TECHNISCHE UNIVERSITÄT MÜNCHEN

LEHRSTUHL FÜR GENOMORIENTIERTE BIOINFORMATIK

STABILITY AND ACCURACY ANALYSIS OF A FEATURE SELECTION ENSEMBLE FOR BINARY CLASSIFICATION IN BIOMEDICAL DATASETS

URSULA NEUMANN

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Volker SIEBER

Prüfer der Dissertation:

1. Prof. Dr. Dmitrij FRISCHMANN
2. Prof. Dr. Dominik HEIDER

Die Dissertation wurde am 25.10.2017 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 17.01.2018 angenommen.

"All models are wrong, but some are useful."

GEORGE BOX

Abstract

In the last few decades, major advances have been made in the way we collect and generate data in all aspects of life. At the same time, the technical approaches for analyzing and interpreting datasets have improved. The intersection of these trends is referred to as *Big Data* and this plays an important role in precision medicine and other fields of computer science.
Feature selection algorithms have contributed to considerable progress in coping with the growing amount of machine readable information. They can identify a minimal set of features that are relevant for developing highly accurate prediction models. Thus, feature selection simplifies the interpretability as well as computability of big datasets. Many different feature selection methods already exist. Previous studies have shown that some of these are biased, depending on the feature type and dataset quality. In this thesis, an ensemble consisting of eight different feature selection methods (EFS) is introduced. An ensemble of learning algorithms has the advantage of being able to address the biases of the individual approaches. Additionally, EFS provides a cumulative quantitative feature ranking. EFS was applied on several biomedical datasets. Different feature subset selections resulting from the EFS ranking were evaluated on three popular prediction models, namely logistic regression, random forest and support vector machine. In most cases, a significant improvement of the prediction performance could be achieved when compared to the same models constructed with all features.
The EFS approach and the evaluations were implemented as an R package *EFS* as well as a web-based application. A quantitative feature ranking and a cumulative barplot of the feature's importance values are provided as the output.

Zusammenfassung

In den letzten Jahrzehnten wurden bedeutende Fortschritte darin gemacht, Daten aus allen Lebensbereichen zu erzeugen und anzusammeln. Zeitgleich verbesserten sich auch die technischen Möglichkeiten diese Datensätze zu analysieren und interpretieren. Die Schnittstelle dieser beiden Entwicklungen wird *Big Data* genannt und spielt eine wichtige Rolle im Bereich der Hochtechnologiemedizin.
Einen beträchtlichen Beitrag zur Bewältigung dieser enorm wachsenden Menge an maschinenlesbaren Informationen brachten die sogenannten *Feature Selection* Algorithmen. Sie bestimmen die minimale Teilmenge von Parametern, die für Vorhersagemodelle mit hoher Genauigkeit relevant sind. Somit vereinfacht eine *Feature Selection* die Interpretierbarkeit, sowie die Berechenbarkeit großer Datensätze. Es existieren bereits mehrere verschiedene *Feature Selection* Methoden. Frühere Studien zeigen jedoch, dass einige dieser Methoden Fehleranfälligkeiten aufgrund von Parametertyp und der Qualität der Datensätze aufzeigen. In dieser Arbeit wird ein Ensemble aus acht verschiedenen *Feature Selection* Methoden (EFS) vorgestellt. Ein Ensemble von Lernalgorithmen hat den Vorteil die Fehleranfälligkeiten von einzelnen Methoden auszugleichen. Zusätzlich liefert EFS eine kumulative, qualitative Rangliste der Parameter. EFS wurde auf mehrere biomedizinische Datensätze angewendet. Verschiedene Parameterteilmengen, die aus der EFS-Rangliste hervor gegangen sind, wurden mittels folgender drei gängiger Vorhersagemodellen evaluiert: *logistische Regression*, *Random Forest* und *Support Vector Machines*. In den meisten Fällen konnte eine signifikante Steigerung der *Vorhersageperformance* erreicht werden. EFS und die Evaluationsmethoden wurden sowohl als R-Paket *EFS*, wie auch als Web-Applikation implementiert. Der *Output* besteht hierbei aus einer quantitativen Parameterrangliste und einem kumulativen Barplot der Werte der *Feature-Importance*.

Acknowledgments

Through my PhD studies at the Straubing Center of Science, I had the greatest privilege to work with kind and inspiring people without whom it would not be possible to finish my thesis. My special appreciation and thanks go to my mentor, Prof. Dr. Dominik Heider, whose guidance and support enabled me to complete this dissertation. I would like to thank him for encouraging my research and for allowing me to grow as a research scientist. I also greatly appreciate the support and supervision of the dissertation by Prof. Dr. Dmitrij Frischmann.

Many thanks also go to my colleagues and friends at the Straubing Center of Science, in particular those at the Chair of Bioinformatics, the Chair of Marketing and Management of Biogenic Resources and the Chair of Business Economics of Biogenic Resources for their great company and the friendly working atmosphere during the last three years.

Finally, I am very grateful to my family and friends for their continuous support and encouragement whilst completing the dissertation. Special thanks to my parents for their constant support and understanding.

Contents

List of Figures

List of Tables

1. Introduction

Data is more than just binary information. It is a set of related features. The process of discovering patterns, trends and anomalies in datasets is called data mining, this is one of the greatest challenges of the information age, involving approaches from the fields of machine learning, statistics and database systems.
In the following chapter, I will give an introduction to this doctoral thesis. First, I will explain the background of this thesis. Then I will introduce the aims and hypothesis and finally, I will give an outlook over the remaining parts of this thesis.

1.1. Background

The term *machine learning* (ML) was coined by Arthur L. Samuel, who was a pioneer in the field of artificial intelligence, in 1959. He came up with the idea that it might be more efficient to teach computers to learn themselves rather than teaching them everything about the world that they should know. Due to the internet, the amount of digital information has increased enormously in recent times. More and more data is being generated, stored and made available for further processing. This may appear useful as it represents the procurement of new information, but apart from that, it also poses new challenges in terms of identifying the significant elements of this information. ML approaches are developing quickly and they are able to manage these vast datasets. They are so pervasive nowadays that we use them a dozen times a day without even knowing it. ML is the current state-of-the-art in developing artificial intelligence. Because the amount of machine readable information is growing immensely, pre-processing filters for big amounts of data are required. In large datasets, there is often the problem that there are many features with only a small number of samples. Building a valid model is negatively impacted by an increasing ratio between the number of features and the sample size [1]. Comprehensive data analysis also includes many irrelevant and redundant features, which "degrade the performance of concept learners both in speed (due to high dimensionality) and predictive accuracy (due to irrelevant information)"

[2]. Therefore, distinguishing between relevant and irrelevant or redundant features is fundamental to ML approaches.

Feature selection (FS) is the process of detecting and selecting a subset of relevant features for the construction of a model. In this process, as much of the irrelevant and redundant information as possible should be identified and removed [3]. Guyon and Elisseeff [4] summarize the main benefits of FS: "facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times, defying the curse of dimensionality to improve prediction performance."
There are already many FS methods [5], most of which are focused on one aspect, for example, dimensionality reduction in microarrays [6] or relevance and redundancy analysis [7]. Ideally, the selected subset is necessary and sufficient to describe the real relationships between dependent and independent features. In reality, there may be more than one optimal subset which describes the real function [8]. In a preceding study together with cardiologists from the University Hospital Duisburg-Essen, we recognized that the Gini-index, when used as the feature selection method, is strongly biased on unbalanced datasets (cf. B.1 in 3). Thus, the idea arose to look for an intersection of multiple subsets proposed by different kinds of FS methods. In other words, implementing an ensemble of FS methods may help to find an optimal subset of features. The individual FS methods in an ensemble are called base selectors.

Preceding studies [9, 10, 11] have shown that ensemble learning approaches may outperform the single methods when several weak methods are combined. There are three main reasons for this:

a) several different but equally optimal hypotheses may exist and an ensemble diminishes the risk of choosing an incorrect hypothesis,

b) learning algorithms calculate different locally optimal outcomes and the ensemble approximates the real function, and

c) the real function may not correspond with one of the hypotheses in the hypothesis space but by aggregating the outputs of the single models, the hypothesis space may be expanded.

In this context, weak FS methods are regarded as those that are unstable with respect to the errorproneness. Unstable methods may perform very badly on new datasets due to overfitting in the training process or because they tend to prefer specific features,

i.e., features of a specific type. There are two different approaches to ensemble learning with respect to feature selection: i) homogeneous ones, which use the same FS method with different training data and ii) heterogeneous ones, which use different FS methods with the same dataset [12]. As I used different base selectors, the resulting ensemble is heterogeneous.

1.2. Aims and Hypothesis

Against this background, this thesis states the following hypothesis:

A single feature selection approach gives less reliable results than an ensemble of different base feature selection methods for binary classifications.

The objective of this study is to develop an ensemble of feature selection methods which outperforms each single feature selection in terms of classification accuracy and stability of the selection. An accurate approach is expected to not show any bias arising from either the quality of the dataset, e.g. because of an imbalance in the class variables, or from the existence of different types of features in the dataset. The required stability can be defined as the variation in feature selection results due to changes in the dataset. Therefore, the ensemble feature selection (EFS) aims to be applicable to every kind of data used in binary classification. This approach should propose a reliable ranking of features and suggest a suitable subset of features, which improves the accuracy of a model which is constructed by all features. In this thesis, existing feature selection methods will be analyzed to determine their deficiencies and to compare them with the EFS by several different evaluation techniques.
Thus, the aim is to develop a universal FS method which performs equally well on all kinds of datasets, independently of the balance of data and type of features.

1.3. Thesis Structure

The rest of this thesis is arranged as follows: In chapter 2, an overview of the **methods** used in the published papers and for the other research is given. Firstly, a deeper insight into the chosen feature selection methods for the EFS approach is provided. Then, the way in which the FS methods were assembled is described, followed by the subset selection criteria. Finally, an evaluation of the methods is made.
In chapter 3, the **results** of the three articles (B.1,B.2, B.3) are presented together with extended abstract (B.4 submitted at the 14th Annual Meeting of the Bioinformatics

Italian Society, 2017 in Cagliari, Italy) and the results of the other investigations.
Chapter 4 includes a **discussion** of the strengths and weaknesses of the methods used and the results of my research. In this chapter, an outlook on future work which is needed is provided together with the conclusion.
In appendix A, all relevant figures are provided and the publications are included in appendix B.

2. Methods

2.1. Base Selectors

In computer science, features are distinct attributes or aspects of an observed process. Machine learning approaches deal with sets of features, which may contain up to hundreds of thousands of features. The purpose of feature selection (FS) methods is to detect features, that are relevant for the prediction of target variables. For clarification purposes,: "[...] if a feature is to be relevant it can be independent of the input data but cannot be independent of the class labels i.e. the feature that has no influence on the class labels can be discarded." [13].

By removing redundant features, both the efficiency of prediction models and the learning performance can improve. In a multiple regression model, the prediction performance can be biased by a high correlation between two or more predictor variables. This phenomenon is called collinearity or multicollinearity, meaning that one feature can be linearly described from the others with a substantial degree of accuracy. To prevent bias in the prediction accuracy of the model, multicollinearity should be avoided by removing the correlated feature subset and representing these data with their best exponent. In terms of interpretation, having less features implies an enhanced comprehensibility of the learning results [14]. Due to the need for feature selection, there are many different methods. Generally, they can be divided into supervised and unsupervised procedures. In supervised learning, a function is inferred from a labelled training dataset. In contrast, unsupervised learning tasks deal with unlabelled datasets. Hence, the output cannot be evaluated in terms of accuracy. A classic example for the latter is clustering. In this thesis, I only considered supervised methods for binary classification. That means that the focus is on training data which can be classified in two groups on the basis of a certain classification rule. In other words, binary classification is a dichotomization of training datasets. There are three different kinds of supervised FS approaches, namely filter, wrapper, and embedded methods[13, 15].

Filter methods are a pre-processing step used to obtain a ranking of relevant features.

Their merit is their simplicity. They are based on ranking criterion and a set threshold, which are used to remove dispensable variables.

Wrapper methods use the quality of the prediction performance as importance measure. Embedded methods select the features within a training process, which saves computation time. However, the wrapper models tend to be computationally more costly than filters [16].

I used implementations in R (`http://www.r-project.org/`) for the different basal feature selection methods.

In the following sections, eight different FS methods are introduced. These were used as the base selectors in the thesis. Three of them are filter methods and include median, Pearson and Spearman correlation. The other methods are embedded methods and include logistic regression, the Gini-index-based, error-rate-based variable importance measures (VIMs) of Breiman's random forest implementation and the error-rate-based and AUC-based VIMs of the conditional random forest implementation.

2.1.1. Median

The simplest method for finding out whether a feature is relevant for a classification is to compare the distributions of the feature's values in both classes. Therefore, a Mann-Whitney-U test (also called Mann-Whitney-Wilcoxon test) is conducted for each feature.

The Mann-Whitney-U test is often used as an alternative to the t test, where the features are not normally distributed. While the t test compares the means, the Mann-Whitney-U test compares the median.

If the shapes of the distribution curves are the same and the only difference between the two classes is a shift in location, there is indeed a difference in the medians. However, the Mann-Whitney-U test can also detect differences in the spread of values, even when the medians are equal. Thus, it is not just a test of medians (cf. [17]), but for reasons of abbreviation and simplification the method is called the median.

The null hypothesis is that the distributions of feature values of the two classes do not differ. Normal distribution is not necessary. The following calculations are incorporated in the Mann-Whitney-U test: Firstly, ranks are assigned to the values of the features, where observations with tied values are assigned the equal averaged ranks. The statistic U is calculated as follows:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2},$$

where R_1 is the sum of the ranks of class one and n_1 is the sample size of class one. This is analogous for U_2. The smaller value of U_1 and U_2 is taken for testing significance. Let us suppose that $U_1 \leq U_2$. The p-value for the corresponding rank-sum test is calculated as follows:

$$p = Pr(r_1 \neq R_1) = 2 \cdot Pr(r_1 \geq R_1),$$

where $Pr(E)$ is the probability of event E and r_1 is the distribution of the rank sum of class one, since $U_1 \leq U_2$.

The Mann-Whitney-U test is very fast to compute and it is not parametric. However, the results are not reliable where there is a relationship between the features. Another negative influence on the accuracy emerges in high-dimensional datasets.

2.1.2. Correlation coefficients

A correlation coefficient describes the extent to which two features are dependent by calculating the strength and the direction of the correlation. Correlation coefficients as a feature selection criteria fall into the category of filter methods. The advantage of this method is the high computational speed. Former studies have shown that correlation based feature selections outperformed wrapper methods on small datasets and delivered comparable results [18, 19].

In big datasets, it often happens that two features are highly correlated with each other, which means that a relationship can be inferred between them. This phenomenon is called collinearity or multicollinearity. It does not reduce the prediction power of the model, but it distorts the feature ranking. Furthermore, irrelevant and redundant features affect the speed of the learning algorithms. To avoid multicollinearity, I implemented the correlation methods as fast-correlation-based filter (FCBF) after [7]. For this, a threshold had to be defined which determined the maximum tolerated correlation among the features. In my experiments, I decided to set the threshold at 0.7, as this is the most frequently used correlation threshold [20]. There are also more restrictive [21] and less restrictive [22] suggestions for the threshold. Taking X_1 and X_2 as two features which have a correlation exceeding the given threshold, the features with the higher correlation with the class variable is the predominant feature. The other feature will be discarded, meaning that it is given an importance value of zero.
Another aspect is how the correlations between features are conducted. A pairwise correlation between all features is very time-consuming. A solution to this is the following: first, correlations of all features with the class variable are calculated. Then, the best

correlated feature is tested against all other features. If a correlation can be detected which exceeds the threshold, that feature is deleted. The next step is to test the second best correlated feature and so on. By not testing the correlations of redundant features, substantial savings in computation time can be achieved.

Of the many which exist the two most popular correlation coefficients are Pearson's product moment and Spearman's rank correlation coefficients. In some respects, these correlation coefficients do not differ significantly. However, their results can be distinguished according to the type of the tested parameter. To cover the whole range of different parameter types, I included both measures in the ensemble.

Pearson correlation

The Pearson product-moment correlation coefficient r evaluates the linear relationship between two features. This is convenient for measuring the correlation between numerical parameters. The Pearson correlation coefficient r is calculated as follows:

$$r = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}.$$

Instead of calculating the covariance and standard deviation of the ranks, these statistics are calculated using the actual parameters X and Y. The formula is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{Y})^2}},$$

where n is the sample size and $\overline{X}$ and $\overline{Y}$ are the sample means.

Spearman correlation

Spearman's *rho* compares the monotonic relationship of the two parameters. Thus, it is often used to find a relationship between ordinal parameters. De La Fuente et al. recommend Spearman rank correlation for biochemical networks, because it does not depend on normality and the linearity of interactions [23]. To calculate Spearman's *rho*, we observe the ranks $rk(x_i)$ and $rk(y_i)$ of the two parameters X and Y of length n.

$$\rho = \frac{Cov(rk(X), rk(Y))}{\sigma_{rk(X)}\sigma_{rk(Y)}},$$

where $\sigma_{rk(X)}$ is the standard deviations of the ranks of X,

$$Cov(rk(X), rk(Y)) = \frac{1}{n}\sum_{i=1}^{n}(rk(X_i)rk(Y_i) - \overline{rk(X)rk(Y)}$$

is the covariance of the ranks of X and Y and $\overline{rk(X)}$ is the mean of the ranks of X. There is a more popular formula, which can only be used, if all ranks appear exactly once:

$$\rho = 1 - 6\sum_{j=1}^{n}\frac{d_i^2}{n(n^2-1)},$$

where $d_i = rk(X_i) - rk(Y_i)$.

2.1.3. Logistic Regression

Regression models are used to calculate the relationship between features. If the dependent variable is categorical and binary, a logistic regression model is used. It estimates probabilities using a logistic function, which has the following formula:

$$f(t) = \frac{1}{1+e^{-t}}$$

This logistic function, also called the antilogit transformation, allows us to go from real input to probabilities, which lie between 0 and 1. Values under 0.5 are allocated to class 0 and values above 0.5 are allocated to class 1. For multiple explanatory variables, t can be expressed as follows:

$$t = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n,$$

where $X_1, \ldots, X_n$ are the n parameters. Logistic regression models the logit-transformed probability as a linear relationship with the dependent variable. To provide comparability of the variables' importances, a z-transformation must be conducted in a preprocessing step:

$$z_{X_i} = \frac{X_i - \overline{X}}{\sigma_X},$$

where $\overline{X}$ is the mean and σ_X the standard deviation of variable X. The β-values can be used as importance measures, because they describe the strength of the relationship towards the dependent variable. The inverse of the logistic function $f(t)$ is the logit

$$g(p) = \ln(\frac{p}{1-p}),$$

where p is a number between 0 and 1. The β-values are calculated as follows:

$$g(f(t)) = \ln\left(\frac{f(t)}{1-f(t)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n.$$

The β-values can have negative values. Thus, the absolute values are used as quantitative importance measures.

2.1.4. Variable Importance Measures embedded in Random Forests

The following four FS methods are incorporated into two different implementations of the random forest (RF) algorithm. RFs are an ensemble of learning methods for classification and regression tasks. They consist of multiple decision trees and, in the case of classifications, their output is the class voted by the majority. Ho [24] criticized the limited complexity of decision trees. Thus, he was the first who developed the idea of randomly selecting features. Breiman extended the RFs with the concept of *bagging* [25] to reduce variances and avoid overfitting. Bagging is an abbreviation for *bootstrap aggregating*, which describes a two-step process. The first one is the bootstrapping, which means randomly generating distinct subsets of the same training dataset. The second step is aggregating the outputs by calculating their average. Breiman's algorithm incorporates two different FS methods, namely the Gini-index-based measure and the permutation error-rate-based measure [26]. In the context of RFs, the FS methods are called variable importance measures (VIMs). Breiman's RF is the most famous and its corresponding R-package is called *randomForest*. It uses the CART (classification and regression trees) algorithm, which only generates binary trees and therefore seeks the optimal binary splitting. However, there are also other implementations of the RF. For example, the conditional random forest incorporated in the R-package *party* [27]. It is based on conditional inference trees and offers two VIMs, which are preferable to those of *randomForest* if there are different types of features: a conditional permutation error-rate [28] and a permuted AUC-based VIM [29]. The difference to CART trees is that in the tree building process, each feature is globally tested for its association with the response, yielding a global p-value. Within this globally selected predictor, the best split is finally chosen. Thus, the splitting is unbiased.

Gini-index-based VIM

The Gini-index (also called Gini coefficient) is a statistical measure, which was developed by Corrado Gini to express unequal distributions [30]. Breiman integrated the Gini impurity into his RF algorithm as a splitting criterion. The Gini impurity measures how often a randomly chosen element would be incorrectly labelled if it was randomly labelled. The random labels correspond with the distribution of labels in the subset. In the formula of the Gini index, the measure of the Gini impurity for binary classification is defined as follows:

$$G = 2p(1-p),$$

where $p = \frac{N_1}{N}$ is the proportion of one of the classes, in this case for response $Y = 1$, and N_1 is the number of units in this class. The Gini impurity is:

$$I_G = G - (\frac{N_L}{N}G_L + \frac{N_R}{N}G_R),$$

where G_R and G_L are the Gini-indexes calculated for the following right and left child nodes and N_L and N_R are the numbers of units in the left and right nodes after splitting. Adding up the decrease in Gini impurity over all of the trees in the random forest for each feature indicates the importance of the variable.

Error-rate-based VIM

For each construction of a tree, a different bootstrap sample is chosen from the original data. Thus, it is not necessary to conduct cross-validation, which would involve dividing the dataset into k subsets and defining iteratively one subset as the test dataset and the other $k-1$ subsets as the training dataset. In RFs, about one-third of the samples are left out of the bootstrap sample and not used in the construction of the tree. This portion is called the out-of-bag (OOB) data. The error-rate-based VIM is computed from permuting the OOB data: first, the prediction error on the OOB data is calculated, then the same is done after randomly permuting each feature. The difference between both error-rates is averaged over all trees and normalized by their standard deviation (except the standard deviation is zero).

Conditional error-rate-based VIM

The underlying mathematics in this error-rate-based VIM is the same as in the previous VIM. Strobl et al [31] pointed out that correlations between features have a severe

effect on random forest VIMs. The classical error-rate-based VIM overestimates the importance of the correlated features, which may be due to the preference for correlated features in the early splits. Taking X_i as the tested feature on the dependent variable Y and $Z = X_1, \ldots, X_{j-1}, X_{j-1}, \ldots, X_m$ the remaining features. A positive value of the classical error-rate-based VIM signifies a correlation of X_i and either Y or Z. An unbiased VIM should only consider the correlation between X_i and Y. Therefore, the remaining features are grouped and X_i is only permuted in those fixed groups $Z = z_j$. Unbiased conditional inference tree [32] are used as the underlying permutation grid.

AUC-based VIM

The AUC-based permutation VIM [29] is closely related to the conditional error-rate-based VIM. However, instead of using the error rate, the prediction accuracy is measured by the area under the curve (AUC) [33]. Analogous to the error-rate the AUC is computed for each tree after and before permuting a feature. In general, the AUC is used to evaluate the ability of classification models to correctly distinguish between different classes. In this case, the dependent variables $X_1, \ldots, X_n$ are considered as the units to be predicted, rather than the samples $i = 1, \ldots, n$ [34]. The AUC of a tree is an estimator of class probabilities, i.e. it estimates the probability of each observation belonging to either class 0 or to class 1.

2.2. Ensemble Feature Selection

The idea of ensemble learning is widely used in the field of machine learning. The high levels of interest in ensemble learning methods is based on the assumption that several models obtain more reliable results than any one individual model. An ensemble increases the performance compared to single methods, if the individual classifiers are both accurate and make their errors on different parts of the input space [35].

Therefore, we utilized the eight base selectors introduced above. All of them have distinct biases and benefits according to the types of features, the degree of imbalance or the size of datasets.
Paper B.2 explains how the base selectors are integrated and how their results are normalized to a common range to ensure comparability.

2.3. Subset selection criteria

After analyzing feature importance and compiling a ranking of features, a subset selection is required. The selected feature subset should simplify the comprehensibility of the relationships within the dataset and optimize the calculation speed as well as the performance of the successive prediction models. From EFS, a continuous measure of importances is obtained. The next step is to identify a cut-off point above which features are considered to improve prediction performance. There are different methods for that task, for example, the mean method. It calculates the mean of all importance values from EFS as being the cut-off point. The mean measure is implemented in the R-package EFS. Another technique would be to arrange the features in ascending order and identify the largest difference from one feature to another. As mentioned in B.4, this method is not suitable for datasets with an exponentially growing importance curve, because it will only select the most important feature. Therefore, I developed a new subset selection method: the $\frac{\pi}{4}$-rotation B.4.

2.4. Evaluation Methods

All evaluations of the prediction performance were made using an area under the curve (AUC) analysis of the receiver operating characteristic (ROC) curves. ROC curves illustrate the performance of binary classifiers by plotting the true positive rate (TPR)

against the false positive rate (FPR) at various threshold settings. The underlying prediction models of the ROC curves were either logistic regression, random forest or a support vector machine (SVM). The main ideas of logistic regression and random forest have already been described. Thus, I will only elaborate on SVM in the following subsection.
Comparing two AUCs of the ROC curve is done using a roc-test after DeLong et al. from the R-package *pROC* [36].
The EFS R-package, described in paper B.3, provides certain additional evaluation tools including permutation tool, which permutes the class variable in order to compare the prediction performance with random guessing, an importance variance measure, and the Jaccard-index [37], which tests the stability of the importance values of multiple EFS iterations.

2.4.1. Support Vector Machine

Support vector machines are supervised learning models for regression and classification analysis. In the case of binary classification, the goal of the SVM is to separate the two classes by a function, which is generated through the training dataset. The aim is to separate the data using something similar to a hyperplane. However, in many datasets, it is not possible to divide the classes linearly. Therefore, additional dimensions are added, a process which is described by a scalar product $K(x, y) = \langle \phi(x), \phi(y) \rangle$.According to Mercer's theorem [38], this scalar product has to be a positive-definite kernel, to ensure that the structures are retained. In other words, the kernel matrices should only have non-negative Eigenvalues. The use of a positive-definite kernel ensures that the optimization problem will be convex and the solution will be unique. Choosing the right kernel can be tedious. Therefore, I tested the four most common kernels on all datasets and chose the one which performed the best. The tested kernels include radial basis, linear, polynomial and the Bessel function. Linear kernels only allow lines or hyperplanes to be identified. The simplest kernel function is given by

$$K(x, y) = x^T y + c,$$

where c is a constant. The Gaussian kernel is also called radial basis kernel and enables identification of circles (or hyperspheres):

$$K(x, y) = \frac{||x - y||^2}{2\sigma^2},$$

where the adjustable parameter σ plays a crucial role in the performance of the kernel. The polynomial kernel can model feature conjunctions up to the order d of the polynomial:

$$K(x, y) = (\alpha x^T y + c)^d,$$

the adjustable parameter α is the slope and c is a constant. The polynomial kernels are well suited for normally distributed datasets. The Bessel kernel is defined as:

$$K(x, y) = \frac{J_{v+1}(\sigma||x - y||)}{||x - y||^{-n(v+1)}},$$

where J is the Bessel function of first kind [39]. All of them are provided in the `ksvm` function of the *kernlab* R-package.

3. Results

3.1. Paper I

Baars T, Neumann U, Jinawy M, Hendricks S, Sowa JP, Kälsch J, Riemenschneider M, Gerken G, Erbel R, Heider D, and Canbay A. **In Acute Myocardial Infarction Liver Parameters are Associated with Stenosis Diameter.** Medicine. 2016; 95(6): e2807 (Appendix B.1)

3.1.1. Brief Introduction

This paper examines the statistical coherence of the liver parameter and the stenosis diameter in patients with acute myocardial infarction. The retrospective single-centre study includes a cohort of 437 patients, who underwent coronary angiography in the catheterization laboratory of the West German Heart and Vascular Centre Essen, University Hospital Essen, after suffering an acute myocardial infarction. The observed parameters were of different type and range. There were socio-demographic and serum parameters, which were either numeric or categorical. Most of the categorical parameters were dichotomous. Due to the imbalance of the class variable stenosis diameter, models were conducted with a 100-fold leave-one-out cross-validation. Missing values were imputed by mean imputation.

The author of this thesis contributed the statistical analysis of the importance of the parameters via the Gini-index-based VIM, which is embedded in the random forest algorithm. The importance was calculated based on 100 individual RF models.

3.1.2. Study Findings

The results of the Gini-index-based VIM appeared to be biased in such a way that categorical parameters with fewer categories, namely the dichotomous parameters, seem to get lower importance values by default.

For example, the medical practitioners considered the gender to be an important prediction parameter, but according to the Gini-index this was 0. Further investigations revealed that the Gini-index is known to compute a lower importance for features with fewer categories [40, 41]. Thus, dichotomous features are evaluated to be irrelevant for the prediction model. Besides gender, a low importance was calculated for the dichotomous parameters Diabetes mellitus, (N-)STEMI, dislipidemia, predisposition, and smoking, which had three categories, including smoker, former smoker, and non-smoker (cf. boxplot on page 4 of appendix B.1).

3.1.3. Conclusion

The Gini-index-based VIM incorporated in the random forest algorithm produces unreliable results for categorical parameters. Therefore, a comparison with other feature selection methods was needed. Subsequently, I applied the following EFS method from paper B.2 on the dataset from paper B.1, which gave the gender parameter higher importance (0.34 with an average importance of 0.34) (cf. figure A.1).

3.2. Paper II

Neumann U, Riemenschneider M, Sowa JP, Baars T, Kälsch J, Canbay A, and Heider D. **Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach.** BioData Mining. 2016; 9(1), 36. (Appendix B.2).

3.2.1. Brief Introduction

The article introduces and analyzes the eight different FS methods mentioned in chapter 2. Furthermore, the author of this thesis developed an ensemble framework of these FS methods by normalizing their outputs to a common scale. The so-called ensemble feature selection (EFS) was tested on six different datasets. As a follow-up to the study described in paper B.1, I retrieved the *MI-Mortality* dataset from the West German Heart and Vascular Centre Essen, University Hospital Essen. The datasets *Fibrosis* and *FLIP* were provided by the Department of Gastroenterology and Hepatology of the University Hospital Duisburg-Essen. The remaining datasets *SPECTF, Sonar*, and *WBC* were provided by UCI Machine Learning Repository [42].

To evaluate the feature ranking, logistic regressions were performed for the following different feature subsets and the AUCs of their ROC curves were also analyzed:

a) features, which were above the average EFS importance,

b) features, which were above the average importance of the AUC-based VIM,

c) all features.

A leave-one-out cross-validation (LOOCV) was conducted, although LOOCV is known to give inflated variance estimation [43]. However, in this case, it was only used to assess the method's performance.

The author of this thesis developed the EFS framework, performed the data analysis and wrote the manuscript together with co-authors MR and DH.

3.2.2. Study Findings

The resulting ROC curves are shown on page 10 of appendix B.2. The AUCs of the subsets a) and b) were compared using the method of DeLong et al. [36]. The performance of the best EFS ranked features increased in all datasets, however, the improvement was not always significant: *MI-Mortality* (p=0.228), *Fibrosis* (p=0.273), *FLIP* (p=0.254), *SPECTF* (p=0.444), *Sonar* (p=0.2), and *WBC* (p=0.02). Additionally, a comparison via Venn diagrams was made between the feature subsets a) and b)(cf. Venn diagrams on page 8 of appendix B.2). The results showed, that no distinct trend is obvious according to the size of the subsets.

3.2.3. Conclusion

In conclusion, it can be said that the prediction performance of logistic regression could be enhanced by the EFS method and there is no particular preference for any type of parameter in the feature subset selected by EFS.

3.3. Paper III

Neumann U, Genze N, and Heider D. **EFS: An Ensemble Feature Selection Tool implemented as R-package and Web-Application.** BioData Mining. 2017; 10(1), 21. (Appendix B.3)

3.3.1. Brief Introduction

The third publication is a technical report on the implementation of the EFS method as an R-package and its corresponding web-application on `http://efs.heiderlab.de`. The R package EFS is divided into three functions, namely `ensemble_fs`, `barplot_fs` and `efs_eval`.

The author of this thesis implemented the R-Package together with NG, implemented the web application and drafted the manuscript.

3.3.2. Study Findings

The first function `ensemble_fs` includes an ensemble of the eight feature selection methods introduced in paper B.2. The user can select an individual subset of these selections and their summed results will be normalized to a range between 0 and 1. The output is a $m \times n$-table, where n is the number of features and m is the number of selected methods to be conducted. The `barplot_fs` function displays the results of `ensemble_fs` in a barplot. In `efs_eval`, several measures are offered to evaluate the accuracy and stability of the subset selection made by EFS. The user can choose between an evaluation of the prediction performance via the ROC curve of an underlying LR model or measures of stability, either as the Jaccard index or a permutation test. The class variable is randomly permuted to see if there is a significant difference in performance when compared to guessing. The web-application is a pared-down version of the R-package developed by NG with R-shiny. It contains the `ensemble_fs` and a simplified version of `barplot_fs`. Evaluations are not available on the web-application.
The resulting barplots from `barplot_fs` with `order = TRUE` of the the six datasets of paper B.2 are shown in figures A.2 to A.7.

3.3.3. Conclusion

The EFS method can improve prediction accuracy and simplify the interpretability by reducing the number of features. So far, the method was only tested on small datasets. For future research, the method should be tested on larger datasets with more features than the number of samples. With the web-application, we supply a pared-down version suitable for practitioners who are not used to dealing with R.

3.4. Paper IV

Neumann U and Heider D.: **Assessment of Subset Selection Criteria of Quantitative Feature Selection Methods.** Proceedings of the 14th Annual Meeting of the Bioinformatics Italian Society, 2017 in Cagliari, Italy, submitted (Appendix B.4)

3.4.1. Brief Introduction

Feature selection (FS) methods are an important pre-processing step in prediction models. They distinguish between features which are relevant for prediction and those which can be neglected. By applying the EFS method, we obtain a quantitative feature ranking. approach.EFS uses the mean as an integrated subset selection criterion. However, there are also other possibilities to define a subset selection criterion, such as selecting the best 15% or the best 10% of all features. A more conservative cut-off point is the elbow of the importance curve. In other words, we locate the point on the curve where the slope exceeds 45 degrees: first, a rotation of the curve by -45 (i.e. $-\frac{\pi}{4}$ degrees) is conducted followed by a minimum search.

The author of this thesis developed the $\frac{\pi}{4}$-framework, performed the data analysis and drafted the manuscript.

3.4.2. Study Findings

We analyzed two big datasets *Ad* and *Arcene* with 1.430 and 79.360 features respectively. The datasets were obtained from the UCI Machine Learning Repository [42]. We could observe that the curve of feature importance values grew exponentially (cf. figures A.8 and A.9). Tables 3.1 and 3.2 show the results of the assessment of subset selection criteria. In general, it turned out that having less features results in better performance due to the reduction of noise. The $\frac{\pi}{4}$-rotation is the most conservative method, i.e., it selects the smallest number of features. Figures A.10 to A.13 show the according ROC curves.

3.4.3. Conclusion

For datasets with exponential curves of importances, the $\frac{\pi}{4}$-rotation method provides an improvement in prediction performance with using the random forest model compared to more liberal methods like the mean. In further studies, the performance should be tested with other datasets with large numbers of parameters.

Table 3.1.: **Evaluation of Arcene data.** AUC values of a logistic regression model and a random forest model.

Method	AUC from LR	AUC from RF	Nr of features
Mean	66.7%(60.0...80.0)	89.9%(80.0...100.0)	5038
best 15%	79.9%(70.0...90.0)	91.9%(90.0...100.0)	1488
best 10%	79.9%(70.4...89.4)	90.9%(85.2...96.6)	992
$\frac{\pi}{4}$-rotation	62.5%(51.5...73.6)	92.5%(87.3...97.8)	374

Table 3.2.: **Evaluation of Ad data.** AUC values of a logistic regression model and a random forest model.

Method	AUC from LR	AUC from RF	Nr of features
Mean	55.2%(50.0...60.0)	98.4%(100.0...100.0)	613
best 15%	69.9%(70.0...70.0)	98.4%(100.0...100.0)	214
best 10%	91.9%(89.6...94.1)	98.5% (97.6...99.3)	143
$\frac{\pi}{4}$-rotation	95.2%(93.5...97.0)	98%(97.1...98.9)	53

3.5. Further Results

Following paper B.2, I tested two larger dataset named *Arcene* and *Ad* from the UCI Machine Learning Repository with the R-package described in paper B.3. The evaluation using logistic regression gave poor ROC curve results (figures A.14 and A.15), which indicates that it was not a linear problem. The evaluation was conducted with random forest as the prediction model. The ROC curves for all of the eight smaller datasets are shown in figures A.16 to A.23. Meanwhile, the random forest evaluation test is also incorporated in the EFS R-package. Table 3.3 shows all results for the ROC evaluation test based on an RF model using all features and another model using only above-average-EFS features. The accuracy of the datasets *Fibrosis*, *Sonar*, and *Arcene* improved significantly with a significance level of 5%.

The difference in prediction performance between a model with all features and a model with only the above-average-EFS features was also tested using a ROC test, which was constructed on a support vector machine (SVM) prediction. The results are shown in table 3.4. We can observe that for the *MI*, *Fibrosis*, and *Arcene* datasets, the improvement in accuracy was significant, with a significance level of 5%. The kernels were selected based on the best AUCs. The following SVM kernels were tested on all datasets: radial basis, linear, polynomial, and Bessel.

Taking the accuracy analysis with logistic regression from paper B.2 (Table 3) into account, nearly every dataset showed an improvement in prediction performance de-

Table 3.3.: **Random Forest AUC evaluation.** Comparison of ROC curves from RF model with all features and RF model with features with importance values over the mean as calculated by EFS.

Dataset	p-value of ROC-test
MI-Mortality	0.17
Fibrosis	0.037
FLIP	0.491
SPECTF	0.456
Sonar	0.018
WBC	0.125
Arcene	0.005
Ad	0.250

Table 3.4.: **Support vector machine AUC evaluation.** Comparison of ROC curves from SVM model with all features and SVM model with features with importance values over the mean as calculated by EFS.

Dataset	kernel	p-value of ROC-test
MI-Mortality	Bessel	<0.001
Fibrosis	radial basis	0.039
FLIP	Bessel	0.098
SPECTF	linear	0.106
Sonar	radial basis	0.481
WBC	radial basis	0.123
Arcene	linear	0.005
Ad	polynomial	0.279

pending on the underlying model, except for *FLIP* and *Ad*.

In paper B.4, I studied possible subset selection tools. After testing the $\frac{-\pi}{4}$-rotation method on the two large datasets *Arcene* and *Ad*, I also tested it on the smaller datasets from paper B.2. Table 3.5 shows the resulting p-values for the roc-tests with logistic regression and RF as the underlying models after DeLong [36].

Table 3.5.: **Subset selection criteria.** Comparison between importance mean and minimum of importances of $\frac{-\pi}{4}$-rotation method.

Dataset	mean threshold	$\frac{-\pi}{4}$ threshold	p (LR)	p (RF)
MI-Mortality	5	8	0.881	0.179
Fibrosis	7	6	0.486	0.598
FLIP	5	5	1	0.759
SPECTF	19	12	0.8	0.335
Sonar	24	10	0.556	0.006 (mean is better)
WBC	10	7	0.949	0.427
Arcene	5038	69	0.004	0.012
Ad	613	15	<0.001	0.065 (mean is better)

4. Discussion

4.1. Summary

An ensemble of feature selection methods aims to improve subsequent data analysis and model construction through ranking and selecting suitable subsets of features, thereby reducing dimensionality and simplifying the model. Therefore, the objective of this thesis is to create an ensemble of FS methods which outperforms each single approach with respect to stability and accuracy. Dietterich [9] stated in 2000 that this assertion is commonly valid for machine learning approaches. For FS methods, the statement was proven by Saeys [8]. The novelty of EFS is the concrete model consisting of eight quantitative FS methods with different strengths and weaknesses, with these being combined in a cumulative feature ranking. The aim was to increase diversity in the base selectors and stability in the ranking of features. In addition, it contains several evaluation methods and is available as both an R-package and a web-application.

The demand for such an approach emerged following analysis of the cardiological dataset from paper B.1. A binary classification via random forest was conducted and the feature importance values were calculated using the Gini index. This measure showed a strong discrimination against binary features in contrast to continuous ones, a typical behavior confirmed by former studies [40, 41]. Thus, we aimed to find a better method for biomarker discovery. Medical practitioners require easy to handle and comprehensible techniques to be able to identify important features from a large quantity of features by ranking them in order of relevance. The need for such techniques does not only appear in precision medicine but rather in all fields which work with high-dimensional data.

Different evaluations showed that the EFS approach could more or less accomplish the objectives of my study on various datasets. However, there also appeared to be some limitations on the chosen methods.

4.2. Discussion of Methods

In this thesis, the chosen methods can be divided into three different categories: base selectors which are implemented in an ensemble, techniques to distinguish between relevant and redundant features as well as to define a subset of selected features, and the approaches used for evaluation.

In the following section, I will first discuss the method of ensemble learning, then the choice of base selectors, followed by the subset selection techniques, and finally the evaluation methods.

4.2.1. Ensemble method

In recent decades, ensemble learning methods have drawn increased attention in the machine learning field. They are considered to enhance the robustness [8] of the selected feature subsets. Robust feature selection allow the domain experts to have more confidence in their results and their interpretations. Previous studies have shown that different feature subsets may yield equal results [5]. Therefore, an ensemble of feature selection techniques may help with choosing a more stable subset, which is a combination of the proposals of several different FS methods. There are two kind of ensemble learning methods, namely homogeneous and heterogeneous methods. Two well-known homogeneous ensemble approaches are boosting [44] and bagging [25], which have shown higher accuracy when using decision trees [45] and neural networks [46] as classification algorithms. The idea of both methods is to obtain different training sets for each classifier through a resampling procedure. However, in the case of feature selection, the ensemble of classifiers is used to vote for features rather than class labels. An example of an ensemble of FS method is the SVM ensemble by Kim et al. [47], which is an ensemble of support vector machines (SVMs). Ensembles of FS are applied to biomedical data to detect biomarkers [48]and to conduct text mining tasks [49]. EFS does not resample the same method in random training set variations. As a heterogeneous measure, it contains eight different FS methods for binary classifications, with each differing in their bias to error-proneness. In addition, the implementations of EFS give the users the ability to choose which FS methods they want to use.

There are different possibilities for combining the results of the base selectors using so called aggregators. Seijo-Pardo et al. [12], for example, introduced the SVM-Rank. Also, voting systems, where each base selector has one vote, are possible aggregators. I decided to preserve the metric scale of the single rankings and add it up to a cumulative

ranking. Thus, all FS methods have the same weighting in the ranking process.

4.2.2. Base Selectors

Nowadays, a big variety of FS methods exist with many different variations [4, 5, 50]. They have been developed to handle different tasks, with some being widely applicable and others only having special applications. In this study, eight different FS methods for binary classification were used. In the following section, I will discuss these methods and explain why they were chosen.

The median FS method involves the Mann-Whitney-U-test, for which homogeneity of variance is a precondition. If this criteria has not been met, the underlying test should be the Brunner-Munzl test[51]. However, a previous study has shown that the differences in results are not significant, but only slightly better for equal sample sizes. For varying sample sizes, the Brunner-Munzel test performs better [52]. Therefore, a preceding subsampling would be appropriate. In practice, the test data should be pre-processed to satisfy variance homogeneity. MWU only tests the equality of medians if the distribution of the two classes are of similar shape and equal scales. This is called the pure shift model, which does not appear very often in real datasets. Several studies address this topic and recommend only using the MWU test if variances are equal [53, 54], or at least, their ratio should not exceed 1.5 [55].

Correlations, of which there are different variations, are fast filter methods created to detect causal dependencies. I implemented a version of the correlation coefficients called fast-correlation-based filter (FCBF) after Yu and Liu [7]. This measure prevents multicollinearity by removing those features where there is a high correlation between them. This might be criticized as practitioners are interested in all features which are relevant to the class variable when ranking features. However, multicollinearity can lead to an overestimation of the accuracy of subsequent prediction models. In this work, the ranking functions is a preprocessing step for binary classification and therefore multicollinearity can be avoided.

Logistic regressions are widely used to detect the relationships between variables and a binary class variable. The weights of the regression are embedded importance measures [56]. They reflect the relevance of features for the prediction performance and thus allow us to neglect the features with very small weights. LRs have the disadvantage that they are only able to identify linear relationships between features.

Originally, the Gini-index was developed to describe the dispersion of wealth or income of populations. Therefore, it was intended to be a measure of inequality of metric values.

Previous articles have described its bias with categorical variables [57, 58]. They detected a downward-error for small sample sizes as well as a bias for binary features or features with less categories. The reason for including the Gini-index in the EFS approach is that in predictive medicine, it is still the state of the art in variable importance measures for random forests [41].

It may be possible to criticize four of the eight FS methods in my EFS approach, which are embedded in an RF algorithm. Due to my research on random forest prediction, I recognized the differences in the results of variable importance measures from RF. The next step was comparing it with the other four standard FS methods. We noticed marked differences in the importance values, cf. figures A.24 to A.29. Interestingly, the two measures from the conditional forest did not show big differences in the feature ranking. Janitza et al. [29] state that the AUC-based measure only performs better than the error-rate-based one for unbalanced datasets, i.e., for datasets with different class sizes. In our case, we prevent unbalanced data by sub-sampling.
Another limitation on the EFS method is its slowness and computational costs. The aim of ensemble learning approaches is to reduce computational time by combining multiple weak classifiers to obtain a reliable classifier which outperforms each single one. The enhancement should affect accuracy as well as computational velocity. Due to both FS measures embedded in the conditional random forest, the EFS became computationally expensive. For future studies, there is a potential for optimization.

4.2.3. Subset Selection

There are two different ways to determine the importance of features [7] and therefore a subset of relevant features. On the one hand there is the individual evaluation, which returns a feature ranking by assigning an importance value to each feature individually. Subset evaluation, on the other hand, is an iterative process in which successive feature subsets are tested according to an optimality criterion until the optimal subset is found. The latter has the advantage of being able to of remove redundant features. However, searching through all possible feature subsets is computationally inefficient and thus not suitable for high-dimensional data. Thus, I implemented an individual evaluation method into EFS. The cumulative ranking of all base selectors is used for the underlying importance values. After getting a feature ranking, the next step is to find a decision procedure on the relevance of the features. Traditional thresholds are fixed percentages of features, e.g., the best 10% or 25%. When choosing such a percentage, the number of features for each individual dataset and the distribution of importance

values play an important role. Another straightforward threshold criterion is to take the mean of the importance values, as was done in papers B.2 and B.3. For datasets with only a few features, the method worked well. However, in bigger datasets the mean importance value could be very low due to the presence of many irrelevant features. Therefore, the subsets of features are too big for a precise prediction. In paper B.4, a novel approach is introduced which improves the subset selection for importance values ascending exponentially, namely the $\frac{-\pi}{4}$-rotation method. This method searches for the global minimum after rotating $\frac{-\pi}{4}$. Therefore, it detects the point where the difference in importance values of the increasingly ordered features exceeds 45 degrees. The disadvantage of this method is that it only works for features with exponentially growing importance. If the importance values form a logarithmic curve, the mean measure would be more appropriate.

4.2.4. Evaluation Methods

In papers B.2 and B.3, the evaluation of prediction accuracy was performed by comparing the ROC curves of an underlying logistic regression with a leave-one-out cross-validation (LOOCV). Afterwards, I also evaluated a random forest model and support vector machines (SVMs) with the ROC curves to detect non-linear relations in the datasets.

Saeys et al. [8] also used the Jaccard-index to evaluate the stability of subset selection. They suggested using a Spearman rank correlation to check the similarity in feature rankings and a Pearson correlation for the feature weighting and more specifically for the importance values. In addition to that, I implemented a permutation test to check if the retrieved AUCs are better than making a guess. Furthermore, the variances of importances are calculated after bootstrapping the class variable.

All evaluation tools are intended for the users to be able to examine the stability of the feature selection and the accuracy of a model constructed using EFS feature subsets.

4.3. Discussion of Results

Although the underlying model has an effect on the ROC curves, nearly each dataset showed enhanced performance when models with all features were compared with ones with only the above-average-EFS features, except in the case of *FLIP* and *Ad*. In the case of *Ad*, the prediction accuracy is very high when all features are taken in both the RF (AUC: 98.8% with confidence interval $[100\%, 100\%]$) and the SVM (AUC: 96.4%

with confidence interval $[95.1\%, 97.8\%]$) models. Therefore, a significant improvement is nearly impossible. The *FLIP* dataset, on the other hand, has a very low prediction performance for all the tested models. Thus, the dataset is assumed to be inappropriate for making good predictions, even when the subset of features is optimized.

In general, big datasets include many features which can be neglected as predictors [59]. Thus, the need for better cut-off algorithms to distinguish the relevant features from the irrelevant features emerges. The results of EFS from six datasets are discussed in paper B.2. Subsequently, I tested the EFS on two bigger datasets, namely *Ad* and *Arcene*. The EFS output provides a ranking of features. In both datasets, the curve of the calculated importance values from EFS is similar to an exponential function. By taking the mean of importance as a cut-off point for relevance, figures A.14 and A.15 show no significant enhancement of the AUC values compared to constructing the model with all features. In the case of *Ad*, the performance is significantly worse. Here, the underlying prediction model is a logistic regression. The means of importance values of *Arcene* and *Ad* are in both cases 0.12 and the maximum importance values are 0.84 and 0.96 respectively. Figures A.8 and A.9 show that in both cases, there are many features with little importance, which reduce the mean value. Provided there is a reliable feature ranking approach, a criteria for subset selection should be able to detect the elbow of the curve of importance values. The novel $\frac{-\pi}{4}$-procedure detects the point where the slope of the importance values curve exceeds 45 degrees. By applying this threshold to the datasets *Ad* and *Arcene*, figures A.10 and A.11 show an increase in accuracy. The underlying model for the ROC curves was again a logistic regression. However, conducting the same analysis with random forest as the underlying model does not show significant difference between the mean and the $\frac{-\pi}{4}$-cutoff for *Ad* (cf. figure A.13). The reason for this is the same as mentioned above: the AUC of the features with importance above the mean is already very high (98.2% with confidence interval $[100\%, 100\%]$). On the opposite, figure A.12 shows a significant enhancement in performance with a p-value of 0.012. As can be seen, two datasets are not sufficient to make a valid statement. Therefore, more high-dimensional datasets will have to be tested.

4.4. Future Prospects and Conclusion

I have already mentioned that some future work is needed to improve the speed of the EFS algorithm. The computational time could be reduced by appropriate parallelization. Another approach would be to replace the two methods implemented in the

conditional random forest with weaker and therefore faster FS methods. For this, future investigations on a variety of FS method are needed. The EFS implementations are not able to recognize the form of the importance values curve. Therefore, it cannot decide which subset selection criteria to choose: either the mean or the $\frac{-\pi}{4}$-measure, meaning that the user has to make a manual decision. There is still a lot to be done on this in future to obtain an integrated subset selection. The results show some enhancement in prediction performance with the $\frac{-\pi}{4}$-cut-off. More high-dimensional datasets would have to be tested to validate the approach.
So far, EFS is only applicable for binary classifications. Extensions to general classification or regressions would require more effort but could be feasible.

In conclusion, it can be said that the EFS method brought the expected reduction in dimension in the form of a quantitative ranking of features. With the $\frac{-\pi}{4}$ threshold a promising method has been developed for finding a threshold between relevant and redundant features. The reliability of the EFS feature ranking is very high. Although the results could not always improve the prediction performance, a significant improvement could be shown in most cases.

Bibliography

[1] AB Brahim and M Limam. Ensemble feature selection for high dimensional data: a new method and a comparative study. *Advances in Data Analysis and Classification*, pages 1–16, 2017.

[2] K Kira and LA Rendell. The feature selection problem: Traditional methods and a new algorithm. In *AAAI*, volume 2, pages 129–134, 1992.

[3] MA Hall and G Holmes. Benchmarking attribute selection techniques for discrete class data mining. *IEEE Transactions on Knowledge and Data Engineering*, 15(6):1437–1447, 2003.

[4] I Guyon and A Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[5] Y Saeys, I Inza, and P Larrañaga. A review of feature selection techniques in bioinformatics. *Springer*, 23(19):2507–2517, 2007.

[6] ZM Hira and DF Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, 2015, 2015.

[7] L Yu and H Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–124, 2004.

[8] Y Saeys, T Abeel, and Y Van de Peer. *Robust feature selection using ensemble feature selection techniques*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[9] TG Dietterich. Ensemble methods in machine learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, MCS '00, pages 1–15. Springer-Verlag, 2000.

[10] LI Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. John Wiley & Sons, 2004.

[11] YH Yang, Y Xiao, and MR Segal. Identifying differentially expressed genes from microarray experiments via statistic synthesis. *Bioinformatics*, 21(7):1084–1093, 2004.

[12] B Seijo-Pardo, I Porto-Díaz, V Bolón-Canedo, and A Alonso-Betanzos. Ensemble feature selection: Homogeneous and heterogeneous approaches. *Knowledge-Based Systems*, 118:124–139, 2017.

[13] G Chandrashekar and F Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16 – 28, 2014.

[14] L Yu and H Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pages 856–863, 2003.

[15] R Kohavi and GH John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.

[16] P Langley et al. Selection of relevant features in machine learning. In *Proceedings of the AAAI Fall symposium on relevance.*, volume 184, pages 245–271, 1994.

[17] A Hart. Mann-whitney test is not just a test of medians: differences in spread can be important. *BMJ: British Medical Journal.*, 323(7309):391, 2001.

[18] MA Hall. *Correlation-based feature selection for machine learning.* PhD thesis, Department of Computer Science, Waikato University, New Zealand., 1999.

[19] AG Karegowda, AS Manjunath, and MA Jayaram. Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, 2(2):271–277, 2010.

[20] CF Dormann, J Elith, S Bacher, C Buchmann, G Carl, G Carré, JR Marquéz, B Gruber, B Lafourcade, PJ Leitão, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1):27–46, 2013.

[21] N Suzuki, DH Olson, and EC Reilly. Developing landscape habitat models for rare amphibians with small geographic ranges: a case study of siskiyou mountains salamanders in the western usa. *Biodiversity and Conservation*, 17:2197 –2218, 2008.

[22] J Elith, CH Graham, RP Anderson, M Dudík, S Ferrier, A Guisan, RJ Hijmans, F Huettmann, JR Leathwick, A Lehmann, J Li, LG Lohmann, BA Loiselle, G Manion, C Moritz, M Nakamura, Y Nakazawa, JM Overton, A Townsend Peterson, SJ Phillips, K Richardson, R Scachetti-Pereira, RE Schapire, J Soberón, S Williams, MS Wisz, and NE Zimmermann. Novel methods improve prediction of speciesâĂŹ distributions from occurrence data. *Ecography*, 29(2):129–151, 2006.

[23] A De La Fuente, N Bing, I Hoeschele, and P Mendes. Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics*, 20(18):3565–3574, 2004.

[24] TK Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282, 1995.

[25] L Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

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

[27] T Hothorn, K Hornik, and A Zeileis. party: A laboratory for recursive part(y)itioning, 2006.

[28] C Strobl, T Hothorn, and A Zeileis. Party on! a new, conditional variable-importance measure for random forests available in the party package. *The R Journal*, 1(2):14–17, 2009.

[29] S Janitza, C Strobl, and AL Boulesteix. An auc-based permutation variable importance measure for random forests. *BMC Bioinformatics*, 14:119, 2013.

[30] C Gini. Variabilità e mutabilità. *Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Rome: Libreria Eredi Virgilio Veschi*, 1912.

[31] C Strobl, AL Boulesteix, T Kneib, T Augustin, and A Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307, 2008.

[32] T Hothorn, K Hornik, and A Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, 2006.

[33] M Pepe. *The statistical evaluation of medical tests for classification and prediction.* USA: Oxford University Press, 2004.

[34] S Janitza, G Tutz, and AL Boulesteix. Random forests for ordinal response data: prediction and variable selection,. Technical Report Technical Report 174, Department of Statistics, University of Munich., 2014.

[35] DW Opitz and JW Shavlik. Actively searching for an effective neural network ensemble. *Connection Science*, 8:337–354, 2010.

[36] ER DeLong, DM DeLong, and DL Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

[37] P Jaccard. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[38] AJ Smola and B Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[39] GN Watson. *A treatise on the theory of Bessel functions*. Cambridge university press, 1995.

[40] M Sandri and P Zuccolotto. A bias correction algorithm for the gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, 17(3):611–628, 2008.

[41] AL Boulesteix, S Janitza, J Kruppa, and IR König. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):493–507, 2012.

[42] M Lichman. UCI machine learning repository, 2013.

[43] T Hastie, R Tibshirani, and J Friedman. *Elements of Statistical Learning*. 2009.

[44] Y Freund and RE Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[45] E Bauer and R Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, 36(1):105–139, 1999.

[46] DW Opitz and R Maclin. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11:169–198, 1999.

[47] HC Kim, S Pang, HM Je, D Kim, and SY Bang. Constructing support vector machine ensemble. *Pattern Recognition*, 36(12):2757–2767, 2003.

[48] Y Piao, M Piao, K Park, and KH Ryu. An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. *Bioinformatics*, 28(24):3306–3315, 2012.

[49] S Van Landeghem, T Abeel, Y Saeys, and Y Van de Peer. Discriminative and informative features for biomolecular text mining with ensemble feature selection. *Bioinformatics*, 26(18):i554–i560, 2010.

[50] V Bolón-Canedo, N Sánchez-Maroño, and A Alonso-Betanzos. A review of feature selection methods on synthetic data. *Knowledge and information systems*, 34(3):483–519, 2013.

[51] E Brunner and U Munzel. The nonparametric behrens-fisher problem: asymptotic theory and a small-sample approximation. *Biometrical Journal*, 42:17–25, 2000.

[52] MW Fagerland and L Sandvik. The wilcoxon–mann–whitney test under scrutiny. *Statistics in medicine*, 10(10):1487–1497, 2009.

[53] BK Moser, GR Stevens, and CL Watts. The two-sample t test versus satterthwaite's approximate f test. *Communications in Statistics-Theory and Methods*, 18(11):3963–3975, 1989.

[54] Penfield DA. Choosing a two-sample location test. *Journal of Experimental Education*, 62(4):343–360, 1994.

[55] Zimmerman DW. Failure of the mann-whitney test: a note on the simulation study of gibbons and chakraborti. *Journal of Experimental Education*, 60(4):359–364, 1991.

[56] S Ma and J Huang. Regularized roc method for disease classification and biomarker selection with microarray data. *Bioinformatics*, 21(24):4356–4362, 2005.

[57] G Deltas. The small-sample bias of the gini coefficient: results and implications for empirical research. *Review of economics and statistics*, 85(1):226–234, 2003.

[58] C Strobl, AL Boulesteix, A Zeileis, and T Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(25):1–21, 2007.

[59] AL Blum and P Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1-2):245–271, 1997.

A. Figures

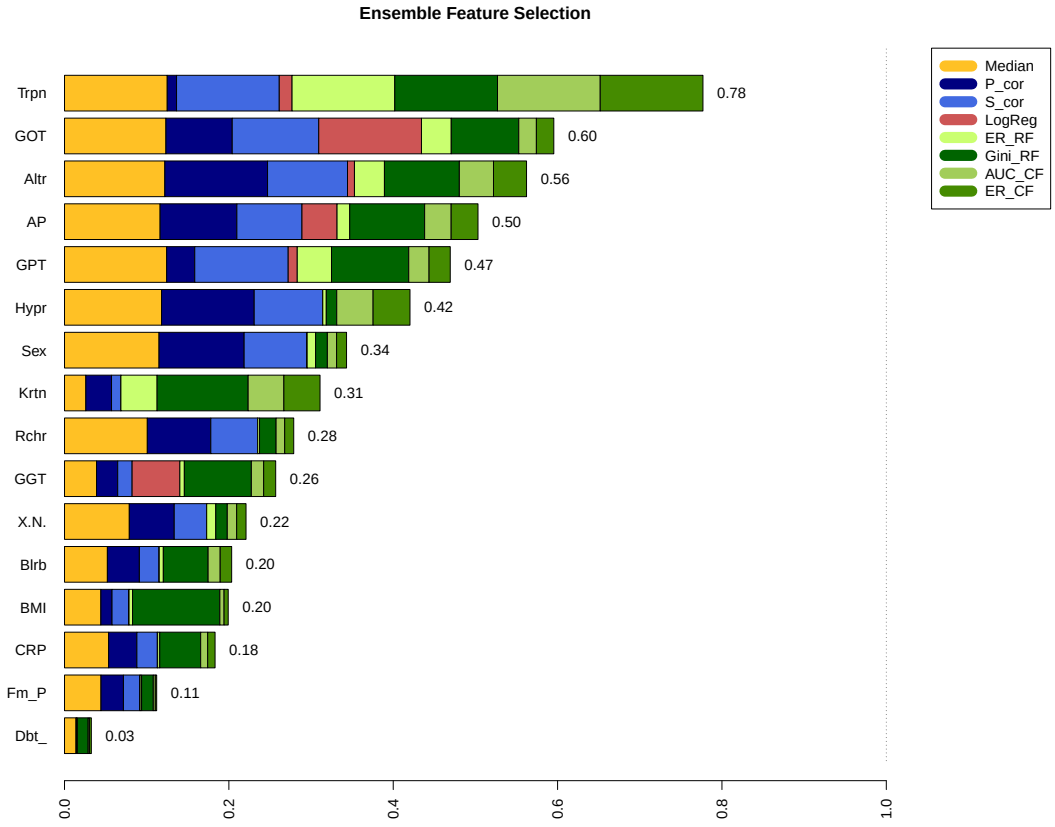

Figure A.1.: Cumulative barplot of liver parameters associated with stenosis diameter

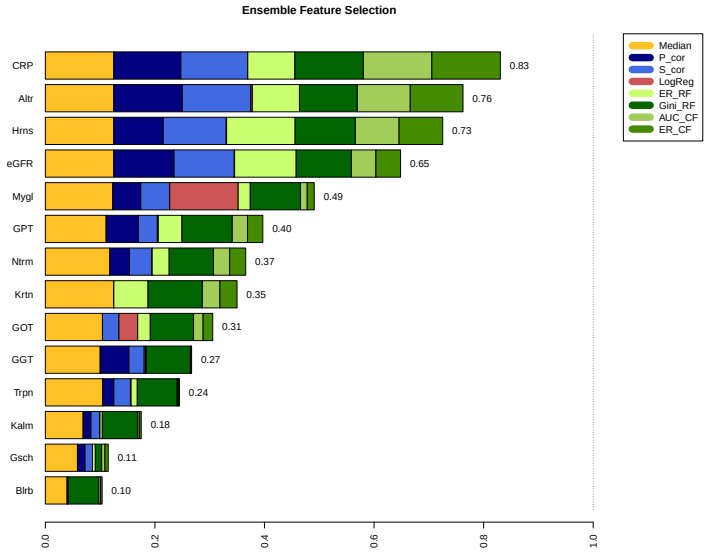

Figure A.2.: Cumulative barplots of MI-Mortality datset

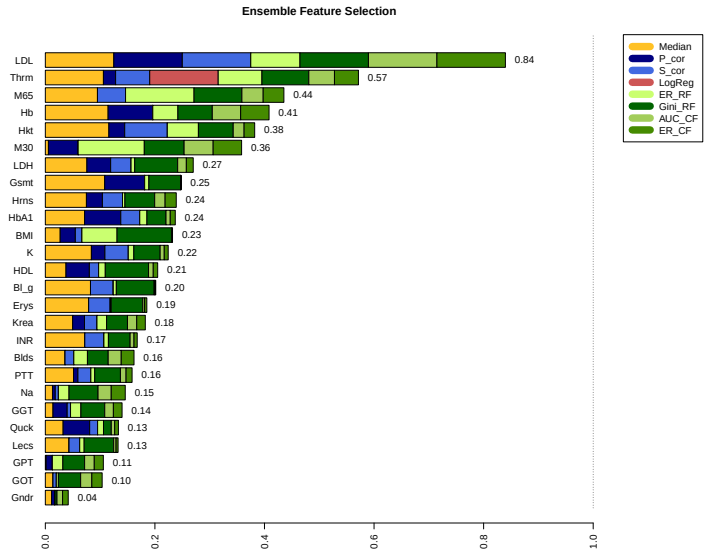

Figure A.3.: Cumulative barplots of Fibrosis datset

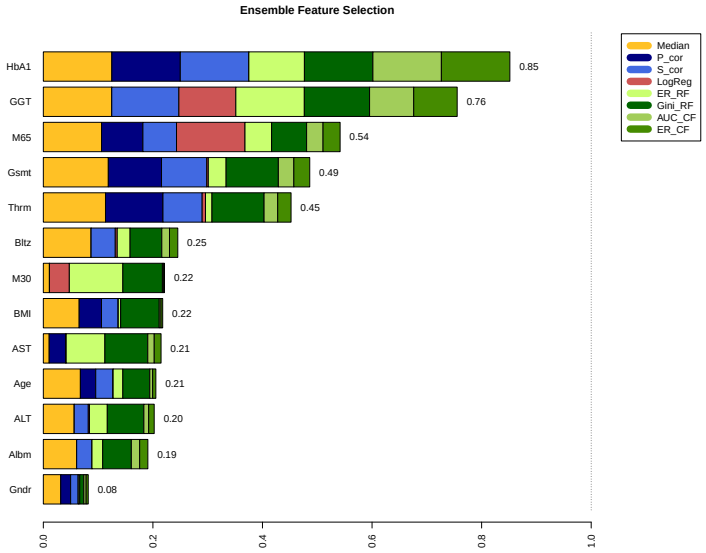

Figure A.4.: Cumulative barplots of FLIP datset

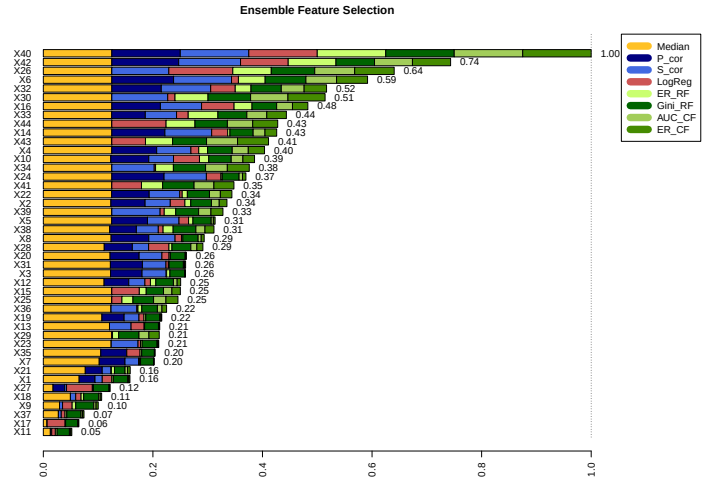

Figure A.5.: Cumulative barplots of SPECTF datset

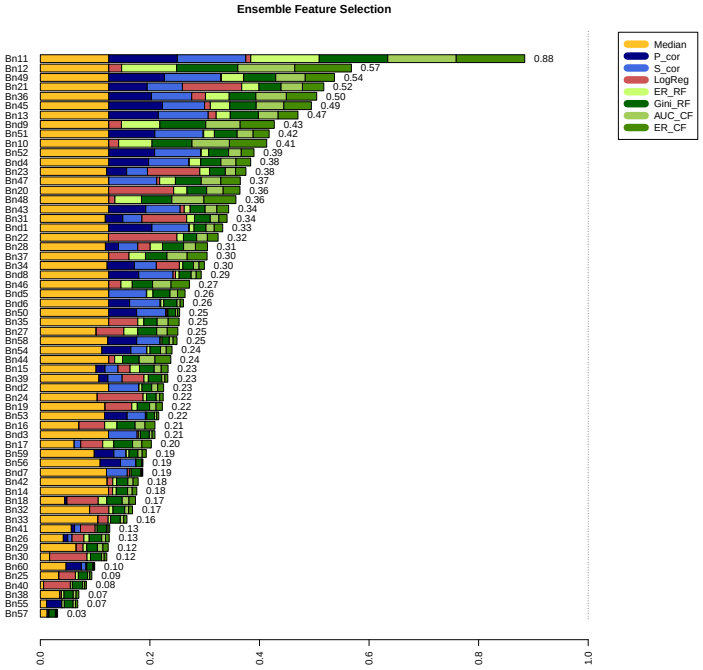

Figure A.6.: Cumulative barplots of Sonar datset

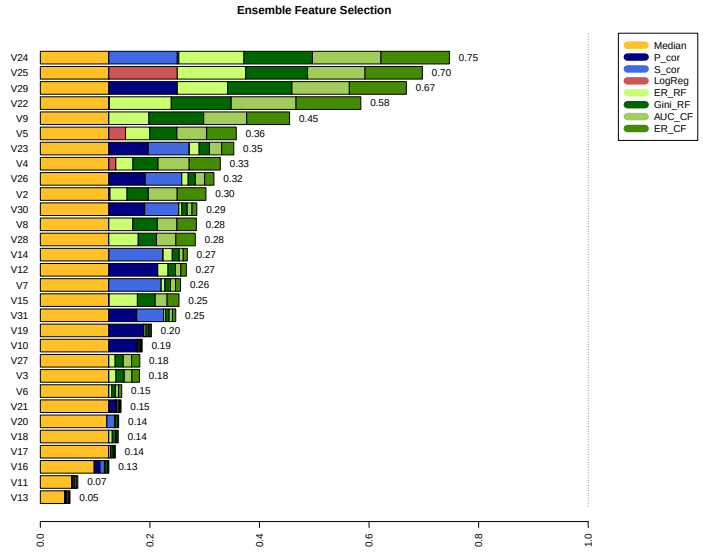

Figure A.7.: Cumulative barplots of WBC datset

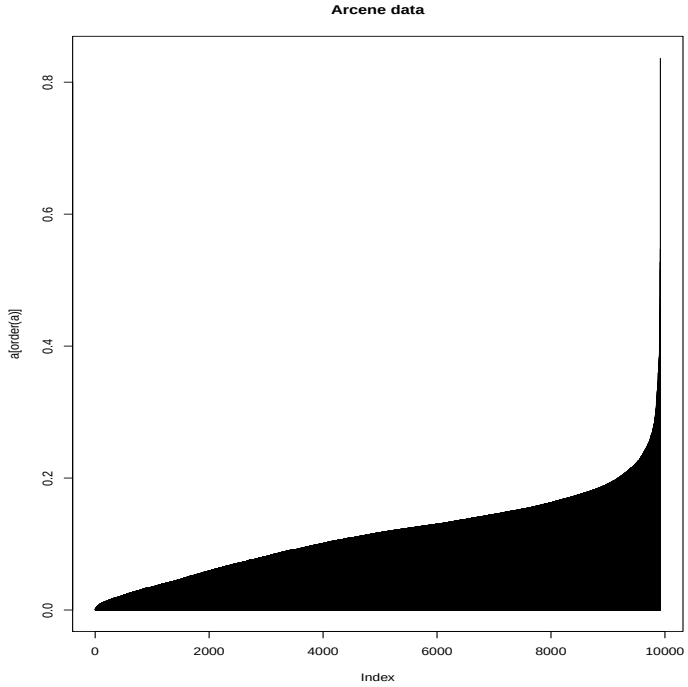

Figure A.8.: Barplot of importance values of the Arcene dataset

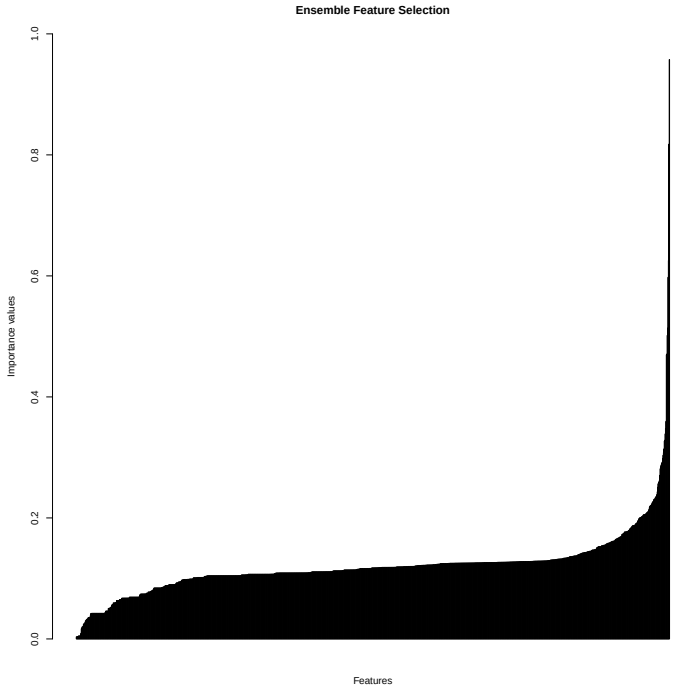

Figure A.9.: Barplot of importance values of the Ad dataset

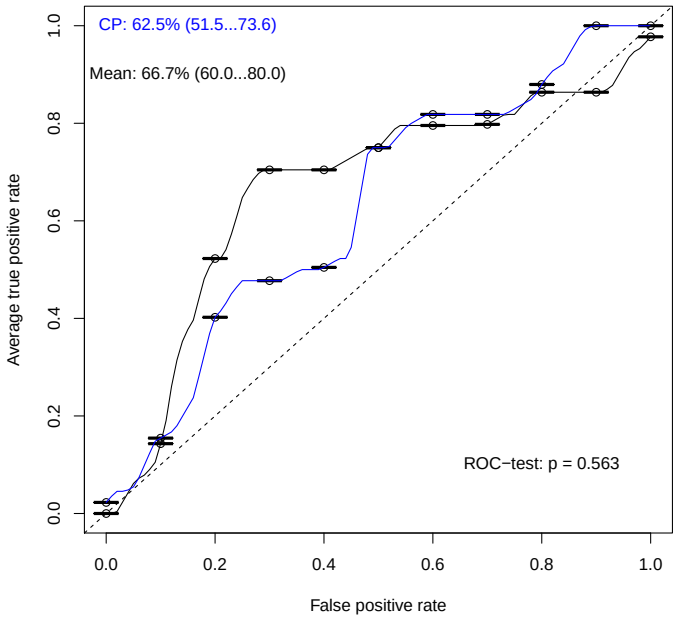

Figure A.10.: ROC curves of logistic regression models with above-average-EFS features and features over the $\frac{\pi}{4}$-cutoff of Arcene dataset.

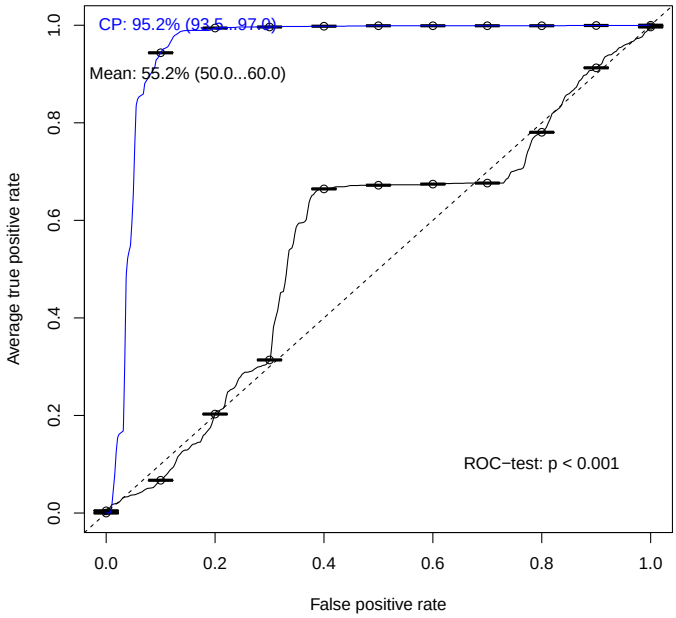

Figure A.11.: ROC curves of logistic regression models with above-average-EFS features and features over the $\frac{\pi}{4}$-cutoff of Ad dataset.

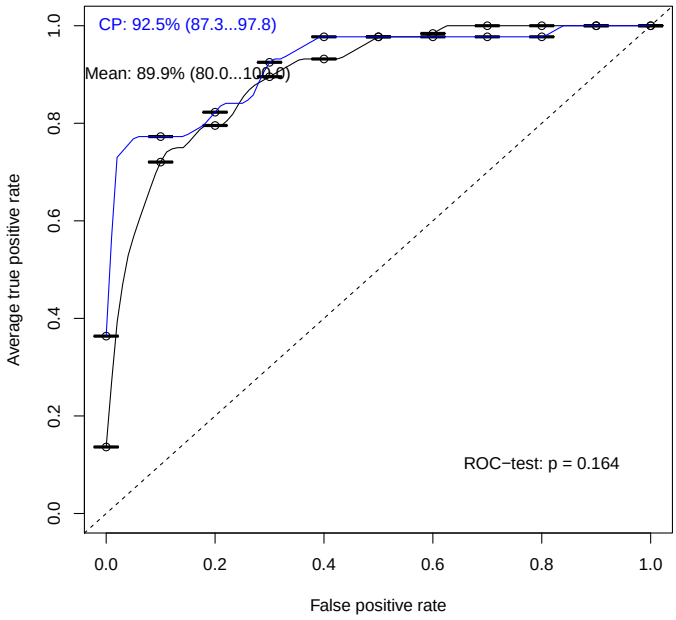

Figure A.12.: ROC curves of logistic regression models with above-average-EFS features and features over the $\frac{\pi}{4}$-cutoff of Arcene dataset.

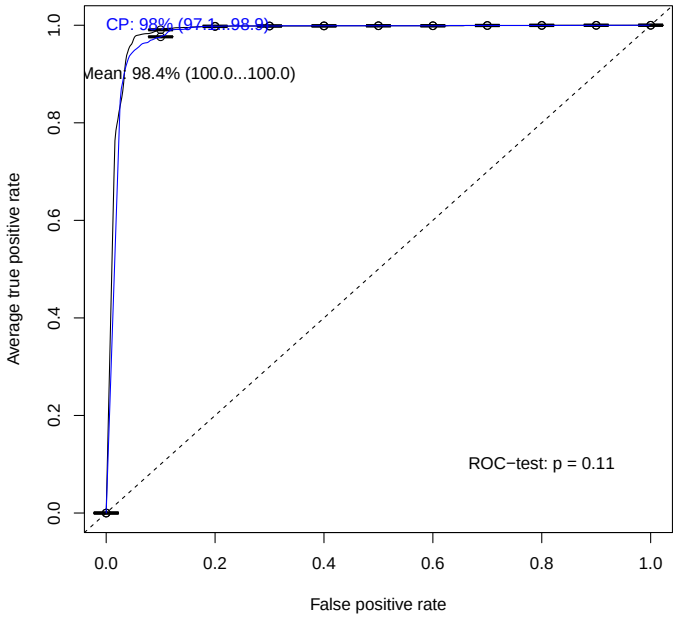

Figure A.13.: ROC curves of logistic regression models with above-average-EFS features and features over the $\frac{\pi}{4}$-cutoff of Ad dataset.

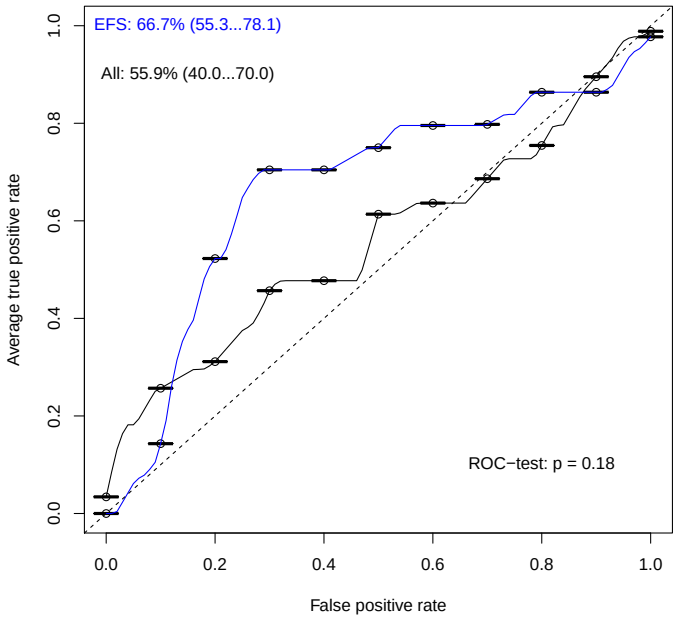

Figure A.14.: ROC curves of logistic regression models with all features and features with importance values over the mean by EFS of Arcene dataset

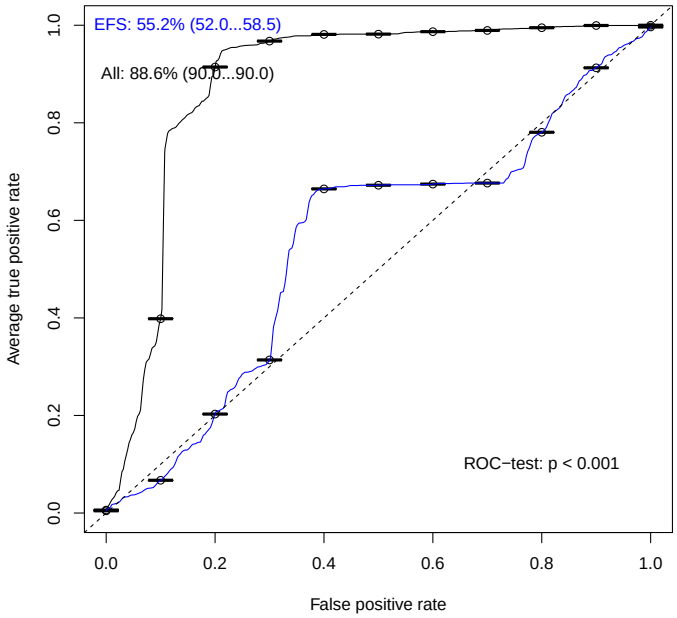

Figure A.15.: ROC curves of logistic regression models with all features and features with importance values over the mean by EFS of Ad dataset

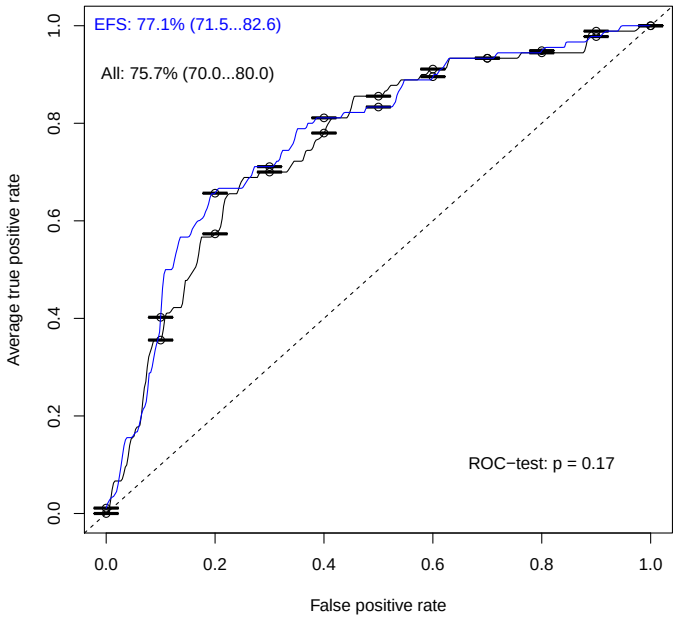

Figure A.16.: ROC curves of RF models with all features and above-average-EFS features of MI dataset.

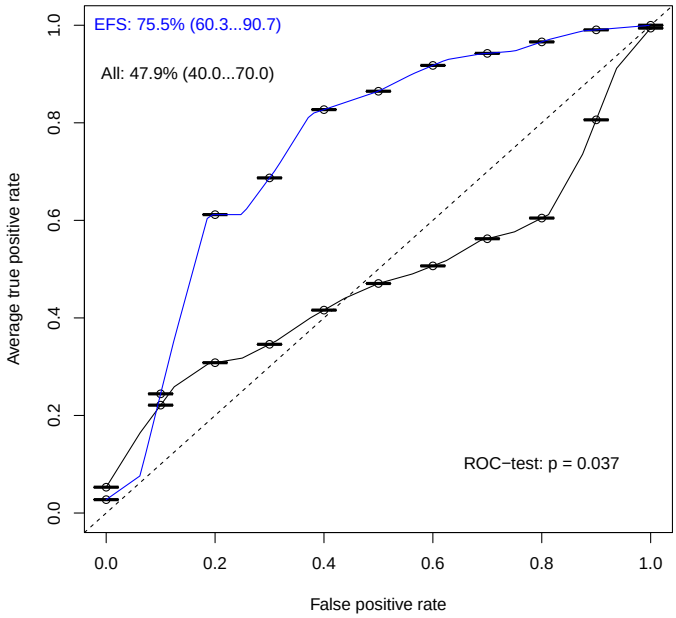

Figure A.17.: ROC curves of RF models with all features and above-average-EFS features of Fibrosis dataset.

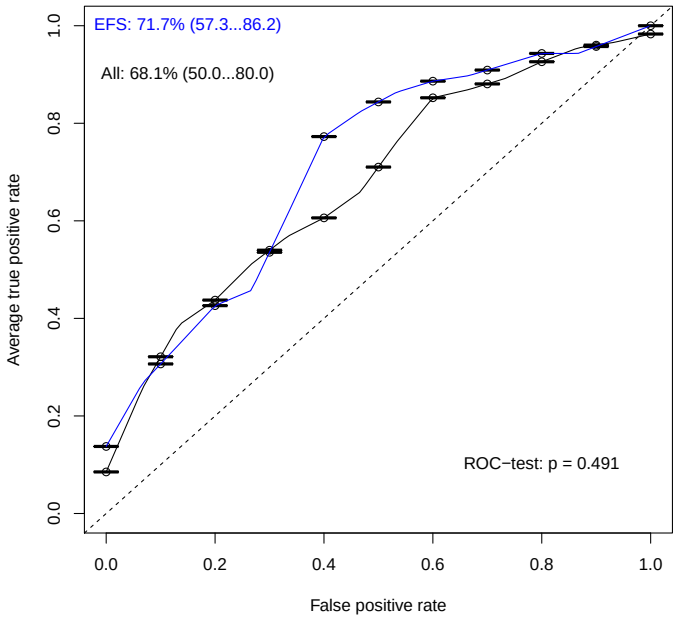

Figure A.18.: ROC curves of RF models with all features and above-average-EFS features of FLIP dataset.

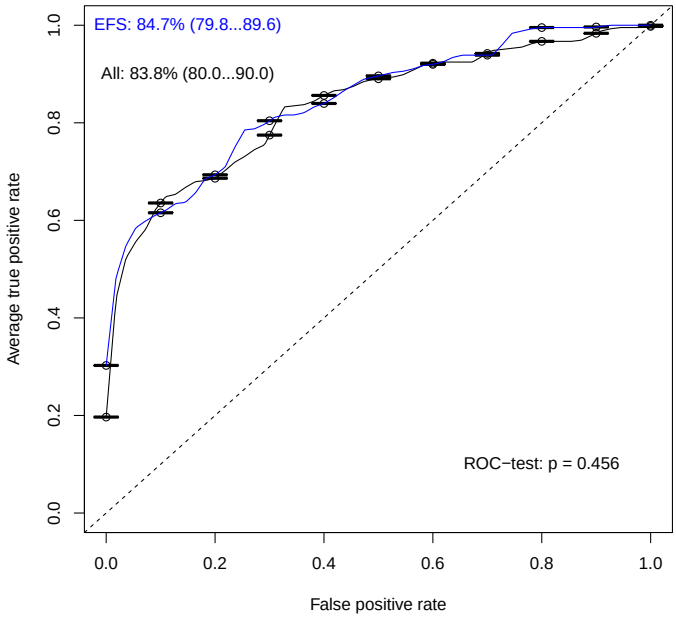

Figure A.19.: ROC curves of RF models with all features and above-average-EFS features of SPECTF dataset.

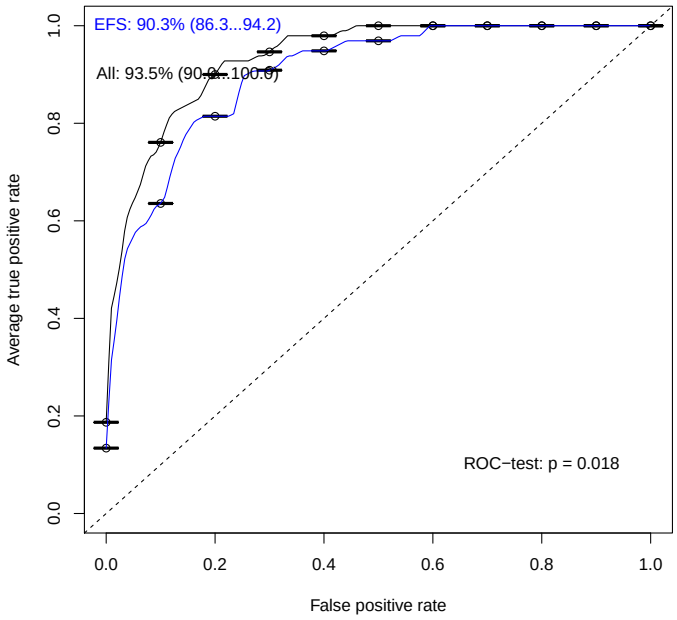

Figure A.20.: ROC curves of RF models with all features and above-average-EFS features of Sonar dataset.

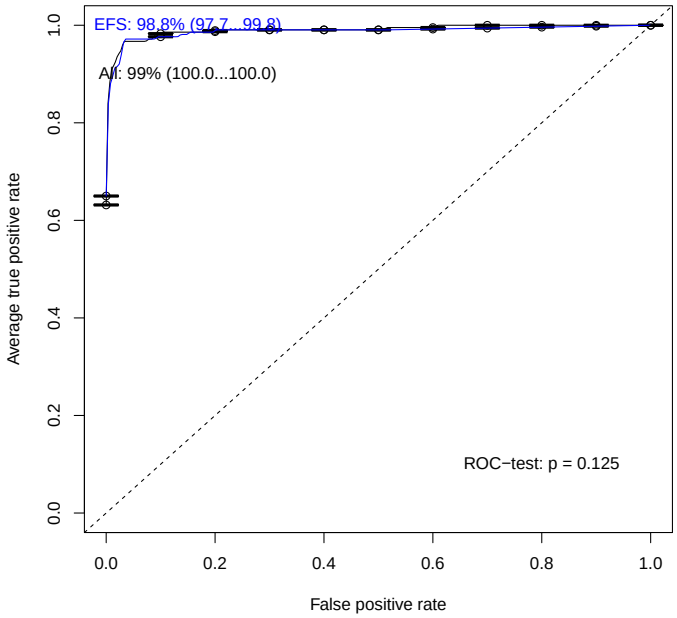

Figure A.21.: ROC curves of RF models with all features and above-average-EFS features of WBC dataset.

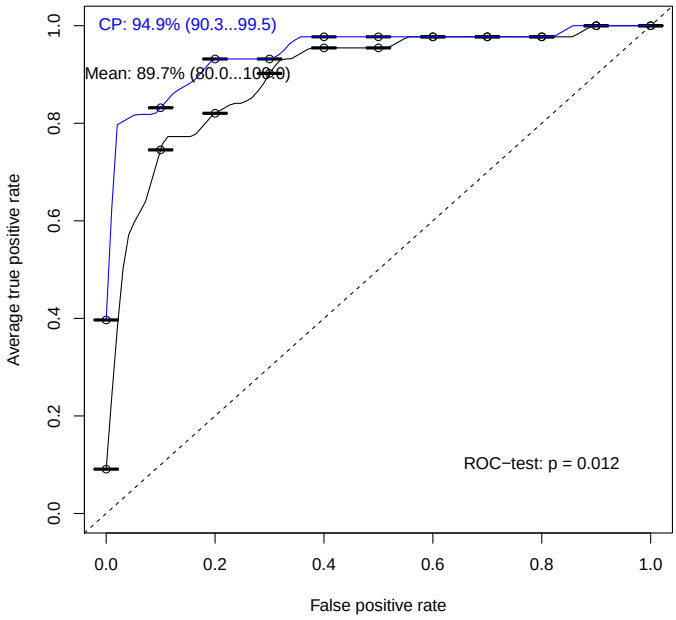

Figure A.22.: ROC curves of RF models with all features and above-average-EFS features of Arcene dataset.

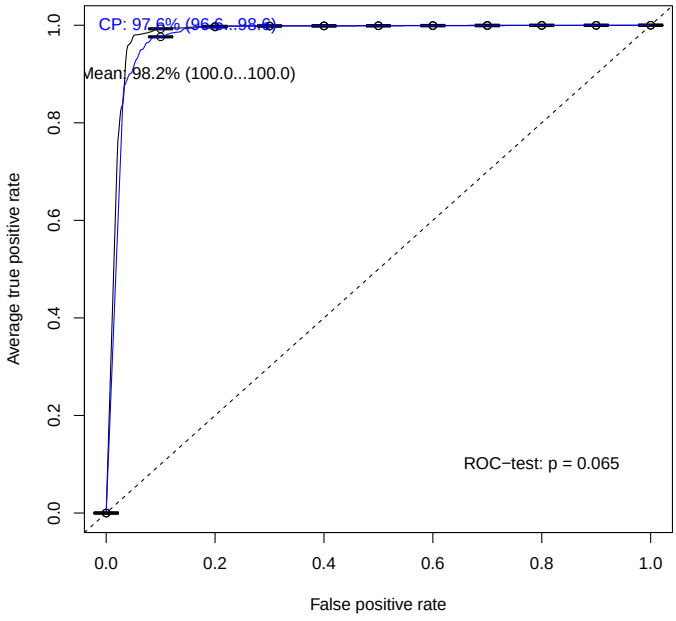

Figure A.23.: ROC curves of RF models with all features and above-average-EFS features of Ad dataset.

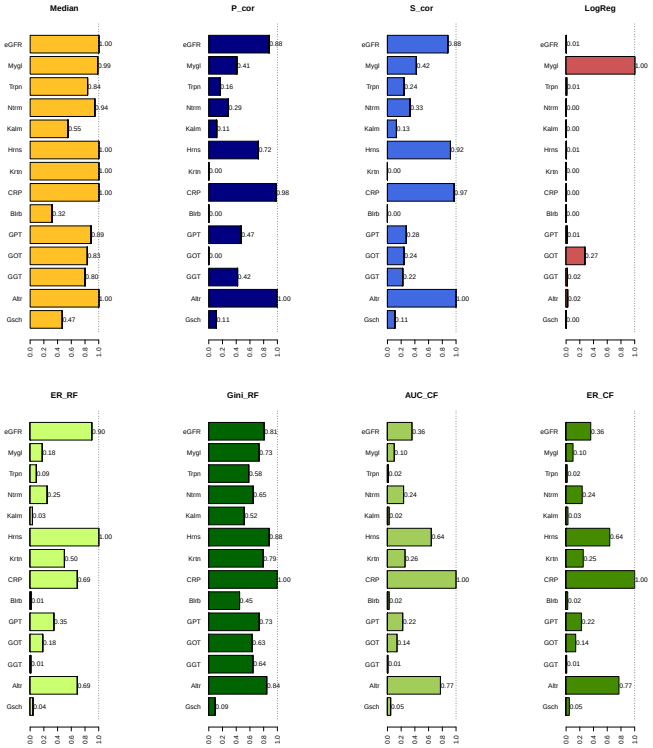

Figure A.24.: Single barplots of MI-Mortality datset

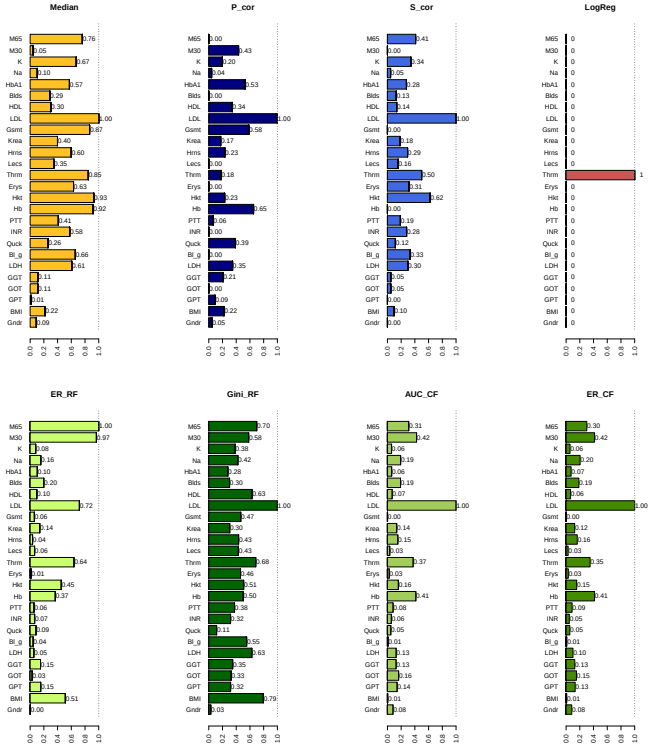

Figure A.25.: Single barplots of Fibrosis datset

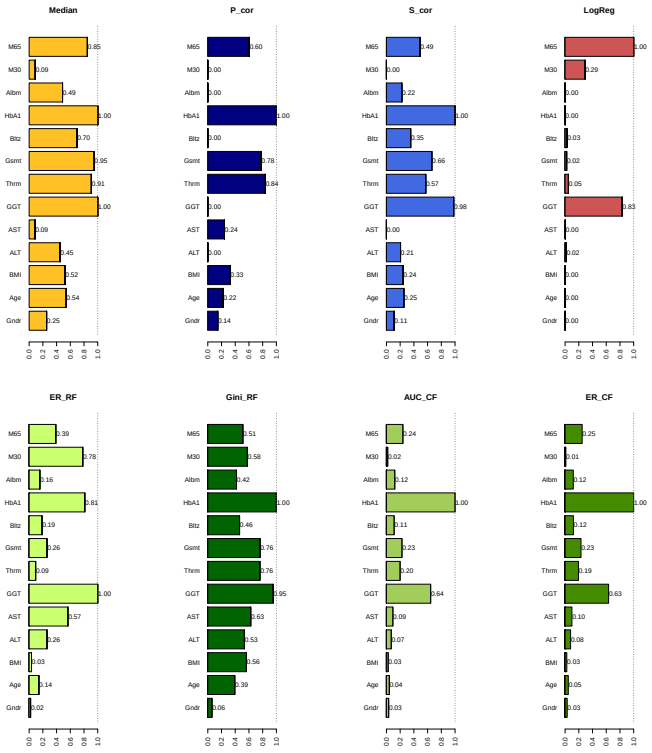

Figure A.26.: Single barplots of FLIP datset

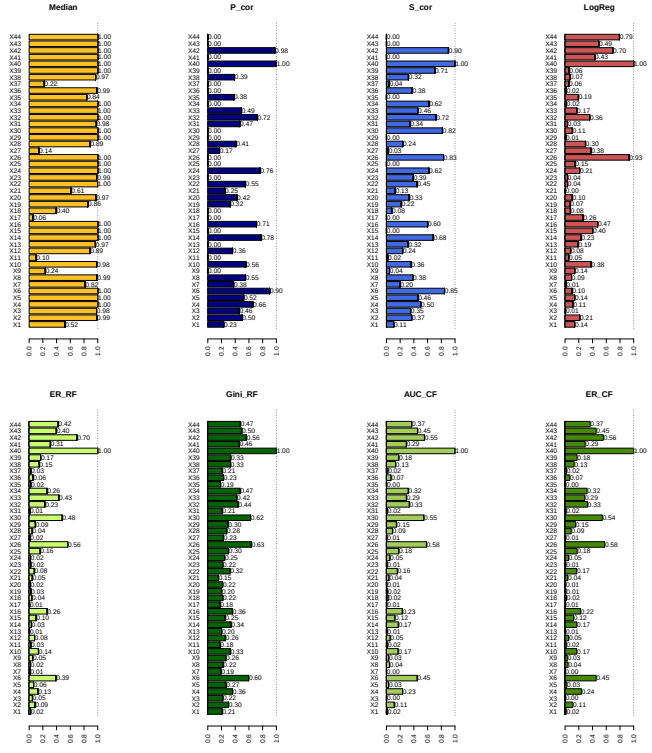

Figure A.27.: Single barplots of SPECTF datset

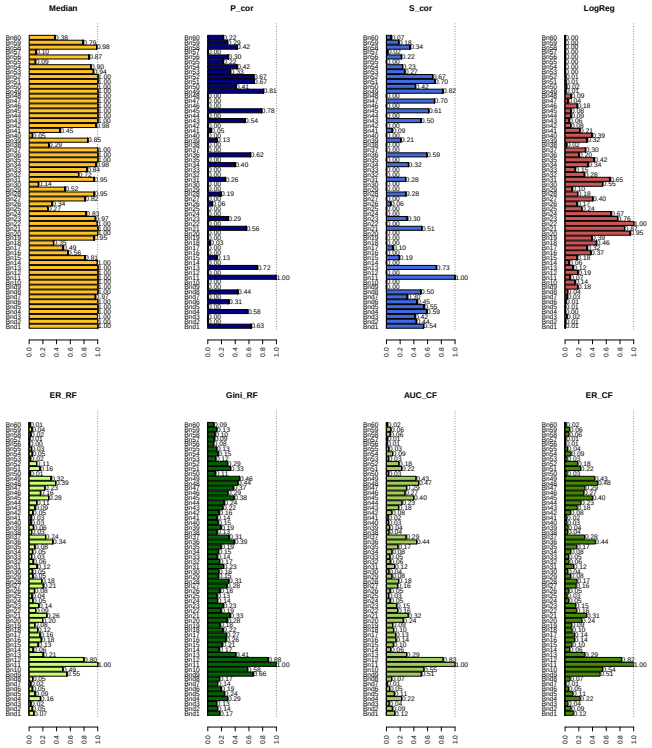

Figure A.28.: Single barplots of Sonar datset

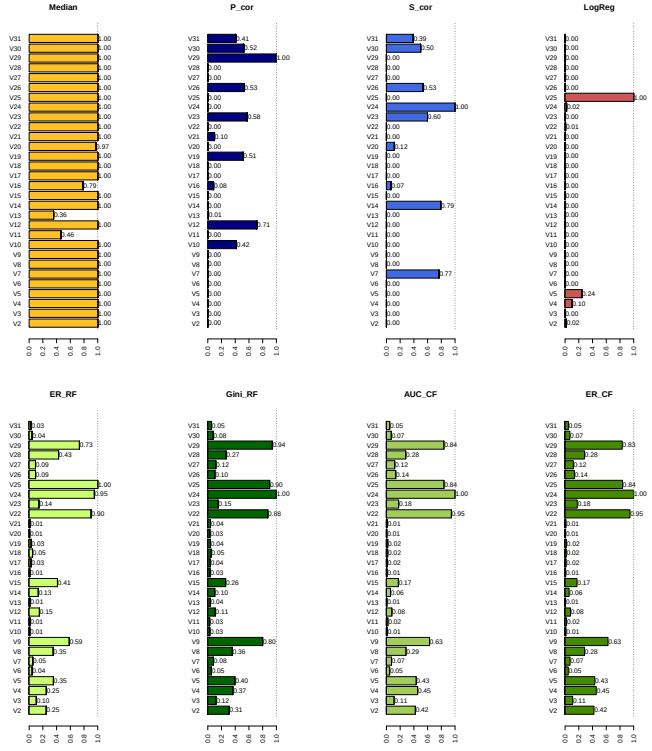

Figure A.29.: Single barplots of WBC datset

B. Papers

B.1. Paper I

Medicine®

OBSERVATIONAL STUDY

OPEN

In Acute Myocardial Infarction Liver Parameters Are Associated With Stenosis Diameter

Theodor Baars, MD, Ursula Neumann, MSc, Mona Jinawy, cand.med., Stefanie Hendricks, cand.med., Jan-Peter Sowa, PhD, Julia Kälsch, MD, Mona Riemenschneider, PhD, Guido Gerken, MD, Raimund Erbel, MD, Dominik Heider, PhD, and Ali Canbay, MD

Abstract: Detection of high-risk subjects in acute myocardial infarction (AMI) by noninvasive means would reduce the need for intracardiac catheterization and associated complications. Liver enzymes are associated with cardiovascular disease risk. A potential predictive value for liver serum markers for the severity of stenosis in AMI was analyzed.

Patients with AMI undergoing percutaneous coronary intervention (PCI; n = 437) were retrospectively evaluated. Minimal lumen diameter (MLD) and percent stenosis diameter (SD) were determined from quantitative coronary angiography. Patients were classified according to the severity of stenosis (SD ≥ 50%, n = 357; SD < 50%, n = 80). Routine heart and liver parameters were associated with SD using random forests (RF). A prediction model (M10) was developed based on parameter importance analysis in RF.

Age, alkaline phosphatase (AP), aspartate aminotransferase (AST), and MLD differed significantly between SD ≥ 50 and SD < 50. Age, AST, alanine aminotransferase (ALT), and troponin correlated significantly with SD, whereas MLD correlated inversely with SD. M10 (age, BMI, AP, AST, ALT, gamma-glutamyltransferase, creatinine, troponin) reached an AUC of 69.7% (CI 63.8–75.5%, $P < 0.0001$).

Routine liver parameters are associated with SD in AMI. A small set of noninvasively determined parameters can identify SD in AMI, and might avoid unnecessary coronary angiography in patients with low risk. The model can be accessed via http://stenosis.heiderlab.de.

(*Medicine* 95(6):e2807)

Abbreviations: ALT = alanine aminotransferase, AMI = acute myocardial infarction, AP = alkaline phosphatase, AST = aspartate aminotransferase, AUC = Area under the Curve, BMI = body mass index, CI = confidence interval, CRP = C-reactive protein, CVD = cardiovascular diseases, DT = decision tree, GGT = gamma-glutamyltransferase, HDL = high-density lipoprotein, LDL = low-density lipoprotein, MLD = minimal lumen diameter, NAFLD = nonalcoholic fatty liver disease, PCI = percutaneous coronary intervention, QCA = quantitative coronary analysis, RF = random forest, RLD = reference lumen diameter, ROC = receiver operation characteristics, SD = stenosis diameter, SEM = standard error of the mean.

Editor: Daniel Gotthardt.
Received: November 2, 2015; revised: January 8, 2016; accepted: January 20, 2016.
From the Department for Cardiology, West German Heart and Vascular Centre Essen, University Hospital, University Duisburg-Essen, Essen, Germany (TB, MJ, SH, RE); Department of Bioinformatics, Straubing Center of Science, University of Applied Science Weihenstephan-Triesdorf, Straubing, Germany (UN, MR, DH); and Department of Gastroenterology and Hepatology, University Hospital, University Duisburg-Essen (J-PS, JK, GG, AC), Essen, Germany.
Correspondence: Ali Canbay, Professor of Medicine, Department of Gastroenterology and Hepatology, University Hospital Essen, Hufelandstr. 55, 45122 Essen, Germany (e-mail: ali.canbay@uni-due.de).
Supplemental Digital Content is available for this article.
The authors have no conflicts of interest to declare.

ISSN: 0025-7974
DOI: 10.1097/MD.0000000000002807

INTRODUCTION

Cardiovascular diseases (CVD) and acute myocardial infarction (AMI) are responsible for about 17.5 million of worldwide deaths per year and are the leading cause of death globally.[1] The prognosis of patients surviving the AMI depends on the amount of myocardium that undergoes irreversible injury, that is, the infarct size.[2] Early reperfusion is the gold standard for therapy in AMI and the only way to reduce infarct size. However, reperfusion can induce an additional damage as reperfusion injury.[3] To identify the extent of stenosis in coronary vessels and thus the necessity for a percutaneous coronary intervention (PCI) in AMI, a cardiac catheterization has to be performed. This procedure is invasive and not without risk for the patients (eg, aortic dissection, aneurysm, arrhythmia, etc.).[4] Assessment of the stenosis severity with noninvasive means, that is, based on serum markers of cardiovascular injury, would spare patients without the need for an intervention.

Besides traditional cardiovascular risk factors, clinical studies indicated a potential link between liver disease, primarily nonalcoholic fatty liver disease (NAFLD), and CVD. NAFLD is by now accepted as hepatic manifestation of metabolic syndrome[5] and is associated with insulin resistance,[6] type 2 diabetes,[7,8] and CVD.[9,10] NAFLD patients may also have a higher prevalence of subclinical atherosclerosis, independent of the established cardiovascular risk factors.[11] To assess subclinical atherosclerosis potent noninvasive procedures are available, such as carotid intima media thickness measurement, brachial artery flow mediated dilatation, and arterial stiffness.[12,13] Using coronary imaging, such as multislice computed tomography, studies have also shown, that NAFLD was significantly related to lipid core[14] calcified plaques.[15–17] Apart from this, elevation of common markers of liver injury [gamma-glutamyltransferase (GGT), alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (AP), and bilirubin] are associated with the risk of CVD.[18–21] However, it is currently unknown, if liver enzyme concentrations are associated with the severity of a stenosis in AMI.

In the present retrospective study, we have focused on patients with AMI undergoing PCI and aimed to predict the necessity for a PCI with minimally invasive measures. Due to the emergency situation of AMI, quantitative coronary analysis (QCA) remained the only coronary imaging method to assess the severity of the stenosis. To evaluate the use of possible

noninvasive predictors in AMI we compared patients with AMI undergoing PCI with a percent stenosis diameter (SD) $\geq$ 50% and with SD $<$ 50% using QCA. Correlations were calculated for demographic data and classic serum parameters for heart and liver diseases with the SD using a machine learning approach (random forests). Based on those (noninvasive) parameters which were identified as most important by the random forests, a prediction model was developed to identify high-risk subjects in need for PCI.

MATERIALS AND METHODS

Ethics Statement

The study protocol conformed to the ethical guidelines of 1975 Declaration of Helsinki and was approved by the Institutional Review Board (Ethik-Kommission der Medizinischen Fakultät der Universität Duisburg-Essen; Germany; reference number 15-6356-BO). Due to the retrospective nature of the study, the Institutional Review Board cancelled the need for written informed consent.

Study Design and Sample Acquisition

In a retrospective single-center study, a cohort of 437 patients was enrolled [body mass index (BMI): 27.8 ± 0.2 kg/m^2; age: 64.8 ± 0.6; males/females: 306 (70.0%)/131 (30.0%)]. All patients had an AMI and underwent coronary angiography in the catheterization laboratory of the West German Heart and Vascular Centre Essen, University Hospital Essen between January 2009 and June 2014. AMI was defined as a troponin value above the 99th percentile of the upper reference level and either with a ST-segment elevation or new left bundle-branch block on the electrocardiogram (STEMI, n = 176)[22] or without an ST-segment elevation or new left bundle-branch block on the electrocardiogram (NSTEMI, n = 261).[23] Patients were classified according to their calculated SD in 2 groups: SD $\geq$ 50% (n = 357) and SD $<$ 50% (n = 80). Serum parameters were determined in the central laboratory unit of the University Hospital Essen (Department of Clinical Chemistry and Laboratory Medicine) by standardized methods. Exclusion criteria were a high-grade aortic valve disease, cardiomyopathy, cardiac allograft vasculopathy, endocarditis, hypertensive emergency, myocarditis, pericarditis, tachyarrhythmia absoluta by atrial fibrillation, coronary vasospasm, and survival of sudden cardiac death. In addition, patients with a coronary artery disease after coronary artery bypass graft were excluded from the present study (see Supplementary Figure 1, http://links.lww.com/MD/A688). Patients with noncardiac reasons of troponin elevation were excluded: acute neurological disease (including stroke or cerebral hemorrhage), acute pulmonary embolism, aortic dissection, diseases like amyloidosis, sarcoidosis, or scleroderma, inflammatory myopathies (ie, polymyositis, dermatomyositis), sepsis, and patients, who were on cardiotoxic medication (adriamycin, 5-flurouracil, herceptin).

Quantitative Coronary Angiography

All patients received oral acetylsalicylic acid (100 mg/day) and underwent PCI. All measurements were performed at the Angiography Core Laboratory at the West German Heart and Vascular Centre Essen, University Hospital, University Duisburg-Essen. Coronary angiography was performed using the femoral approach and 6 or 8 F guiding catheters. Stenosis severity was quantified using off-line caliper measurements (QCA-MEDIS, Leiden, NL) before stent implantation.[24] The diameter of the catheter tip was measured with digital calipers and used for image calibration. The reference lumen diameter (RLD) and the most narrow point (ie, minimal lumen diameter (MLD)) were calculated. The SD was defined as follows:

$$\mathrm{SD} = \frac{\mathrm{RLD} - \mathrm{MLD}}{\mathrm{RLD}} \times 100$$

Dataset and Statistics

The dataset (437 patients) included the sociodemographic parameters sex and age, BMI, the dichotomous variables STEMI/NSTEMI, smoker/nonsmoker, diabetes mellitus, dyslipidemia, family predisposition, as well as the serum parameters AP, GGT, AST, ALT, C-reactive protein (CRP), bilirubin, creatinine, and troponin. For 302 cases (245 with SD $\geq$ 50%, 57 with SD $<$ 50%), also information about total cholesterol levels, high-density lipoprotein (HDL), low-density lipoprotein (LDL), and triglyceride levels were available. We refer to these 302 cases as dataset 2.

Predictive Modeling

Statistical data analyses were performed with R (http://www.r-project.org/). All data are presented as mean $\pm$ standard error of the mean (SEM) unless specified otherwise. Missing values were imputed by mean imputation. Correlation analysis was performed using Spearman correlation coefficient *r*.

For building up predictive models, random forests (RFs) implemented in the *randomForest* package of R were used. An RF is an ensemble learning method that can be used for classification as well as regression,[25] which has gained popularity in the recent years.[26–28] RFs are classifiers consisting of a collection of decision trees (DTs) that are combined via majority vote. When using the trained RF for prediction, an unseen instance was assigned to the positive class voted for by at least 50% of the trees. Besides being highly accurate classifiers, RFs can be used to estimate variable importance, for example, by measuring the mean decrease in Gini impurity. The importance was calculated based on 100 individual RF models. Due to the imbalance in the dataset, we also build models with repeated (100 times) subsampling.

For evaluation of the classifier performance, a 100-fold leave-one-out cross-validation scheme was used and the receiver operation characteristics (ROC) curve and the corresponding area under the curve (AUC) were calculated with pROC.[29] The 95% confidence interval (CI) was computed with 2000 stratified bootstrap replicates. The ROC curve was built by plotting the sensitivity against the specificity for every possible cut-off between the 2 classes. The significance P_σ of the AUC values was calculated based on permutation tests (n = 1000). P_U values for comparison between classifiers are based on the Mann–Whitney *U* test.

RESULTS

Patient Characteristics and Basic Parameters

Detailed data of the included patients can be found in Table 1, comprising distribution of demographic parameters (age, BMI) as well as standard parameters of heart (troponin), renal (creatinine), liver damage (AP, GGT, AST, ALT, bilirubin), and risk factors (smoking, diabetes, etc.). The data set was divided into 2 groups, 1 containing patients with a SD $\geq$ 50% and 1 containing patients with a SD $<$ 50%. This cut-off usually indicates necessity for an intervention (ie,

TABLE 1. Demographic and Basic Parameters of the Patient Cohort

Parameter	Stenosis Diameter ≥50% (n = 80)	Stenosis Diameter <50% (n = 357)	*P*
Age	64.12 ± 0.09	68.03 ± 0.17	0.0157
BMI	27.76 ± 0.05	27.73 ± 0.11	0.966
AP	80.78 ± 0.26	72.88 ± 0.33	0.031
GGT	55.06 ± 0.72	56.56 ± 1.66	0.9123
AST	121.42 ± 1.72	64.18 ± 1.24	0.0077
ALT	59.23 ± 1.33	53.48 ± 2.1	0.7558
Bilirubin	0.63 ± 0.03	0.58 ± 0.05	0.3436
CRP	2.17 ± 0.14	1.65 ± 0.28	0.2161
Creatinine	1.36 ± 0.04	1.42 ± 0.11	0.6732
Troponin	11.37 ± 0.46	13.01 ± 1.88	0.8138
STEMI (n)	26 (32.5%)	150 (42%)	0.1309*
NSTEMI (n)	54 (67.5%)	207 (58%)	
Ex-/smoker	18 (22.8%)/22 (27.8%)	74 (21.1%)/122 (34.9%)	0.4875*
Type 2 diabetes	28 (35.4%)	121 (34.6%)	0.8957*
Hypercholesterolemia	70 (88.6%)	273 (78%)	0.0421*
Familial predisposition	30 (38%)	105 (30%)	0.1809*

Data are presented as mean ± SEM.

ALT = alanine transaminase, AP = alkaline phosphatase, AST = aspartate transaminase, BMI = body mass index, CRP = C-reactive protein, GGT = gamma-glutamyl transferase.

* Comparison between patients with astenosis diameter 50% and with a stenosis diameter <50% was done by Student t test (not marked) or Fisher's exact test /Chi-square, respectively (*).

balloon dilation or placement of a stent). For both datasets only a small number of values had to be imputed. For dataset 1 (without HDL, LDL, cholesterol, and triglycerides) the following numbers of values were missing and imputed: BMI: 25 cases (5.7%), smoker/nonsmoker, diabetes mellitus, dyslipidemia, family predisposition (each): 8 cases (1.8%), AP: 52 cases (11.9%), GGT: 1 case (0.2%), bilirubin: 35 cases (8%), CRP: 14 cases (3.2%), creatinine: 6 cases (1.4%), and troponin: 4 cases (0.9%). For dataset 2 (with HDL, LDL, cholesterol, and triglycerides) the following numbers of values were missing and imputed: BMI: 7 cases (2.3%), AP: 14 cases (7.9%), GGT: 1 case (0.3%), bilirubin: 21 cases (6.9%), and CRP: 7 cases (2.3%). Patients with a SD ≥ 50% were younger and had significantly higher AP and AST levels, whereas BMI, ALT, GGT, bilirubin, CRP, troponin, and creatinine did not differ between the groups (Table 1).

Angiographic Data of Minimal Lumen Diameter (MLD), Reference Lumen Diameter (RLD), and Stenosis Diameter (SD) Before Stent Implantation

The RLD did not differ between patients with a SD ≥ 50 (2.84 ± 0.74 mm) and those with a SD < 50 (2.92 ± 0.78 mm) before stent implantation. The group separation by SD resulted in an artificially significant difference in MLD (SD ≥ 50: 0.55 ± 0.46 mm; SD < 50: 1.93 ± 0.67 mm; $P < 0.0001$) as well as SD (SD ≥ 50: 80.4 ± 0.8%; SD < 50: 34.6 ± 0.5%; $P < 0.0001$).

Stenosis Diameter and Noninvasively Determined Parameters Correlate Significantly

Significant correlations between SD and age ($r = -0.1112$, $P = 0.02$), AST ($r = 0.2606$, $P < 0.0001$), ALT ($r = 0.2099$, $P < 0.0001$), and troponin ($r = 0.3104$, $P < 0.0001$) were found. As expected, MLD correlated inversely with SD before stent implantation (Supplementary Figure 2, http://links.lww.com/MD/A688). Other noninvasive parameters did not exhibit significant correlations with the SD (Table 2).

High Importance of Liver Serum Parameters for Prediction of Stenosis Diameter

As serum liver markers were significantly correlated to the SD, an RF importance analysis for classification into SD ≥ 50 and SD < 50 was performed. A high importance (≥10) of the variables age (12.16 ± 0.03), BMI (13.40 ± 0.03), as well as the serum parameters AP (15.81 ± 0.04), GGT (10.41 ± 0.02), AST (10.98 ± 0.02), ALT (11.84 ± 0.03), creatinine (13.40 ± 0.03), and troponin (15.20 ± 0.03) were identified. The serum

TABLE 2. Correlation of Noninvasively Determined Parameters With Diameter Stenosis

Parameter	*R*	*P*
Age	−0.1112	0.02
BMI	0.0502	0.3092
AP	0.0857	0.0931
GGT	0.0068	0.8872
AST	0.2606	<0.0001
ALT	0.2099	<0.0001
Bilirubin	0.0922	0.0648
CRP	0.0711	0.1442
Creatinine	−0.0917	0.057
Troponin	0.307	<0.0001

Spearman correlation coefficient r of parameter x with stenosis diameter.

ALT = alanine transaminase, AP = alkaline phosphatase, AST = aspartate transaminase, BMI = body mass index, CRP = C-reactive protein, GGT = gamma-glutamyl transferase.

parameters bilirubin and CRP showed a relatively high importance (importance >5), while other parameters did not exhibit a high importance for the RF classification and thus were excluded from further analyses (Figure 1).

New Diagnostic Model for Prediction of Stenosis Diameter

Based only on parameters with an importance ≥10 (age, BMI, AP, GGT, AST, ALT, creatinine, troponin) a prediction model (M10) was developed. M10 reached an AUC of 69.7% CI 63.8–75.5% ($P_\sigma < 0.0001$) (Figure 2). Addition of the less important parameters (bilirubin, CRP), did not improve the model (M5) in terms of AUC. In fact, the AUC value of M5 was slightly, but significantly lower ($P_\sigma < 0.0001$, $P_U = 0.0015$). The AUC gives an overview of the general performance of a classifier. Certain specificities and corresponding sensitivities can be read out directly from the ROC curve, for example, the sensitivity of M10 at a specificity of 90% and 95% is 38.5% and 32.1%, respectively. Repeated subsampling did not improve the prediction performance of the classifiers. In dataset 2, the univariate analyses essentially led to the same results. However, age ($P = 0.0745$) was no longer significantly correlated with SD. This may be due to the smaller number of patients in dataset 2. Total cholesterol levels, HDL, LDL, and triglyceride levels were not significantly correlated with the SD and no significant differences between SD ≥ 50 and SD < 50 were found. Next, we trained RF models on dataset 2: (i) with the parameters of M10, and (ii) with parameters from M10 and total cholesterol levels, HDL, LDL, and triglyceride levels. We found no significant differences in terms of AUC between (i) and (ii).

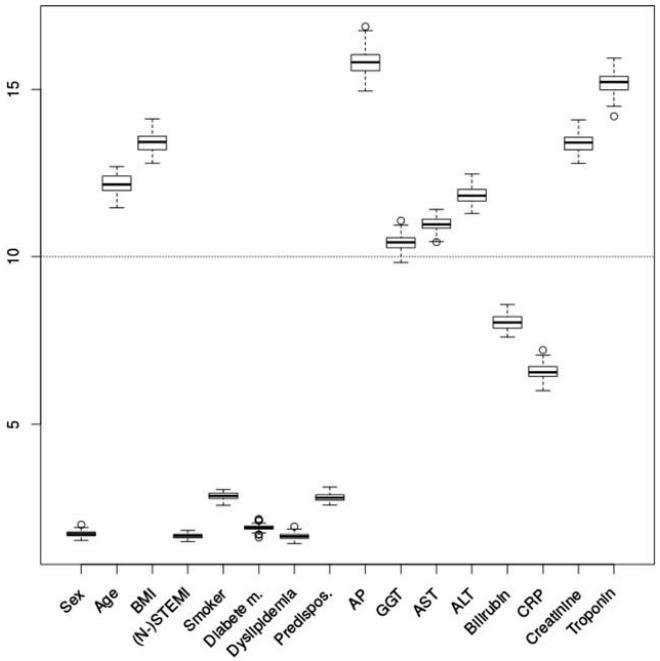

FIGURE 1. Importance analysis. The *y*-axis show the estimated importance by mean decrease in Gini impurity for the different parameters (*x*-axis). ALT = alanine transaminase, AP = alkaline phosphatase, AST = aspartate transaminase, BMI = body mass index, CRP = C-reactive protein, GGT = gamma-glutamyl transferase, NSTEMI = non-ST elevation myocardial infarction, STEMI = ST elevation myocardial infarction.

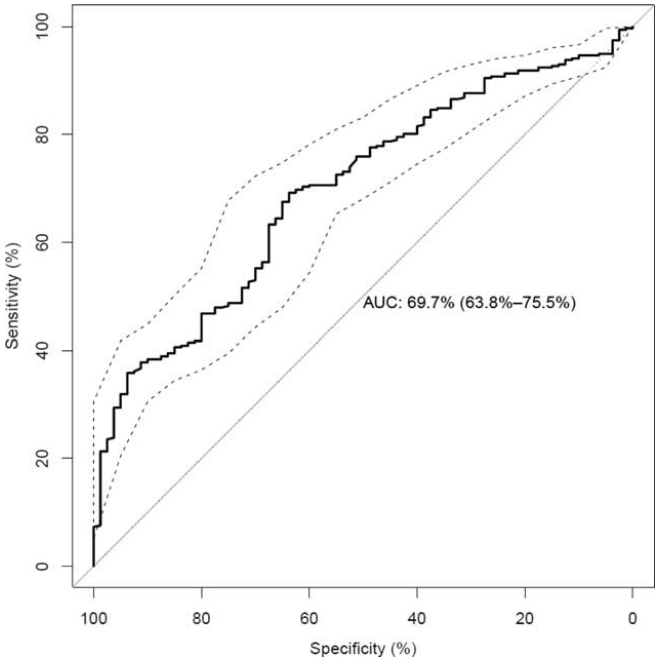

FIGURE 2. Performance of prediction model (M10) for the prediction of stenosis diameter. On the *y*-axis the sensitivity and on the *x*-axis the specificity is shown. The ROC curve is shown as a bold solid line. The AUC of M10 is 69.7% (CI 63.8–75.5%). The confidence interval is shown as dashed lines. The dotted line marks the performance of random guessing.

DISCUSSION

In the present retrospective study, we have investigated the correlation of demographic data and specific serum parameters with the SD in patients with AMI undergoing PCI using a machine learning approach. Based on parameters with an importance of ≥10 in the RF analyses, a prediction model with an AUC of 69.7% was developed. Besides age and troponin, liver transaminases (ALT and AST), and AP were identified as highly important for discrimination of patients with SD ≥ 50 and SD < 50.

The extent and severity of stable CVD is reported to be lower in younger populations,[30] and the incidence of ST segment elevation myocardial infarction increases linearly with the age in men, but exponentially in women.[31] In contrast, in the presented cohort age was inversely correlated to SD, implying that younger patients had more severe stenosis. In line with previous studies troponin was associated with a higher SD in the present study. Elevated levels of troponin are associated with coronary artery calcification in patients with mild CVD[32,33] as well as in non-ST-elevation myocardial infarction.[34]

It was also possible to confirm an association between the liver transaminases (ALT, AST) and the severity of the SD.[19] In a previous study a positive correlation was observed between liver transaminases and the severity of CVD in women, but not in men.[35] Saely et al[36] also investigated the possible association of ALT and AST with angiographically determined CVD and the presence of metabolic syndrome. A significant association was identified between ALT and ALT/AST ratio with metabolic syndrome, but there was no association between liver transaminases and angiographically determined CVD. This might be due to the currently set normal ranges, which might be too high to detect ongoing liver injury in a metabolic setting.[37] Apart from liver diseases reduced arterial perfusion or congestion due to (acute) cardiac failure can also affect liver serum parameters.

Within our cohort 100 patients (23%) with AST or ALT concentrations >100 U/L were found, though only 12 of these exhibited signs of right heart burden in echocardiography (Supplementary Table 1, http://links.lww.com/MD/A688). Though, the majority of patients with elevated serum liver enzymes exhibited either confirmed liver disease or signs of metabolic syndrome, which would suggest a NAFLD-type liver injury. Taken together works of other groups and our data suggest that liver damage precedes cardiac manifestations in metabolic syndrome in most cases.

In the present study, AP was identified as the variable with the highest importance to predict SD ≥ 50 or SD < 50. However, this was only detected by the RF approach as the univariate analysis was not able to show a significant association between AP and the SD. AP has been known as predictor of mortality for patients with CVD, who already underwent successful PCI with drug-eluting stent,[38] and for those, who survived AMI.[39] AP is associated with CVD risk in elderly men[40] and correlated with the severity of CVD.[41] A possible mechanistic explanation for the high importance of AP in the presented classification could be related to AP acting as a regulator of vascular calcification.[42] Shantouf et al[20] found a significant association between high AP and the coronary artery calcification score in a cohort of 137 maintenance hemodialysis patients. AP and CVD could also be connected via inflammatory processes,[39] which may derive from adipose tissue inflammation observed in obese patients with metabolic syndrome.[43,44] This is supported by association of AP with CRP observed in previous studies.[40,45] Thus, AP levels may reflect inflammation of hepatic origin, as CRP is also mainly derived from the liver. In vascular disease, atherosclerosis is associated with inflammatory processes, and in advanced atherosclerotic plaque also increased serum AP were found.[45]

Study Limitations

The presented study has some limitations that need to be addressed in future works. Besides the relatively small number of patients in the whole cohort, one limiting aspect is the different size of subgroups. The cohort consisted of 357 patients with a SD ≥ 50% and the reference group with a SD < 50% consisted of only 80 samples. To reduce the bias in the classification, a subsampling approach was performed, which randomly selects a subset out of the larger group in similar size as the smaller group (n = 80). However, there were no significant differences in the results between the subsampled and the initial computations. Due to the retrospective nature of the study, there were some missing values, especially within the parameters HDL, LDL, triglyceride, and total cholesterol.

To assess atherosclerosis and the SD in coronary vessels a range of methods is available for patients with subclinical or stable CVD. These methods comprise different noninvasive approaches (carotid intima media thickness measurement, brachial artery flow mediated dilatation, arterial stiffness, and multislice computed tomography) and invasive procedures (intravascular ultrasound, QCA). Due to the special emergency situation of AMI and the retrospective nature of the present study, QCA remained the only coronary imaging to assess the severity of the SD. QCA has important limitations: Only lumen, but not the coronary vessel wall can be visualized. The extent of atheroma within the vessel wall is not reliably determined by standard angiographic techniques. Moreover, lumen size is a relatively crude measure of atherosclerotic disease, especially in patients with only mild stenotic lesions.[46] Though, in the presented setting (emergency) no other coronary imaging method was feasible. This may change, when noninvasive parameters could predict the severity of stenosis and thus enable a larger time frame until PCI needs to be performed in some AMI cases with less severe stenosis.

Another limitation occurring due to the retrospective nature in a cardiologic emergency setting is a lack of information on liver diseases. Only serum parameters as surrogate indicators for liver injury were available. While our data hint to a connection of liver injury and severity of AMI in metabolic syndrome the exact association cannot be inferred from this study. When reviewing the medical records of the patient cohort only 6 individuals with established/ liver disease (3 with NAFLD, 3 with alcoholic fatty liver disease) were identified.

CONCLUSIONS

Taken together, age and troponin, but also the classic liver enzymes AST and ALT were significantly correlated with the diameter stenosis in patients with AMI undergoing PCI. This adds to the proposed close link between liver and CVD, especially in metabolic syndrome. Moreover, it was possible to build a predictive model from age, BMI, and 6 noninvasively determined serum parameters to classify patients for a SD of < or ≥50. By using a sensitivity cutoff of 90%, the false negative rate is only 10%. The corresponding specificity of the model is 27%. Thus, those patients that have a severe stenosis are reliably detected with our model, however taking into account a moderate number of patients that would undergo catheterization without a clinical need. We implemented a webserver at http://stenosis.heiderlab.de using the aforementioned sensitivity cutoff of 90%, which can be used to predict SD.

Liver parameters may be relevant factors to predict the severity of stenosis in AMI and to identify high-risk subjects in a noninvasive way, sparing patients with less severe stenosis the dangerous procedure of cardiac catheterization and coronary angiography.

ACKNOWLEDGMENT

We would like to thank the Dr. Heinz-Horst-Deichmann Foundation, which provided funding for parts of the presented study.

REFERENCES

1. WHO. Cardiovascular diseases (CVDs). WHO at http://www.who.int/mediacentre/factsheets/fs317/en/. Accessed May 19, 2015
2. Burns RJ, Gibbons RJ, Yi Q, et al. The relationships of left ventricular ejection fraction, end-systolic volume index and infarct size to six-month mortality after hospital discharge following myocardial infarction treated by thrombolysis. *J Am Coll Cardiol.* 2002;39:30–36.
3. Heusch G. Reduction of infarct size by ischaemic post-conditioning in humans: fact or fiction? *Eur Heart J.* 2012;33:13–15.
4. Levine GN, Bates ER, Blankenship JC, et al. 2011 ACCF/AHA/SCAI Guideline for Percutaneous Coronary Intervention: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines and the Society for Cardiovascular Angiography and Interventions. *Circulation.* 2011;124:e574–e651.
5. Marchesini G, Brizi M, Bianchi G, et al. Nonalcoholic fatty liver disease: a feature of the metabolic syndrome. *Diabetes.* 2001;50:1844–1850.
6. Gastaldelli A, Kozakova M, Højlund K, et al. Fatty liver is associated with insulin resistance, risk of coronary heart disease, and early atherosclerosis in a large European population. *Hepatology.* 2009;49:1537–1544.

7. Adams LA, Waters OR, Knuiman MW, et al. NAFLD as a risk factor for the development of diabetes and the metabolic syndrome: an eleven-year follow-up study. *Am J Gastroenterol.* 2009;104:861–867.
8. Targher G, Bertolini L, Poli F, et al. Nonalcoholic fatty liver disease and risk of future cardiovascular events among type 2 diabetic patients. *Diabetes.* 2005;54:3541–3546.
9. Targher G, Arcaro G. Non-alcoholic fatty liver disease and increased risk of cardiovascular disease. *Atherosclerosis.* 2007;191:235–240.
10. Kim D, Choi S-Y, Park EH, et al. Nonalcoholic fatty liver disease is associated with coronary artery calcification. *Hepatology.* 2012;56:605–613.
11. Blackett PR, Sanghera DK. Genetic determinants of cardiometabolic risk: a proposed model for phenotype association and interaction. *J Clin Lipidol.* 2013;7:65–81.
12. Ozturk K, Uygun A, Guler AK, et al. Nonalcoholic fatty liver disease is an independent risk factor for atherosclerosis in young adult men. *Atherosclerosis.* 2015;240:380–386.
13. Oni ET, Agatston AS, Blaha MJ, et al. A systematic review: burden and severity of subclinical cardiovascular disease among those with nonalcoholic fatty liver; should we care? *Atherosclerosis.* 2013;230:258–267.
14. Akabame S, Hamaguchi M, Tomiyasu K-I, et al. Evaluation of vulnerable coronary plaques and non-alcoholic fatty liver disease (NAFLD) by 64-detector multislice computed tomography (MSCT). *Circ J.* 2008;72:618–625.
15. Santos RD, Nasir K, Conceição RD, et al. Hepatic steatosis is associated with a greater prevalence of coronary artery calcification in asymptomatic men. *Atherosclerosis.* 2007;194:517–519.
16. Assy N, Djibre A, Farah R, et al. Presence of coronary plaques in patients with nonalcoholic fatty liver disease. *Radiology.* 2010;254:393–400.
17. Chen C-H, Nien C-K, Yang C-C, et al. Association between nonalcoholic fatty liver disease and coronary artery calcification. *Dig Dis Sci.* 2010;55:1752–1760.
18. Ruttmann E, Brant LJ, Concin H, et al. Gamma-glutamyltransferase as a risk factor for cardiovascular disease mortality: an epidemiological investigation in a cohort of 163,944 Austrian adults. *Circulation.* 2005;112:2130–2137.
19. Masoudkabir F, Karbalai S, Vasheghani-Farahani A, et al. The association of liver transaminase activity with presence and severity of premature coronary artery disease. *Angiology.* 2011;62:614–619.
20. Shantouf R, Kovesdy CP, Kim Y, et al. Association of serum alkaline phosphatase with coronary artery calcification in maintenance hemodialysis patients. *Clin J Am Soc Nephrol.* 2009;4:1106–1114.
21. Mahabadi AA, Lehmann N, Möhlenkamp S, et al. Association of bilirubin with coronary artery calcification and cardiovascular events in the general population without known liver disease: the Heinz Nixdorf Recall study. *Clin Res Cardiol.* 2014;103:647–653.
22. Levine GN, Bates ER, Blankenship JC, et al. 2015 ACC/AHA/SCAI Focused Update on Primary Percutaneous Coronary Intervention for Patients With ST-Elevation Myocardial Infarction: an Update of the 2011 ACCF/AHA/SCAI Guideline for Percutaneous Coronary Intervention and the 2013 ACCF/AHA Guideline for the Management of ST-Elevation Myocardial Infarction: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Society for Cardiovascular Angiography and Interventions. *Circulation.* 2015doi:10.1161/CIR.0000000000000336.
23. Amsterdam EA, Wenger NK, Brindis RG, et al. 2014 AHA/ACC guideline for the management of patients with non-ST-elevation acute coronary syndromes: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation.* 2014;130:e344–e426.
24. Haude M, Caspari G, Baumgart D, et al. Comparison of myocardial perfusion reserve before and after coronary balloon predilatation and after stent implantation in patients with postangioplasty restenosis. *Circulation.* 1996;94:286–297.
25. Breiman L. Random forests. *Mach Learn.* 2001;45:5–32.
26. Sowa J-P, Heider D, Bechmann LP, et al. Novel algorithm for non-invasive assessment of fibrosis in NAFLD. *PLoS ONE.* 2013;8:e62439.
27. Heider D, Senge R, Cheng W, et al. Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. *Bioinformatics.* 2013;29:1946–1952.
28. Heider D, Appelmann J, Bayro T, et al. A computational approach for the identification of small GTPases based on preprocessed amino acid sequences. *Technol Cancer Res Treat.* 2009;8:333–341.
29. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics.* 2011;12:77.
30. Khan HU, Khan MU, Noor MM, et al. Coronary artery disease pattern: a comparison among different age groups. *J Ayub Med Coll Abbottabad.* 2014;26:466–469.
31. Kytö V, Sipilä J, Rautava P. Gender, age and risk of ST segment elevation myocardial infarction. *Eur J Clin Invest.* 2014;44:902–909.
32. Jung HH, Ma KR, Han H. Elevated concentrations of cardiac troponins are associated with severe coronary artery calcification in asymptomatic haemodialysis patients. *Nephrol Dial Transplant.* 2004;19:3117–3123.
33. Laufer EM, Mingels AMA, Winkens MHM, et al. The extent of coronary atherosclerosis is associated with increasing circulating levels of high sensitive cardiac troponin T. *Arterioscler Thromb Vasc Biol.* 2010;30:1269–1275.
34. Qadir F, Farooq S, Khan M, et al. Correlation of cardiac troponin I levels (10 folds upper limit of normal) and extent of coronary artery disease in non-ST elevation myocardial infarction. *J Pak Med Assoc.* 2010;60:423–428.
35. Adibi P, Sadeghi M, Mahsa M, et al. Prediction of coronary atherosclerotic disease with liver transaminase level. *Liver Int.* 2007;27:895–900.
36. Saely CH, Vonbank A, Rein P, et al. Alanine aminotransferase and gamma-glutamyl transferase are associated with the metabolic syndrome but not with angiographically determined coronary atherosclerosis. *Clin Chim Acta.* 2008;397:82–86.
37. Kälsch J, Bechmann LP, Heider D, et al. Normal liver enzymes are correlated with severity of metabolic syndrome in a large population based cohort. *Sci Rep.* 2015:513058.
38. Park J-B, Kang D-Y, Yang H-M, et al. Serum alkaline phosphatase is a predictor of mortality, myocardial infarction, or stent thrombosis after implantation of coronary drug-eluting stent. *Eur Heart J.* 2013;34:920–931.
39. Tonelli M, Curhan G, Pfeffer M, et al. Relation between alkaline phosphatase, serum phosphate, and all-cause or cardiovascular mortality. *Circulation.* 2009;120:1784–1792.
40. Wannamethee SG, Sattar N, Papcosta O, et al. Alkaline phosphatase, serum phosphate, and incident cardiovascular disease and total mortality in older men. *Arterioscler Thromb Vasc Biol.* 2013;33:1070–1076.
41. Sahin I, Karabulut A, Gungor B, et al. Correlation between the serum alkaline phosphatase level and the severity of coronary artery disease. *Coron Artery Dis.* 2014;25:349–352.
42. O'Neill WC, Sigrist MK, McIntyre CW. Plasma pyrophosphate and vascular calcification in chronic kidney disease. *Nephrol Dial Transplant.* 2010;25:187–191.

43. Hotamisligil GS. Inflammation and metabolic disorders. *Nature*. 2006;444:860–867.

44. Wree A, Kahraman A, Gerken G, et al. Obesity affects the liver—the link between adipocytes and hepatocytes. *Digestion*. 2011;83:124–133.

45. Webber M, Krishnan A, Thomas NG, et al. Association between serum alkaline phosphatase and C-reactive protein in the United States National Health and Nutrition Examination Survey 2005–2006. *Clin Chem Lab Med*. 2010;48:167–173.

46. Ballantyne CM, Raichlen JS, Nicholls SJ, et al. Effect of rosuvastatin therapy on coronary artery stenoses assessed by quantitative coronary angiography: a study to evaluate the effect of rosuvastatin on intravascular ultrasound-derived coronary atheroma burden. *Circulation*. 2008;117:2458–2466.

B.2. Paper II

Neumann *et al. BioData Mining* (2016) 9:36
DOI 10.1186/s13040-016-0114-4

BioData Mining

RESEARCH Open Access

Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach

Ursula Neumann[1,2,3], Mona Riemenschneider[1,2], Jan-Peter Sowa[4], Theodor Baars[5], Julia Kälsch[4], Ali Canbay[4] and Dominik Heider[1,2,3*]

*Correspondence:
d.heider@wz-straubing.de
[1]Department of Bioinformatics, 94315, Straubing, Germany
[3]Wissenschaftszentrum Weihenstephan, Technische Universität München, 85354, Freising, Germany
Full list of author information is available at the end of the article

Abstract

Motivation: Biomarker discovery methods are essential to identify a minimal subset of features (e.g., serum markers in predictive medicine) that are relevant to develop prediction models with high accuracy. By now, there exist diverse feature selection methods, which either are embedded, combined, or independent of predictive learning algorithms. Many preceding studies showed the defectiveness of single feature selection results, which cause difficulties for professionals in a variety of fields (e.g., medical practitioners) to analyze and interpret the obtained feature subsets. Whereas each of these methods is highly biased, an ensemble feature selection has the advantage to alleviate and compensate for such biases. Concerning the reliability, validity, and reproducibility of these methods, we examined eight different feature selection methods for binary classification datasets and developed an ensemble feature selection system.

Results: By using an ensemble of feature selection methods, a quantification of the importance of the features could be obtained. The prediction models that have been trained on the selected features showed improved prediction performance.

Keywords: Machine learning, Feature selection, Ensemble learning, Biomarker discovery, Random forest

Background

In the fields of predictive medicine as well as molecular diagnostics the need for simplification of datasets with many parameters frequently emerges. Therefore, approaches are necessary, which can identify important parameters (sometimes also referred to as features, independent variables, or predictor variables). Such quantifiable parameters that allow diagnostic validity are called biomarkers. In 2001, the Biomarkers Definitions Working Group of the American National Institute of Health defined a biomarker as "a characteristic that is objectively measured and evaluated as an indication of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention" [1]. Examples for biomarkers are serum parameters, genetic markers, or socio-demographic markers.

The detection of biomarkers can be conducted by computer-assisted approaches, namely feature selection (FS) methods. A great variety of FS techniques already exist. In general, these approaches can be separated into: filter methods, wrapper methods, and embedded methods. The first one is independent of any prediction model and therefore shows an advantage in regards of computation time compared to the other approaches. Filter methods use weighting measures, such as correlation coefficients [2] or mutual information [3]. The wrapper methods are computationally intensive, but in turn provide better accuracy compared to filter methods [4]. This type of approach occurs outside the model construction, however it uses the outcome as a guideline. The third type, the embedded methods, is an alternative to wrapper methods. These approaches combine the advantages of both methods stated above, namely the low computational costs and an adequate accuracy. This is due to the fact that the process of feature selection is already part of the model construction. There are three main criteria a feature selection method should meet, namely reliability, validity, and reproducibility. Methods that display these characteristics are called stable. Based on the definition of biomarkers, non-generalizable features are not considered to be reliable markers. There are several factors that can cause instability of the feature selection, e.g., the complexity of multiple biomarkers, a small-n-large-p-problem, or when the algorithm simply ignores stability [5, 6]. Thus, feature selection results have to be treated with care. For example, the Gini-index is widely used in predictive medicine, but it has also been demonstrated to deliver instable results due to unbalanced datasets [7, 8]. To counteract instability of feature selection methods, we developed an ensemble feature selection (EFS) method, which compensates biases of single FS. The idea of ensemble methods is already widely used in learning algorithms [9]. In this article we will introduce eight FS methods and our quantifying EFS method. We evaluated our EFS method compared to the state-of-the-art method AUC-FS with regard to the prediction performance in subsequent classification based on six different datasets. Furthermore, we compared the results with prediction models without pre-selection of features.

Methods

With the development of the EFS method we take advantage of the benefits of multiple feature selection methods and combine their normalized outputs to a quantitative ensemble importance. The key features of our EFS method are:

1. The combination of widely known and extensively tested feature selection methods.
2. The balance of biases by using an ensemble.
3. The evaluation of EFS.

Eight different feature selection methods have been used for the EFS approach. Since random forests have drawn increased attention in the field of predictive medicine, four of the chosen feature selection methods are embedded in a random forest algorithm. Further, we considered the outcome of a logistic regression (i.e., the coefficients) as another embedded method as well as the filter methods median, Pearson-, and Spearman-correlation.

We used implementations in **R** (http://www.r-project.org/) for the different basal feature selection methods. Before we go into details a general data setting is introduced:
Let vectors $X_i = (x_{1,i}, \ldots, x_{N,i})$ be the prediction variables for $i \in \{1, \ldots M\}$ and

$Y = (y_1, \ldots, y_N)$ be the response variable. Altogether, a data matrix of size $N \times M + 1$ is received, where N is the number of samples and M is the number of prediction variables.

Random forest

Random forests (RFs) are ensemble learning methods for classifications and regressions consisting of multiple decision trees [10]. RFs have been shown to give highly accurate predictions on biological [11–13] and biomedical data [14, 15]. There are different implementations of the RF algorithm in R available, which offer diverse feature selection methods. In the context of RFs, these feature selection methods are called variable importance measures (VIMs). We integrated two of the available implementations of RFs into our EFS method: (i) the RF method adapted from Breiman [10], which uses the CART (classification and regression tree) algorithm for individual node decisions, implemented in the R package *randomForest* and (ii) the cforest [16] implementation from the R-package *party*, because of its promising AUC score VIM. In RF approaches, randomness is gained by the general technique of bootstrap aggregating, also called bagging, meaning that for the tree building process only a subset of the data samples are chosen with replacement. We used 1000 decision trees in both RFs. In order to get robust results, we averaged the VIMs over 100 RF models.

The raw variable importance score for X_i is given by the average over the set of all decision trees $t \in \{1, \ldots, T\}$ in the RF:

$$\widehat{VI}_{X_i} = \frac{1}{T} \sum_{t=1}^{T} \widehat{VI}_{X_i}(t).$$

In addition, we define an indicator function $I(A)$ by:

$$I(A) = \begin{cases} 1, & \text{if the argument} A \text{is fulfilled,} \\ 0, & \text{otherwise.} \end{cases}$$

Gini-index

The Gini-index is the sum of products between different class proportions over all classes for each variable, which is in the case of a binary classification:

$$G = 2p(1-p),$$

where $p = \frac{N_1}{N}$ is the proportion of one of the classes, in this case for response $Y = 1$, and N_1 is the number of units in this class.
The Gini-index G defines a measure d_{ij} of the decrease in heterogeneity at node j:

$$d_{ij} = G - (\frac{N_L}{N} G_L + \frac{N_R}{N} G_R),$$

where G_R and G_L respectively are the Gini-indexes calculated for the following right and left nodes and N_L and N_R are the numbers of units in the left and right node after splitting. With this measure the variable importance for X_i in tree t is defined as:

$$VI_{X_i}(t) = \sum_{j \in J} d_{ij} I(X_i \textit{ splits at node } j).$$

For deeper insights in the functionality of the Gini VIM we refer to [7].

Mean accuracy error-rate-based VIM

The mean accuracy error-rate-based VIM uses the out-of-bag (OOB) data. The OOB consists of the subset of all samples which are not used for the construction of decision trees: For each tree, the prediction error on the OOB portion of the data is recorded (error rates for classification, mean square errors for regression). This process is repeated after permuting each predictor variable. The difference between both is averaged over all trees, and normalized by the standard deviation of the differences, except the standard deviation is zero. For each tree t, we get the following formula:

$$\widehat{VI}_{X_i}(t) = \frac{1}{|B(t)|} \sum_{j \in B(t)} I(y_j = p_j) - I(y_j = p_{j,\pi i}),$$

where p_i is the RF prediction of the response variable, πi is the permutation of the values in the i-th variable and $B(t)$ is the OOB data for tree t.

Conditional error-rate-based VIM

In principle, the underlying mathematical model for the conditional error-rate-based VIM is the same as for the mean accuracy error-rate-based VIM. The conditional VIM takes biases in variable importance into account, which are generated by a correlation of the tested X_i with the other prediction variables.
For $Z = X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_M$ we calculate

$$\widehat{VI}_{X_i}(t) = \frac{1}{|B(t)|} \sum_{j \in B(t)} I(y_j = p_j) - I(y_j = p_{j,\pi i|Z}),$$

where $B(t)$ is the OOB data for tree t. In other words, the variable X_i is permuted, while Z is fixed at $Z = z := (cp_1, \ldots, cp_{i-1}, cp_{i+1}, \ldots, cp_M)$, consisting of the cut points for each variable in Z, which are defined through the partition of the feature space of X_i induced by the current tree t.

AUC-based VIM

In contrast to the aforementioned VIMs, the AUC-based VIM does not employ the error-rate, but instead uses the Area Under the Curve (AUC). It is calculated as the integral of the Receiver Operating Characteristic (ROC) curve, which is received by mapping the sensitivity against specificity for every possible cut-off between the two classes.

In contrast to error-rate-based methods, which give more weight to the majority class, the AUC does not favor any class. In previous studies the AUC was shown to be a particularly appropriate VIM for unbalanced data settings and should be considered as the state-of-the-art model [17, 18]. The AUC-score is an estimator for the probability that a randomly chosen sample from class $Y = 1$ receives a higher class probability for class $Y = 1$ than a randomly chosen sample from class $Y = 0$. The variable importance for each tree t is calculated as:

$$\widehat{VI}_{X_i}(t) = AUC_i - AUC_{\pi i}$$

where AUC_i and $AUC_{\pi i}$ respectively are the AUCs computed from the OOB observations in tree t before respectively after permuting the values of predictor X_i.

Logistic regression

Even though RFs have become very popular, it is not totally understood why the algorithm acts in its specific way. An embedded feature selection method, which is understood in more details, is the weighting system (i.e., coefficients) of the logistic regression. For feature selection, we access the model's coefficients, i.e., the $\beta-$values of the regression equation. It should be noted that the range of features can strongly differ. Due to this fact, the β-coefficients of parameters are not directly comparable. To provide comparability of the variables' importances, we conducted a z-transformation:

$$z_X = \frac{X - \overline{X}}{s_X},$$

where $\overline{X}$ is the mean and s_X the standard deviation of variable X, respectively. Through standardization by z-transformation, the mean of β-coefficients become zero with a standard deviation of 1, thus assuring that the features all have the same domain. Subsequently, the values are ordered according to their absolute values in decreasing order.

Correlation coefficient

The correlation between any two features can be described as the quantification of the extent of statistical dependence between them, which can be quantified by different correlation coefficients. We used the approach of [19] to select features that are highly correlated with the dependent variable, but show only low correlation with other predictors. We used a threshold for the correlation between the predictor variables of $p = 0.7$. In order to avoid collinearity a threshold of 0.7 is most frequently used [20], although recommendations for more restrictive (e.g., 0.4 [21]) and less restrictive (e.g., 0.8 [22]) thresholds exist. In our study, we adopted two correlation coefficients, namely the Pearson product-moment correlation and the Spearman rank correlation coefficient.

Pearson

For any two features X and Y with samples $j = 1, \ldots, n$, the Pearson product-moment correlation coefficient is defined as

$$r_{XY} = \frac{\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \overline{x})^2 \sum_{j=1}^{n}(y_j - \overline{y})^2}},$$

where $\overline{x}$ and $\overline{y}$ are the sample means of X and Y.

Spearman

For the Spearman rank correlation coefficient we observe the sample's ranks $rk(x_i)$ and $rk(y_i)$ of the features $X = (x_1, \ldots, x_n)$ and $Y = (y_1 \ldots, y_n)$ and compute

$$\rho = 1 - 6\sum_{j=1}^{n} \frac{d_i^2}{n(n^2 - 1)},$$

where $d_i = rk(x_i) - rk(y_i)$.

Median

For the median feature selection, we used a Mann-Whitney-U test [23] comparing the positive and negative class of the response variable Y. The test evaluates following hypothesis: Since med_0 and med_1 are the medians of the negative and positive class of a predictor variable, the null hypothesis for each predictor variable is defined as:

$$H_0 : med_0 = med_1.$$

The resulting p-values of the Mann-Whitney-U test are used as scoring system for the feature selection. Thus, a smaller p-value indicates a higher importance.

Ensemble feature selection

Feature selection methods as a preprocessing step for supervised learning algorithms provide several benefits, such as reduced computational costs (e.g., training time, storage requirements), but also improved prediction performance. However, different feature selection methods provide different subsets of features. Hence they give rise to sample selection bias. In general, the aim of supervised learning algorithms is to find a suitable hypothesis which makes the best prediction for a particular problem. Improvements can be achieved by combining multiple hypotheses instead of testing only one. This is the main concept of ensemble learning methods. Ensemble techniques are widely used in machine learning algorithms to achieve higher stability. The RF algorithm is an example for bootstrap aggregating [24]. This technique combines several prediction models using a randomly drawn subset of the training data. Another type of ensemble learning methods are boosting algorithms, which merge several weak classifiers to a stronger one. The most popular implementation is AdaBoost [25].

In the current study, we developed a stable feature selection procedure, which is based on the idea of ensemble learning. For our EFS method we integrated eight different feature selection methods and normalized all individual outputs to a common scale, an interval from 0 to 1. Thereby we ensure the comparability between different FS methods and conserve the distances of importance between one feature to another. This normalization is achieved in two different ways: For all feature selections, except for the median, the absolute value of the FS method output is a value which illustrates the increase of importance. By dividing through the maximum value we get values between 1 and 0:

$$imp_{X_i} = \frac{\beta_i}{\max(\beta_m)_{m \in M}}.$$

In the case of the median FS we receive a p-value for each feature X_i, which is normalized as follows:

$$imp_{X_i} = 1 - p_i + \min(p_i).$$

For the four RF based VIMs, we computed 100 repetitions and averaged the importance for each feature. This procedure guarantees a higher robustness of the feature importance and the selected subset.

We evaluated the selected subsets by using a logistic regression model with a leave-one-out cross validation (LOOCV) to avoid overfitting. LOOCV is known to give inflated variance estimation [26], but in our study we used the LOOCV only for comparing the

different methodologies. The EFS system selects those parameter that have a higher importance than the mean importance:

$$imp_{X_i} > \overline{imp_{X_M}},$$

where $\overline{imp_{X_M}}$ symbolizes the mean of all variable importances. Alternatively, the median or Q3 could be used as well, however, both would lead to a fixed number of selected parameter irrespective of their relevance for the subsequent classification model.

The logistic regression model based on the EFS-selected features was then compared to logistic regression models either trained on all features and on features selected by the AUC-based VIM, which is considered to be one of the state-of-the-art methods for feature selection. We examined the AUC-values of the ROC curves with ROCR [27]. Additionally, the improvement in performance between the AUC-based VIM, the EFS subset, and the model without feature selection is measured by a comparison of the AUCs via the method of DeLong et al. [28].

Datasets

To evaluate our EFS method, we used six different datasets. An overview of the datasets is given in Table 1.

The first dataset *MI-Mortality* was provided by the Clinic for Cardiology, West German Heart and Vascular Centre Essen of the University Hospital Duisburg-Essen. It consists of 14 socio-demographic and serum parameters from 406 patients. The purpose of this study was to examine which parameters are important for the mortality prediction after treatment on myocardial infarction. The data was collected during a follow-up study of [29].

The Department of Gastroenterology and Hepatology of the University Hospital Duisburg-Essen provided the datasets *Fibrosis* [30] and *FLIP*, which again consist of socio-demographic and serum parameters. Both deal with different scores to predict fibrosis.

SPECTF is a dataset from the UCI Machine Learning Repository [31]. It describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. The class-variable is distinguishing between normal (=0) and abnormal (=1).

The *Sonar* dataset has also been retrieved from the UCI Machine Learning Repository and obtained by bouncing sonar signals off a metal cylinder or rock at various angles and under various conditions. The prediction model should be able to distinguish between rocks and metal cylinders.

In the *WBC* dataset a classification between benign and malignant tumors in breast cancer samples is intended. Benign tumors are not cancerous, thus these samples are class 0. Malignant tumor samples are class 1.

Table 1 Overview of datasets. Number of features after removing samples with missing values

Dataset	No. of Samples	No. of Features	Categorical	Numeric
MI-Mortality	406	14	7	7
Fibrosis	101	26	7	19
FLIP	103	13	6	7
SPECTF	267	44	44	0
Sonar	208	60	0	60
WBC	569	30	0	30

In order to reduce the number of missing values in the datasets, features with more than 20% missing values were discarded. Additionally, columns with zero variance were removed.

Results

Selected features

The number of selected features from EFS and AUC-FS varies for each dataset. The Gini FS method is known to prefer categorical variables with many categories and disregards potential important binary prediction variables [32]. In contrast to the Gini FS, we could observe that the variable type did not play a decisive role for the importance. Through aggregating different FS methods into an ensemble, biases of individual methods are compensated.

In Fig. 1 Venn diagrams are shown, illustrating the feature subsets derived from the AUC-FS and EFS, respectively. The Venn diagrams show no distinct trend for the number of features that were selected by the respective method, i.e., in some datasets EFS selects more features than the AUC-FS, while in other datasets it is the other way around.

For the *Fibrosis* data the selected subset of AUC-FS contains eight features, whereas the EFS subset consists of only seven. Five features have been selected by both methods, while the other features are disjoint. The *WBC* dataset yielded a similar result. Both methods selected a subset of ten features, with eight features being selected by both methods. The

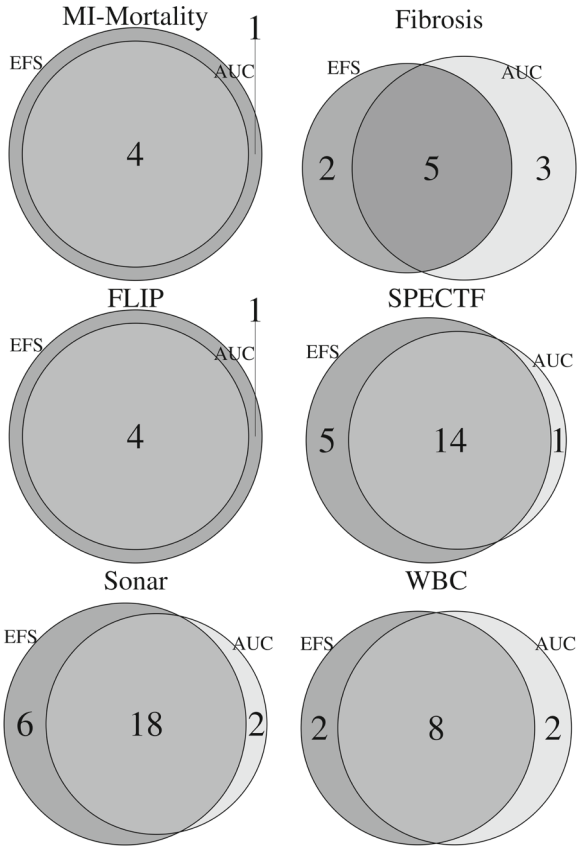

Fig. 1 Venn diagrams. Comparison of feature subsets retrieved from AUC importance and EFS importance

results of the *MI-Mortality* data and *FLIP* data are similar: EFS selected a subset of five features while AUC-FS returned four features, which all are contained in the EFS selected subset. The datasets of the *SPECTF* resp. *Sonar* studies also deliver analogous subset schemes. The major part of selected features are chosen by both FS methods (14 and 18, respectively). Our EFS method considered five and six additional features, while the AUC-FS selected one and two additional features, which do not occur in the intersection of both subsets.

The EFS selected more features than the AUC-FS in four out of six cases, however the percentages of selected features out of all possible prediction variables ranged from 26.9 to 43.2% (cf. Table 2).

Performance evaluation

In order to evaluate our EFS method in comparison to the AUC-FS, we used a logistic regression model with LOOCV. Additionally, we trained a logistic regression model without feature selection. Table 3 summarizes the results for all datasets. The resulting ROC curves are shown in Fig. 2.

For each dataset, the resulting model trained on the EFS selected subset of features performed superior compared to the models trained either on the AUC-FS selected features or on all features without selection.

However, the EFS showed a significantly higher AUC value only for the dataset WBC. For all other datasets, the AUCs were higher for the EFS compared to the AUC-FS as well, however the results were not significant: MI-Mortality ($p = 0.228$), Fibrosis ($p = 0.273$), FLIP ($p = 0.254$), SPECTF ($p = 0.444$), Sonar ($p = 0.2$), and WBC ($p = 0.02$).

The model using the EFS selected features showed significant higher AUC values compared to the model trained without feature selection for all datasets except MI-Mortality and FLIP ($p = 0.201$ and $p = 0.971$, respectively). Taken together, throughout all datasets we can observe an enhancement of performance by using the EFS method, although it is not significant in all datasets.

Additionally, we evaluated the robustness of our EFS approach by using permutation tests [33, 34]. To this end, the logistic regression models are compared to models that are trained on randomly permuted class labels. P-values for all datasets were less than 0.001.

Moreover, we evaluated the stability of the EFS approach in terms of selected features. To this end, we evaluated the variance of the importance of the five most important features using a 10-fold cross-validation of the datasets repeated 10 times. Furthermore, we used the Jaccard-index [35] as a stability score, described by the following formula:

$$J(S_1, \ldots, S_n) = \frac{|S_1 \cap \ldots \cap S_n|}{|S_1 \cup \ldots \cup S_n|},$$

Table 2 Types of selected features. Evaluation of the selected features subsets of AUC-FS and EFS

Dataset	AUC-FS selected	EFS selected	EFS/all in %	Numeric*	Categorical*
MI-Mortality	4	5	35.7	3	2
Fibrosis	8	7	26.9	5	3
FLIP	4	5	38.5	3	2
SPECTF	15	19	43.2	0	19
Sonar	20	24	40.0	24	0
WBC	10	10	33.3	9	1

*refers to the EFS selected features

Table 3 Results on datasets

Dataset	All [CI]	AUC-FS [CI]	EFS [CI]	AUC-FS vs. EFS*	all vs. EFS**
MI-Mortality	0.758 [0.700, 0.800]	0.757 [0.704, 0.811]	0.776 [0.725, 0.826]	0.228	0.201
Fibrosis	0.493 [0.300, 0.600]	0.681 [0.537, 0.824]	0.746 [0.617, 0.874]	0.273	**0.018**
FLIP	0.759 [0.600, 0.900]	0.723 [0.582, 0.863]	0.761 [0.633, 0.890]	0.254	0.971
SPECTF	0.807 [0.700, 0.900]	0.856 [0.811, 0.901]	0.865 [0.821, 0.910]	0.444	4.68e-4
Sonar	0.792 [0.700, 0.900]	0.840 [0.787, 0.894]	0.862 [0.813, 0.911]	0.200	**0.009**
WBC	0.611 [0.600, 0.700]	0.987 [0.977, 0.998]	0.991 [0.981, 1.000]	**0.020**	**1.21e-41**

Column 1 to 3 are AUCs values of all features, selected by AUC-FS and by the EFS with confidential intervalls in brackets. The last two columns show the *p*-values of the comparison by the method of [28]. The function compares the AUC of the ROC curves of (*) the AUC-FS and EFS method and (**) of all parameters and EFS outcome. Statistical significant *p*-values are printed in bold

where $S_1, \ldots, S_n$ are different subsets of features. Thereby, a Jaccard-index close to 1 represents a high similarity of feature subsets. It turned out that EFS gives highly stable results with variances of the importance values less than 0.0235. Moreover, the Jaccard-index of the selected features by EFS was 1 for all data sets. Table 4 shows all variances of the importance and the corresponding boxplots can be found in the Additional file 1.

Discussion

Feature selection methods have been studied for several decades (e.g., [36]). There are already many publications [37–41] on how to improve the performance of feature selection methods.

We provide an ensemble feature selection tool to conduct a feature selection for binary classification, which showed promising performance on all datasets. In contrast to ensemble methods of previous studies [42–44], the aim of this work was to combine filter

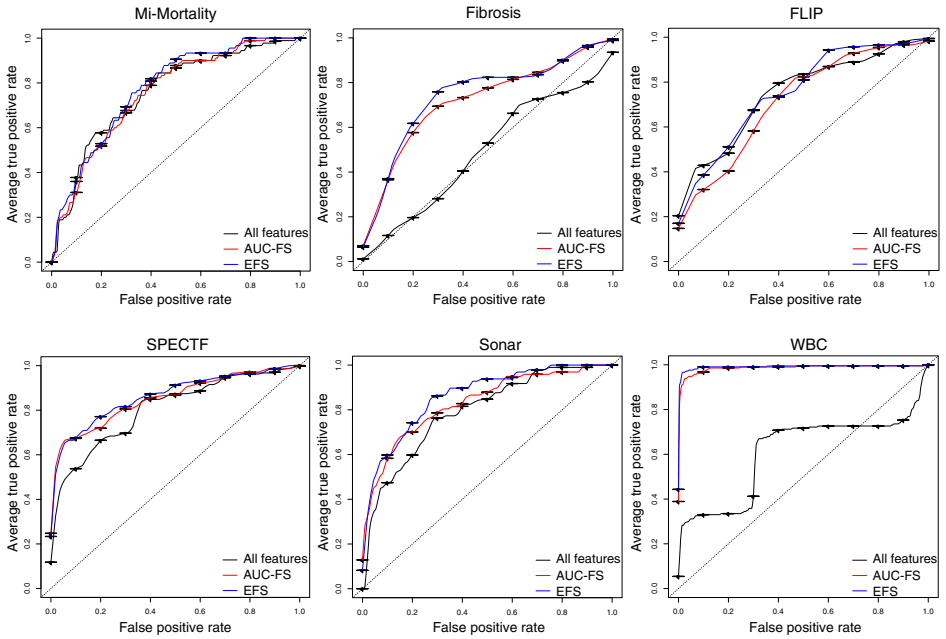

Fig. 2 Performance of logistic regression models. On the y-axis the sensitivity and on the x-axis the specificity is shown. Three ROC curves are shown per dataset: of all features (*solid*), the AUC-FS selected (*dashed*) and the EFS selected (*twodashed*) features. The *dotted line* marks the performance of random guessing

Table 4 Variance of feature importances. Variance of the five most important features of a 10-fold cross-validation

Dataset	Variance #1	Variance #2	Variance #3	Variance #4	Variance #5
MI-Mortality	0.001759124	0.004694053	0.004904828	0.003720571	0.001580310
Fibrosis	0.003124527	0.008085472	0.019901386	0.009202372	0.019804508
FLIP	0.006604973	0.011325453	0.014731007	0.023499884	0.020140657
SPECTF	0.000380482	0.014946809	0.011520607	0.005807655	0.002880478
Sonar	0.003887830	0.001792209	0.003004598	0.003115140	0.002680274
WBC	0.001071784	0.001769331	0.002912278	0.000387555	0.001096465

and embedded methods. Due to their focus on predictions, embedded methods usually attain a higher prediction performance, whereas the advantage of filter methods are low computational cost and low complexity. By using ensembles, the advantages of both strategies can be combined and individual biases are alleviated. Concerning the enhanced approximation of embedded methods, we excluded wrapper methods from our study.

The cforest method requires more time than any other component of the EFS algorithm, thus calculations of datasets with hundreds of thousands of features would take up a lot of CPU time. A workload saving alternative would be a reduction of the repetition rate of the RF algorithms, in particular of the cforest algorithms. However, in turn this will negatively affect the VIM's robustness. In our computations the repetition rate was set to 100 and the average variable importance was reported. Since, there is no generalization on how many repeats are necessary to get a robust result.

The evaluation of feature subsets depicted in the Venn diagrams reflects that in four out of six cases our EFS method selects more features than the AUC-FS. We assume that the reason for this phenomenon is based on the importance weighting system of the AUC-FS. As threshold for the decision which features are considered to be the most important ones, the respective mean over all importance values was taken. If there are only a few features lying above average, this might be an indication that the values of those features which are considered important are overestimated compared to the non-selected features. Thus the mean increases and less features reach that threshold. Alternatively, the opposite case could be true, meaning in one or more of the other feature selection methods the assigned importance values hardly differ. This in turn has an alleviating factor on the importance values of our ensemble of feature selection methods.

In the current study, we used the logistic regression method to analyze the performance of our EFS. For binary classification, logistic regression is the statistical method of choice, in particular in the field of predictive medicine [45]. It has the ability of detecting possible causal relationships between features. By conducting a z-transformation on the whole dataset the relationships become easy to interpret via the β-coefficients. Although the logistic regression model has many advantages, the prediction performance could be improved by using other predictive models in future studies. To get a broader and more generalizable rating for the results of our EFS method, an evaluation by methods such as support vector machines or RFs could additionally be conducted.

The output of all individual feature selection methods is normalized and summed up to our EFS result using the same weighting for all methods. However, there are more possibilities how the ensemble importance of features can be calculated, such as majority vote or by a weighting system. A weighting system could consider the individual robustness of each FS method, whereas a majority vote does not provide comparability between the

Table 5 Quantity of selected features. Number of selected features of our EFS method with and without the AUC-FS

Dataset	EFS	EFS without AUC-FS	Intersection
MI-Mortality	5	5	5
Fibrosis	7	9	7
FLIP	5	5	5
SPECTF	19	20	19
Sonar	24	24	24
WBC	10	11	9

importance of features. This issue could be solved by a weighted majority vote. For more details on fusion methods we refer to [9].

We determined several thresholds for the computation, namely the number of repetitions of the RF algorithms (100 times), the threshold of missing values (20%), and the correlation threshold between the dependent variables (0.7). In some data cases varying these thresholds might yield a better performance. However, for comparability reasons we used fixed thresholds for all datasets.

We also examined the subsets of features selected by the EFS method without the AUC-FS to estimate the influence of the AUC-FS. The selected features are essentially the same (cf. Table 5). In three datasets the subsets are slightly larger, which supports our theory on the overestimating effect of the AUC-FS on relevant feature's importance.

By the stability-test we proofed, that the EFS method is a stable and reliable approach for binary classification.

Conclusion

In the current study, we could show the advantages of our EFS method for binary classification data, namely the robustness and stability of feature ranking and subset selection. The evaluation of prediction performance via ROC curves of a logistic regression model showed an improvement of the prediction based on the EFS selected features compared to all features on every tested dataset.

Further investigations on the topic of enhancing feature selection methods will be conducted in future. Firstly, we will evaluate our EFS method on high-dimensional data, such as data retrieved from microarray or next-generation sequencing analyses. So far we used datasets with less than 600 samples and a maximum of 60 features. Secondly, in future studies we would like to investigate how our method deals with multiple classes instead of binary classification. Therefore, it will be necessary to substitute the median feature selection method with an appropriate alternative. Another interesting application will be the extension on regression models where classes are replaced by continuous values. Another direction of our future work on EFS methods will concern the composition of our FS method set. By combining feature selection algorithms the accuracy will improve by the expense of increased complexity. Using an ensemble of several simple methods can gain a higher accuracy than one complex method (cf. [9]). Due to this theory, an evaluation is needed on which FS methods are mandatory to gain a maximum accuracy.

Additional file

Additional file 1: Boxplots of five most important features in bootstrapping analyses. (JPEG 199 kb)

Abbrevations
AUC: Area under the curve; CART: Classification and regression tree; CPU: Central processing unit; EFS: Ensemble feature selection; FS: Feature selection; LOOCV: Leave-one-out cross validation; OOB: Out-of-bag; RF: Random forest; ROC: Receiver operating characteristic; VIM: Variable importance measure

Acknowledgements
We are grateful to the UCI Machine Learning Repository for granting access to a great variety of datasets.

Funding
This work was supported by the Deichmann Foundation, which had no role in study design, collection, analysis, and interpretation of data, and in writing the manuscript. This work was supported by the German Research Foundation (DFG) and the Technische Universität München within the funding programme Open Access Publishing.

Availability of data and material
The datasets *SPECTF*, *Sonar* and *WBC* in this article are available in the UCI Machines Learning repository, http://archive.ics.uci.edu/ml. Other datasets of this study are available from University Hospital Duisburg-Essen but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of University Hospital Duisburg-Essen.

Author' contributions
UN developed the EFS framework and performed data analyses. JPS, TB, JK, and AC participated in designing the evaluation and in selecting the medical datasets. UN, MR, and DH wrote the manuscript. DH supervised the study. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
[1]Department of Bioinformatics, 94315, Straubing, Germany. [2]University of Applied Science, Weihenstephan-Triesdorf, 85354, Freising, Germany. [3]Wissenschaftszentrum Weihenstephan, Technische Universität München, 85354, Freising, Germany. [4]Department of Gastroenterology and Hepatology, University Hospital, University Duisburg-Essen, 45122, Essen, Germany. [5]Clinic for Cardiology, West German Heart and Vascular Centre Essen, University Hospital, University Duisburg-Essen, 45122, Essen, Germany.

Received: 23 June 2016 Accepted: 27 October 2016
Published online: 18 November 2016

References
1. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Therap. 2001;69(3):89–95.
2. Hall M. Correlation-based feature selection for machine learning. 1999. PhD thesis, Department of Computer Science, Waikato University, New Zealand.
3. Peng H, Long F, Ding C. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27(8):1226–38.
4. Kohavi R, John GH. Wrappers for feature subset selection. Artif Intell. 1997;97:273–324.
5. Jain A, Zongker D. Feature selection: evaluation, application, and small sample performance. IEEE Trans Pattern Anal Mach Intell. 1997;19(2):153–8.
6. He Z, Yu W. Stable feature selection for biomarker discovery. Comput Biol Chem. 2010;34:215–25.
7. 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.
8. Boulesteix AL, Janitza S, Kruppa J, KÃűnig IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdisciplinary Rev Data Mining Knowl Discov. 2012;2(6):493–507.
9. Kuncheva LI. Combining Pattern Classifiers: Methods and Algorithms. Hoboken: John Wiley & Sons; 2004.
10. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
11. Heider D, Hauke S, Pyka M, Kessler D. Insights into the classification of small gtpases. Adv Appl Bioinform Chem. 2010;3:15–24.
12. van den Boom J, Heider D, Martin SR, Pastore A, Mueller JW. 3-phosphoadenosine 5-phosphosulfate (paps) synthases, naturally fragile enzymes specifically stabilized by nucleotide binding. J Biol Chem. 2012;287(21):17645–55.
13. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, Wels M, van Hijum SA. Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? Brief Bioinform. 2013;14(3):315–26.
14. Dybowski JN, Riemenschneider M, Hauke S, Pyka M, Verheyen J, Hoffmann D, Heider D. Improved bevirimat resistance prediction by combination of structural and sequence-based classifiers. BioData Mining. 2011;4:26.

15. Riemenschneider M, Senge R, Neumann U, Hüllermeier E, Heider D. Exploiting hiv-1 protease and reverse transcriptase cross-resistance information for improved drug resistance prediction by means of multi-label classification. BioData Mining. 2016;9:10.
16. Hothorn T, Hornik K, Zeileis A. Party: A Laboratory for Recursive Part(y)itioning. http://CRAN.R-project.org/.
17. Calle M, Urrea V, Boulesteix LA, Malats N. Auc-rf: A new strategy for genomic profiling with random forest. Hum Heredity. 2011;72(2):121–32.
18. Janitza S, Strobl C, Boulesteix AL. An auc-based permutation variable importance measure for random forests. BMC Bioinformatics. 2013;14:119.
19. Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res. 2004;5:1205–24.
20. Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carr G, Marquz JRG, Gruber B, Lafourcade B, Leito PJ, Mnkemller T, McClean C, Osborne PE, Reineking B, Schrder B, Skidmore AK, Zurell D, Lautenbach S. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography. 2013;36(1):27–46.
21. Suzuki N, Olson DH, Reilly EC. Developing landscape habitat models for rare amphibians with small geographic ranges: a case study of siskiyou mountains salamanders in the western usa. Biodiversity Conserv. 2008;17:2197–218.
22. Elith J, Graham CH, Anderson RP, Dudk M, Ferrier S, Guisan A, Hijmans RJ, Huettmann F, Leathwick JR, Lehmann A, Li J, Lohmann LG, Loiselle BA, Manion G, Moritz C, Nakamura M, Nakazawa Y, Overton JM, Townsend Peterson A, Phillips SJ, Richardson K, Scachetti-Pereira R, Schapire RE, Sobern J, Williams S, Wisz MS, Zimmermann NE. Novel methods improve prediction of species distributions from occurrence data. Ecography. 2006;29(2):129–51.
23. Bauer DF. Constructing confidence sets using rank statistics. J Am Stat Assoc. 1972;67:687–90.
24. Breiman L. Bagging predictors. Mach Learn. 1996;24(2):123–40.
25. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1997;55(1):119–39.
26. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York: Springer; 2011.
27. Sing T, Sander O, Beerenwinkel N, Lengauer T. Rocr: visualizing classifier performance in r. Bioinformatics. 2005;21(20):3940–1.
28. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics. 1988;44(3):837–45.
29. Baars T, Neumann U, Jinawy M, Hendricks S, Sowa JP, Klsch J, Riemenschneider M, Gerken G, Erbel R, Heider D, Canbay A. In acute myocardial infarction liver parameters are associated with stenosis diameter. Medicine. 2016;95(6):2807.
30. Sowa JP, Heider D, Bechmann LP, Gerken G, Hoffmann D, Canbay A. Novel algorithm for non-invasive assessment of fibrosis in nafld. PLOS ONE. 2013;8(4):62439.
31. Lichman M. UCI Machine Learning Repository. 2013. http://archive.ics.uci.edu/ml.
32. Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics. 2007;8(25):1–21.
33. Barbosa E, Rttger R, Hauschild AC, Azevedo V, Baumbach J. On the limits of computational functional genomics for bacterial lifestyle prediction. Brief Funct Genomics. 2014;13:398–408.
34. Sowa JP, Atmaca, Kahraman A, Schlattjan M, Lindner M, Sydor S, Scherbaum N, Lackner K, Gerken G, Heider D, Arteel GE, Erim Y, Canbay A. Non-invasive separation of alcoholic and non-alcoholic liver disease with predictive modeling. PLOS ONE. 2014;9(7):101444.
35. Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl Inform Syst. 2007;12:95–116.
36. Mucciardi AN, Gose EE. A comparison of seven techniques for choosing subsets of pattern recognition properties. IEEE Trans Comput. 1971;9:1971910231031.
37. Almuallim H, Dietterich TG. Learning with many irrelevant features. 1991547–552. Proceedings of the Ninth National Conference on Artificial Intelligence,San Jose, CA: AAAI Press.
38. Doak J. An evaluation of feature-selection methods and their application to computer security (technical report cse-92-18). Davis: University California, Department of Computer Science. 1992.
39. Caruana RA, Freitag D. Greedy attribute selection. In: Proceedings of the Eleventh Inter-national Conference on Machine Learning. New Brunswick, NJ: Morgan Kaufmann; 1994. p. 28–36.
40. Kononenko I. On biases in estimating multi-valued attributes. Montreal; 1995. p. 1034–1040.
41. Blum AL, Langleyb P. Selection of relevant features and examples in machine learning. Artif Intell. 1997;97(1–2): 245–71.
42. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics. 2010;26(3):392–8.
43. Piao Y, Piao M, Park K, Ho Ryu K. An ensemble correlation-based gene selection algorithm for cancer classification with gene expression data. Bioinformatics. 2012;28(24):3306–15.
44. van Landeghem S, Abeel T, Saeys Y, van de Peer Y. Discriminative and informative features for biomolecular text mining with ensemble feature selection. Bioinformatics. 2010;26(18):554–60.
45. Bagley SC, Whiteb H, Golomb BA. Logistic regression in the medical literature:standards for use and reporting, with particular attention to one medical domain. J Clin Epidemiol. 2001;54:979–85.

B.3. Paper III

Neumann *et al. BioData Mining* (2017) 10:21
DOI 10.1186/s13040-017-0142-8

BioData Mining

SOFTWARE ARTICLE Open Access

EFS: an ensemble feature selection tool implemented as R-package and web-application

Ursula Neumann[1,2,3], Nikita Genze[1] and Dominik Heider[1,2,3*]

*Correspondence:
d.heider@wz-straubing.de
[1]Straubing Center of Science, Schulgasse 22, 94315 Straubing, Germany
[3]Wissenschaftszentrum Weihenstephan, Technische Universität München, 85354 Freising, Germany
Full list of author information is available at the end of the article

Abstract

Background: Feature selection methods aim at identifying a subset of features that improve the prediction performance of subsequent classification models and thereby also simplify their interpretability. Preceding studies demonstrated that single feature selection methods can have specific biases, whereas an ensemble feature selection has the advantage to alleviate and compensate for these biases.

Results: The software EFS (Ensemble Feature Selection) makes use of multiple feature selection methods and combines their normalized outputs to a quantitative ensemble importance. Currently, eight different feature selection methods have been integrated in EFS, which can be used separately or combined in an ensemble.

Conclusion: EFS identifies relevant features while compensating specific biases of single methods due to an ensemble approach. Thereby, EFS can improve the prediction accuracy and interpretability in subsequent binary classification models.

Availability: EFS can be downloaded as an R-package from CRAN or used via a web application at http://EFS.heiderlab.de.

Keywords: Machine learning, Feature selection, Ensemble learning, R-package

Background

In the field of data mining, feature selection (FS) has become a frequently applied pre-processing step for supervised learning algorithms, thus a great variety of FS techniques already exists. They are used for reducing the dimensionality of data by ranking features in order of their importance. These orders can then be used to eliminate those features that are less relevant to the problem at hand. This improves the overall performance of the model because it addresses the problem of overfitting. But there are several reasons that can cause instability and unreliability of the feature selection, e.g., the complexity of multiple relevant features, a small-n-large-p-problem, such as in high-dimensional data [1, 2], or when the algorithm simply ignores stability [3, 4]. In former studies, it has been demonstrated that a single optimal FS method cannot be obtained [5]. For example, the Gini-coefficient is widely used in predictive medicine [6, 7], but it has also been demonstrated to deliver unstable results in unbalanced datasets [8, 9]. To counteract instability and therewith unreliability of feature selection methods, we developed an FS procedure for binary classification, which can be used, e.g., for random clinical trials. Our new

approach ensemble feature selection (EFS) [10] is based on the idea of ensemble learning [11, 12], and thus is based on the aggregation of multiple FS methods. Thereby a quantification of the importance scores of features can be obtained and the method-specific biases can be compensated. In the current paper we introduce an R-package and a web server based on the EFS method. The user of the R-package as well as the web application can decide which FS methods should be conducted. Therewith, the web server and the R-package can be applied to perform an ensemble of FS methods or to calculate an individual FS score.

Implementation

We used existing implementations in **R** (http://www.r-project.org/) for our package EFS. The following section will briefly introduce our methodology. For deeper insights please refer to [10]. Our EFS currently incorporates eight feature selection methods for binary classifications, namely median, Pearson- and Spearman-correlation, logistic regression, and four variable importance measures embedded in two different implementations of the random forest algorithm, namely *cforest* [9] and *randomForest* [13].

Median

This method compares the positive samples (class = 1) with negative samples (class = 0) by a Mann-Whitney-U Test. The resulting p-values are used as a measure of feature importance. Thus, a smaller p-value indicates a higher importance.

Correlation

We used the idea of a fast correlation based filter of of Yu and Liu [14] to select features that are highly correlated with the dependent variable, but show only low correlation with other features. The fast correlation based filter eliminates features with high correlation with other features to avoid multicollinearity. The eliminated features get an importance value of zero. Two correlation coefficients, namely the Pearson product-moment and the Spearman rank correlation coefficient were adopted and their p-values were used as importance measure.

Logistic regression

The weighting system (i.e., β-coefficients) of the logistic regression (LR) is another popular feature selection method. As preprocessing step a Z-transformation is conducted to ensure comparability between the different ranges of feature values. The β-coefficients of the resulting regression equation represent the importance measure.

Random forest

Random forests (RFs) are ensembles of multiple decision trees, which gain their randomness from the randomly chosen starting feature for each tree. There are different implementations of the RF algorithm in R available, which offer diverse feature selection methods. On the one hand we incorporated the *randomForest* implementation based on the classification and regression tree (CART) algorithm by Breiman [13]. The *cforest* implementation from the party package, on the other hand, uses conditional trees for the purpose of classification and regression (cf. [15]). In both implementations an error-rate-based importance measure exists. The error-rate-based methods measure the difference

before and after permuting the class variable. Due to their dependency on the underlying trees, results are varying for both error-rates. The *randomForest* approach also provides an importance measure based on the Gini-index, which measures the node impurity in the trees. Whereas in *cforest* an AUC-based variable importance measure is implemented. The AUC (area under the curve) is the integral of the receiver operating characteristics (ROC) curve. The AUC-based variable importance measure works to the error-rate-based one, but instead of computing the error rate for each tree before and after permuting a feature, the AUC is computed.

Ensemble learning

The results of each individual FS methods are normalized to a common scale, an interval from 0 to $\frac{1}{n}$, where n is the number of conducted FS methods chosen by the user. Thereby we ensure the comparability of all FS methods and conserve the distances between the importance of one feature to another.

R-package

The EFS package is included in the Comprehensive R Archive Network (CRAN) and can be directly downloaded and installed by using the following R command:

```
install.packages("EFS")
```

In the following, we introduce EFS's three functions `ensemble_fs`, `barplot_fs` and `efs_eval`. A summary of all commands and parameters is shown in Table 1.

Table 1 Method overview

Command	Parameters	Information
ensemble_fs	data	object of class data.frame
	classnumber	index of variable for binary classification
	NA_threshold	threshold for deletion of features with a greater proportion of NAs
	cor_threshold	correlation threshold within features
	runs	amount of runs for randomForest and cforest
	selection	selection of feature selection methods to be conducted
barplot_fs	name	character string giving the name of the file
	efs_table	table object of class matrix retrieved from ensemble_fs
efs_eval	data	object of class data.frame
	efs_table	table object of class matrix retrieved from ensemble_fs
	file_name	character string, name which is used for the two possible PDF files.
	classnumber	index of variable for binary classification
	NA_threshold	threshold for deletion of features with a greater proportion of NAs
	logreg	logical value indicating whether to conduct an evaluation via logistic regression or not
	permutation	logical value indicating whether to conduct a permutation of the class variable or not
	p_num	number of permutations; default set to a 100
	variances	logical value indicating whether to calculate the variances of importances retrieved
		from bootstrapping or not
	jaccard	logical value indicating whether to calculate the Jaccard-index or not
	bs_num	number of bootstrap permutations of the importances
	bs_percentage	proportion of randomly selected samples for bootstrapping

The R-package EFS provides three functions

`ensemble_fs`

The main function is `ensemble_fs`. It computes all FS methods which are chosen via the `selection` parameter and gives back a table with all normalized FS scores in a range between 0 and $\frac{1}{n}$, where n is the number of incorporated feature selection methods. Irrelevant features (e.g., those with too many missing values) can be deleted.

```
ensemble_fs(data, classnumber,
            NA_threshold, cor_threshold,
            runs, selection)
```

The parameter `data` is an object of class `data.frame`. It consists of all features and the class variables as columns. The user has to set the parameter `classnumber`, which represents the column number of the class variable, i.e., the dependent variable for classification. `NA_threshold` represents a threshold concerning the allowed proportion of missing values (NAs) in a feature column. The default value is set to 0.2, meaning that features with more than 20% of NAs are neglected by the EFS algorithm. The `cor_threshold` parameter is only relevant for the correlation based filter methods. It determines the threshold of feature-to-feature correlations [14]. The default value of `cor_threshold` is 0.7. The results of RF-based FS methods vary due to the randomness of their underlying algorithms. To obtain reliable results, the RF methods are conducted several times and averaged over the number of runs. This parameter, namely `runs`, is set to 100 by default. The user can select the FS methods for the EFS approach by using the `selection` parameter. Due to the high computational costs of the RFs, the default selection is set to

```
selection = c(TRUE, TRUE, TRUE, TRUE, TRUE, TRUE, FALSE, FALSE),
```

meaning that the two FS methods of the conditional random forest are not used by default.

`barblot_fs`

The `barblot_fs` function sums up all individual FS scores based on the output of `ensemble_fs` and visualizes them in an cumulative barplot.

```
# Create a cumulative barplot based on the output from EFS
barplot_fs(name, efs_table)
```

The `barplot_fs` function uses the output of the `ensemble_fs` function, namely the `efs_table`, as input. The parameter `name` represents the filename of the resulting PDF, which is saved in the current working directory.

`efs_eval`

The `efs_eval` function provides several tests to evaluate the performance and validity of the EFS method. The parameters `data`, `efs_table`, `file_name`, `classnumber` and `NA_threshold` are identical to the corresponding parameters in the `ensemble_fs` function:

```
efs_eval(data, efs_table, file_name,
         classnumber , NA_threshold,
         logreg = TRUE,
         permutation = TRUE, p_num,
         variances = TRUE, jaccard = TRUE,
         bs_num, bs_percentage).
```

Performance evaluation by logistic regression

The performance of the EFS method can automatically be evaluated based on a logistic regression (LR) model, by setting the parameter `logreg = TRUE`. `efs_eval` uses an LR model of the selected features with a leave-one-out cross-validation (LOOCV) scheme, and additionally trains an LR model with all available feature in order to compare the two LR models based on their ROC curves and AUC values with ROCR [16] and pROC based on the method of DeLong et al. [17]. A PDF with the ROC curves is automatically saved in the working directory.

Permutation of class variable

In order to estimate the robustness of the resulting LR model, permutation tests [18, 19] can be automatically performed, by setting the parameter `permutation = TRUE`. The class variable is randomly permuted `p_num` times and logistic regression is conducted. The resulting AUC values are then compared with the AUC from the original LR model using a Student's t-Test. By default, `p_num` is set to 100 permutations.

Variance of feature importances

If the parameter `variances` is `TRUE` an evaluation of the stability of feature importances will be conducted by a bootstrapping algorithm. The samples are permuted for `bs_num` times and a subset of the samples (`bs_percentage`) is chosen to calculate the resulting feature importances. By default, the function chooses 90% of the samples and uses 100 repetitions. Finally, the variances of the importance values are reported.

Jaccard-index

The Jaccard-index measures the similarity of the feature subsets selected by permuted EFS iterations:

$$J(S_1, \ldots, S_n) = \frac{|S_1 \cap \ldots \cap S_n|}{|S_1 \cup \ldots \cup S_n|},$$

where S_i is the subset of features at the i-th iteration, for $i = 1, \ldots, n$. The value of the Jaccard-index varies from 0 to 1, where 1 implies absolute similarity of subsets. If `jaccard = TRUE` is set, the Jaccard-index of the subsets retrieved from the bootstrapping algorithm is calculated.

Availability and requirements

The package is available for R-users under the following requirements:

- **Project name:** Ensemble Feature Selection
- **Project home page (CRAN):** http://cran.r-project.org/web/packages/EFS
- **Operating system (s):** Platform independent
- **Programming language:** R ($\geq$ 3.0.2)
- **License:** GPL ($\geq$ 2)
- **Any restrictions to use by non-academics:** none

Due to the high relevance of our EFS tool for researchers who are not very familiar with R (e.g., medical practitioners), we also provide a web application at

http://EFS.heiderlab.de. It contains the functions `ensemble_fs` and `barplot_fs`. Therefore no background knowledge in R is necessary to use our new EFS software.

Results

The dataset SPECTF has been obtained from the UCI Machine Learning Repository [20] and is used as an example. It describes diagnosing of cardiac Single Proton Emission Computed Tomography (SPECT) images. The class-variable represents normal (= 0) and abnormal (= 1) results and can be found in the first column of the table of the file SPECTF.csv at the UCI repository. In general, the EFS approach accepts all types of data, i.e., all types of variables, except categorical variables. These variables have to be transformed to dummy variables in advance. Data has to be combined in a single file with one column indicating the class variable with 1 and 0, e.g., representing patients and control samples, or, e.g., positive and negative samples. After loading the dataset, we compute the EFS and store it in the variable "efs":

```
library(EFS)
# Loading dataset in environment
efsdata <- read.table("SPECTF.csv", sep = ";")
# Start feature selection
efs <- ensemble_fs(data = efsdata, classnumber = 1,
                   NA_threshold = 0.2, cor_threshold = 0.7,
                   runs = 100, selection = rep(TRUE, 8))
```

The results can be visualized by the `barplot_fs` function:

```
# Create a cumulative barplot based on the output from efs
barplot_fs("SPECTF", efs)
```

The output is a PDF named "SPECTF.pdf". Figure 1 shows this cumulative barplot, where each FS method is given in a different color. Various methods to evaluate the stability and reliability of the EFS results are conducted by the following command:

```
# Create a ROC Curve based on the output from efs
eval_tests <- efs_eval(data = efs_data, efs_table = efs,
                       file_name = "SPECTF",
                       classnumber = 1, NA_threshold = 0.2,
                       logreg = TRUE,
                       permutation = TRUE, p_num = 100,
                       variances = TRUE, jaccard = TRUE,
                       bs_num = 100, bs_percentage = 0.9)
```

The user retrieves two PDF files. Firstly, the resulting ROC curves of the LR test ("SPECTF_ROC.pdf") including the p-value, according to Fig. 2. The p-value clearly shows that there is a significant improvement in terms of AUC of the LR with features selected by the EFS method compared the LR model without feature selection. Additionally, Fig. 3 shows the file "SPECTF_Variances.pdf", in which boxplots of the importances retrieved from the bootstrapping approach are given. The calculated variances can be accessed in the eval_tests output. A low variance implies that the importance of a feature is stable and reliable.

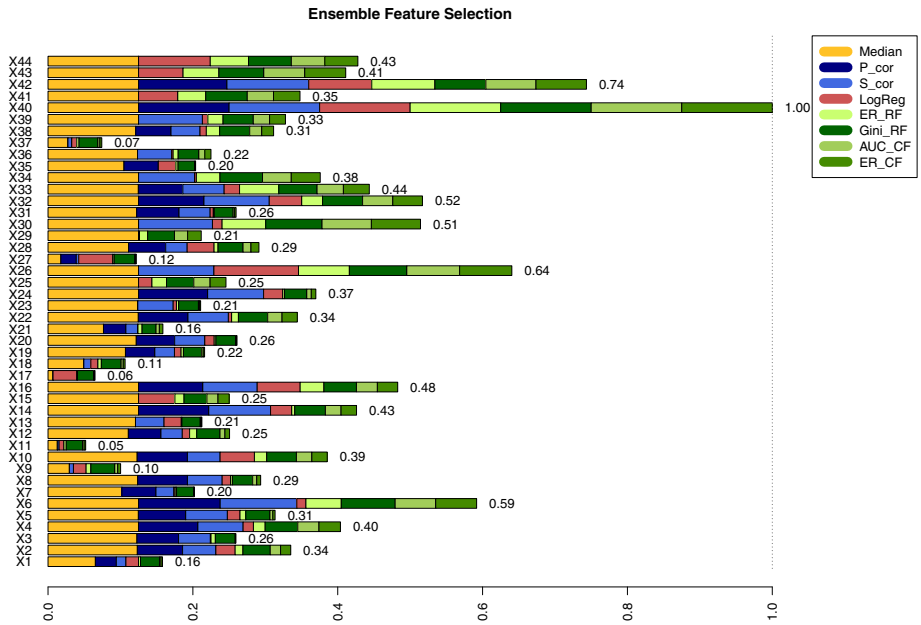

Fig. 1 Cumulative barplot retrieved from `barplot_fs` function of R-package EFS

An additional example is provided in the documentation of the R-package on a dataset consisting of weather data from the meteorological stations in Frankfurt(Oder), Germany in February 2016.

Conclusion

The EFS R-package and the web-application are implementations of an ensemble feature selection method for binary classifications. We showed that this method can improve the prediction accuracy and simplifies the interpretability by feature reduction.

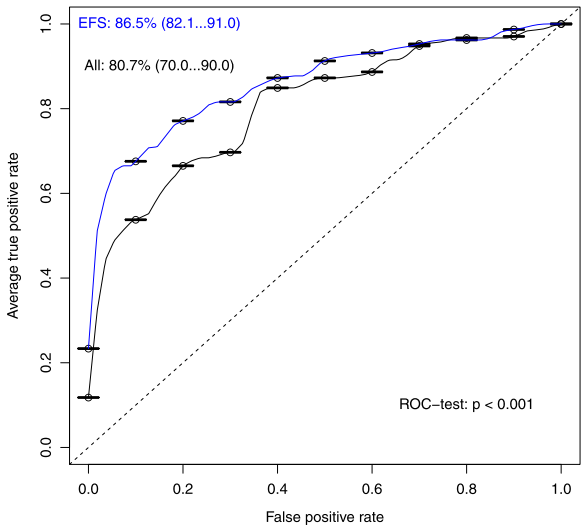

Fig. 2 Performance of LR model. On the *y-axis* the average true positive rate (i.e., sensitivity) and on the *x-axis* the false positive rate (i.e., 1-specificity) is shown. Two ROC curves are shown: of all features (*black*) and the EFS selected features (*blue*). The *dotted line marks* the performance of random guessing

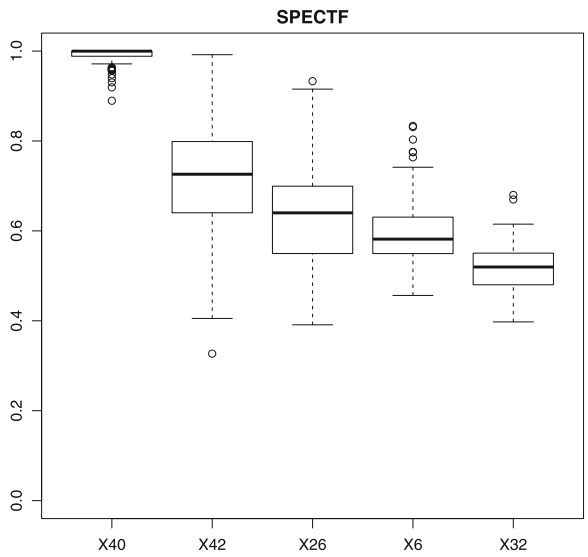

Fig. 3 Boxplot of importances retrieved from the bootstrapping algorithm

Abbreviations
AUC: Area under the curve; CART: Classification and regression tree; CRAN: Comprehensive R archive network; EFS: Ensemble feature selection; FS: Feature selection; LR: logistic regression; LOOCV: leave-one-out cross validation; RF: random forest; ROC: receiver operating characteristic

Acknowledgments
We are grateful to the UCI Machine Learning Repository for granting access to a great variety of datasets.

Funding
This work was supported by the German Research Foundation (DFG) and the Technische Universität München within the funding program Open Access Publishing and the Deichmann Foundation, which had no role in study design, collection, analysis, and interpretation of data, and in writing the manuscript.

Availability of data and materials
The dataset *SPECTF* in this article is available in the UCI Machines Learning repository, http://archive.ics.uci.edu/ml.

Authors' contributions
UN and NG have implemented the R-package. UN has implemented the web application and drafted the manuscript. DH designed and supervised the study. DH revised the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Publisher's Note

Author details
[1]Straubing Center of Science, Schulgasse 22, 94315 Straubing, Germany. [2]University of Applied Science, Weihenstephan-Triesdorf, 85354 Freising, Germany. [3]Wissenschaftszentrum Weihenstephan, Technische Universität München, 85354 Freising, Germany.

Received: 23 November 2016 Accepted: 12 June 2017
Published online: 27 June 2017

References

1. Dybowski JN, Heider D, Hoffmann D. Structure of hiv-1 quasi-species as early indicator for switches of co-receptor tropism. AIDS Res Ther. 2010;7:41.
2. Pyka M, Hahn T, Heider D, Krug A, Sommer J, Kircher T, Jansen A. Baseline activity predicts working memory load of preceding task condition. Hum Brain Mapp. 2013;34(11):3010–22.
3. Jain A, Zongker D. Feature selection: evaluation, application, and small sample performance. IEEE Trans Pattern Anal Mach Intell. 1997;19(2):153–8.
4. He Z, Yu W. Stable feature selection for biomarker discovery. Comput Biol Chem. 2010;34:215–25.
5. Yang YHY, Xiao Y, Segal MR. Identifying differentially expressed genes from microarray experiments via statistic synthesis. Bioinformatics. 2004;21(7):1084–93.
6. Leclerc A, Lert F, Cecile F. Differential mortality: Some comparisons between england and wales, finland and france, based on inequality measures. Int J Epidemiol. 1990;19(4):1001–10.
7. Llorca J, Delgado-Rodríguez M. Visualising exposure-disease association: the lorenz curve and the gini index. Med Sci Monit. 2002;8(10):193–7.
8. 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.
9. Boulesteix AL, Janitza S, Kruppa J, König IR. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. Wiley Interdiscip Rev Data Min Knowl Discov. 2012;2(6):493–507.
10. Neumann U, Riemenschneider M, Sowa JP, Baars T, Kälsch J, Canbay A, Heider D. Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection. BioData Min. 2016;9(1):36.
11. Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer-Verlag; 2008. p. 313–25.
12. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics. 2010;26(3):392–8.
13. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
14. Yu L, Liu H. Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res. 2004;5:1205–24.
15. Strobl C, Boulesteix AL, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forest. BMC Bioinforma. 2006;9(307):1–11.
16. Sing T, Sander O, Beerenwinkel N, Lengauer T. Rocr: visualizing classifier performance in r. Bioinformatics. 2005;21(20):3940–41.
17. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics. 1988;44(3):837–45.
18. Barbosa E, Röttger R, Hauschild AC, Azevedo V, Baumbach J. On the limits of computational functional genomics for bacterial lifestyle prediction. Brief Funct Genomics. 2014;13:398–408.
19. Sowa JP, Atmaca Ö, Kahraman A, Schlattjan M, Lindner M, Sydor S, Scherbaum N, Lackner K, Gerken G, Heider D, Arteel GE, Erim Y, Canbay A. Non-invasive separation of alcoholic and non-alcoholic liver disease with predictive modeling. PLOS ONE. 2014;9(7):101444.
20. Lichman M. UCI Machine Learning Repository. 2013. http://archive.ics.uci.edu/ml.

B.4. Paper IV

Assessment of Subset Selection Criteria of Quantitative Feature Selection Methods

Ursula Neumann and Dominik Heider

Straubing Center of Science,
Petersgasse 18, 94315 Straubing, Germany
Department of Mathematics and Computer Science, University of Marburg
Hans-Meerwein-Str. 6, 35032 Marburg, Germany
u.neumann@gmx.de,d.heider@uni-marburg.de

Motivation

Feature selection plays a crucial role as a preprocessing step of data mining. Former studies revealed the importance of feature selection methods to improve the performance and efficiency of algorithms in pattern recognition, classification, and regression [1, 2]. The methods are designed to distinguish features which are relevant for a prediction model from those which are negligible. By that, an efficient subset selection criterion is indispensable to set a cutpoint between relevant and irrelevant features. In general, we can distinguish between quantitative and qualitative feature rankings. The latter provide a binary decision as output, classifying the features to be relevant or redundant. In contrast, a quantitative feature ranking estimates the relevance for each feature as a quantitative value. Ordering these values in an ascending order an importance curve is obtained. Based on this ranking, a cutpoint must be identified, which selects a subset of highly relevant features. In this work, we used the ensemble feature selection (EFS)[4] as underlying feature ranking approach.EFS uses the mean as an integrated subset selection criterion. For small datasets, the EFS method outperforms each single method [3]. However, in large datasets, i.e., with more than 1000 features, smaller subsets seem to be less prone to overfitting and thus lead to better prediction results in subsequent prediction models. There exist also more conservative subset selection criteria, such as selecting the best 15% or the best 10% of all features. Another method for finding a suitable subset is the detection of the cutpoint with the highest slope, i.e., the point with the highest increase of importance. For datasets with an exponential curve of importance values, this method selects only the feature with the highest importance. Based on these findings, we developed a feature subset selection method: the $\frac{\pi}{4}$-rotation.

Methods

For the assessment of subset selection methods we chose four different approaches. First, the mean value of the feature's relevance value, which is integrated in the EFS ranking. It identifies the features above-average as relevant,

which is most common and a liberal cutpoint. Unfortunately, many redundant features have importance values below average, which levels down the mean. Besides the mean, we evaluated two percentage-based cutoffs, namely the best 15% and the best 10% features. Moreover, we used the $\frac{\pi}{4}$-rotation method, which is a more sophisticated method. The idea behind the $\frac{\pi}{4}$-rotation is to draw a curve of the ascending importance values received from the EFS algorithm (c.f. panel A) and B) of figure 1) and to detect the cutpoint where the smoothed curve exceeds a slope of 45 degrees. Features which lie above this cutpoint are considered to be important.

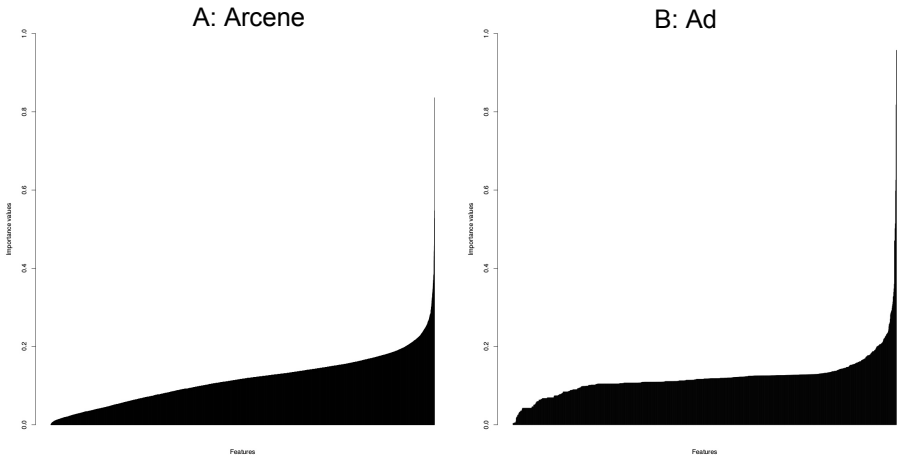

Fig. 1. Feature importance values in ascending order of A) Arcene dataset and B) Ad dataset.
ROC curves of logistic regression models with all features and features with importance values over the mean by EFS of C) Arcene dataset and D) Ad dataset.

It might happen that there are several points in which the slope fulfills the requirement to be over 45 degree. Therefore, we rotate the curve by -45 degrees (i.e., $-\frac{\pi}{4}$ in radians) and seek for a global minimum. If the minimum is the last variable of the range ordered by ascending importance, no distinct leap at the curve of importance values exists. In this case, the $\frac{\pi}{4}$-rotation method cannot be applied.

Results

We analyzed two big datasets *Ad* and *Arcene* with 1430 and 79360 features received from the UCI Machine Learning Repository [5]. Both possess an exponential curve of importance values. The results from the evaluation via ROC curves of logistic regression models as well as random forest models are shown

in tables 1 and 2. It turned out that the pi/4-rotation is the most conservative method, i.e., it selects the smallest number of features. Its subset selection has a high performance in both datasets. The highest AUCs with the logistic regression models based on the Arcene dataset are found with the percentage-based criteria. In comparison, the random forest results showed the best AUC with the $\frac{\pi}{4}$-rotation, but not significantly higher than the AUC obtained from the mean-selection (p-value = 0.164). In the Ad dataset, the AUCs of the logistic regression models are negatively correlated with the number of selected features. In case of the random forest model, there are no significant differences in the AUCs.

Significance was calculated with a roc-test by the method of DeLong et al. [6], which compares the ROC curves retrieved from the mean subset with the ROC curve from the $\frac{\pi}{4}$-rotation cutpoint subset.

Table 1. Evaluation of Arcene data. AUC values of a logisic regression model and a random forest model.

Method	AUC from LR	AUC from RF	Nr of features
Mean	66.7%(60.0...80.0)	89.9%(80.0...100.0)	5038
best 15%	79.9%(70.0...90.0)	91.9%(90.0...100.0)	1488
best 10%	79.9%(70.4...89.4)	90.9%(85.2...96.6)	992
$\frac{\pi}{4}$-rotation	62.5%(51.5...73.6)	92.5%(87.3...97.8)	374

Table 2. Evaluation of Ad data. AUC values of a logisic regression model and a random forest model.

Method	AUC from LR	AUC from RF	Nr of features
Mean	55.2%(50.0...60.0)	98.4%(100.0...100.0)	613
best 15%	69.9%(70.0...70.0)	98.4%(100.0...100.0)	214
best 10%	91.9%(89.6...94.1)	98.5% (97.6...99.3)	143
$\frac{\pi}{4}$-rotation	95.2%(93.5...97.0)	98%(97.1...98.9)	53

References

1. Kohavi R., John G.H.: Wrappers for feature subset selection. Artificial Intelligence, 97(1), 273-324 (1997)
2. Chandrashekar G., Sahin F.: A survey on feature selection methods. Computers & Electrical Engineering 40(1), 16-28 (2014)
3. Neumann U., Riemenschneider M., Sowa J.P., Baars T., Kaelsch J., Canbay A., Heider D.: Compensation of feature selection biases accompanied with improved predictive performance for binary classification by using a novel ensemble feature selection approach. Biodata Mining, 9,36 (2016)
4. Neumann U., Genze N., Heider D.: EFS: an ensemble feature selection tool implemented as R-package and web-application. Biodata Mining, 10,21 (2017)
5. Lichman, M.: UCI Machine Learning Repository (2013), http://archive.ics.uci.edu/ml
6. DeLong ,E.R., DeLong, D.M., Clarke-Pearson, D.L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, 44, 837-845 (1988)