Cantor *et al. BioData Mining* (2024) 17:34
https://doi.org/10.1186/s13040-024-00388-8

BioData Mining

RESEARCH **Open Access**

Knowledge-slanted random forest method for high-dimensional data and small sample size with a feature selection application for gene expression data

Erika Cantor[1*], Sandra Guauque-Olarte[2†], Roberto León[3], Steren Chabert[4,5,6] and Rodrigo Salas[4,5,6†]

†Sandra Guauque-Olarte and Rodrigo Salas contributed equally to this work.

*Correspondence: erika.cantor@javeriana.edu.co

[1] Department of clinical epidemiology and biostatistics, Pontificia Universidad Javeriana, Bogotá 110221, Colombia
[2] Department of basic sciences and oral medicine, Universidad Nacional de Colombia, Bogotá 16486, Colombia
[3] Department of Computer Science, Universidad Técnica Federico Santa María, Santiago de Chile 8940897, Chile
[4] School of Biomedical Engineering, Universidad de Valparaiso, Valparaíso 2360102, Chile
[5] Millennium Science Initiative Intelligent Healthcare Engineering, Santiago de Chile 7820436, Chile
[6] Center of Interdisciplinary Biomedical and Engineering Research for Health - MEDING, Universidad de Valparaiso, Valparaíso 2360102, Chile

Abstract

The use of prior knowledge in the machine learning framework has been considered a potential tool to handle the curse of dimensionality in genetic and genomics data. Although random forest (RF) represents a flexible non-parametric approach with several advantages, it can provide poor accuracy in high-dimensional settings, mainly in scenarios with small sample sizes. We propose a knowledge-slanted RF that integrates biological networks as prior knowledge into the model to improve its performance and explainability, exemplifying its use for selecting and identifying relevant genes. knowledge-slanted RF is a combination of two stages. First, prior knowledge represented by graphs is translated by running a random walk with restart algorithm to determine the relevance of each gene based on its connection and localization on a protein-protein interaction network. Then, each relevance is used to modify the selection probability to draw a gene as a candidate split-feature in the conventional RF. Experiments in simulated datasets with very small sample sizes ($n \leq 30$) comparing knowledge-slanted RF against conventional RF and logistic lasso regression, suggest an improved precision in outcome prediction compared to the other methods. The knowledge-slanted RF was completed with the introduction of a modified version of the Boruta feature selection algorithm. Finally, knowledge-slanted RF identified more relevant biological genes, offering a higher level of explainability for users than conventional RF. These findings were corroborated in one real case to identify relevant genes to calcific aortic valve stenosis.

Keywords: Prior knowledge, Random forest, Gene selection, High-dimensional, Feature selection, Explainability, Protein-protein interaction, RNA-Seq

Introduction

The main goal of gene expression analysis is to identify genes that could determine the risk and prognosis of a particular disease among thousands of other genes when only a subset of them is biologically important [1]. This can be performed only based on the information collected (data-driven) or involving some prior knowledge, such as encapsulated knowledge through protein-protein interaction networks (PPI) [2]. Ensemble methods, such as random forest (RF), have gained popularity because they allow identifying relevant genes (or features) using variable importance measures [3]. Despite the RF model represents a flexible non-parametric approach with several properties, in high-dimensional settings *large P-variables small N-sample size*, RF may provide poor accuracy, especially if complex variable interactions (e.g., gene-gene) exist in small sample sizes because similar information gain coefficients are likely to be obtained [4–6]. In addition, the curse of dimensionality (COD) is an inherent problem in analyzing high-dimensional spaces, leading to two main effects called data sparsity and distance concentration. Both effects make it more demanding to find similarities and patterns between samples, affecting the performance of the models, especially in classification tasks [7]. To handle the COD effects, the use of prior knowledge and dimensionality reduction techniques with feature selection have been explored.

The current literature has described that the use of biological information could improve the feature selection during the analysis of genomics data and some attempts to involve prior biological knowledge have been made using RF algorithm [2, 8, 9]. Guan et al. [10] proposed a knowledge-based guided regularized RF (Know-GRRF) that performs a regularized RF using a penalty coefficient for each feature from different domains, deriving a composite score between 0 and 1 (higher biological relevance). Know-GRRF allows the identification of a subset of relevant and irrelevant features after multiple runs, achieving a better performance compared to regularized RF and Lasso logistic regression [10]. However, the composite score used in the application made by Guan et al. [10] was not computed using the information accumulated in biological networks. Furthermore, Guan et al. [10] only performed experiments considering large sample sizes (100, or 200) with a small number of input features (maximum 1000), so the actual performance of Know-GRRF on extreme high-dimensional data is unknown. Another approach, called NetBiTE (Network-based Biased Tree Ensembles) for drug sensitivity biomarker identification, uses prior knowledge through a probabilistic bias weight distribution constructed with the information from a biological network using random walk with restart (RWR). NetBiTE modifies the selection probability of each feature to split a node in RF regression, without implementing a mechanism to identify relevant features or genes reporting better accuracy compared to RF, XGBoost, and linear regression [11].

In this study, we developed a comprehensive knowledge-slanted tree ensemble that allows the integration of biological networks as prior knowledge into the model to improve its performance and explainability. The RF algorithm is the focus of analysis in scenarios with very small sample sizes ($n \leq 30$) for gene selection or classification problems. A RWR algorithm is used to determine the relevance of each feature (gene) in the RF model by modifying the selection probability to be in the subset of features to split a node. The last approach is called knowledge-slanted RF, which represents a way to prioritize features and mitigate the COD effects. In addition, extensive simulation studies are performed to identify the conditions to obtain its best performance. Similar to this research, Ghosh & Cabrera [12] recently developed a data-driven approach called Enriched Random Forest in which the algorithm of conventional RF is modified by applying weighted random sampling for each feature to choose the eligible subset for splitting at each node. However, the weights are computed based on the ability of each feature to discriminate between groups in training data through filter methods based on hypothesis testing, which represents a data-driven method, whereas the proposed RF is slanted by prior knowledge.

In biological applications, it is more important to understand the mechanisms related to the event of interest than to build the best predictive model based on a black box method [13]. In the RF algorithm, there are two main variable importance measures (VIM) developed by Breiman L [14], the mean decrease in impurity importance (MDI) and the mean decrease in accuracy (MDA). While MDI is based on the decrease of impurities achieved at each node using a specific variable, MDA measures the importance based on predictive accuracy. However, empirical analysis have demonstrated that these measures are influenced by masking effects, the correlation between the input features, and the number of categories [15–18]. To overcome this drawback, Nembrini et al. [19] proposed the *actual impurity reduction (AIR)* in which a modification of the conventional Gini MDI is made seeking to quantify the bias generated by certain sources. AIR measure decomposes the Gini MDI into two parts: the true importance of each feature and the impurity reduction achieved by chance as a consequence of recursive partitioning during RF construction [19, 20].

In order to face the problem of feature selection in knowledge-slanted RF, a new evaluation approach is proposed to identify relevant features in the knowledge-slanted RF in order to involve over- or under-representation of features. Basically, this new assessment is a mixture of a wrapper algorithm and a measure of importance with the Gini MDA using the conventional Boruta selection feature algorithm with the use of the AIR measure[21]. In general, Boruta creates shadows as a permuted version of the original features and then compares the VIM values between them to define whether a specific contribution is significant to the model. In genomic applications, other feature selection methods, such as ReliefF or minimal-redundancy-maximum-relevancy, have been widely used. However, Boruta algorithm has proven to be more reliable in identifying relevant genes mainly in high-dimensional dataset, which has attracted attention among available feature selection algorithms [22–24]. The knowledge-slanted RF framework is implemented in the *kslboruta* package of the R software to facilitate its application to real-world problems.

Methods

The proposed approach, called knowledge-slanted RF, is a combination of two stages as an attempt to implement a knowledge-guided supervised learning approach. In the first stage, prior knowledge represented by graphs (e.g., PPI networks) is translated by running an RWR algorithm in order to identify which variables could be relevant to the model. In the second stage, this information is involved in the conventional RF, prioritizing the selection of some features or genes during the RF construction.

Knowledge-slanted random forest

Suppose there are $i\text{–}th$ samples with $i = 1, ..., m$, for which the feature vectors $\{x_{ij}\}_{j=1}^{p}$ and the label y_i are measured. The goal is to find a classification model that allows to predict the label and identify which features are most relevant for the classification task.

Formally, knowledge-slanted RF predicts y_i by combining the results of tree-structure classifiers $f_t(x)_{t=1}^{T}$. In conventional RF, for each node within each tree, a subset of features denoted *mtry*, is randomly selected with equal probability ($1/p$), while in knowledge-slanted RF the *mtry* subset of variables is selected with probability $\{\mathbf{p_j}\}_{j=1}^{p} \in (0,1)$. More specifically, $\mathbf{p_j}$ is determined based on the prior knowledge that is represented through a PPI network. This is relevant because the genes associated with a specific disease share similar functions and tend to be located in neighboring regions on the PPI network, which helps to identify new disease-related genes and perform candidate-gene prioritization. The structure of a PPI network is equivalent to an undirected weighted graph $G = (V, E)$, where nodes $i, j \in V$ correspond to each gene, and edges or connections $(i, j) \in E$ are weighted with a weight matrix W that represents the strength of the relationship between each pair of edges or connections. Consequently, W is known as the weighted precision matrix of G, where all entries in W are in (0,1).

For gene prioritization, an RWR algorithm is applied to rank the genes on the PPI network, allowing that the random walker can move from i node to a randomly neighbor node or goes back to the initial node with a back-probability $\theta \in (0,1)$. RWR simulates a random walker that explores the PPI network from node i to node j using a transition probability matrix $A = WD^{-1}$, where D is a diagonal matrix with elements $d_{ij} = \sum_j w_{ij}$. RWR Algorithm can be represented by equation 1, where $\mathbf{p}$ is the converged probability of each node or gene j being a candidate-gene. At each step s, the RWR algorithm updates the probability $\mathbf{p}^{(s)}$ that the walker is at a specific node (or gene) at step s, until convergence is reached for a given threshold $\|\mathbf{p}^{(s+1)} - \mathbf{p}^{(s))}\|_{L_1} \leq \delta. \theta \in (0,1)$ represents the probability of returning to a specific set of nodes in each interaction of the algorithm, called seed nodes, and the vector $\mathbf{p}^{(0)}$ is the initial distribution probability.

$$\mathbf{p}^{(s+1)} = (1-\theta)A^T\mathbf{p}^{(s)} + \theta\mathbf{p}^{(0)}, \tag{1}$$

In knowledge-slanted RF, the selection probability is modified using the probabilities obtained after executing the RWR algorithm with $\mathbf{p}^{(s+1)}$. Therefore, the most informative genes can be selected in the first steps of the algorithm. A relevant aspect of the RWR algorithm is the setting of seed notes at $s = 0$ to allow the random walker to visit all nodes, using a higher probability of being at a specific node at time $s = 0$ for relevant features. Consequently, the seed nodes must be selected using features (or genes) that have been established with a statistically significant difference between the labels of the classification task in the literature. The algorithm of knowledge-slanted RF is shown in Algorithm 1. The proposed ensemble is constructed using the Classification and Regression Tree (CART) algorithm with the Gini index as a splitting rule and the bootstrap aggregation procedure.

Algorithm 1 Knowledge-slanted Random Forest algorithm

```
Input: A training set $\mathcal{D} = (x_1, y_1), ..., (x_m, y_m), mtry \in \{1, ..., p\}$, number of trees $T > 0$, and prior knowledge represented by weight matrix $W$.
Output: Prediction of Knowledge-slanted RF model.
procedure RWR
    $A \leftarrow$ Column normalized Adjacency Matrix of prior knowledge (e.g., PPI network).
    $\mathbf{p}^{(s+1)} \leftarrow$ Selection probability.
    while $\|\mathbf{p}^{(s+1)} - \mathbf{p}^{(s))}\|_{L_1} \geq \delta$ do
        $\mathbf{p}^{(s+1)} = (1-\theta)A^T\mathbf{p}^{(s)} + \theta\mathbf{p}^{(0)}$
    end while
    return $\mathbf{p}^{(s+1)}$
end procedure
procedure KNOWLEDGE-SLANTED RF
    for t=1,....T do
        Draw a sample with replacement of size N $\mathcal{D}^t$.
        Select a weighted random subset with $mtry$ features $\{x_j\} \subset 1, ..., p$ using $\mathbf{p}^{(s+1)}$.
        Construct $f_t(x)$ using Gini index for measuring node impurity.
    end for
    Compute the predicted value from the ensemble of trees $\hat{F}_{RF}(x) = majority\ vote\{f_t(x)\}_1^T$.
    return $\hat{F}_{RF}(x)$
end procedure
```

Boruta approach for the knowledge-slanted

Boruta feature selection algorithm is a flexible method for finding all relevant variables [21]. To summarize, Boruta replicates the original set of features X, generating a permuted version denoted by X^*, known as the shadow subset. Then, Boruta built a model using $X \cup X^*$ and compares the relevance of the original version of x_j with the maximum score among shadow features (MSF) X^*, using any VIM measure. If $VIM(x_j) > MSF$, then, the feature x_j stores a hit. After performing a minimum of iterations *nruns*, the total of hits stored for each feature (H_j) is tested using a two-sided test of equality based on the binomial distribution in order to compare the observed value with respect to the expected number of hits in *nruns* iterations, thus is $H_j \sim Binomial(p = 0.50, n = nruns)$.

Using the central idea of the Boruta algorithm and AIR measure, an extended approach was developed to ensure that each variable's relevance will be evaluated, taking into account the under- or over-representation during the construction of the knowledge-slanted RF based on $\mathbf{p_j} \in (0, 1)$. The procedure is summarized in the following steps:

1. **Step 1:** Run a Knowledge-slanted RF using $X \cup X^*$ preserving that each pair x_j and x_j^* have the same level of over and under-representation according to the original prior weight ($\mathbf{p_j}$) of x_j. The new prior weights are determined by:

   $$\mathbf{p_j^*} = R_j/(2p),$$

   where, $R_j = \mathbf{p_j}/(1/p)$.
2. **Step 2:** Calculate the maximum of the AIR measure [19] among the shadow features (MSF).

3. **Step 3:** Assign a hit to every feature $H_j = 1$ if the respective AIR_j from each x_j is greater than MSF.
4. **Step 4:** Perform a two-sided test to compare the accumulated hits for each variable in each *nrun*. Rejected or accepted based on the results from a binomial test.
5. **Step 5:** Remove the original x_j and its respective shadow x_j^* when the status is rejected. Repeat the process until all features get status or the maximum number of iterations is reached.

A multiple testing adjustment can be done using Bonferroni, Benjamini, and Hochberg (BH), or Benjamini and Yekutieli (BY) methods [25, 26]. The Boruta approach for the knowledge-slanted RF is implemented in the *kslboruta* package of the R software. The AIR measure is calculated with the Gini MDI using the following equation:

$$AIR(x_j) = MDI(x_j) - MDI(x_j^*),$$

where, $MDI(x_j^*)$ is the estimator of the impurity reduction of x_j achieved by chance and therefore, $AIR(x_j)$ quantifies the true importance of x_j.

Experimental evaluation

To simulate the prior information represented by graphs,more specifically the PPI network, a Gaussian Graphical Model was assumed $G \sim N(0, \Omega = \Sigma^{-1})$ with the precision matrix Ω, which means that the conditional independence relationship between the nodes follows a multivariate Gaussian distribution. Because $\Omega = BB^T + L$, the precision matrix was generated by a random sparse lower triangular matrix (B) and a random diagonal matrix (L) using the algorithm proposed by [27]. To control the sparsity grade of Ω, B was divided into three submatrix of equal size (I_1, I_2, I_3). In each submatrix I_k, a random number of nonzero connections with probability η was established, and then, these connections were sampled from $U \sim (0, 1)$ due to STRING scores ranging from 0 to 1 [28]. Similarly when $(i, j) \notin I_k$, a random number of connections with nonzero values were defined with probability $extra_\eta = \eta/5$. The diagonal values of L were sampled from $U \sim (10^{-3}, 5 \times 10^{-3})$. Finally, a matrix Ω was obtained, and the adjacency matrix of G was retrieved by setting all diagonal values of Ω to 0. Thus, the weighted adjacency matrix W was generated, which represented the prior knowledge stored in a biological PPI network. When an element of W was less than 0.20 ($w_{ij} < 0.20$), its value was set to zero. According to [29], the number of interactions in a PPI network is approximately 10^{-3}, so we consider four cases: when the proportion of non-zero connections in W is $10^{-5}, 10^{-4}$, 0.02 and 0.1.

Synthetic data was simulated assuming $X \sim N(0, \Sigma)$ and $\Sigma = f(G)$ to describe the existing relationship between genes or features based on information from the simulated PPI network. We generate 100 datasets with $p = 1000$, each with a training sample of size $n = 15$, $n = 30$, and $n = 100$. An additional one-third of the sample size was generated to use as a test sample. To create a binary result y with two classes, each data set was divided into two subsets representing a group of cases and a group of controls. For the case group, 100 genes or features were randomly selected from the set of all features (1, 2, ..., p) and then, their means were modified and sampled from $(-2, -1, -0.5, 0.5, 1, 2)$ to define small, medium, and large effects. Therefore, only 100 features were considered as relevant for the classification task. Using the above strategy, easy and medium classification scenarios were generated (Fig. 1). In the medium scenario an additive noise $\mathcal{E} \sim N(0, I_p)$ for each feature was considered. A third case study simulating a hard classification scenario was achieved, generating y through a probit model. Thus, the binary response y for each *i–th* sample was generated by:

$$P(y=1|X) = \Phi(\beta_0 + X\beta),$$

on each generated dataset, we apply three models: knowledge-slanted RF, Conventional RF, and logistic Lasso regression. For RF models, *mtry* parameter was equal to $\sqrt{p}$ and *ntree* $= 500$. All RF models were grown to maximum depth with a minimum terminal node size of 1, using the Gini index to measure impurity at each node. The *ranger* package was used for all simulations [30]. Lasso was implemented using *glmnet* R package, which fits the model via penalized maximum likelihood [31]. The penalty parameter $\lambda \in [0, 1]$ was set with the cross-validation method by *cv.glmnet* function. The performance of the models was measured in the test sample, through prediction error (PE), F1 score, sensitivity (S), and specificity (E).

In the knowledge-slanted RF, the seed nodes in the RWR algorithm were the relevant features used in the generation of y in the three scenarios. We investigated the influence of prior knowledge on the performance of knowledge-slanted RF selecting a proportion of the (q) relevant features and $(1-q)$ non-relevant features as seed nodes to obtain a new probability of selection for each feature (or gene). To study the behavior of AIR measure and modified Boruta under the knowledge-slanted RF, we performed simulations using a sample size of $n = 30$, $p = 5000$, and 100 relevant features. Since the knowledge-slanted RF combines RWR algorithm with RF model and the resulting probabilities of selection or prior weights $\{\mathbf{p}\}_{j=1}^{p} \in (0, 1)$ of each feature may depend on the selected seed nodes to initialize RWR algorithm, it investigates the effects of choosing the seed nodes using 50 relevant features and 50 false relevant in the behavior of VIM. Knowledge-slanted RF and conventional RF were constructed by combining 5000 trees and using a size of $mtry = \sqrt{p}$ for a medium classification task, with a proportion of non-zero connections between the features X of $\eta = 10^{-4}$.

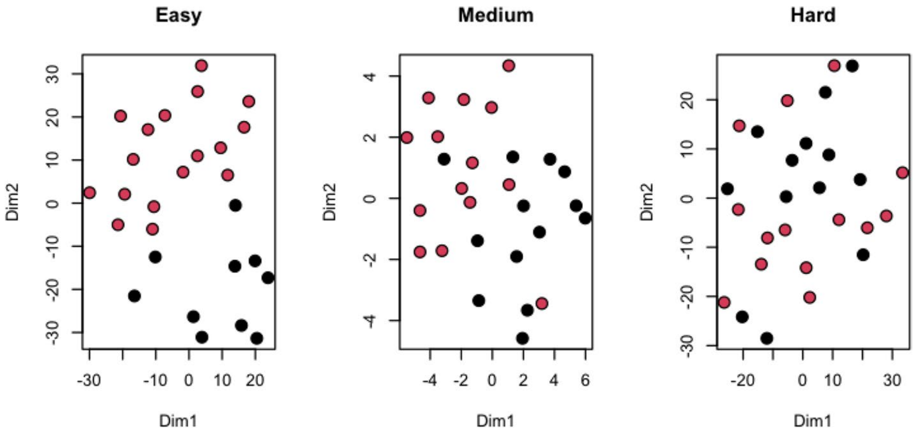

Fig. 1 T-distributed Stochastic Neighbor Embedding representation of the simulated three scenarios with $n = 30$, $p = 1000$ y $\eta = 10^{-4}$. Red points correspond to cases

Results

Performance comparison on simulated data

As can be seen in Figs. 2 and 3, the results on the simulated datasets showed that knowledge-slanted RF reported good metrics on easy and medium classification tasks, outperforming conventional RF with the highest accuracy. While in hard classification tasks, its behavior was limited and similar to that reported by conventional RF. As the proportion of non-zero connections (η) between nodes increased, the performance of knowledge-slanted RF was better than the other approaches. The performance of the RF algorithm with and without involving prior knowledge was better than Lasso regression.

The detailed results on simulated datasets are shown in Tables 1, 2, and 3 for a sample size of 15, 30 y 100, respectively. Knowledge-slanted RF worked best in small sample sizes ($n \leq 30$), reporting good results in terms of PE, S, E, and F1 score. Although, the mentioned tables displayed the findings only for $\eta \leq 10^{-4}$, the observed performance did not vary for η to $10^{-5}, 0.02$ and 0.1. The effect of prior knowledge was irrelevant when the sample size reached 100 observations during the training process.

As mentioned in the Methods section, the strategy evaluated for identifying relevant features in the knowledge-slanted RF is the extended Boruta approach with and without correction for multiple testing. Findings after applying this strategy are reported in Table 4. The AIR measure under knowledge-slanted framework reported a high percentage (40.6% ± 6.4%) of relevant features among the ones erroneously flagged as seed nodes, but with no real effect on the result ($\beta = 0$), so they are false discoveries. With the extended Boruta approach, the percentage of false discoveries decreased compared to the unmodified AIR measure for knowledge-slanted RF, ranging from 24.2% (Bonferroni adjustment) to 32.9% (simple Boruta version). The percentage of relevant features was always higher in features labeled as seed nodes than in features not labeled as seed nodes.

There was a tendency to detect a higher percentage of relevant features as the effect size increased for those with a true effect on y. It also observed large variability in the results for features with small ($\beta = 0.5$) and moderate ($\beta = 1$) effect sizes. Among variables with a small effect, the performance between the AIR measure and the extended Boruta was similar ($\beta = 0.5$). When the effect size increases to moderate ($\beta = 1$) and strong ($\beta = 2$), the extended Boruta showed a tendency to identify a higher percentage of relevant features compared to the AIR measure, even if the features were set as non-seed nodes at the beginning of the algorithm (Table 4).

Performance on real data: calcific aortic valve stenosis

Calcific aortic valve stenosis (CAVS) is a fatal disease and there is no pharmacological treatment to prevent the progression of CAVS. The objective of this subsection is to identify genes potentially implicated with CAVS in patients with congenital bicuspid aortic valve (BAV) and tricuspid aortic valve (TAV) in comparison with patients having normal valves, using a knowledge-slanted RF. CAVS dataset was obtained from the primary study performed by [32] and approved by the ethics committee of the Institut universitaire de cardiologie et de pneumologie de Québec, Laval University, Quebec, Canada. Written informed consent was obtained from all participants. The

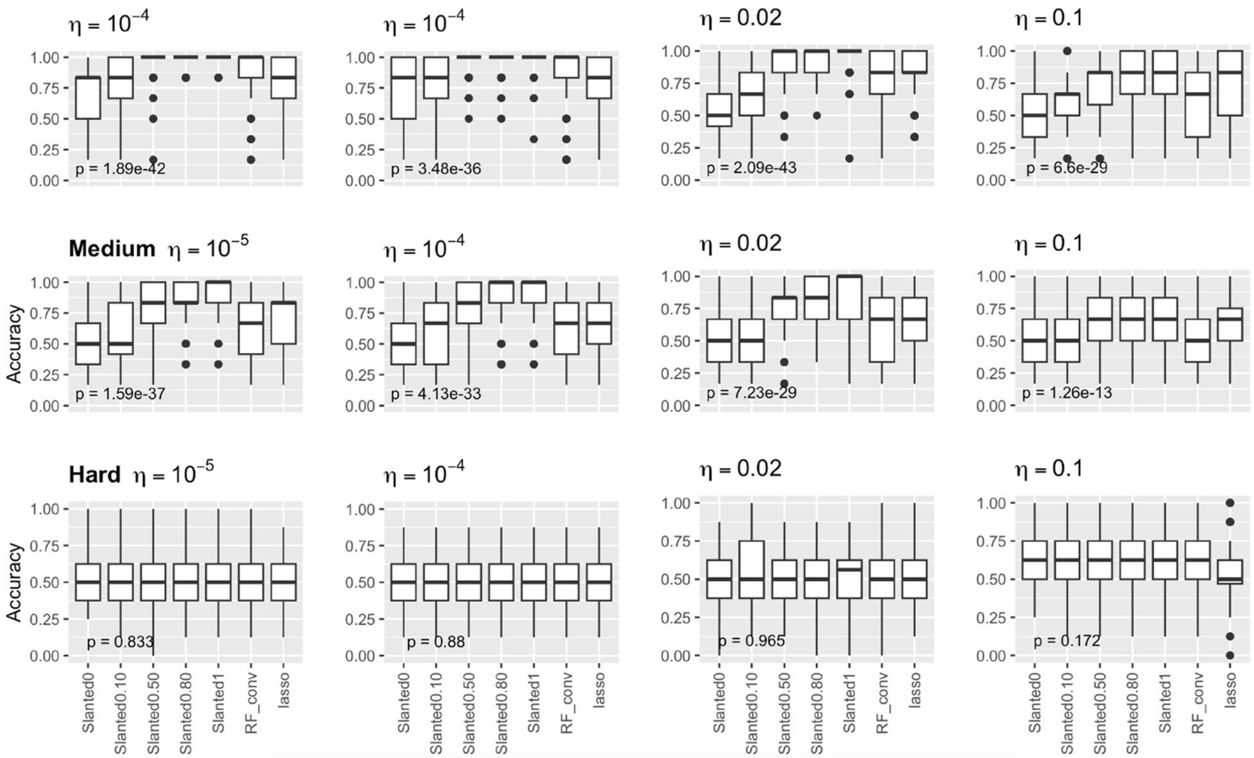

Fig. 2 Comparison of accuracy in binary classification tasks with a sample size of 15 and 1000 features ($p = 1000$) in easy, medium, and hard scenarios. Different proportions (0, 0.10, 0.50, 0.80) of seed nodes (relevant features) were evaluated in the knowledge-slanted RF

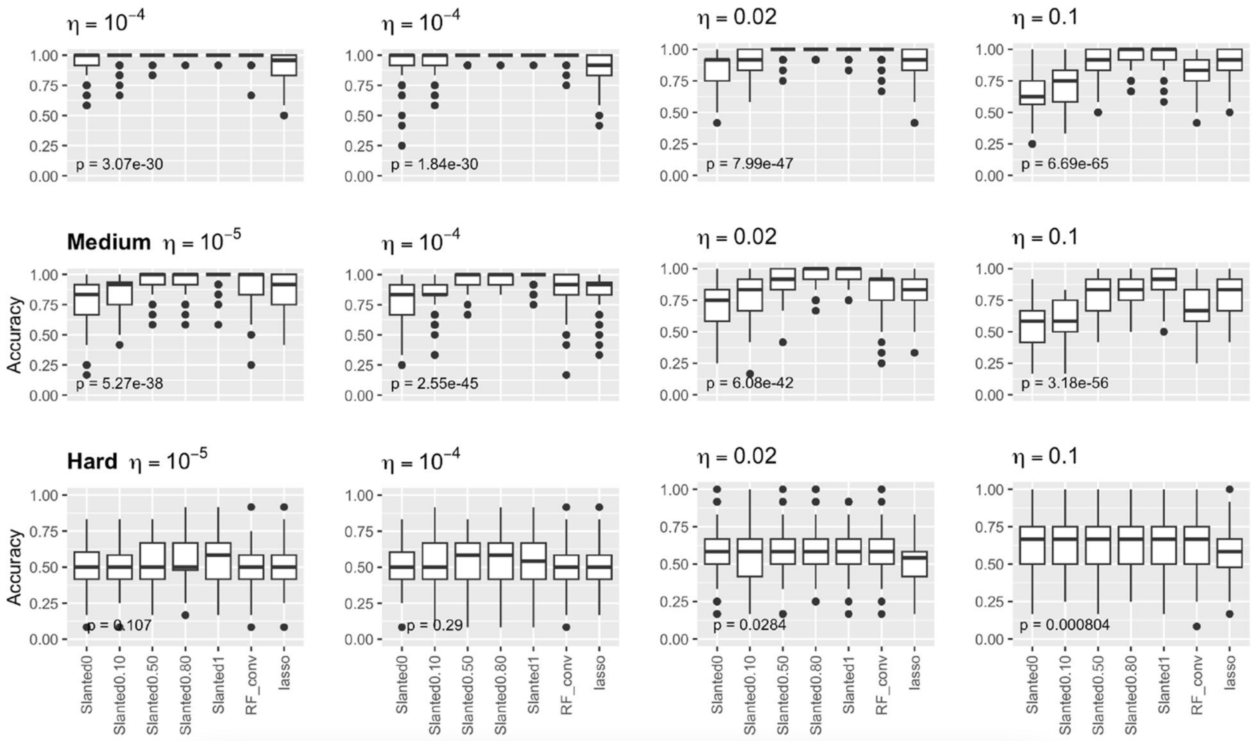

Fig. 3 Comparison of accuracy in binary classification tasks with a sample size of 30 and 1000 features ($p = 1000$) in easy, medium, and hard scenarios. Different proportions (0, 0.10, 0.50, 0.80) of seed nodes (relevant features) were evaluated in the knowledge-slanted RF

Table 1 Comparison of models with $p = 1000$ in easy, medium and hard classification scenarios using $\eta = 10^{-4}$ and $n = 15$

	PE		S		E		F1	
Method	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Easy scenario								
Lasso	0.16	0.2	0.71	0.35	0.98	0.11	0.76	0.33
Conv. RF	0.11	0.2	0.94	0.19	0.92	0.21	0.88	0.21
Slanted (0)	0.28	0.24	0.8	0.31	0.19	0.3	0.71	0.28
Slanted (0.1)	0.19	0.22	0.85	0.25	0.87	0.24	0.8	0.24
Slanted (0.5)	0.03	0.08	0.98	0.07	0.98	0.09	0.97	0.1
Slanted (0.8)	0.01	0.03	1	0.03	0.99	0.04	0.99	0.04
Slanted (1)	0.01	0.03	1	0	0.99	0.04	0.99	0.04
Medium scenario								
Lasso	0.3	0.25	0.61	0.35	0.88	0.21	0.63	0.32
Conv. RF	0.38	0.27	0.74	0.37	0.7	0.37	0.6	0.3
Slanted (0)	0.48	0.23	0.65	0.38	0.61	0.4	0.5	0.27
Slanted (0.1)	0.42	0.23	0.7	0.35	0.66	0.37	0.57	0.27
Slanted (0.5)	0.23	0.23	0.84	0.28	0.84	0.26	0.77	0.24
Slanted (0.8)	0.15	0.17	0.87	0.22	0.92	0.18	0.85	0.18
Slanted (1)	0.11	0.17	0.92	0.18	0.93	0.18	0.89	0.17
Hard scenario								
Lasso	0.49	0.18	0.38	0.38	0.68	0.39	0.34	0.31
Conv. RF	0.47	0.16	0.63	0.43	0.4	0.43	0.47	0.31
Slanted (0)	0.48	0.17	0.62	0.42	0.4	0.4	0.47	0.31
Slanted (0.1)	0.48	0.17	0.61	0.41	0.41	0.42	0.47	0.31
Slanted (0.5)	0.47	0.18	0.64	0.41	0.42	0.42	0.5	0.29
Slanted (0.8)	0.47	0.17	0.6	0.42	0.44	0.42	0.46	0.32
Slanted (1)	0.45	0.17	0.61	0.42	0.47	0.42	0.48	0.32

main characteristic of the data structure is described in [32] and [33]. Knowledge-slanted RF was applied with the extended Boruta approach using as input features the expression data of 15,191 genes from 8 controls, 10 BAV, and 9 TAV cases. To increase the probability that a relevant feature will be selected to split a node, the model was built with 5000 trees using a $mtry = \sqrt{15,191}$. A conventional RF with AIR measure was also fitted for comparative purposes and a Leave-one-out cross-validation was implemented due to the limited sample size.

Results in the CAVS dataset can be found in Fig. 4 and Table 5. It is clear that the conventional RF with AIR method identified a greater number of relevant genes than the extended Boruta approach in the knowledge-slanted RF with and without correction for multiple testing; all five approaches simultaneously identified 330 genes. However, genes obtained from knowledge-slanted RF ranked better in RWR based on PPI information with a median position ranging from 548 to 569 compared to 3820 from conventional RF, so users could more easily interpret the results from knowledge-slanted RF because the prediction can be attributed mainly to associated genes that could participate in important molecular mechanisms according to the information from PPI network. Additionally, when only relevant genes were used as

Table 2 Comparison of models with $p = 1000$ in easy, medium and hard classification scenarios using $\eta = 10^{-4}$ and $n = 30$

	PE		S		E		F1	
Method	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Easy scenario								
Lasso	0.10	0.13	0.82	0.22	1.00	0.00	0.88	0.17
Conv. RF	0.01	0.03	0.99	0.04	1.00	0.02	0.99	0.03
Slanted (0)	0.08	0.13	0.91	0.17	0.97	0.10	0.92	0.15
Slanted (0.1)	0.04	0.09	0.95	0.11	0.98	0.07	0.96	0.09
Slanted (0.5)	0.01	0.03	0.99	0.04	1.00	0.02	0.99	0.03
Slanted (0.8)	0.00	0.02	0.99	0.02	1.00	0.00	1.00	0.01
Slanted (1)	0.00	0.01	1.00	0.01	1.00	0.01	1.00	0.02
Medium scenario								
Lasso	0.13	0.13	0.78	0.21	0.98	0.06	0.84	0.16
Conv. RF	0.10	0.15	0.87	0.19	0.96	0.13	0.89	0.15
Slanted (0)	0.23	0.19	0.72	0.26	0.90	0.20	0.74	0.22
Slanted (0.1)	0.16	0.14	0.81	0.20	0.92	0.14	0.83	0.16
Slanted (0.5)	0.05	0.07	0.94	0.11	0.98	0.05	0.95	0.07
Slanted (0.8)	0.03	0.04	0.97	0.07	0.99	0.04	0.97	0.04
Slanted (1)	0.03	0.05	0.97	0.08	0.99	0.03	0.97	0.06
Hard scenario								
Lasso	0.49	0.15	0.28	0.33	0.78	0.30	0.27	0.29
Conv. RF	0.49	0.16	0.56	0.41	0.46	0.42	0.45	0.30
Slanted (0)	0.48	0.15	0.54	0.38	0.48	0.38	0.45	0.28
Slanted (0.1)	0.48	0.16	0.56	0.38	0.47	0.40	0.47	0.28
Slanted (0.5)	0.46	0.15	0.57	0.39	0.50	0.37	0.48	0.29
Slanted (0.8)	0.46	0.16	0.58	0.38	0.51	0.38	0.49	0.28
Slanted (1)	0.45	0.16	0.57	0.38	0.54	0.38	0.49	0.29

input features, knowledge-slanted RF outperformed conventional RF to discriminate between BAV, TAV, and control groups (Table 5).

For knowledge-slanted RF using different adjusting methods in the extended Boruta approach, the performance of the models was similar, but the BY adjustment selected more interconnected genes in the PPI network, offering the same results and higher biological explainability. Further evidence of the higher level of explainability in the knowledge-slanted RF compared to the conventional RF was the lower depth found in the trees constructed with the relevant genes identified by our approach using any adjustment for multiple comparisons (Table 5).

Discussion and conclusion

This research contains the development and application of an analysis framework for incorporating prior knowledge into the RF algorithm, called Knowledge-slanted RF, with the main objective of achieving better accuracy and explainability for classification tasks in high dimensional data with a small number of samples ($n \leq 30$). In addition, it addresses the problem of identifying and selecting features (or genes) that are relevant to an event of interest or outcome. Although the entire development was intended for genetic and genomic field due to the high probability of producing

Table 3 Comparison of models with $p = 1000$ in easy, medium and hard classification scenarios using $\eta = 10^{-4}$ and $n = 100$

	PE		S		E		F1	
Method	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Easy scenario								
Lasso	0.05	0.07	0.90	0.14	1.00	0.00	0.94	0.09
Conv. RF	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
Slanted (0)	0.00	0.01	1.00	0.01	1.00	0.01	1.00	0.01
Slanted (0.1)	0.00	0.01	1.00	0.01	0.99	0.02	1.00	0.01
Slanted (0.5)	0.00	0.00	1.00	0.01	1.00	0.00	1.00	0.00
Slanted (0.8)	0.00	0.00	1.00	0.00	1.00	0.00	1.00	0.00
Slanted (1)	0.00	0.00	1.00	0.01	1.00	0.00	1.00	0.00
Medium scenario								
Lasso	0.04	0.05	0.93	0.10	1.00	0.01	0.96	0.06
Conv. RF	0.00	0.01	1.00	0.01	1.00	0.01	1.00	0.01
Slanted (0)	0.02	0.03	0.98	0.04	0.98	0.04	0.98	0.03
Slanted (0.1)	0.02	0.03	0.98	0.04	0.98	0.04	0.98	0.03
Slanted (0.5)	0.01	0.01	0.99	0.02	0.99	0.02	0.99	0.02
Slanted (0.8)	0.00	0.01	1.00	0.01	1.00	0.01	1.00	0.01
Slanted (1)	0.00	0.01	1.00	0.01	1.00	0.01	1.00	0.01
Hard scenario								
Lasso	0.47	0.09	0.31	0.23	0.77	0.21	0.36	0.22
Conv. RF	0.44	0.10	0.63	0.31	0.46	0.33	0.55	0.21
Slanted (0)	0.46	0.09	0.61	0.30	0.46	0.31	0.54	0.19
Slanted (0.1)	0.44	0.09	0.62	0.30	0.47	0.31	0.55	0.19
Slanted (0.5)	0.42	0.09	0.65	0.27	0.49	0.29	0.59	0.16
Slanted (0.8)	0.41	0.09	0.66	0.27	0.50	0.29	0.59	0.17
Slanted (1)	0.39	0.09	0.68	0.25	0.52	0.28	0.62	0.15

high-dimensional structures, our approach can be applied in several fields if prior knowledge is available to guide the RF algorithm.

Knowledge-slanted RF is a combined approach that leverages the advantages of RWR algorithm and conventional RF. In biological contexts, the RWR algorithm is often employed to search for unknown genes based on existing connections to known genes that are referred to as seed nodes [34, 35]. From the RWR results it is possible to infer novel genes if the probability of visiting that gene (node) by the random walker is high. In this work, the resulting RWR probabilities guided the construction of the RF model. Simulation results infer that good performance of the knowledge-slanted RF remains when at least 50% of the features (genes) indicated as seed nodes are behind the true mechanism of y.

Similar to [11], we showed that the RF algorithm, when prior knowledge is incorporated to modify the feature selection probability during the construction of tree ensembles, outperforms the conventional RF, which uses an equal selection probability for each feature. The integration of gene interaction data could offer a better prediction performance, specifically when class overlap exists (e.g., TAV vs. BAV). In scenarios with easily separable classes (e.g., TAV/BAV vs Control) or larger sample sizes (e.g., sample size equal to 100 observations), we believe that the use of prior

Table 4 Proportion of relevant features identified using the extended Boruta approach for the knowledge-slanted RF

		Non-seed node		**Seed node**	
Effect	**Method**	**Mean**	**SD**	**Mean**	**SD**
$\beta = 0$	AIR	0.05	0.00	0.41	0.06
	Boruta	0.04	0.00	0.02	0.02
	Boruta Bonferroni	0.02	0.01	0.24	0.06
	Boruta BH	0.04	0.01	0.31	0.07
	Boruta BY	0.03	0.01	0.28	0.06
$\beta = 0.5$	AIR	0.08	0.06	0.49	0.13
	Boruta	0.09	0.07	0.47	0.12
	Boruta Bonferroni	0.05	0.05	0.37	0.13
	Boruta BH	0.07	0.06	0.46	0.13
	Boruta BY	0.07	0.07	0.41	0.13
$\beta = 1$	AIR	0.23	0.10	0.73	0.12
	Boruta	0.25	0.12	0.76	0.11
	Boruta Bonferroni	0.17	0.09	0.69	0.11
	Boruta BH	0.24	0.12	0.75	0.11
	Boruta BY	0.21	0.10	0.73	0.12
$\beta = 2$	AIR	0.74	0.10	0.98	0.04
	Boruta	0.79	0.11	0.99	0.02
	Boruta Bonferroni	0.69	0.11	0.98	0.05
	Boruta BH	0.79	0.10	0.99	0.03
	Boruta BY	0.75	0.10	0.99	0.04

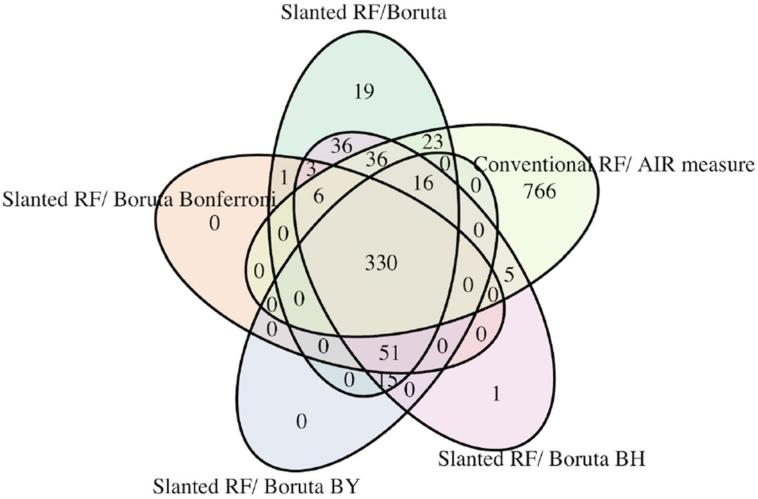

Fig. 4 Venn diagrams showing the number of identified relevant features in the four version of knowledge-slanted RF and conventional RF for CAVS Dataset

knowledge would not be useful to achieve better performance because the algorithm can learn directly from the data as shown in [33].

Due to the large number of genes (features), the process of network reconstruction through exhaustive search may require high computational costs, which is a limitation of this method. Based on simulation studies, knowledge-slanted RF could be better

Table 5 Performance of knowledge-slanted RF and conventional RF using only identified relevant genes as input features for CAVS dataset

Method	Number of relevant genes	Rank in RWR			Accuracy	Depth	
		P25	Median	P75		Mean	SD
Slanted Boruta	536	298	569	896	0.704	1.602	0.613
Slanted Boruta Bonferroni	391	260	566	876	0.741	1.208	0.416
Slanted Boruta BY	412	268	548	875	0.741	1.222	0.430
Slanted Boruta BH	499	297	574	897	0.704	1.242	0.433
Conv. RF	1181	750	3820	9798	0.592	1.542	0.563

than conventional RF under the following conditions: 1) there is a low or moderate similarity in feature space between classes in high-dimensional datasets ($p >> n$), 2) the sample size is limited ($n \leq 30$), and, 3) the prior information is correctly specified.

Identifying features involved in outcome prediction has been the main way in which machine learning models have provided researchers with explanations of how decisions are made within the model [36]. Consequently, the design of an approach to assess the importance of features for knowledge-slanted RF was a critical stage and an important problem to address in this work. The performance of knowledge-slanted RF with extended Boruta was tested in specific situations that may be more realistic for biological applications, assuming that half of the seed nodes are erroneous, leading to biases. Overall, the extended Boruta approach with the AIR measure allowed a decrease in the proportion of false features identified as relevant to y, as well as an increase in the identification of true relevant features.

The extended Boruta approach was constructed using the original method proposed by [21]. The original version of the Boruta algorithm identifies relevant features by setting an importance threshold determined by the highest importance among the mimic features or shadows generated from the original data. Thus, in each run of Boruta, the method generates an RF model and compares whether each original feature offers an importance value higher than the threshold, storing a hit. In the extended Boruta approach for knowledge-slanted RF, the relevance of each feature (or gene) obtained from the RWR is preserved by assigning equal prior weight to the original feature and its respective shadow. This allows the threshold from the shadow features to be generated while preserving the selection probability of each original feature (gene) in the RF construction, so comparability is fair. By definition of knowledge-slanted RF, it is hoped that this method will identify a greater number of relevant features among seed node features even if they do not have a significant relationship on y. For this reason, the simulations showed a higher percentage of false relevant features among the seed nodes with no effect on y when the proposed RF was implemented with the AIR measure. This limitation was mitigated with the extended Boruta approach.

Boruta originally determines the relevance of a feature after n runs of the algorithm by comparing the observed cumulative hits with respect to the expected value using a binomial exact test. Because Boruta performs p multiple tests at once, the use of adjustment for multiple comparisons was evaluated with Bonferroni, BH, or BY methods [25, 26]. Although the overall type I error rate was adequately controlled in all versions of the extended Boruta approach, the Bonferroni method achieved better control of the false discovery rate among seed features. These results are explained by the fact that the Bonferroni correction controls the family-wise error rate (FWER) or type I error and is a conservative method that does not reject many hypotheses. On the other hand, BH and BY adjustments were proposed to control the false discovery rate (FDR), which can be interpreted as the expected proportion of false positives among the rejected hypotheses. In the knowledge-slanted RF framework, Boruta extended with BH and BY adjustments increased the proportion of relevant features identified compared to the AIR measure, improving its performance as the effect size (β) between X and y increases. Both adjustment methods are recommended in biomedical studies, as they can lead to a higher probability of identifying causal features compared to FWER methods, especially when a large number of multiple tests are performed [37, 38].

Regarding the level of explainability achieved with the knowledge-slanted RF framework, the results of the models built on the CAVS dataset evidenced that this approach prefers to select highly interconnected genes (features) compared to conventional RF. This can lead to the successful identification of biologically relevant genes for a specific disease or event of interest, which represents the main advantage of the proposed RF. In addition, this framework allows the prioritization of biological features based on prior knowledge, which helps address COD by reducing the possibility of selecting redundant features that are correlated with true features.

Authors' contributions

All authors designed the study. EC implemented the algorithms and wrote the first draft of the manuscript. RS, SC, RL supervised the research and made the required updates to the methods. SQ provided the clinical expertise and helped with the interpretation of the data. All authors read and approved the final manuscript.

Funding

This work was supported by the National Agency for Research and Development (ANID)/Scholarship Program/DOCTORADO BECAS CHILE/2019:Grant N. 21190261. The authors acknowledge the support given by Chilean ANID FONDECYT research grants number 1221938 and 1231268, and ANID—Millennium Science Initiative Program *ICN*2021_004.

Availability of data and materials

Algorithms can be found at https://github.com/ErikaCantor/kslboruta.

Data availability

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 29 July 2024 Accepted: 2 September 2024
Published online: 10 September 2024

References

1. McDermaid A, Monier B, Zhao J, Liu B, Ma Q. Interpretation of differential gene expression results of RNA-seq data: review and integration. Brief Bioinforma. 2019;20(6):2044–54.

2. Crawford J, Greene CS. Incorporating biologicadoi. Curr Opin Biotechnol. 2020;63:126–34.
3. Efron B. Prediction, Estimation, and Attribution. J Am Stat Assoc. 2020;115(530):636–55.
4. Bommert A, Sun X, Bischl B, Rahnenführer J, Lang M. Benchmark for filter methods for feature selection in high-dimensional classification data. Comput Stat Data Anal. 2020;143:106839.
5. Speiser JL, Miller ME, Tooze J, Ip E. A comparison of random forest variable selection methods for classification prediction modeling. Expert Syst Appl. 2019;134:93–101.
6. Deng H, Runger G. Gene selection with guided regularized random forest. Pattern Recogn. 2013;46(12):3483–9.
7. Hall P, Pittelkow Y, Ghosh M. Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. J R Stat Soc Ser B (Stat Methodol). 2008;70(1):159–73.
8. Nelson W, Zitnik M, Wang B, Leskovec J, Goldenberg A, Sharan R. To embed or not: network embedding as a paradigm in computational biology. Front Genet. 2019;10:381.
9. Seifert S, Gundlach S, Junge O, Szymczak S. Integrating biological knowledge and gene expression data using pathway-guided random forests: a benchmarking study. Bioinformatics. 2020;36(15):4301–8.
10. Guan X, Runger G, Liu L. Dynamic incorporation of prior knowledge from multiple domains in biomarker discovery. BMC Bioinformatics. 2020;21(2):1–10.
11. Oskooei A, Manica M, Mathis R, Martínez MR. Network-based biased tree ensembles (NetBiTE) for drug sensitivity prediction and drug sensitivity biomarker identification in cancer. Sci Rep. 2019;9(1):1–13.
12. Ghosh D, Cabrera J. Enriched random forest for high dimensional genomic data. IEEE/ACM Trans Comput Biol Bioinforma. 2021;19(5):2817–28.
13. Shmueli G. To explain or to predict? Stat Sci. 2010;25(3):289–310.
14. Breiman L. Random Forests. Mach Learn. 2001;45:5–32.
15. Sutera A. Importance measures derived from random forests: characterization and extension. 2021. arXiv preprint arXiv:2106.09473. https://doi.org/10.48550/arXiv.2106.09473.
16. Louppe G. Understanding random forests: From theory to practice. 2014. arXiv preprint arXiv:1407.7502. https://doi.org/10.48550/arXiv.1407.7502.
17. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics. 2007;8(1):1–21.
18. Nicodemus KK. On the stability and ranking of predictors from random forest variable importance measures. Brief Bioinforma. 2011;12(4):369–73.
19. Nembrini S, König IR, Wright MN. The revival of the Gini importance? Bioinformatics. 2018;34(21):3711–8.
20. Sandri M, Zuccolotto P. A bias correction algorithm for the Gini variable importance measure in classification trees. J Comput Graph Stat. 2008;17(3):611–28.
21. Kursa MB, Rudnicki WR. Feature selection with the Boruta package. J Stat Softw. 2010;36:1–13.
22. Sun Y, Zhang Q, Yang Q, Yao M, Xu F, Chen W. Screening of gene expression markers for corona virus disease 2019 through Boruta_MCFS feature selection. Front Public Health. 2022;10:901602.
23. Maurya NS, Kushwah S, Kushwaha S, Chawade A, Mani A. Prognostic model development for classification of colorectal adenocarcinoma by using machine learning model based on feature selection technique boruta. Sci Rep. 2023;13(1):6413.
24. Degenhardt F, Seifert S, Szymczak S. Evaluation of variable selection methods for random forests and omics data sets. Brief Bioinform. 2019;20(2):492–503.
25. Benjamini Y, Hochberg Y. Multiple hypotheses testing with weights. Scand J Stat. 1997;24(3):407–18.
26. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001;29(4):1165–88.
27. Giraud C, Huet S, Verzelem N. Graph selection with GGMselect. Stat Appl Gene Mole Biol. 2012;11(3):1–50.
28. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43(D1):D447–52.
29. Mi Z, Guo B, Yin Z, Li J, Zheng Z. Disease classification via gene network integrating modules and pathways. R Soc Open Sci. 2019;6(7):190214.
30. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw. 2017;77(1):1–17.
31. Tay JK, Narasimhan B, Hastie T. "Elastic Net Regularization Paths for All Generalized Linear Models." J Stat Soft. 2023;106(1):1–31.
32. Guauque-Olarte S, Droit A, Tremblay-Marchand J, Gaudreault N, Kalavrouziotis D, Dagenais F, et al. RNA expression profile of calcified bicuspid, tricuspid, and normal human aortic valves by RNA sequencing. Physiol Genomics. 2016;48(10):749–61.
33. Cantor E, Salas R, Rosas H, Guauque-Olarte S. Biological knowledge-slanted random forest approach for the classification of calcified aortic valve stenosis. BioData Min. 2021;14:1–11.
34. Erten S, Bebek G, Koyutürk M. Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. J Comput Biol. 2011;18(11):1561–74.
35. Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet. 2008;82(4):949–58.
36. Watson DS. Conceptual challenges for interpretable machine learning. Synthese. 2022;200(2):65.
37. Benjamini Y. Discovering the false discovery rate. J R Stat Soc Ser B (Stat Methodol). 2010;72(4):405–16.
38. Glickman ME, Rao SR, Schultz MR. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. J Clin Epidemiol. 2014;67(8):850–7.

Publisher's Note