable importance of predictor variable X_j is computed as $1 - \hat{F}_0(\widehat{VI}_j^{HO})$.

Table 1 Overview over high-dimensional genomic data sets used for the studies

Data set	No. predictor variables, p	No. observations, n	Source
Prostate cancer	6033	102	Singh et al. (2002)
Breast cancer	4869	77	van't Veer et al. (2002)
Leukemia	7129	72	Golub et al. (1999)
Colon cancer	2000	62	Alon et al. (1999)
Embryonal tumor	7129	60	Pomeroy et al. (2002)

In this paper, we use the hold-out version of the classical permutation variable importance which uses the difference in error rates before and after randomly permuting the values of the considered variable. The proposed testing procedure is very general in the sense that hold-out versions of different permutation-based variable importance measures might be used, such as the conditional permutation importance of [Strobl et al. \(2008\)](#), the AUC-based importance of [Janitza et al. \(2013\)](#), or the importance measures for ordinal responses considered in [Janitza et al. \(2016\)](#).

It is important to note that if one wants to use a different measure, say, the conditional importance of [Strobl et al. \(2008\)](#), the hold-out version of this measure should be computed, that is, the variable importance should be computed using the splitting procedure described in Sect. 2.3.

The new testing approach is implemented in the R package *vita*, which is based on the R package *randomForest* ([Liaw and Wiener 2002](#)). Currently, only the hold-out version of the classical variable importance measure is implemented. The R package *vita* also contains an implementation of the testing approach of [Altmann et al. \(2010\)](#).

3 Simulation studies

Since the new testing approach is suitable for high-dimensional genomic data, we only consider settings with large numbers of predictor variables and high signal-to-noise ratios. There is common consensus in the literature that it is very difficult—if not impossible—to simulate realistic complex data structures which capture all the patterns and sources of variability that are generated by a real biological system. Therefore we based our studies on five high-dimensional genomic data sets from real world applications (see Table 1 for an overview). These data sets were often used by various authors for binary classification purposes (e.g., [Díaz-Uriarte and De Andres 2006](#); [Dettling and Bühlmann 2003](#); [Tan and Gilbert 2003](#)). A brief description of the data sets is given in the Online Resource of ADAC. Note that no pre-selection of data sets based on the results was done, instead we report the results of all data sets that we analyzed, as has been recommended by [Boulesteix \(2015\)](#).

To study the properties of our test, we have to know which of the variables are relevant and which are not. In other words, we have to know the truth, which we can never know from real world data. Therefore, in our studies we used the design matrix of the real world data sets, but the response vector was generated anew according to a

Table 2 Overview of performed studies which differ in the inclusion of predictor variables with effect and in the presence of correlations between predictor variables

	Predictor variables with effect	Correlations between predictor variables
Study I	No	Yes
Study II	Yes	Yes
Study III	Yes	No

specified relation. Three different studies were performed. Table 2 gives an overview of the three studies.

In Study I, none of the predictor variables of a data set has an effect and there are correlations between predictor variables. In Studies II and III, some of the predictor variables have an effect on the response. While Study II includes correlated variables, in Study III all predictor variables are independent of each other.

We tested our new testing procedure and the naive testing procedure using Studies I, II and III. To obtain stable results we performed the computations for 500 repetitions of each study. Due to computational reasons, we performed only 200 repetitions of each study for Altmann's approach. We used the permutation importance defined in Eq. (2) for computing p values according to Altmann's approach. This enables a fair comparison of our new approach, which is based on the permutation variable importance measure, and Altmann's approach. We always computed p values for both approaches (non-parametric and parametric). Altmann et al. (2010) point out that a Kolmogorov–Smirnov test might be used to choose the most appropriate distribution for the parametric approach. In our studies we adhere to Algorithm 1 (outlined in the Supplement to Altmann et al. 2010), which uses a Gaussian distribution with mean and variance estimated by the arithmetic mean and sample variance, respectively. The number of permutations, S , in the permutation-based approach of Altmann and his colleagues should be chosen so that it is large enough. For the parametric approach the recommendation of Altmann et al. (2010) is a value S between 50 and 100. No recommendations were given for the non-parametric method. We always used a large value $S = 500$ in the studies to exclude the possibility that the performance of Altmann's approach may be related to a suboptimal choice of the number of permutations, S .

Each study is described in more detail in the following. The R codes implemented in our studies are available as Online Resource of ADAC.

3.1 Study I

The first study reflects scenarios where all predictor variables are pure noise. We used the design matrix and the response vector of the real data applications and modified this data. To destroy associations between the response vector and the design matrix we permuted the elements of the response vector. In this modified data, associations between predictor variables and the response are only due to chance. Note that the design matrix was not modified and correlations between predictor variables were preserved.

3.2 Study II

In the second study we simulated a scenario in which some of the variables have an effect on the response. We again used the design matrix of the real data applications since these reflect realistic correlation patterns but we generated a new response vector. This allows for a complex data scenario, but at the same time we have the information which of the variables are relevant.

The binary response Y for an observation with covariate vector $\mathbf{x}^\top = (x_1, x_2, \dots, x_p)$ was generated from a logistic regression model with success probability

$$P(Y = 1|\mathbf{x}) = \frac{\exp(x_1\beta_1 + x_2\beta_2 + \dots + x_p\beta_p)}{1 + \exp(x_1\beta_1 + x_2\beta_2 + \dots + x_p\beta_p)},$$

with p denoting the total number of predictor variables in the considered data set (see Table 1). The coefficients $\beta_1, \beta_2, \dots, \beta_p$ were chosen as follows: First we randomly drew j_1, j_2, \dots, j_{100} without replacement from the set $\{1, 2, \dots, p\}$ to define which of the variables have an effect on the response and should therefore be selected as relevant by a variable importance testing procedure. The corresponding coefficients $\beta_{j_1}, \beta_{j_2}, \dots, \beta_{j_{100}}$ were subsequently drawn from the set $\{-3, -2, -1, -0.5, 0.5, 1, 2, 3\}$, while ensuring that effects in the set were drawn equally often. All other coefficients were set to zero. Thus, there are 100 variables with an effect and $p - 100$ variables without any effect.

Although standardization is not necessary for the application of random forests in general, we standardized the design matrix in order to make effects comparable across variables of different scales.

3.3 Study III

This study includes only uncorrelated predictor variables. We used the design matrix of the real data sets and permuted the values within each predictor variable independently to create uncorrelated predictor variables. As with Study II, 100 of the variables were supposed to have an effect on the response. The approach for generating the response and for deciding which variables have an effect is exactly the same as described for Study II.

3.4 Parameter settings

We performed analyses under different parameter settings to see if the choice of parameters affects the results. All studies (Studies I, II, III) were performed

- for two different values for the parameter `mtry`: `mtry` = \sqrt{p} and `mtry` = $\frac{p}{5}$, with p denoting the number of predictor variables;
- for two different numbers of predictor variables. We used either a very large number of candidate predictors, namely that from the design matrices of the real data applications (see Table 1), or a subset of $p = 100$ predictor variables randomly

- drawn from the design matrices. In the studies with large predictor numbers, 100 variables had an effect, and in the studies with a subset of $p = 100$ predictor variables only 20 variables had an effect (only relevant to Study II and III);
- for two different sets from which the coefficients of relevant predictor variables were randomly drawn. Therefore, these sets determine the effects of relevant predictor variables (only relevant to Study II and III). The first set was chosen as $\{-3, -2, -1, -0.5, 0.5, 1, 2, 3\}$, and the second set with smaller effects was $\{-1, -0.8, -0.6, -0.4, -0.2, 0.2, 0.4, 0.6, 0.8, 1\}$. Since results were very similar for the two sets, only those for the effect set $\{-3, -2, -1, -0.5, 0.5, 1, 2, 3\}$ are shown.

The number of trees in the random forest was always set to 5000. We used subsamples of size $\lceil 0.632n \rceil$ to construct trees, with n denoting the number of total observations (Strobl et al. 2007). All other parameters not mentioned here were set to the default values so that trees were grown to maximal depth.

3.5 Evaluation criteria

One important aspect that was investigated in the studies is the statistical power of the testing approaches. The statistical power is generally defined as the probability of rejecting the null hypothesis, given that the null hypothesis is false. In this paper the null hypothesis for a variable X_j was defined as $VI_j \leq 0$ (cf. Sect. 2.5). If the null hypothesis is rejected for X_j , there is evidence that X_j is relevant. The statistical power of the testing approaches was explored by computing the fraction of variables with p value below $\alpha = 0.05$ of those that have an effect. Note that there are variables with different effect strengths in Studies II and III; the absolute effect strengths are 0.5, 1, 2, 3, or 0.2, 0.4, 0.6, 0.8, 1 in the alternative setting. For power considerations we computed the proportion of variables with p value below $\alpha = 0.05$ within each subset of variables with the same absolute effect.

The second important aspect concerns the validity of the testing approaches. The type I error of a test is defined as the probability of rejecting the null hypothesis, given that the null hypothesis is true. A test is valid if its type I error does not exceed the significance level α . In our studies we investigated if the testing procedures control the type I error by computing the fraction of variables with p value below $\alpha = 0.05$ among those variables that are not relevant. For this purpose we had to know which variables are not relevant. In Study I, none of the variables has an effect and thus none is relevant. In Study III, exactly those variables whose regression coefficient is zero are not relevant. In Study II, however, due to the correlation between the variables it is difficult to assess which variables are not relevant: predictor variables that do not have an “own” effect (i.e., those with coefficient of zero) but are correlated with variables that have an effect, might significantly improve the trees’ predictive abilities. Therefore in Study II, the regression coefficients cannot be used to judge which variables are not relevant, because variables with coefficients of zero can also be relevant. Thus, only Study I and III can be used for investigating the type I error.

In addition to type I error and power investigations, we inspected two further related issues. The first issue concerns the assumption of the new and the naive testing proce-

dures that under the null hypothesis the variable importance distribution is symmetric around zero. We empirically assessed if this is the case for the variable importance measures introduced in Sects. 2.2 and 2.3 by plotting the distribution of variable importance scores observed in Study I, where none of the variables is relevant. If we observe an asymmetric distribution or a distribution which is shifted along the x -axis, we expect the testing procedure to have a systematically too high or too low type I error.

The second issue concerns the discrimination between relevant and non-relevant variables by their importance scores. A testing procedure will have low statistical power if it is based on a variable importance measure that does not discriminate well between relevant and non-relevant variables. Thus we inspected the discriminative ability to see if the hold-out variable importance may be used in a testing procedure. We considered the classical permutation importance as “gold standard” and compared its discriminative ability to that of the hold-out importance. For these investigations we used Study III because, in contrast to Study II, we know which variables are relevant and which are not. The area under the curve was used as a measure for discriminative ability. Let the set of predictor variable indices, $B = \{1, \dots, p\}$, be partitioned into the disjoint sets $B = B_0 \cup B_1$, where B_0 represents the non-relevant variables and B_1 represents the relevant variables. The area under the curve is defined by

$$AUC = \frac{1}{|B_0||B_1|} \sum_{i \in B_0} \sum_{j \in B_1} \{I(\widehat{V}I_i < \widehat{V}I_j) + 0.5I(\widehat{V}I_i = \widehat{V}I_j)\} \quad (8)$$

where $|B_l|$ denotes the cardinality of B_l with $l \in \{0, 1\}$, and $I(\cdot)$ denotes the indicator function (see, e.g., Pepe 2004). Note that the area under the curve is usually used for evaluating the ability of a method (which may be for example a diagnostic test or a prediction model) to correctly discriminate between observations with binary outcomes (often diseased versus healthy). In our studies, in contrast, the area under the curve is computed considering the predictor variables X_1, \dots, X_p as the units to be predicted (as relevant or non-relevant variables) rather than the observations $i = 1, \dots, n$. The area under the curve here corresponds to an estimate of the probability that a randomly drawn relevant variable has a higher importance score than a randomly drawn non-relevant variable. An AUC value of 1 means that each of these relevant variables receives a higher importance score than any non-relevant variable, thus indicating perfect discrimination by the importance measure. An AUC value of 0.5 means that a randomly drawn relevant variable receives a higher importance score than a randomly drawn non-relevant variable in only half of the cases, indicating no discriminative ability by the importance measure.

4 Results

Before inspecting the type I error and the statistical power of the testing procedures, the properties of the variable importance measures are studied in this section. The naive and the new testing procedures rely on the assumption of a symmetric null distribution. Thus, in the first part of this section we empirically assess if the VIMs described in

Sects. 2.2 and 2.3 meet this assumption. Moreover, we explore whether the VIM based on cross-validation strategies yields reliable predictor rankings and might be used as an alternative to the classical out-of-bag based VIM. In the second part of this section, the type I error of the three testing approaches is investigated based on the results of Study I. Finally, in the last two parts of this section the statistical power is explored using Studies II and III and information on the runtime is given.

4.1 Properties of the classical and the cross-validation based permutation importance

4.1.1 Null distribution

Figure 2 shows the null variable importance distributions for the hold-out variable importance (left panel) and the classical variable importance (right panel) for the settings with large predictor space and m_{try} set to \sqrt{p} . Results are very similar for $m_{\text{try}} = \frac{p}{5}$, but there is a higher number of variables with importance score of zero (Online Resource Figure 1). The higher number of zero importance scores for higher values of m_{try} results from the fact that with large m_{try} the set of variables considered for a split contains a large number of variables that compete with each other and some of the variables are never selected and obtain importance scores of zero. With a smaller number of trees the number of variables that are never selected for a split further increases (results not shown).

The null distribution of the hold-out variable importance seems to be symmetric around zero and thus seems to satisfy the presumption of a symmetric null distribution. In contrast to that, the null distribution of the classical variable importance is not totally symmetric. In the studies with $p = 100$ this asymmetry is much more apparent (Online Resource Figure 9): All distributions are clearly positively skewed showing that a large fraction of variables have small negative importance scores, while smaller fractions of variables have large positive importance scores. The null distribution of the cross-validated variable importance looks very similar for $k \geq 3$ (see Online Resource Figures 1, 9). In contrast, the null distribution of the fold-specific variable importance is nearly symmetric around zero (results not shown). This seems to be contradictory since the cross-validated variable importance is the average of fold-specific variable importances. Further inspection of the simulation results reveals that this effect is possibly due to the overlap of random forests. For $k \geq 3$ the same observations are used for creating the random forests of several folds. For example, if we had three sets, S_1, S_2, S_3 , the first random forest is constructed using S_2 and S_3 , the second random forest is constructed using S_1 and S_3 , and the third random forest is based on S_1 and S_2 . Each pair of random forests have some part of the observations in common. For example, the first and the second random forests are both based on observations from set S_3 . The variables have similar predictive abilities for the sets $S_2 \cup S_3$ (on which the first random forest is trained) and $S_1 \cup S_3$ (on which the second random forest is trained). If high values for a variable X_j speak in favour of class 1 in the subset $S_2 \cup S_3$, then in the subset $S_1 \cup S_3$ high values for X_j will also speak in favour of class 1—even if there is, in reality, no association between X_j and the class membership.

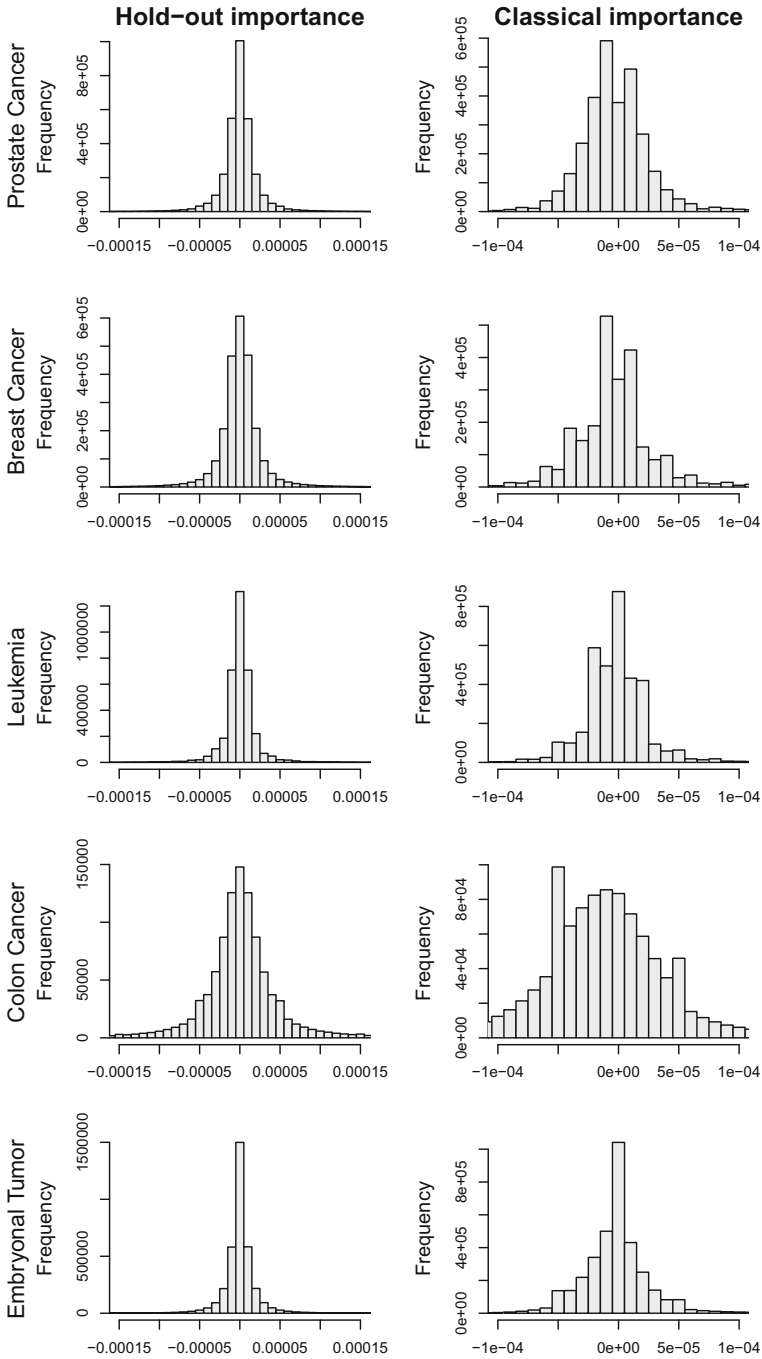


Fig. 2 Variable importance null distribution when using the classical permutation variable importance measure and the hold-out permutation variable importance measure and setting m_{try} to \sqrt{p} (default value). Distributions are shown for all variables and 500 repetitions of Study I (no relevant variables)

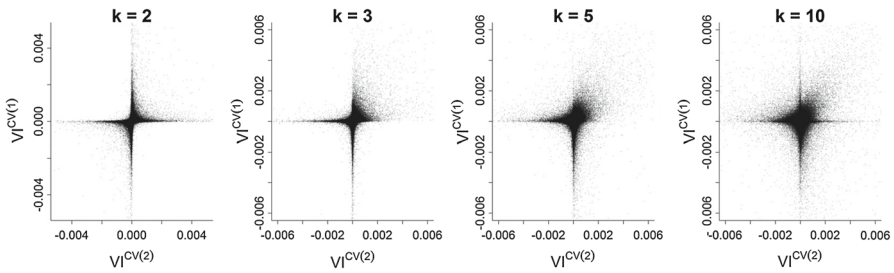


Fig. 3 Fold-specific variable importance for the first fold plotted against fold-specific variable importance for the second fold for all variables X_1, \dots, X_{2000} of Study I (no relevant variables; $m\tau_{XY} = \frac{p}{5}$) for the colon cancer data with $k = 2, k = 3, k = 5$ and $k = 10$. Five-hundred repetitions of Study I were performed, yielding a total of 2000×500 points shown in each *plot*

The two random forests then often select the same predictor variables for a split even in settings without any associations.

Thus, for $k \geq 3$ the same few variables will always obtain high fold-specific importance scores as can also be seen from empirical studies. In Fig. 3 the fold-specific variable importance scores for the first two folds (for the colon cancer data) are plotted against each other for different values of k . The fold-specific variable importance computed for 500 repetitions of Study I (no relevant variables) with $m\tau_{XY}$ set to $\frac{p}{5}$ are shown. Results for $m\tau_{XY} = \sqrt{p}$ are shown in Online Resource Figure 2. Similar results are obtained for the other data sets and when using only a subset of $p = 100$ predictor variables (not shown). For $k \geq 3$ (2nd, 3rd and 4th plot) we clearly observe the phenomenon just described: There are some variables which have large positive fold-specific variable importance scores for both folds resulting in a large cross-validated variable importance score. In contrast, there are not as many variables with negative fold-specific variable importances for both folds. From that it is clear that the cross-validated variable importance has a skewed null distribution.

Due to the relatedness of the classical variable importance and the cross-validated variable importance for $k = n$, the classical variable importance is affected in the same manner. In fact, the overlap between the sub-random forests which are used by the classical variable importance measure to derive predictions, might be even more pronounced. This is because the same trees are used in the sub-random forests (cf. Sect. 2.3.3). In contrast to that, the cross-validated variable importance involves the computation of k different random forests that are not based on the same trees, although the trees across the k random forests are very similar.

The hold-out variable importance, in contrast, is not affected in the same manner. Here the data is partitioned into the sets S_1 and S_2 . Each set—and correspondingly each observation within the set—is used for the construction of one random forest. The first random forest uses S_2 and the second random forest uses S_1 , resulting in two random forests which are completely independent of each other. The selection of variables for a split in the second random forest is thus independent of which variables have been selected in the first random forest. Therefore the mechanisms described for $k \geq 3$ do not apply for $k = 2$. This is also supported by the empirical results in Fig. 3 and Online Resource Figure 2 (first plot), where we observe an

a comparable number of variables with negative fold-specific variable importance scores for both folds as variables with positive fold-specific importance scores for both folds. Although, we note a substantially higher number of variables with both negative or positive fold-specific importance scores than variables with one negative and one positive fold-specific importance score. This might be explained by the fact that the variable importance for the first random forest is computed using observations from set S_1 , that have been used for the construction of the second random forest, and vice versa. A positive correlation might therefore be expected between the fold-specific importance scores. However, this has no effect on the symmetry of the null distribution of the hold-out variable importance.

To conclude, it was shown that the hold-out variable importance has a symmetric null distribution, while the classical importance and the cross-validated variable importance do not have a symmetric distribution. From those results we would expect that the new testing approach controls the type I error exactly, while the naive testing approach does not.

4.1.2 Discriminative ability

Figure 4 shows the discriminative ability of the classical and the hold-out variable importance for Study II (left) and Study III (right). Results are shown when using the default `mtry` value. The discriminative ability is measured in terms of the area under the curve (cf. Sect. 3.5). The hold-out variable importance and the classical variable importance have very similar discrimination ability. For Study III the performance of the hold-out importance is slightly better than the performance of the classical permutation importance. The results with `mtry` = $\frac{p}{5}$ are very similar (Online Resource Figure 3), and a slightly better performance of the hold-out importance can be observed

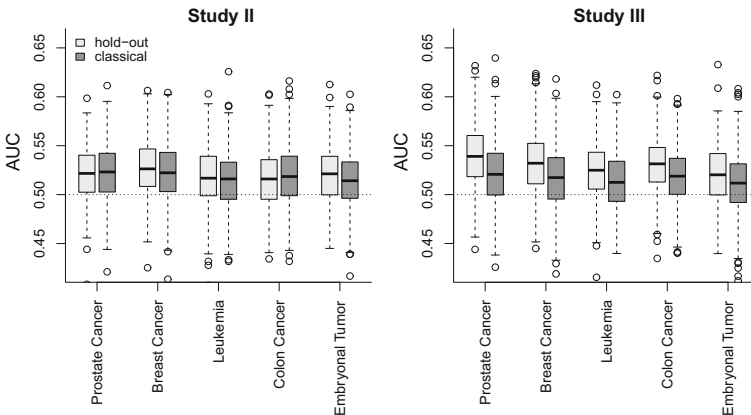


Fig. 4 Discriminative ability of the hold-out permutation variable importance measure and the classical permutation variable importance measure. Discriminative ability is measured by the area under the curve for 500 repetitions of Studies II (with relevant variables and correlations) and III (with relevant variables and without any correlations between predictor variables), and `mtry` was always \sqrt{p} . Values of 0.5 indicate no discriminative ability (*horizontal dotted line*)

in both, Study II and III. The results for the predictor space reduced to $p = 100$ are in line with these findings and are shown in Online Resource Figures 12 and 13. We therefore consider the hold-out variable importance a good measure to reflect the relevance of variables. The cross-validated variable importances with $k \geq 3$ have similar discriminative ability, too (results not shown). As with the classical variable importance, when computing the hold-out and cross-validated importance each observation is used for tree construction and for variable importance computation. In contrast to that, the fold-specific variable importance defined in Eq. (4) uses one part of the observations only for tree construction and the other part for variable importance computation. By building an average of fold-specific importances we make sure that all information is used for tree construction and for variable importance computation.

To summarize, according to these studies the hold-out importance does not have a worse discriminative ability than the classical variable importance and thus might be used as reasonable alternative to the classical importance. The hold-out importance, in addition, is symmetric around zero for variables not associated with the response—a criterion that is not fulfilled for the classical and the cross-validated variable importance. This motivates the use of this measure in the proposed testing procedure.

4.2 Type I error

The type I errors of all approaches are investigated using Study I and are depicted in Fig. 5. The type I errors of the new testing procedure are always close to the significance level $\alpha = 0.05$, indicating that the test does not systematically reject the null hypothesis too often or too rare. The studies with a subset of $p = 100$ give similar results (Online Resource Figure 10). These findings are in line with the results in Sect. 4.1.1 where it was shown that the null distribution of the hold-out variable importance is nearly symmetric around zero.

The results for the naive approach are also in line with the findings from Sect. 4.1.1. As expected, the type I error of the naive approach is systematically different from 0.05. More precisely, in the studies with large predictor numbers the naive approach always gives slightly too large type I errors if `mtreey` is set to the default value and too small type I errors if `mtreey` is $\frac{p}{5}$ (Fig. 5). In the studies with a smaller predictor number ($p = 100$), the type I errors are always close to 0.1 for both large and small `mtreey` values (Online Resource Figure 10). Therefore the naive approach should only be used with caution.

The non-parametric approach of Altmann et al. (2010) always gives type I errors close to 0.05 for both the studies with large and smaller ($p = 100$) predictor numbers. The type I error for the parametric approach is always considerably smaller than 0.05 in the studies with large predictor numbers, indicating that the parametric approach is too conservative in settings with large predictor numbers. In the studies with $p = 100$, in contrast, the type I error is much closer to 0.05. The variability in type I errors was smaller for Altmann's approach than for the new and the naive testing procedures. The variability increased for all testing approaches in settings with the predictor space reduced to $p = 100$.

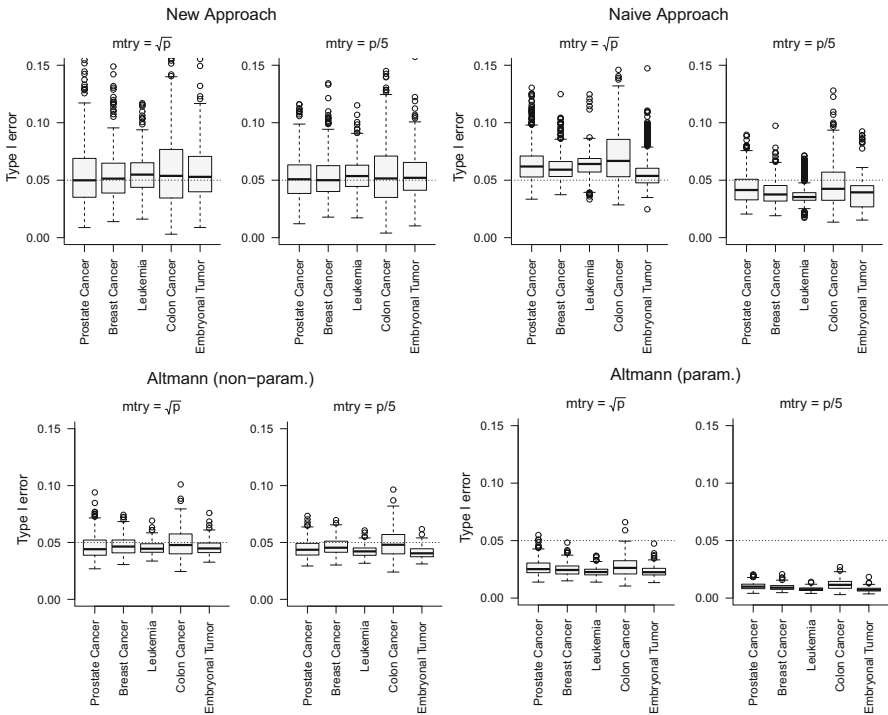


Fig. 5 Type I error in Study I for the new testing approach (which uses the hold-out permutation variable importance measure), the naive testing approach (which uses the classical permutation variable importance measure), the non-parametric Altmann approach and the parametric Altmann approach assuming a normal distribution. Hypothesis tests were performed at significance level $\alpha = 0.05$ (dotted horizontal line)

4.3 Statistical power

4.3.1 Study III

Figure 6 shows the proportion of variables with p value below 0.05 averaged over 500 (200 for Altmann’s approach, resp.) repetitions of Study III when the full predictor space is used. The proportions are computed among variables with the same absolute effect size of 0.5, 1, 2 and 3, respectively. In addition, the proportion of variables with p value below 0.05 among variables without effect (i.e., those variables X_j for which $\beta_j = 0$) are shown. For all testing procedures the proportion of variables with p value below 0.05 increases with increasing absolute effect size. This shows that variables with larger absolute effects are more easily identified. The parametric approach of Altmann et al. (2010) consistently has the smallest power. The non-parametric approach of Altmann and the new and the naive testing approaches have similar performance. However, the new approach has slightly higher statistical power than the non-parametric approach of Altmann, especially in settings with $mtry = \frac{p}{5}$. For $mtry = \sqrt{p}$ the naive approach has a slightly higher number of variables with p value below 0.05 than the other two approaches for both, non-relevant (i.e., $\beta_j = 0$)

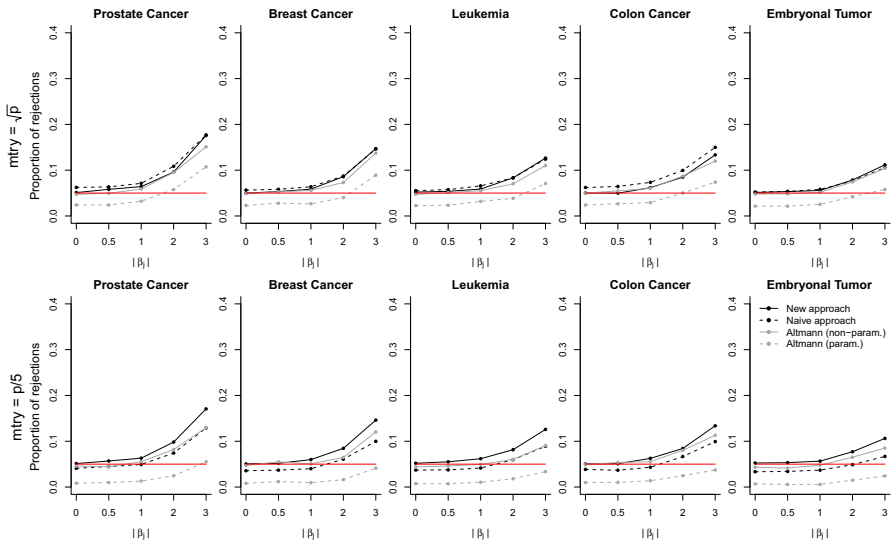


Fig. 6 Proportion of rejected null hypotheses among predictor variables X_j with specified absolute effect size $|\beta_j| \in \{0, 0.5, 1, 2, 3\}$. The mean proportions over 500 (200 for Altmann’s approach, resp.) repetitions of Study III (with relevant variables and without any correlations between predictor variables) are shown when using our new approach, the naive approach, the non-parametric Altmann approach and the parametric Altmann approach assuming a normal distribution, with $mtry$ set to \sqrt{p} (upper panel) and $\frac{p}{5}$ (lower panel). The horizontal line represents the 5% significance level

and relevant (i.e., $\beta_j \neq 0$) variables. In contrast, for $mtry = \frac{p}{5}$ the naive approach has fewer variables with p value below 0.05.

The results are in line with the results in Sect. 4.2, where it was shown that the type I error is smallest for the non-parametric approach of Altmann, and is higher (lower) for the naive approach than for the new approach if $mtry$ was set to the default value, \sqrt{p} (a large value, $\frac{p}{5}$). We conclude from the results that the new testing approach has the best performance in the settings with large numbers of predictor variables because it consistently has the highest power while preserving the type I error. However, the statistical power of all testing procedures was low.

In the studies with a subset of $p = 100$ predictor variables, we observed much higher statistical power for all approaches (Fig. 7). The naive approach does not preserve the type I error in the settings with reduced predictor space. This can be seen when inspecting the proportion of rejections among predictor variables X_j with $\beta_j = 0$ in Fig. 7. The same can be seen from the results of Study I (Online Resource Figure 10). In contrast to that, the new testing approach preserves the type I error and has similar—and on average even slightly higher—statistical power than the non-parametric and parametric approaches of Altmann.

The results presented so far show the mean performance over all repetitions of Study III. Thus, there is no information on the variability in the selected number of variables with effect. Further inspection reveals, however, that the variabilities for the naive approach, the new approach and the non-parametric approach of Altmann are similar (see Online Resource Figures 4–8, 14–18). The variability for the parametric

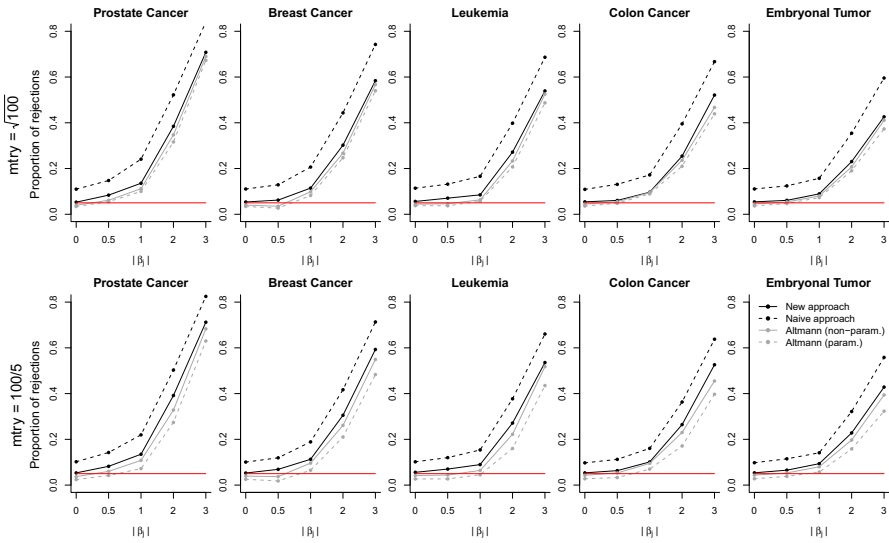


Fig. 7 Proportion of rejected null hypotheses among predictor variables X_j with specified absolute effect size $|\beta_j| \in \{0, 0.5, 1, 2, 3\}$. The mean proportions over 500 (200 for Altmann’s approach, resp.) repetitions of Study III with reduced predictor space (with relevant variables and without any correlations between predictor variables) are shown when using the new approach, the naive approach, the non-parametric Altmann approach and the parametric Altmann approach assuming a normal distribution, with $m\text{tr}_Y$ set to $\sqrt{100}$ (upper panel) and $\frac{100}{5}$ (lower panel). The horizontal line represents the 5% significance level

approach of Altmann, in contrast, is smaller, which is due to the fact that the approach is very conservative and selects only few variables.

4.3.2 Study II

The results for Study II are shown in Fig. 8. The proportion of variables with p value below 0.05 is largest when using the new testing approach. Thereafter, the proportion decreases bit by bit for the naive testing approach, the non-parametric approach of Altmann and the parametric approach of Altmann. The approaches of Altmann identify far less variables as relevant than the naive and the new testing procedures. With the parametric approach, the proportion of variables with p value below 0.05 is very low, especially if $m\text{tr}_Y$ is set to \sqrt{p} . It is even lower than 0.05, indicating that the parametric approach of Altmann is too conservative. This is not the case for the non-parametric approach of Altmann.

In many settings the proportion of identified variables X_j with $\beta_j = 0$ is very large and greatly exceeds 0.05. This is attributable to the correlations between the variables. From the construction of the naive and the new testing approach, variables which do not have an “own” effect, but are correlated to variables with effect, may be considered as relevant as long as they improve the trees’ prediction accuracy. Therefore, even variables that do not have a direct influence on the response are very often identified by the two procedures—but still not as often as variables with direct influence. In contrast to that, it is not clear if Altmann’s approach is also supposed

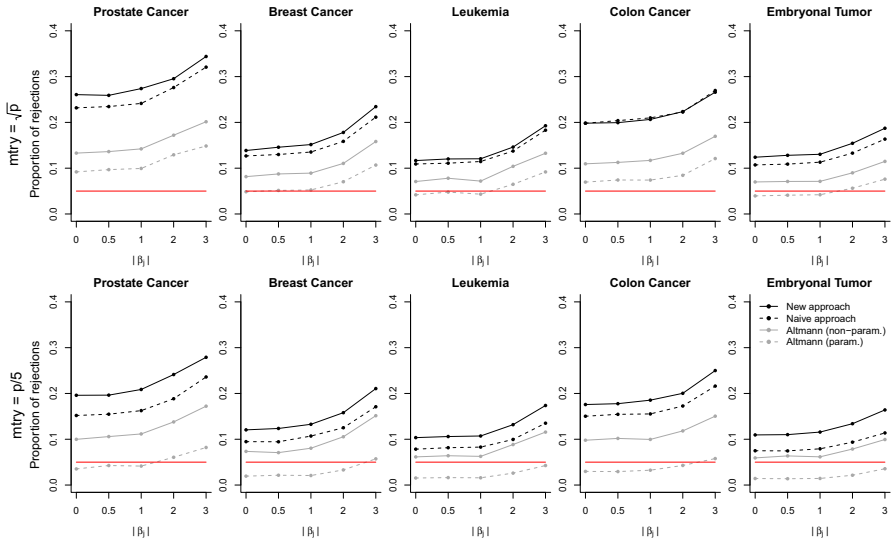


Fig. 8 Proportion of rejected null hypotheses among predictor variables X_j with specified absolute effect size $|\beta_j| \in \{0, 0.5, 1, 2, 3\}$. The mean proportions over 500 (200 for Altmann’s approach, resp.) repetitions of Study II (with relevant variables and with correlations between predictor variables) are shown when using our new approach, the naive approach, the non-parametric Altmann approach and the parametric Altmann approach assuming a normal distribution, with $mtry$ set to \sqrt{p} (upper panel) and $\frac{p}{5}$ (lower panel). The horizontal line represents the 5% significance level

to select variables that do not have a direct influence but correlate with variables that have an effect. Therefore, it is not possible to evaluate which testing approach has better performance in settings with correlated predictor variables.

When based on the conditional importance of Strobl et al. (2008), we expect the testing procedures to not as often select variables that are only associated with the response through their correlation to influential variables.

4.4 Runtime

We have seen from our studies that the approach of Altmann et al. (2010) and the novel approach are both valid testing procedures that preserve the type I error and have similar statistical power. The motivation for the development of a new testing procedure was that the existing approaches, like the approach of Altmann and his colleagues, are based on permutation testing strategies and thus involve the computation of many random forests. The novel testing approach, in contrast, involves the computations of variable importance scores from only two random forests. These two random forests are grown on samples of size $n/2$, which takes less time than if they were grown on samples of size n . Table 3 shows the runtimes for deriving the hold-out variable importance scores using the function `CVPVI` of the R package `vita`. The runtimes are an average of the runtimes for 20 repetitions of Studies I, II and III, in which each random forest consists of 5000 trees that were grown to full size. Note that,

Table 3 Runtimes (in minutes) for computing variable importance scores through the hold-out variable importance measure (function `CVPVI`) and the classical permutation variable importance measure (function `randomForest`)

	$mtry = \sqrt{p}$			$mtry = \frac{p}{5}$		
	Study I	Study II	Study III	Study I	Study II	Study III
Prostate cancer						
Hold-out	6.90	6.25	6.49	6.56	6.00	6.37
Classical	1.62	1.38	1.66	1.86	1.53	1.86
Breast cancer						
Hold-out	3.86	3.79	3.84	3.78	3.79	3.82
Classical	0.94	0.89	0.93	1.01	0.99	0.98
Leukemia						
Hold-out	5.47	5.45	5.45	5.17	5.29	5.30
Classical	1.22	1.13	1.23	1.32	1.29	1.35
Colon cancer						
Hold-out	1.37	1.41	1.35	1.38	1.31	1.31
Classical	0.28	0.26	0.28	0.29	0.28	0.30
Embryonal tumor						
Hold-out	4.74	4.61	4.59	4.64	4.50	4.62
Classical	0.93	0.94	0.96	1.03	1.02	1.04

The average runtime over 20 replications of Study I, Study II and Study III are shown

in contrast to random forest construction and variable importance computation, the operations performed to derive null distributions and p values are very fast and were thus neglected.

The runtime of the approach of Altmann is linearly related to the number of permutations performed: with S permutations, one has to construct $S + 1$ random forests (one additional random forest for the original unpermuted data), and variable importance scores have to be obtained for each random forest. The runtime of Altmann’s approach can therefore be approximated by multiplying the runtime for one single random forest construction and variable importance computation with $S + 1$. Table 3 shows the runtime for the construction of a single random forest (based on n observations and 5000 trees) and the derivation of the variable importance scores. Multiplying this number by 50 gives a lower bound for the runtime of Altmann’s approach. This is according to the recommendation of Altmann et al. (2010) who propose using a value for S between 50 and 100. However, if one wants to derive the p values in a non-parametric way much more than 100 permutations are needed. Large numbers of permutations are needed in this case, especially if multiple testing adjustments are planned (Phipson and Smyth 2010).

The runtime of Altmann’s approach is about $S/4$ to $S/5$ times as large as the runtime of our novel approach. In the studies presented in this paper, we set S to 500, yielding a 125 to 100 times faster computation when using the novel approach. In the considered settings, the number of candidate predictors ranges from 2000 to 7000 and the sample

sizes lie below 100. The difference in runtimes is expected to be far larger with large datasets including several thousands of observations and predictors. Growing a single random forest might last several days in such settings (Wright and Ziegler 2016), and permutation test procedures might become impractical. Note that both testing approaches were implemented based on the R package `randomForest`, and different runtimes are obtained when using a different package.

5 Discussion

During the last years, several approaches have been developed for hypothesis testing based on the variable importance implemented in random forests (see Hapfelmeier and Ulm 2013, and references therein). The existing approaches are computationally demanding and require the repeated computation of random forests. In this paper, we presented a fast implementation of a variable importance test that tests if a predictor variable significantly improves the trees' predictive ability. In all our studies the new testing procedure preserved the type I error and successfully identified at least as many relevant predictor variables as the testing approach of Altmann et al. (2010). However, our studies were restricted to classification tasks. Further studies are needed to assess if the new variable importance test can also be applied to settings with numeric response. This also involves theoretical investigations on the properties of the variable importance measure, which is of utmost importance to both classification and regression settings. There is a lack of theoretical results, which may be explained by the fact that "the permutation method [...] is complex and proved too difficult to analyse theoretically in any detail" (Ishwaran 2008). Future studies might aim to fill this gap.

Our testing procedure is based on a slightly modified version of the permutation variable importance, whose null distribution was empirically shown to be symmetric around zero. We gave an intuitive explanation of why this importance measure, contrary to the classical out-of-bag based version, follows a symmetric distribution. Further studies could address this issue from a more theoretical perspective. The classical permutation variable importance, in contrast, has a skewed null distribution and thus seems inappropriate for the application of our testing procedure. In our studies the testing approach based on the classical permutation importance worked quite well for settings with huge predictor numbers, but did not preserve the type I error in settings with fewer ($p = 100$) predictor variables. Thus it should be used with caution. We therefore strongly recommend the use of our testing procedure which is based on a modified version of the permutation variable importance. This approach has consistently been shown to precisely preserve the type I error in our studies.

The new testing approach is, however, not applicable to any high-dimensional data set. We expect that it may perform poorly if only a few non-positive importance scores are observed. If there are only a few variables with negative importance score or importance score of zero, the approximation of the variable importance null distribution might be too imprecise and might lead to inaccurate p values. In the most extreme setting (100 predictor variables in total and correlations between predictor variables), we observed on average about 70 non-positive importance scores (for the

prostate cancer data even only 40). However, our approach still worked surprisingly well. Nevertheless, in settings (1) with small predictor numbers (below 200), or (2) with very strong correlations between predictor variables, or (3) with high expected signal-to-noise ratio, we recommend that users look closely at the number of non-positive importance scores. If this number is small, we recommend that users be careful when using our testing approach because it is not clear if a small number of non-positive importance scores is sufficient to derive p values. In such cases one should consider the computationally more demanding alternatives, such as the approach by [Altmann et al. \(2010\)](#).

When deriving p values in a non-parametric way, that is without making any distributional assumptions, the testing approach of [Altmann et al. \(2010\)](#) showed almost the same statistical power as the new approach. In contrast, if a parametric distribution for the variable importance of unrelated variables was used, the permutation-based approach of Altmann and his colleagues was very conservative in our studies and had much smaller statistical power than our approach. This suggests that the poor performance is related to the assumed parametric distribution for the importance scores of unrelated variables. In our studies, we used the normal distribution for modelling the variable importance distribution of unrelated variables. Studies indicate that the assumption of a normal distribution is not reasonable due to the skewness of the distribution of null importance scores (data not shown). Researchers who apply the approach of [Altmann et al. \(2010\)](#) to high-dimensional data should therefore consider alternative distributions or approximate the null distribution in a non-parametric way.

Our new testing procedure focuses on the identification of predictor variables which significantly improve the random forest's predictive ability. The permutation variable importance measure, by its definition, reflects the improvement in predictive ability if a variable is used for making the prediction. Thus, there is a monotone relationship between the value of the variable importance and the p value derived from our testing approach: predictor variables with higher importance scores obtain smaller p values. This must not necessarily be the case with permutation-based approaches. This is obvious as [Altmann et al. \(2010\)](#) state that their approach corrects for the bias in the Gini importance measure which ranks, for example, variables with many categories higher than variables with fewer categories. In this case a re-sorting of variables occurs when computing p values from the Gini importance based on the proposed permutation procedure.

The detection of relevant variables through a testing procedure or variable selection strategy relies on the adequacy of the importance score of reflecting the relevance of a variable. Recent studies have shown that there is high instability of variable importance scores, especially in typical high-dimensional data settings with small numbers of observations and huge numbers of variables, and that this instability can greatly be reduced by increasing the number of trees ([Wang et al. 2016](#)). Thus, a sufficiently high number of trees is needed for the novel testing approach, and the default value for the number of trees in the existing random forests software (usually `ntree = 500`) is not appropriate for high-dimensional data and should by no means be used for high-dimensional settings. So far there are no general recommendations on the number of trees, but studies like that of [Wang et al. \(2016\)](#) might be a good starting point for the choice. Often the number of trees is chosen as a compromise between

computational burden and precision. Very fast implementations of random forests, which are available nowadays, such as that of [Wright and Ziegler \(2016\)](#), allow the user to construct reasonably large numbers of trees so that we recommend using several thousands of trees when using random forests or their importance measures in high-dimensional settings.

Overall, the statistical power of all testing procedures was low in our studies with huge predictor numbers. The power of the variable importance measure to discriminate between relevant and non-relevant variables was poor, too. The approach of [Altmann et al. \(2010\)](#) which showed high power in other studies ([Molinario et al. 2011](#); [Hapfelmeier and Ulm 2013](#)) also had very low statistical power in our studies. This discrepancy is possibly related to the fact that the existing studies included only a few variables while our studies are based on several thousands of variables. [Molinario et al. \(2011\)](#) for example focused on candidate-gene studies and considered only a few dozens of variables. When repeating our studies with a subset of 100 variables the statistical power substantially increased, and the variable importance discriminated much better between relevant and non-relevant variables. This suggests that the issue of detecting relevant features by variable importance measures is much more difficult for genome-wide association studies including hundreds of thousands to millions of features than for candidate-gene studies that include only a few hundreds of features.

References

- Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci* 96:6745–6750
- Altmann A, Toloşi L, Sander O, Lengauer T (2010) Permutation importance: a corrected feature importance measure. *Bioinformatics* 26:1340–1347
- Boulesteix A-L (2015) Ten simple rules for reducing overoptimistic reporting in methodological computational research. *PLoS Comput Biol* 4:e1004191
- Boulesteix AL, Bender A, Bermejo JL, Strobl C (2012) Random forest Gini importance favours SNPs with large minor allele frequency: assessment, sources and recommendations. *Brief Bioinform* 13:292–304
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32
- Breiman L, C. A (2008) Random forests. http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm
- Detting M, Bühlmann P (2003) Boosting for tumor classification with gene expression data. *Bioinformatics* 19:1061–1069
- Díaz-Uriarte R, De Andres SA (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinform* 7:3
- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA et al (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286:531–537
- Gregorutti B, Michel B, Saint-Pierre P (2013) Correlation and variable importance in random forests. arXiv preprint [arXiv:1310.5726](https://arxiv.org/abs/1310.5726)
- Hapfelmeier A, Ulm K (2013) A new variable selection approach using random forests. *Comput Stat Data Anal* 60:50–69
- Hothorn T, Hornik K, Zeileis A (2006) Unbiased recursive partitioning: a conditional inference framework. *J Comput Graph Stat* 15:651–674
- Huynh-Thu VA, Saeys Y, Wehenkel L, Geurts P (2012) Statistical interpretation of machine learning-based feature importance scores for biomarker discovery. *Bioinformatics* 28:1766–1774
- Ishwaran H (2007) Variable importance in binary regression trees and forests. *Electron J Stat* 1:519–537

- Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. *Ann Appl Stat* 2:841–860
- Janitza S, Strobl C, Boulesteix AL (2013) An AUC-based permutation variable importance measure for random forests. *BMC Bioinform* 14:119
- Janitza S, Tutz G, Boulesteix A-L (2016) Random forest for ordinal responses: prediction and variable selection. *Comput Stat Data Anal* 96:57–73
- Kim H, Loh W-Y (2001) Classification trees with unbiased multiway splits. *J Am Stat Assoc* 96:589–604
- Liaw A, Wiener M (2002) Classification and regression by randomForest. *R News* 2:18–22
- Louppe G, Wehenkel L, Suter A, Geurts P (2013) Understanding variable importances in forests of randomized trees. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) *Advances in neural information processing systems*, pp 431–439
- Molinaro AM, Carriero N, Bjornson R, Hartge P, Rothman N, Chatterjee N (2011) Power of data mining methods to detect genetic associations and interactions. *Hum Hered* 72:85–97
- Nicodemus K (2011) Letter to the editor: on the stability and ranking of predictors from random forest variable importance measures. *Brief Bioinform* 12:369–373
- Nicodemus K, Malley J (2009) Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics* 25:1884–1890
- Pepe M (2004) *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, USA
- Phipson B, Smyth G (2010) Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. *Stat Appl Genet Mol Biol* 9:1544–6115
- Polak P, Karličić R, Koren A, Thurman R, Sandstrom R, Lawrence MS, Reynolds A, Rynes E, Vlahoviček K, Stamatoyannopoulos JA et al (2015) Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* 518:360–364
- Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C et al (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* 415:436–442
- Prosperi MC, Marinho S, Simpson A, Custovic A, Buchan IE (2014) Predicting phenotypes of asthma and eczema with machine learning. *BMC Med Genomics* 7:S7
- Reif DM, Motsinger-Reif AA, McKinney BA, Rock MT, Crowe J, Moore JH (2009) Integrated analysis of genetic and proteomic data identifies biomarkers associated with adverse events following smallpox vaccination. *Genes Immun* 10:112–119
- Schwarz DF, König IR, Ziegler A (2010) On safari to random jungle: a fast implementation of random forests for high-dimensional data. *Bioinformatics* 26:1752–1758
- Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP et al (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* 1:203–209
- Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A (2008) Conditional variable importance for random forests. *BMC Bioinform* 9:307
- Strobl C, Boulesteix AL, Zeileis A, Hothorn T (2007) Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinform* 8:25
- Strobl C, Malley J, Tutz G (2009) An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychol Methods* 14:323–348
- Strobl C, Zeileis A (2008) Danger: high power!—exploring the statistical properties of a test for random forest variable importance. In: Brito P (ed) *Proceedings of the 18th international conference on computational statistics*. Porto, Portugal (CD-ROM), Springer, Heidelberg, pp 59–66
- Szymczak S, Holzinger E, Dasgupta A, Malley JD, Molloy AN, Mills JL, Brody LC, Stambolian D, Bailey-Wilson JE (2016) r2VIM: a new variable selection method for random forests in genome-wide association studies. *BioData Min* 9:7
- Tan AC, Gilbert D (2003) Ensemble machine learning on gene expression data for cancer classification. *Appl Bioinform* 2:S75–S83
- Tang R, Sinnwell JP, Li J, Rider DN, de Andrade M, Biernacka JM (2009) Identification of genes and haplotypes that predict rheumatoid arthritis using random forests. *BMC Proc* 3:S68
- van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT et al (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415:530–536

- Wang H, Yang F, Luo Z (2016) An experimental study of the intrinsic stability of random forest variable importance measures. *BMC Bioinform* 17:60
- Wang-Sattler R, Yu Z, Herder C, Messias AC, Floegel A, He Y, Heim K, Campillos M, Holzapfel C, Thorand B et al (2012) Novel biomarkers for pre-diabetes identified by metabolomics. *Mol Syst Biol* 8:615. doi:[10.1038/msb.2012.43](https://doi.org/10.1038/msb.2012.43)
- Wright MN, Ziegler A (2016) ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw* (in press)
- Yatsunenkov T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, Magris M, Hidalgo G, Baldassano RN, Anokhin AP et al (2012) Human gut microbiome viewed across age and geography. *Nature* 486:222–227
- Zhu R, Zeng D, Kosorok MR (2015) Reinforcement learning trees. *JASA* 110:1770–1784