Monitoring forest health using hyperspectral imagery: Does feature selection improve the performance of machine-learning techniques?

Monitoring forest health using hyperspectral imagery: Does feature selection improve the performance of machine-learning techniques?

Monitoring forest health using hyperspectral imagery: Does feature selection improve the performance of machine-learning techniques?

Patrick Schratz, Jannes Muenchow,
Eugenia Iturritxa, José Cortés, Bernd Bischl,
and Alexander Brenning

***Abstract*—This study analyzed highly-correlated, feature-rich datasets from hyperspectral remote sensing data using multiple machine and statistical-learning methods. The effect of filter-based feature-selection methods on predictive performance was compared. Also, the effect of multiple expert-based and data-driven feature sets, derived from the reflectance data, was investigated. Feature wise, defoliation at trees (%) was modeled as a function of band reflectance. Variable importance was assessed using permutation-based feature importance.**

With respect to the used models, support vector machine (SVM) outperformed others such as random forest (RF), extreme gradient boosting (XGBoost), lasso (L1) and ridge (L2) regression by at least three percentage points. The combination of certain feature sets showed small increases in predictive performance while no substantial differences between single feature sets were observed. Filter methods helped in certain dataset-learner-filter settings to achieve better predictive performances while ensemble filters did not have a substantial impact on performance. For certain dataset-learner combinations, filter methods helped to improve predictive performance and lower computational runtime.

Permutation-based feature importance estimated features around the red edge to be most important for the models. However, the presence of features in the near-infrared region (800 nm - 1000 nm) was essential to achieve the best performances.

The dataset used in this work was suboptimal with respect to the achieved predictive performances and more training data is needed for more generalizable conclusions. Filter methods have potential to help in high-dimensional situations while still keeping the ability to interpret feature effects of fitted models, which is an essential constrain in environmental modeling studies.

***Index Terms*—hyperspectral imagery, forest health monitoring, machine learning, feature selection, feature effects, model comparison, filter, imaging spectroscopy**

P.Schratz, J.Muenchow, J.Cortés and A.Brenning are with the Department of Geography, GIScience group, Friedrich-Schiller-University of Jena, Germany.

B.Bischl is head of the computational statistics group at the Department of Statistics, Ludwig-Maximilian-University Munich.

E.Iturritxa is with NEIKER Tecnalia, Vitoria-Gasteiz, Arab, Spain.

Monitoring forest health using hyperspectral imagery: Does feature selection improve the performance of machine-learning techniques?

I. Introduction

The use of machine learning (ML) algorithms for analyzing remote sensing data has seen a huge increase in the last decade [1]. This goes in line with the increased availability of remote sensing imagery, especially since the launch of the first Sentinel satellite in the year 2014. At the same time, the implementation and usability of learning algorithms has been greatly simplified with many contributions from the open-source community. Scientists can nowadays relatively easily process large amounts of (environmental) information using various learning algorithms. This makes it possible to extend the benchmark comparison matrix of studies in a semi-automated way, possibly stumbling across unexpected findings of process settings that would have never been tested otherwise [2].

Machine learning methods in combination with remote sensing data are used in many environmental fields such as vegetation cover analysis or forest carbon storage mapping [3], [4]. The ability of predicting to large unknown areas qualifies these tools as a promising toolset for such tasks. One aspect of this research field is to enhance the understanding of biotic and abiotic triggers, for example by analyzing defoliation at trees [5].

Other approaches for analyzing forest health include temporal change detection [6] or describing the current health status of forests on a stand level [7]. In such studies, the defoliation of trees serves as a proxy for forest health by describing the impact of biotic and abiotic pest triggers [7], [8].

Vegetation indices have shown the potential to provide valuable information when analyzing forest health [9], [10]. Most vegetation indices were developed with the aim of being sensitive to changes of specific wavelength regions, serving as a proxy for underlying plant processes. However, often enough indices developed for different purposes than the one to be analyzed can help to explain complex relationships. This emphasizes the need to extract as much information as possible from the available input data to generate promising features which can help to understand the modeled relationship. A less known index type which can be derived from spectral information is the normalized ratio index (NRI). In contrast to most vegetation indices, NRIs do not use an expert-based formula following environmental heuristics but instead makes use of a data-driven feature engineering approach by combining (arbitrary) combinations of spectral bands. Especially when working with hyperspectral data, thousands of NRI features can be derived this way.

Despite its popularity in environmental modeling,

there are no studies so far which used machine-learning algorithms in combination with remote sensing data to analyze defoliation on a tree level. This study aims to close this gap by analyzing defoliation at trees in northern Spain using airborne hyperspectral data. The methodology of this study uses of machine-learning methods in combination with feature selection and hyperparameter tuning. In addition, feature importance and feature effects are evaluated. Incorporating the idea of creating data-driven NRIs, this study also discusses the practical problems of high-dimensionality in environmental modeling [11], [12].

Even though ML algorithms are capable of handling highly-correlated input variables, model fitting becomes computationally more demanding, and model interpretation more complex. Feature selection approaches can help to address this issue, reducing possible in the feature space, simplify model interpretability and possibly enhance predictive performance [13].

This study shows how high-dimensional datasets can be handled effectively with machine-learning methods while still being conduct inference on the fitted models. The predictive power of non-linear methods and their ability to handle highly-correlated predictors is combined with common and new approaches for assessing feature importance and feature effects. However, this study clearly focuses on investigating the effects of filter methods and feature set types on predictive performance rather than on interpreting feature effects.

That said, the research questions of this study are the following:

- Do different (environmental) feature sets show differences in performance when modeling defoliation at trees?
- Can predictive performance be substantially improved by combining feature sets?
- How are feature-selection methods influencing the predictive performance of the models?
- Which features are most important and how can these be interpreted in an environmental context?

II. Data and study area

Airborne hyperspectral data with a spatial resolution of one meter and 126 spectral bands was available for four Monterey Pine (*Pinus radiata*) plantations in northern Spain. The trees in the plots suffer from infections of invasive pathogens such as *Diplodia sapinea*, *Fusarium circinatum*, *Armillaria mellea* or *Heterobasidion annosum*, leading to a spread of cankers or defoliation [14], [15]. In-situ measurements of defoliation at trees (serving as a proxy for tree health) were collected to serve as the response variable *defoliation* which ranges from 0 - 100 (in %) (Figure 1). It is assumed that the fungi infect the trees through open wounds, possibly caused by previous hail damage [16]. The dieback of these trees, which are mainly used as timber, causes high economic damages [17].

A. In-situ data

The *Pinus radiata* plots of this study, namely Laukiz1, Laukiz2, Luiando and Oiartzun, are located in the northern part of the Basque Country (Figure 2). Oiartzun has the most observations (n = 559) while Laukiz2 shows the largest area size (1.44 ha). All plots besides Luiando are located nearby the coast (Figure 2). In total 1808 observations are available Laukiz1 = 559, Laukiz2 = 451, Luiando = 301, Oiartzun = 497). The data was surveyed in September 2016.

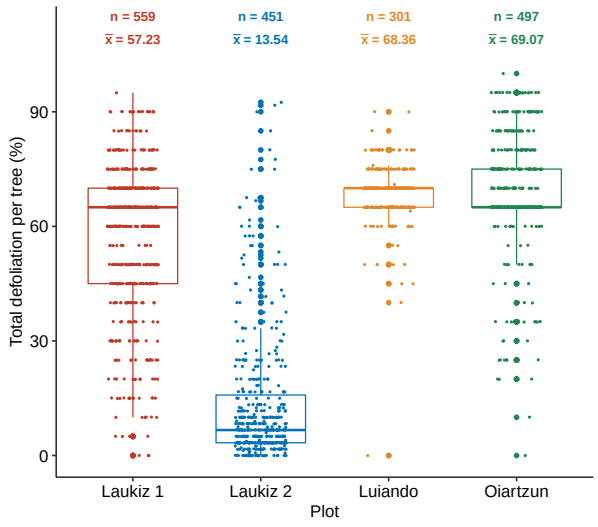

Fig. 1. Response variable *defoliation* at trees for plots Laukiz1, Laukiz2, Luiando and Oiartzun. n corresponds to the total number of trees in the plot, $\bar{x}$ refers to the mean defoliation.

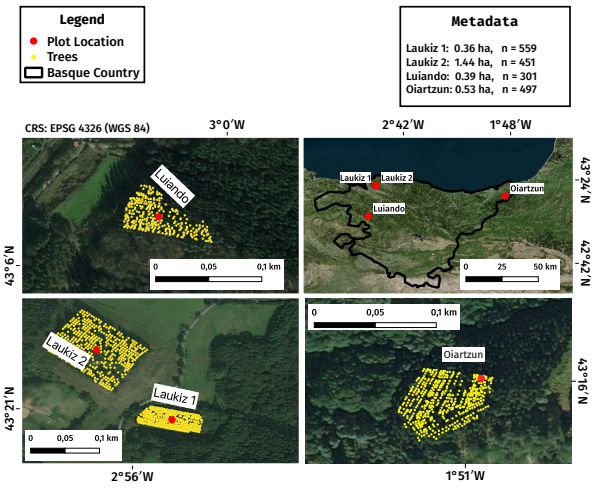

Fig. 2. Information about location, size and spatial distribution of trees for all plots used in this study.

B. Hyperspectral data

The airborne hyperspectral data was acquired during two flight campaigns on September 28th and October 5th 2016, both around noon. The images were taken by an AISAEAGLE-II sensor. All preprocessing steps (geometric, radiometric, atmospheric) were conducted by the Institut Cartografic I Geologic de Catalunya (ICGC). The first four bands were corrupted, leaving 122 bands with valid information. Additional metadata information is available in Table VIII.

TABLE I
SPECIFICATIONS OF HYPERSPECTRAL DATA.

Characteristic	Value
Geometric resolution	1 m
Radiometric resolution	12 bit
Spectral resolution	126 bands (404.08 nm — 996.31 nm)
Correction:	Radiometric, geometric, atmospheric

III. METHODS

A. Derivation of indices

To use the full potential of the hyperspectral data, all possible vegetation indices supported by the R package hsdar (89 in total) as well as all possible NRI combinations were calculated from the reflectances. The following formula was used for the NRI calculation:

$$NRI_{i,j} = \frac{b_i - b_j}{b_i + b_j} \quad (1)$$

where i and j are the respective band numbers.

To account for geometric offsets within the hyperspectral data, which were reported with up to 1 m from ICGC, a buffer of two meters around the centroid of each tree was used when extracting the reflectance values. A pixel was considered to fall into a tree's buffer zone if the centroid of the respective pixel was touched by the buffer. All of those pixels formed the final reflectance value of a single tree and were used as the base information to derive all additional feature sets. In total, $\frac{121*122}{2} = 7471$ NRIs

were calculated due to the first four bands of the sensor being corrupted.

B. Feature selection

The case of a feature-rich dataset comes with several challenges for both model fitting and evaluation.

- Model fitting times increase.
- Noise is possibly introduced into models by highly-correlated variables [18].
- Model interpretation and prediction become more challenging [18].

To reduce the feature space of a dataset, conceptually differing approaches exist: wrapper methods, filters, penalization methods (lasso and ridge) or principal component analysis (PCA) [19]–[22]. In contrast to wrapper methods, filters can be added to the hyperparameter optimization step and have a lower computational footprint. Due to the focus on filter methods, only this sub-group of feature selection methods will be introduced in greater detail in the following subsections.

1) Filter methods: The concept of filters originates from the idea of ranking features using certain heuristics of an algorithm [21]. Some filter methods can only deal with specific types of variables (numeric or nominal). Filters only rank features, they do not decide which covariates to drop or keep [23]. The selection which features to keep for model fitting is usually done within the optimization phase of the model fitting, along with the hyperparameter tuning. Essentially, the number of covariates in the model is treated as a hyperparameter of the model. The goal is to optimize the number of features, after ranking was done, to the point at which the model achieves the best performance.

Besides the concept of choosing a specific filter method to rank variables, studies showed that combining several filters using statistical operations such as 'minimum' or 'mean' are able to enhance the predictive performance of the resulting models, especially when applied to multiple datasets [24], [25]. This approach is referred to as 'ensemble filtering' [26]. Ensemble filters align with the recent rise of the 'ensemble' approach in machine learning which uses the idea of stacking to combine the predictions of multiple models, aiming to enhance predictive performance [27]–[29]. In this work the 'Borda' ensemble filter was applied [25]. For this filter, the final feature order is the sum of all single filters ranks.

single filters' ranks (?)

Filter methods can be grouped into classes: Correlation-based, entropy-based, linear and non-linear methods. Care needs to be taken to not weight certain classes more than others in the ensemble as otherwise the final ranking result will be biased. In this study this was taken care of by checking the rank correlations (Spearman's correlation) of the generated feature rankings of all methods against each other. If filter pairs showed a correlation of 0.9 or higher, only one of the two was included into the ensemble filter, selected at random. By this it was ensured that the ensemble filter composition was not biased towards a certain group of filter methods.

weigh

2) Description of used filter methods: Filter methods can be classified as follows (Table II):

- Univariate/multivariate (scoring based on a single variable / multiple variables).
- Linear/non-linear (usage of linear/non-linear calculations).
- Entropy/correlation (scoring based on derivations of entropy or correlation-based approaches).

The filter 'Information Gain' is only defined for

nominal response variables:

$$H(Class)+H(Attribute)-H(Class,Attribute) \quad (2)$$

where H is the conditional entropy of the response variable (class) or the feature (attribute), respectively. In order to use this method with a numeric response (percentage defoliation at trees), the variable was discretized into equal bins and treated as a class variable. While the number of bins could be treated as a hyperparameter of the filter method, it was decided to use $n_{bin} = 10$ after rank correlations of > 0.9 for different bin sizes were observed during data exploration. **syntax??**

C. Benchmarking design

1) Algorithms: The following learners were used in this work:

- Extreme Gradient Boosting (XGBoost)
- Random Forest (RF)
- Penalized Regression (both L1 (lasso) and L2 (ridge))
- Support Vector Machine (SVM, RBF Kernel)

Random forest and SVM are well established algorithms widely used in (environmental) modeling. Extreme Gradient Boosting (commonly abbreviated as XGBoost) **has shown** showed promising results in benchmark studies in recent years. Penalized regression is a statistical modeling technique capable of dealing with highly-correlated covariates by penalizing the coefficients of the model [36]. Common penalties are 'lasso' (L1) and 'ridge' (L2). Ridge does not remove variables from the model (penalization to zero) but just shrinks them to effectively zero, keeping them in the model. The combination of both penalties is called 'elastic net' but was not used in this work.

TABLE II
LIST OF FILTER METHODS USED IN THIS WORK

Name	Group	Ref.
Linear correlation (Pearson)	univariate, linear, correlation	[30]
Information gain	univariate, non-linear, entropy	[31]
Minimum redundancy, maximum relevance	multivariate, non-linear, entropy	[32]
Carscore	multivariate, linear, correlation	[33]
Relief	multivariate, linear, entropy,	[34]
Conditional minimal information maximization	multivariate, linear, entropy	[35]

2) Feature sets: Three feature sets were used in this study with each representing a different way of feature engineering:

- The raw hyperspectral band information (HR): no feature engineering)
- Vegetation Indices (vegetation index (VI)s): expert-based feature engineering)
- Normalized Ratio Indices (NRIs): data-driven feature engineering)

The idea of splitting the features into different sets originated from the question whether feature-engineered indices from reflectance values have a positive effect on model performance, as for example demonstrated previously in a spectro-temporal setting [37]. Benchmarking learners on these feature sets while keeping all other variables such as model type, tuning strategy and partitioning method constant makes it possible to draw conclusions on their individual impact. However, rather than only looking at these three groups also combinations of such were taken into account:

- HR + VI
- HR + NRI
- HR + VI + NRI

Even though the feature-selection step should be solely left to the filter methods in this study,it was ensured a priori to account for features with a pairwise correlation of 1. Having such features within the data can cause undesired effects during model fit-

ting and feature importance calculation. Hence, after having calculated all pair-wise correlations between features, for pairs which exceeded the threshold of 0.9999999999, the feature with the largest mean absolute correlation across all variables was removed from the dataset. This process was repeated p times, each time calculating a fresh correlation matrix.

This preprocessing step reduced the amount of covariates to 122 (HR), 86 (VI) and 7467 (NRI).

3) Hyperparameter Optimization: Hyperparameters were tuned using model-based optimization (MBO) within a nested spatial cross-validation (CV) [38]–[40]. In MBO, first n randomly chosen hyperparameter settings out of a user defined search space are composed. After these n settings have been evaluated, one new setting, which is going to be evaluated next, is proposed by a fitted surrogate model (by default a kriging method). This strategy continues until a termination criterion, defined by the user, is reached [41], [42].

In this work, an initial design of 30 randomly composed hyperparameter settings in combination with a termination criterion of 70 iterations was used, resulting in a total budget of 100 evaluated hyperparameter settings per fold. The advantage of this tuning approach is the substantial reduction of the tuning budget which is required to find a setting close to the global optimization minimum. MBO may outperform methods that do not use information from previous iterations, such as random search or grid search [43].

To optimize the number of features used for model fitting, the percentage of features was added as a hyperparameter during the optimization stage ([39]). For PCA, the number of principal components was tuned instead. The RF hyperparameter m_{try} was modified into power transformed $p^{m_{try}}$ (ranging from 1 to p, with p being the number of features of the dataset) to work on relative values of the respective dataset's feature count. This was necessary to ensure that m_{try} was not chosen higher than the available number of features left after optimizing the feature percentage during tuning.

4) Spatial resampling: A spatial nested cross-validation on the plot level was chosen to reduce the influence of spatial autocorrelation as much as possible [40], [44]. The root mean square error (RMSE) was chosen as the error measure. Each plot served as one fold within the cross-validation setting, resulting in four iterations in total. For the inner level (hyperparameter tuning), $k-1$ folds were used with k being the number of plots.

In total the benchmarking grid consisted of 156 experiments (6 feature sets $\times$ 3 ML algorithms $\times$ 8 feature-selection methods + 2 $\times$ 6 L1 and L2 models).

D. Feature importance and feature effects

Estimating feature importance for datasets with highly-correlated features is a complicated task. The correlation between covariates makes it challenging to calculate an unbiased estimate for single features [45]. Methods like partial dependence plots (PDP) or permutation-based approaches may produce unreliable estimates in such scenarios because unrealistic situations between covariates are created [45]. Estimating feature importance for machine learning methods is a complicated procedure for which many different approaches, model-specific and agnostic, exist [36], [46], [47]. The development of robust methods which enable an unbiased estimation of feature importance for highly-correlated variables are subject to current research.

In this work permutation-based feature importance and accumulated local effects (ALE) plots (HR only) were calculated to estimate feature importance / effects [45], [48]. With the limitations of both methods in mind when applied to correlated features, the aim was to get a general overview of the feature importance of the hyperspectral bands while trying to avoid an over-interpretation of results. The best-performing algorithm on the HR task (i.e. SVM) was used for the feature importance calculation.

E. Linking feature importance to wavelength regions

For environmental interpretation purposes the ten most important indices of the best performing models of feature sets HR and VI were linked to the spectral regions of the hyperspectral data. The aim was to visualize the most important features along the spectral curve of the plots to better understand which spectral regions were most important for the model.

F. Research compendium

All tasks of this study were conducted using the open-source statistical programming language R [49]. A complete list of all R packages used in this study can be found in linked repositories. Due to space limitations only the selected packages with high impact on this work will be explicitly cited.

The algorithm implementations of the following packages have been used: xgboost [50] (*Extreme Gradient Boosting*), kernlab [51] (Support Vector Machine) and glmnet [52] (penalized regression). The filter implementations of the following packages have been used: praznik [53], FSelectorRcpp [54]. Package mlr [55] was used for all modeling related steps. drake [56] was used for structuring the work and reproducibility. This study is available as a research compendium on Zenodo (10.5281/zenodo.2635403). Besides the availability of code and manuscript sources, a static webpage is available at (https://github.com/pat-s/2019-feature-selection) listing more side-analyses that were carried out during the creation of this study.

IV. Results

A. Predictive performance

Overall, the response variable "tree defoliation" could be modeled with an RMSE of 28 % percentage points (p.p.). SVM showed no differences in RMSE across feature sets whereas other learners (RF, SVM, XGBoost, lasso and ridge) differed up to seven percentage points (Figure 3). Ridge faced major issues in four tasks due to one observation which was predicted off the response scale (i.e. > 100). SVM showed the best overall performance with a difference of around three percentage points to the next best model (RF) (Table V). Performance differences between test folds were large: Predicting on Luiando resulted in an RMSE of 9.0 for learner SVM (without filter) but up to 54.26 p.p. when testing on Laukiz2 (Table VI).

The combination of feature sets showed small increases in performance for some learners. RF and XGBoost scored slightly better on the combined datasets HR-NRI and NRI-VI, respectively, compared to their standalone variants (HR, NRI, VI) (Figure 3). Datasets containing derived features only (VI, NRI) showed no improvement in performance compared to the raw hyperspectral band information (HR). All learners besides SVM showed a substantially worse performance on the VI dataset compared to all others (around five percentage points worse than their respective best performance).

SVM combined with the "Carscore" filter achieved the best overall performance (RMSE of 27.98 p.p.) (Table III). Regression with ridge penalty (L2) showed a high variance when comparing results across tasks: In four out of six tasks (all which VI variables and HR-NRI) the error was enormous (Table IV). For NRI and HR RMSE was 31.16 p.p. and 35.45 p.p., respectively. In all settings for which ridge showed such a high error, only one observation in one fold was predicted which such a high defoliation value (in the millions). The specific observation showed no obvious signs of being anomalous, e.g. extreme values in individual features or in principal components of features. This one outlier caused the error estimates of these folds and the average estimate across all folds to be off scale.

Effects of filter methods on performance differed greatly between algorithms: SVM showed no variation in performance across filters (Figure 4). Using filters for RF showed a substantial increase in performance for all tasks with the exception of VI, for which the difference among all filters was also the smallest (Figure 4). XGBoost showed a high dependency on filtering the data: In 4 out of 6 tasks using no filter resulted in the worst or second worst performance. In contrast, using no filter on dataset NRI resulted in the best performance. XGBoost shows the highest overall differences between filters for a single task: for feature set HR, the range is up to 14 percentage points ("Carscore" vs. "no filter")(Figure 4).

When comparing the usage of filters against using no filter at all, there was only one instance (XGBoost on the NRI task) when a model without filtering scored a better performance than the best filtered one (Figure 4). For SVM, all filters and "no filter" achieved the same performance on tasks VI and NRI even though Figure 3 lists "No Filter" as the best option.

The Borda filter did not achieve the best performance for any learner across any task Figure 5 (Figure 5) (?). For RF and XGBoost it most often ranked within the first 50% with respect to all filters of a specific task. For XGBoost on the VI task, the Borda filter scored the second worst performance.

High differences were observed between the amount of features selected during tuning for the subsequent fitting process. Most features were selected during optimization for plot Laukiz2 and least for Laukiz1 (Table VII). RF used only one feature for plots Luiando and Oiartzun (0.00004 % of 7675) while for Laukiz1 34 and for Laukiz2 230 features were seen as optimal during optimization. In contrast, XGBoost and SVM used in all cases but Laukiz1 (less than 50 features) more than two-third of all available features.

TABLE III
BEST TEN RESULTS FOR ANY TASK/LEARNER/FILTER COMBINATION, SORTED ASCENDING BY RMSE

	Task	Model	Filter	RMSE	SE
1	NRI	SVM	Info Gain	27.99	19.15
2	HR-NRI-VI	SVM	Relief	28.07	19.14
3	VI	SVM	Relief	28.10	19.14
4	HR-NRI-VI	SVM	Car	28.11	19.13
5	HR-NRI	SVM	MRMR	28.12	19.11
6	VI	SVM	Pearson	28.12	19.10
7	HR-NRI	SVM	CMIM	28.12	19.09
8	HR	SVM	Info Gain	28.12	19.12
9	HR	SVM	CMIM	28.12	19.12
10	NRI-VI	SVM	PCA	28.12	19.12

B. Variable importance

1) Permutation-based Variable Importance: The most important features for datasets HR and VI showed an average decrease in RMSE of 1.57 p.p.

TABLE IV
LOWER TEN RESULTS FOR ANY TASK/LEARNER/FILTER COMBINATION, SORTED DECREASING BY RMSE

	Task	Model	Filter	RMSE
1	VI	Ridge-MBO	No Filter	49359394487.65
2	HR-NRI	Ridge-MBO	No Filter	12650121073.66
3	HR-NRI-VI	Ridge-MBO	No Filter	12631934180.91
4	NRI-VI	Ridge-MBO	No Filter	11658468597.68
5	HR	XGBoost	No Filter	46.80
6	VI	XGBoost	Car	46.40
7	VI	XGBoost	MRMR	46.26
8	VI	XGBoost	Borda	46.04
9	VI	XGBoost	No Filter	45.69
10	VI	XGBoost	Pearson	44.61

TABLE V
BEST PERFORMANCE OF EACH LEARNER ACROSS ANY TASK AND FILTER METHOD, SORTED ASCENDING BY RMSE

	Task	Model	Filter	RMSE	SE
1	NRI	SVM	Info Gain	27.99	19.15
2	NRI	RF	Car	30.77	16.86
3	VI	Lasso-MBO	No Filter	31.01	14.71
4	HR-NRI-VI	XGBoost	Borda	31.05	17.01
5	NRI	Ridge-MBO	No Filter	31.16	15.03

TABLE VI
TEST FOLD PERFORMANCES FOR LEARNER SVM ON THE HR DATASET WITHOUT USING A FILTER. FOR EACH ROW, THE MODEL WAS TRAINED ON OBSERVATIONS FROM FOLD ALL OTHER PLOTS AND TESTED ON THE OBSERVATIONS OF THE GIVEN PLOT.

	Plot	RMSE
1	Laukiz1	21.17
2	Oiartzun	28.05
3	Luiando	9.00
4	Laukiz2	54.26

TABLE VII
SELECTED FEATURE PORTIONS DURING TUNING FOR SELECTED LEARNER-FILTER SETTINGS ACROSS FOLDS FOR TASK HR-NRI-VI, SORTED ASCENDING BY RMSE

Learner	Plot	Features (%)	#	RMSE
RF Car	Laukiz1	0.00443	3	27.48
	Oiartzun	0.00004	1	37.20
	Luiando	0.00002	1	36.98
	Laukiz2	0.03037	14	15.39
XGB Borda	Laukiz1	0.00604	4	15.00
	Oiartzun	0.68227	340	36.15
	Luiando	0.99513	300	38.19
	Laukiz2	0.99976	451	29.62
SVM Relief	Laukiz1	0.00017	1	35.50
	Oiartzun	0.71146	354	37.48
	Luiando	0.70672	213	14.89
	Laukiz2	0.94696	428	37.19

(HR, B69) and 1.79 p.p (VI, Vogelmann2) (Figure 6). For both datasets most features among the ten most important ones cluster around a wavelength range of 700 nm - 750 nm (the so called "red edge"). For feature set HR, four features in the infrared region (920 nm - 1000 nm) were identified by the model to be most important (causing a mean decrease in RMSE of around one percentage point). Overall, most features showed only a small importance with average decreases in RMSE below 0.5 percentage points.

2) ALE Plots: The ALE plots show a small relative change compared to their respective mean effects for all chosen features ($<|$ 0.001 $|$) (Figure 7). Most ALE curves show a high dynamic across the value range of each variable, within their respective effect range.

V. DISCUSSION

A. Predictive Performance

The best aggregated performance of this study (SVM + "Info Gain" filter, RMSE 27.99 p.p.) has to be seen in the light of model overfitting (see subsection V-B). Leaving out the performance on Laukiz2 when aggregating results, the mean RMSE would be

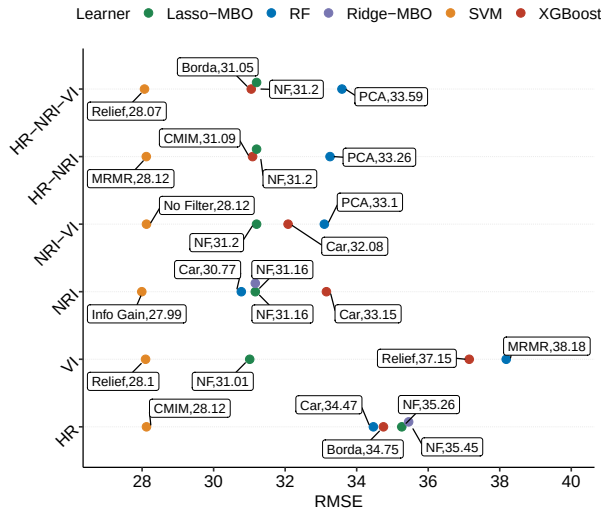

Fig. 3. Predictive performance (RMSE) of models across tasks. Suffix 'CV' denotes that the learner was optimized using internal 10-fold CV while prefix 'MBO' means that Bayesian optimization was used for hyperparameter optimization. Abbreviations on the vertical axis refer to the combinations of feature sets on which each model was scored on. Labels represent the feature selection method (NF = no filter, Car = 'Carscore', Info = 'Information Gain', Borda = 'Borda'). The second value of each label shows the RMSE value of the respective setting.

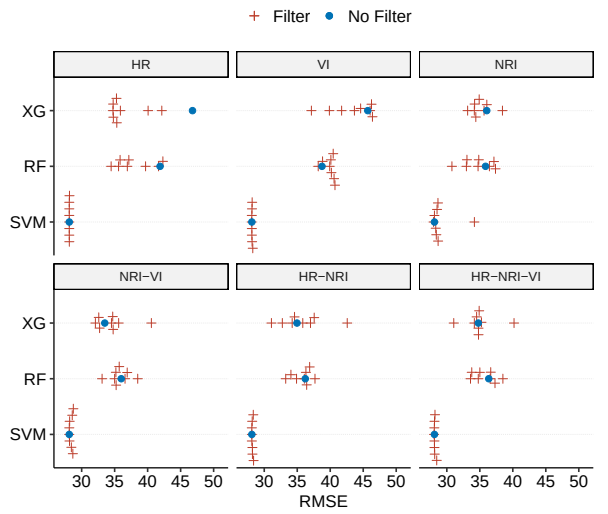

Fig. 4. Model performances in RMSE when using no filter method compared to all other filters across all tasks.

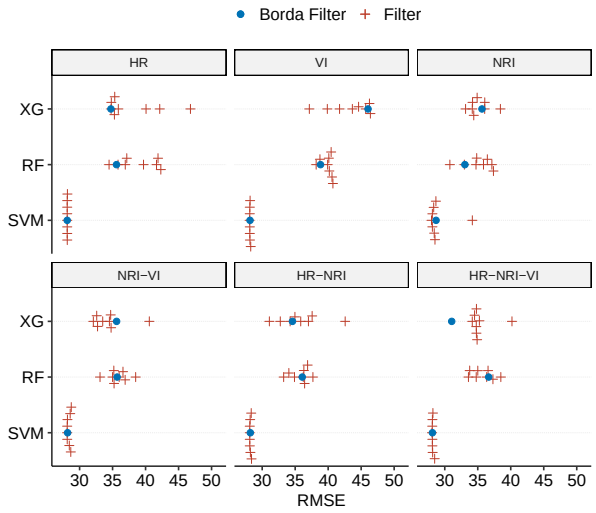

Fig. 5. Predictive performances in RMSE when using the Borda filter method compared to all other filters for each learner across all tasks.

around 19 percentage points. However, leaving out a single plot would also change the prediction results for the other plots because the observations from Laukiz2 would not be available for model training. Due to the apparent presence of model overfitting in this study it can be postulated that more training data representing all health stages of a plot is needed. A model can only make robust predictions if it has learned relationships across the whole range of the response. Hence, care should be taken when predicting to the landscape scale using models fitted on this dataset due to their lack of generalizability caused by suboptimal distributed training data. However, when inspecting the fold level performances, it can be concluded that the model did a decent job predicting defoliation ranging from 50% to 100% but failed for 0% - 50%. This applied to all learners of this study.

1) Model differences: An interesting find is the strength of the SVM algorithm when comparing its predictive performance to its competitors (Table V).

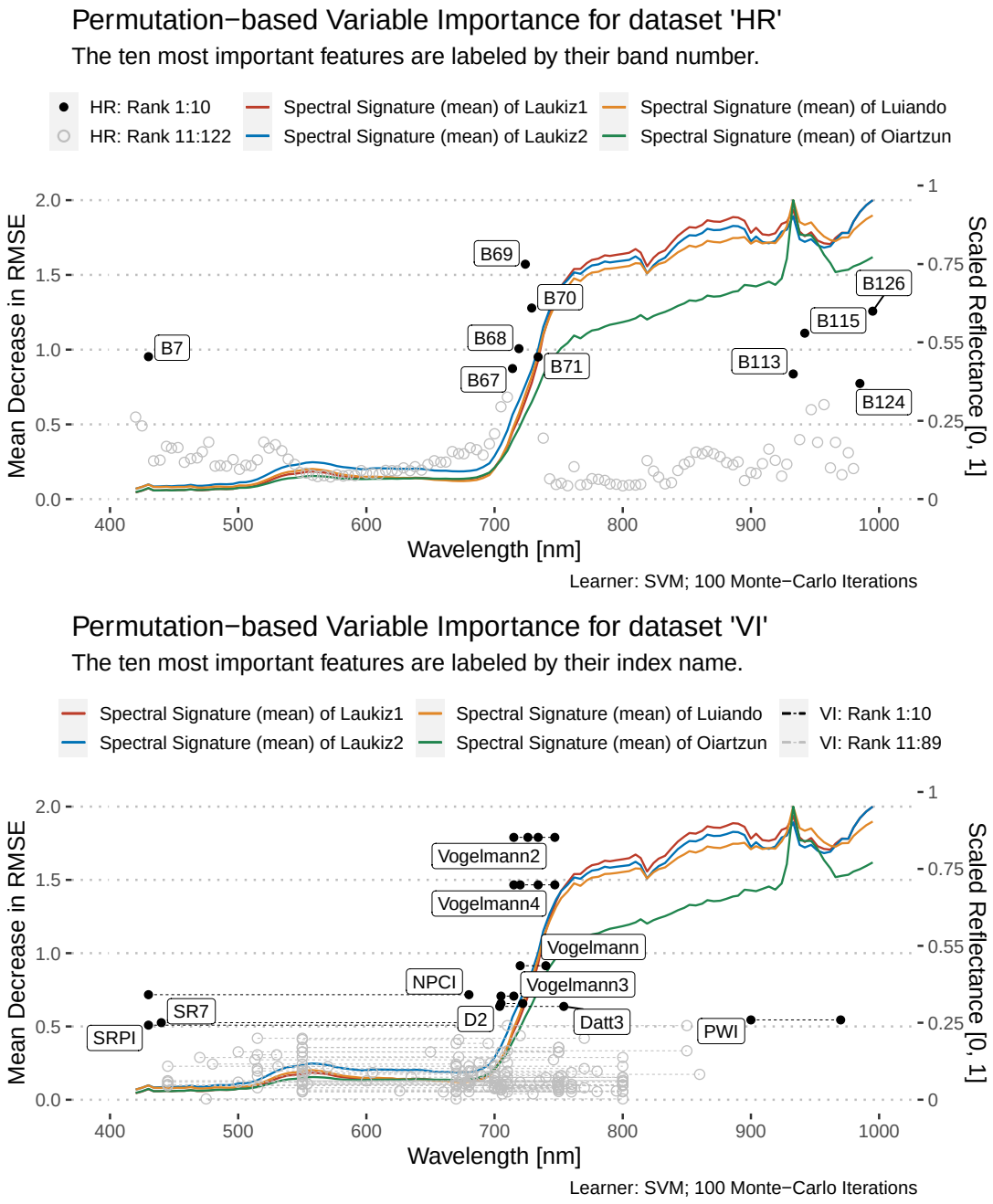

Fig. 6. Variable importance for feature sets HR and VI: Mean decrease in RMSE for one-hundred feature permutations using the SVM learner. The wavelength range on the x-axis matches the range of the hyperspectral sensor (400 nm - 1000 nm). For each dataset, the ten most important features were highlighted as black dots and labeled by name. Grey dots represent features from importance rank 11 to last. The spectral signature (mean) of each plot was added as a reference on a normalized reflectance scale [0, 1] (secondary y-axis). VI features were decomposed into their individual formula parts; all instances being connected via dashed lines. Each VI feature is composed out of at least two instances.

These cluster around a performance of 31 p.p while SVM is able to score about 3 p.p. better than all other methods. However, comparing this finding (both relatively and absolute) to other studies would not be fair since many study design points have an influence on the final result (optimization strategy, data characteristics, feature selection methods, etc.).

Penalized methods showed promising performances, especially when taking runtime into account. When removing features with a correlation of nearly 1, lasso is able to score performances around 31 p.p. and shows, somewhat surprisingly, the best performance on the VI task. The same general conclusion applies to ridge if one discards the problematic results on four tasks. Regarding the problematic predictions of the ridge learner, a careful inspection of the hyperparameter optimization procedure of the fitted models did not reveal any apparent implementation mistakes. One observation turned out to be influential in causing the outlier prediction value even though careful multivariate exploratory data analysis showed no reason for flagging this observation a priori. It was therefore left in the model, but it should be acknowledged that these learners seemed particularly sensitive to the input data.

The upper range of 600 used for hyperparameter nrounds for model XGBoost in this study could be a potential limiting factor with respect to the performance of the datasets including NRI variables (Table VIII). This setting was a compromise between runtime and tuning space extension aimed to work well for most feature sets. It could be that better predictive performances could have been achieved with a nrounds upper limit close to the number of features of the NRI datasets.

2) Feature set differences: One objective which this study aimed to answer was whether expert-based or data-driven feature engineering has a positive influence on model performance. With respect to Figure 3, no overall positive or negative trend was found for all models that related to specific feature sets. The performance of RF and XGBoost on the VI feature set was about six percentage points lower than on others. One reason could be the lack of coverage in the wavelength area between 810 nm and 1000 nm (Figure 6). In addition, for all learners but SVM a better performance was observed when NRI indices were included in the feature set (i.e. NRI-VI, HR-NRI, HR-NRI-VI).

B. Performance vs. plot characteristics

The large differences in RMSE obtained on different test folds can be attributed to model overfitting (Table VI). An RMSE of 54.26 p.p. indicates a complete failure of the model in the prediction stage for this plot (Laukiz2). Laukiz2 differs highly in the distribution of the response variable defoliation compared to all other plots (Figure 1). In the prediction scenario for Laukiz2, the model was trained on data containing mostly medium-high defoliation values and only few low ones. This caused overfitting on the medium-high values, degrading the model's predictive performance in other scenarios. In cases when Laukiz2 was in the training set, the overall aggregated RMSE was reduced by up to 50% with single fold performances as good as 9 p.p. RMSE (with Luiando as test set).

The high differences of selected features per plot plot during tuning give interesting insights into internals of the used models (Table VII). While in most cases, SVM and XGBoost require a substantial portion of all available features to achieve robust predictions, RF is able to achieve the best results with a relatively low amount of features. This fact

leads to reduced computational runtime, especially if model parameters are dependent on the amount of features (e.g. m_{try}). Hence, regardless of the potential advantage of using filters for increased predictive performance, it should be noted that these can have a strong positive effect on runtime, at least for RF in this study.

Ultimately, the results of (Table VII) should be taken with care as they rely on single model-filter combinations and are subject to random variation. More in-depth research is needed to investigate the effect of filters on other criteria than performance (such as runtime), leading to a multi-criteria optimization problem.

C. Feature selection methods

The usefulness of filters with respect to predictive performance in this study varied. While some filters were able to enhance model performances (up to 5 p.p. for RF and XGBoost), some caused a worse performance than using no filter at all Figure 4. Since these negative cases were not caused by a specific filter method, it is recommended to test multiple filters in a study if filters are going to be used. While filters can improve the performance of models, they might be of more interesting in other aspects than performance: reducing variables can reduce computational efforts in high-dimensional scenarios and might enhance the interpretability of models. Here, filters are a lot cheaper than wrapper methods due to their ability of being treated as a hyperparameter during the optimization stage.

The models which used the Borda ensemble method in this study did not score better on average than models which used a single filter or no filter at all. Ensemble methods have higher stability and robustness than single ones and have shown promising results in [25]. Hence, their main advantage are stable performances across datasets with varying characteristics. Single filter methods might yield better model performances on certain datasets but fail on others. The fact that this study used multiple feature sets but only one dataset and tested many single filters could be a potential explanation why in almost all cases (besides XGBoost on task HR-NRI-VI) a single filter outperformed the ensemble filter. However, studies which used ensemble filters are still rare and usually these are not compared against single filters [57]. Therefore no general statement can be made from the results of this study if the use of ensemble filters leads to a substantial better performance than applying one of the various single filter methods. For this, more case studies applying ensemble filter methods are needed. In any case, ensemble filters can be a promising addition to a machine-learning feature-selection portfolio.

PCA used as a filter in this work did neither show the best nor worst result compared to other feature-selection approaches. It was able to reduce model fitting times substantially due to its nature of reducing the number of predictors to a minimum. However, it also removed the possibility to conduct a post-hoc analysis of feature importance by combining predictors into principal components, forcing users to decide upfront what they value more. Since filter scores only need to be calculated once in a benchmark setting, the runtime advantage might in fact be negligible in practice.

D. Quality of the data

The decision to use a buffer of two meters for the extraction of reflectance values was a complex process. Due to the reported geometric offset of up to 1 m within the hyperspectral data, the risk of

assigning a value to an observation which would actually refer to a different observation was reasonably high. By using a buffer of two meters (or more), the probability of including information from other trees into the final value increases, blurring the actual value of the tree observation. However, when using no buffer at all, the difference between single, neighboring pixels might even be higher compared to a smoothed value composed out of a buffer. It was concluded that using a buffer of two meters is a good compromise between the inclusion of information from too many surrounding trees and not accounting for the geometric offset at all. Even though no results showing the influence of multiple buffer values on the extraction were provided, it is hypothesized that the relationships between features would not change substantially, leading to the almost identical model results.

Another point worth discussing is that the exact number of contributing pixels to the final index value of an observation cannot be exactly determined: it depends on the location of the tree within the pixel grid. According to the extract function of the raster package, a pixel is included if its centroid (and not just any part of the grid cell) falls inside the buffer. As the buffer is circular, the total number of contributing pixels of each tree depends on the exact location of a tree within the pixel grid. If a tree observation is located on the border of the plot, some directions of the buffer will contain no values and the subsequent index value will be calculated with fewer pixels than if the tree observation would be located in the middle of the plot.

The R package hsdar was used for the calculation of vegetation indices [58]. All indices that could be calculated with the given spectral range of the data were used. Even though this selection included a large number of available indices, some possibly relevant indices might have been missed by relying on the pre-selection of indices offered by the package.

Overall, the magnitude of uncertainty introduced by the mentioned effects during index derivation cannot be quantified. Such limitations and uncertainties apply to most environmental studies and cannot be completely avoided.

E. Linking feature importance to spectral characteristics

Not surprisingly the most important features for both HR and VI datasets were identified around the red edge of the spectra, specifically in the range of 680 nm to 750 nm.

This area has the highest ability to distinguish between reflectances related to a high density / high healthiness of vegetation and its respective counterpart [59]. However, four out of ten of the most important features of dataset HR are located between 920 nm and 1000 nm. Looking at the spectral curves of the plots, apparent reflectance differences can be observed in this spectral area - especially for plot Oiartzun - which might explain why these features were considered as important by the model.

A possible explanation for the less good performances of most models scored on the VI dataset compared to all other feature sets could be the lack of features covering the area between 850 nm and 1000 nm (Figure 6). The majority of VI features covers the range between 550 nm - 800 nm. Only one index (PWI) covers information in the range beyond 900 nm.

The ALE plots are hard to interpret due to their semi-meaningful absolute values which differ for each band. For example, a reflectance value of 100 has a very different meaning in B67 than in B124 because

for the latter, reflectance values of vegetation are naturally higher. Hence, a reflectance value of 100 would be considered high for B67 but low for B124 with respect to a spectral curve of vegetation. However, no baseline reflectance value for every feature exists since the absolute reflectance values depend on the sensor characteristics. Hence, it is hard to make reasonable interpretations of harsh drops/jumps within curves of certain features, e.g. the one at value 240 for feature B115. ALE plots can be a powerful tool for interpreting feature importance of correlated features but might be limited in interpretability if the absolute feature values are only semi-meaningful.

F. Comparison to other studies

Most other studies analyzing defoliation operated on the plot rather than the tree level. This is due to the low spatial resolution of used satellite products which served as the input data, making a tree-level study infeasible [7], [60], [61].

Studies focusing on tree-level defoliation used ground-level methods such as airborn laser scanning (ALS) or light detection and ranging (LiDAR) [62], [63]. [62] used ordinary least squares (OLS) regression methods while [63] retrieved information from ground-level RGB photos using convolutional neural networks (CNN). However, both did not use spatial CV and [63] no feature selection (FS). [8] used a partial least-squares (PLS) model with high-resolution digital aerial photogrammetry (DAP) to predict cumulative defoliation caused by the spruce budworm. Study results indicated that spectral metrics were found to be most helpful for the model. Incorporating such metrics (both spectral and structural) could be a possible enhancement for future works.

The field of (hyperspectral) remote sensing has a strong focus on using random forest for modeling in recent years [64]. However, in high-dimensional scenarios, tuning parameter m_{try} becomes computational expensive. To account for this and the high-dimensionality in general, studies used feature selection approaches like semi-supervised feature extraction [65], wrapper methods [66]–[68], PCA and adjusted feature selection [69]. However, no study which made use of filter methods in combination with hyperparameter tuning in the field of (hyperspectral) remote sensing could be found. Potential reasons for this gap could be an easier access of applying wrappers methods and a higher general awareness of such compared to filter methods. Applying the filter-based feature selection methodology shown in this study and its related code provided in the research compendium might be a helpful reference for upcoming modeling studies using hyperspectral remote sensing data.

When looking for remote sensing studies which compare multiple models, it turned out that such often operate in a low-dimensional predictor space [70] or use wrapper methods explicitly [68].

[71], [72] are more similar in their methodology but focus on a different response variable (woody cover). [71] used machine learning with ALS data to study dieback of trees for eucalyptus forests. A grid search was used for hyperparameter tuning and forward feature selection (FFS) for variable selection. [72] analyzed woody cover in South Africa using spatial CV and FS approach [73] with a random forest classifier.

In summary, no studies which used filter methods for FS or made use of NRI indices in their work and had a relation to forest health were found. This might relate to the fact that most environmental/ecological datasets are not high-dimensional. In fact, the number of predictors is often less than ten and issues

related to correlations are often solved manually instead of relying on an automated approach.

The bioinformatics field faces high-dimensional datasets more often. Hence more studies using (filter-based) feature-selection approaches can be found for this field [74], [75]. Yet bioinformatics differs conceptually in many ways from environmental modeling and digging deeper just for the sake of finding any similar studies is not a good idea. If a field only rarely faces high-dimensional dataset issues (like environmental modeling) the motivation and expertise of using advanced methods to solve such are rather low. We hope that this work and its methodology raises awareness about the application of filter methods to tackle high-dimensional problems in the environmental modeling field.

VI. Outlook and conclusion

This study analyzed defoliation at trees in northern Spain by using hyperspectral data as input for machine-learning models which used hyperparameter tuning and filter-based feature selection. Substantial differences in performance occurred depending on which feature selection and machine learning methods were combined. SVM showed the most robust behavior across all highly-correlated datasets and was able to predict the response variable of this study substantially better than other methods.

Filter methods were able to improve the predictive performance on datasets in some instances. Their effectiveness depends on the algorithm and the dataset characteristics. Ensemble filter methods did not show a substantial improvement over individual filter methods in this study.

The addition of derived feature sets was in most cases able to improve predictive performance. In contrast, feature sets which focused on only a small fraction of the available spectral range (i.e. dataset VI) showed a worse performance than the ones which covered wider range (400 nm - 1000 nm; HR, NRI). NRIs can be seen as a valuable addition for optimizing predictive performance in vegetation related studies.

Features along the red edge wavelength region were most important for models during prediction. With respect to dedicated vegetation indices, the "Vogelmann" index with all of its derivatives was seen as the most important index for the best performing SVM model. This matches well with the actual purpose of these indices: These were invented to detect defoliation on sugar maple trees (*Acer saccharum Marsh.*) caused by pear thrips (*Taeniothrips inconsequens Uzel*) in 1988 [76]. However, feature importance for highly-correlated features remains a challenging task. Results might be biased and should be taken with care.

The potential of predicting defoliation with the given study design was rather limited with respect to the average RMSE of 27 percentage points scored by the best performing model. More training data covering a wider range of defoliation values in a larger number of forest plantations is needed to train better models which can create more robust predictions.

Appendix A
SVM ALE plots for task HR

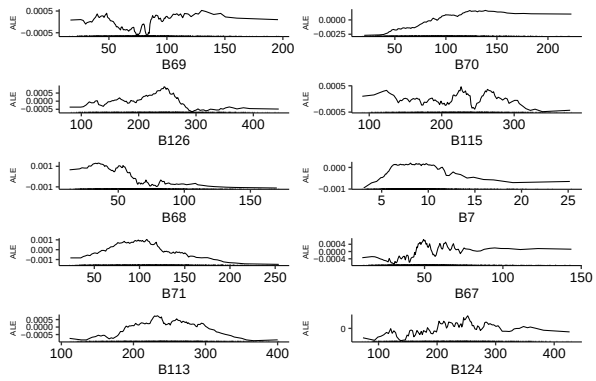

Fig. 7. ALE plots of SVM on dataset HR. The ten most important features from the permutation-based variable importance estimation were used. The y-axis shows the deviation to the mean prediction for each feature, with the mean prediction being centered at zero.

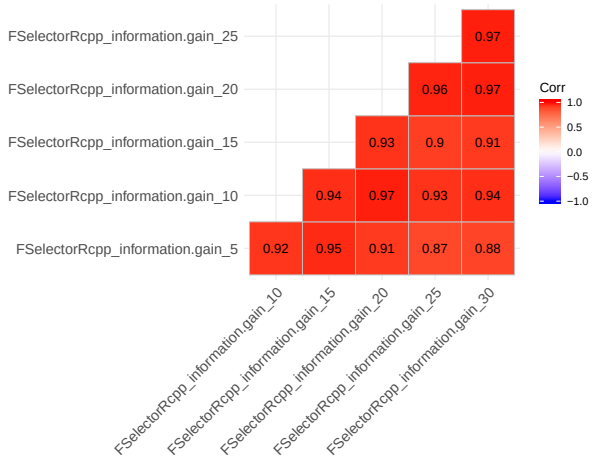

Fig. 9. Spearman correlations of rankings obtained with the information gain filter using different n_{bins} values for discretization of the numeric response.

Appendix B
Correlation among filter methods

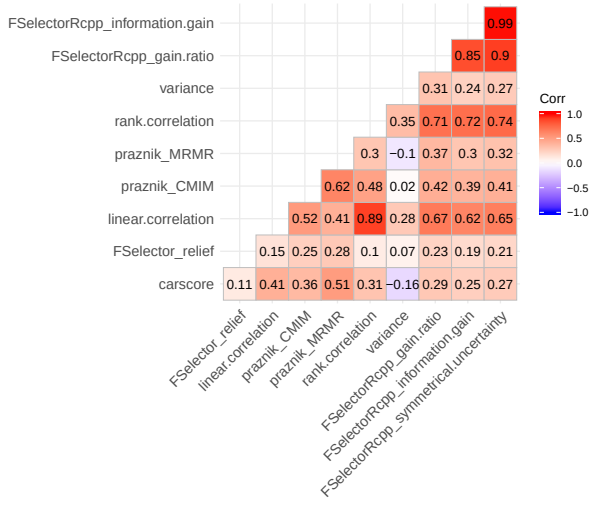

Fig. 8. Spearman correlations of NRI feature rankings obtained with different filters.

Appendix C
Effect of different n_{bins} values on filter 'information gain'

Appendix D
Hyperparameter tuning ranges

TABLE VIII
HYPERPARAMETER RANGES AND TYPES FOR EACH MODEL. HYPERPARAMETER NOTATIONS FROM THE RESPECTIVE R PACKAGES WERE USED.

Model (package)	Hyperparameter	Type	Start	End	Default
RF (ranger)	$m_{try}.power$	int	0	0.5	-
	min.node.size	int	1	10	1
	sample.fraction	dbl	0.2	0.9	1
SVM (kernlab)	C	dbl	2^{-10}	2^{10}	1
	σ	dbl	2^{-5}	2^{5}	1
XGBoost (xgboost)	nrounds	int	10	600	-
	colsample_bytree	dbl	0.3	0.7	1
	subsample	dbl	0.25	1	1
	max_depth	int	1	10	6
	gamma	int	0	10	0
	eta	dbl	0.01	0.6	0.3
	min_child_weight	int	0	20	1

References

[1] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker, "Machine learning in geosciences and remote sensing," *Geoscience Frontiers*, vol. 7, no. 1, pp. 3–10, Jan. 2016.

[2] Y. Ma, H. Wu, L. Wang, B. Huang, R. Ranjan, A. Zomaya, and W. Jie, "Remote sensing big data computing: Challenges and opportunities," *Future Generation Computer Systems*, vol. 51, pp. 47–60, Oct. 2015.

[3] J. Mascaro, G. P. Asner, D. E. Knapp, T. Kennedy-Bowdoin, R. E. Martin, C. Anderson, M. Higgins, and K. D. Chadwick, "A Tale of Two "Forests": Random Forest Machine Learning Aids Tropical Forest Carbon Mapping," *PLOS ONE*, vol. 9, no. 1, p. e85993, Jan. 2014.

[4] M. Urban, C. Berger, T. E. Mudau, K. Heckel, J. Truckenbrodt, V. Onyango Odipo, I. P. J. Smit, and C. Schmullius, "Surface Moisture and Vegetation Cover Analysis for Drought Monitoring in the Southern Kruger National Park Using Sentinel-1, Sentinel-2, and Landsat-8," *Remote Sensing*, vol. 10, no. 9, p. 1482, Sep. 2018.

[5] P. Hawryło o, B. o. Bednarz, P. Wezyk, and M. Szostak, "Estimating defoliation of Scots pine stands using machine learning methods and vegetation indices of Sentinel-2," *European Journal of Remote Sensing*, vol. 51, no. 1, pp. 194–204, Jan. 2018.

[6] K. Zhang, B. Thapa, M. Ross, and D. Gann, "Remote sensing of seasonal changes and disturbances in mangrove forest: A case study from South Florida," *Ecosphere*, p. e01366, 2016.

[7] P. A. Townsend, A. Singh, J. R. Foster, N. J. Rehberg, C. C. Kingdon, K. N. Eshleman, and S. W. Seagle, "A general Landsat model to predict canopy defoliation in broadleaf deciduous forests," *Remote Sensing of Environment*, vol. 119, pp. 255–265, Apr. 2012.

[8] T. R. H. Goodbody, N. C. Coops, T. Hermosilla, P. Tompalski, G. McCartney, and D. A. MacLean, "Digital aerial photogrammetry for assessing cumulative spruce budworm defoliation and enhancing forest inventories at a landscape-level," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 142, pp. 1–11, Aug. 2018.

[9] Y. Jiang, T. Wang, C. A. J. M. de Bie, A. K. Skidmore, X. Liu, S. Song, L. Zhang, J. Wang, and X. Shao, "Satellite-derived vegetation indices contribute significantly to the prediction of epiphyllous liverworts," *Ecological Indicators*, vol. 38, pp. 72–80, Mar. 2014.

[10] J. Adamczyk and A. Osberger, "Red-edge vegetation indices for detecting and assessing disturbances in Norway spruce dominated mountain forests," *International Journal of Applied Earth Observation and Geoinformation*, vol. 37, pp. 90–99, May 2015.

[11] G. V. Trunk, "A Problem of Dimensionality: A Simple Example," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 3, pp. 306–307, Jul. 1979.

[12] H. Xu, C. Caramanis, and S. Mannor, "Statistical Optimization in High Dimensions," *Operations Research*, vol. 64, no. 4, pp. 958–979, Jul. 2016.

[13] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, Jul. 2018.

[14] N. Mesanza, E. Iturritxa, and C. L. Patten, "Native rhizobacteria as biocontrol agents of Heterobasidion annosum s.s. and Armillaria mellea infection of Pinus radiata," *Biological Control*, vol. 101, pp. 8–16, Oct. 2016.

[15] E. Iturritxa, T. Trask, N. Mesanza, R. Raposo, M. Elvira-Recuenco, and C. L. Patten, "Biocontrol of Fusarium circinatum infection of young Pinus radiata trees," *Forests*, vol. 8, no. 2, p. 32, Jan. 2017.

[16] E. Iturritxa, N. Mesanza, and A. Brenning, "Spatial analysis of the risk of major forest diseases in Monterey pine plantations," *Plant Pathology*, vol. 64, no. 4, pp. 880–889, 2014.

[17] R. J. Ganley, M. S. Watt, L. Manning, and E. Iturritxa, "A global climatic risk assessment of pitch canker disease," *Canadian Journal of Forest Research*, vol. 39, no. 11, pp. 2246–2256, Nov. 2009.

[18] Johnstone Iain M. and Titterington D. Michael, "Statistical challenges of high-dimensional data," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 367, no. 1906, pp. 4237–4253, Nov. 2009.

[19] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, "Benchmark for filter methods for feature selection in high-dimensional classification data," *Computational Statistics & Data Analysis*, vol. 143, p. 106839, Mar. 2020.

[20] S. Das, "Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection," in *ICML*, 2001.

[21] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[22] I. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, Apr. 2016.

[23] P. Drotár, J. Gazda, and Z. Smékal, "An experimental comparison of feature selection methods on two-class biomedical datasets," *Computers in Biology and Medicine*, vol. 66, pp. 1–10, Nov. 2015.

[24] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, and Y. Saeys, "Robust biomarker identification for cancer diagnosis with ensemble feature selection methods," *Bioinformatics*, vol. 26, no. 3, pp. 392–398, Feb. 2010.

[25] P. Drotár, M. Gazda, and J. Gazda, "Heterogeneous ensemble feature selection based on weighted Borda count," in *2017 9th International Conference on Information Technology and Electrical Engineering (ICITEE)*, Oct. 2017, pp. 1–4.

[26] T. G. Dietterich, "Ensemble Methods in Machine Learning," in *Proceedings of the First International Workshop on Multiple Classifier Systems*. Springer-Verlag, Jun. 2000, pp. 1–15.

[27] R. Polikar, "Ensemble Learning," in *Ensemble Machine Learning: Methods and Applications*, C. Zhang and Y. Ma, Eds. Boston, MA: Springer US, 2012, pp. 1–34.

[28] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and Robust Automated Machine Learning," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 2962–2970.

[29] V. Bolón-Canedo and A. Alonso-Betanzos, "Ensembles for feature selection: A review and future trends," *Information Fusion*, vol. 52, pp. 1–12, Dec. 2019.

[30] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, Nov. 1901.

[31] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, Mar. 1986.

[32] X.-M. Zhao, "Maximum Relevance/Minimum Redundancy (MRMR)," in *Encyclopedia of Systems Biology*, W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, Eds. New York, NY: Springer New York, 2013, pp. 1191–1192.

[33] V. Zuber and K. Strimmer, "High-Dimensional Regression and Variable Selection Using CAR Scores," *Statistical Applications in Genetics and Molecular Biology*, vol. 10, no. 1, 2011.

[34] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proceedings of the Tenth National Conference on Artificial Intelligence*. AAAI Press, Jul. 1992, pp. 129–134.

[35] F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information," *The Journal of Machine Learning Research*, vol. 5, pp. 1531–1555, Dec. 2004.

[36] T. Hastie, J. Friedman, and R. Tibshirani, *The Elements of Statistical Learning*. Springer New York, 2001.

[37] M. Peña, R. Liao, and A. Brenning, "Using spectrotemporal indices to improve the fruit-tree crop classification accuracy," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 128, pp. 158–169, 2017.

[38] B. Bischl, J. Richter, J. Bossek, D. Horn, J. Thomas, and M. Lang, "mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions," *ArXiv e-prints*, Mar. 2017.

[39] M. Binder, J. Moosbauer, J. Thomas, and B. Bischl, "Multi-Objective Hyperparameter Tuning and Feature Selection using Filter Ensembles," *arXiv:1912.12912 [cs, stat]*, Feb. 2020.

[40] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, "Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data," *Ecological Modelling*, vol. 406, pp. 109–120, Aug. 2019.

[41] F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Sequential model-based optimization for general algorithm configuration," in *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2011, pp. 507–523.

[42] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient global optimization of expensive black-box functions," *Journal of Global Optimization*, vol. 13, no. 4, pp. 455–492, Dec. 1998.

[43] J. Bergstra and Y. Bengio, "Random Search for Hyper-parameter Optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, Feb. 2012.

[44] A. Brenning, "Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest," in *2012 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, Jul. 2012, R package version 2.1.0.

[45] C. Molnar, *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable*, 2019.

[46] J. H. Friedman, "Greedy function approximation: A gradient boosting machine." *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[47] B. M. Greenwell, B. C. Boehmke, and A. J. McCarthy, "A Simple and Effective Model-Based Variable Importance Measure," *arXiv:1805.04755 [cs, stat]*, May 2018.

[48] D. W. Apley and J. Zhu, "Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models," *arXiv:1612.08468 [stat]*, Aug. 2019.

[49] R Core Team, *R: A Language and Environment for Statistical Computing*, Vienna, Austria, 2019.

[50] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 785–794.

[51] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, "Kernlab – An S4 Package for Kernel Methods in R," *Journal of Statistical Software*, vol. 11, no. 9, pp. 1–20, 2004.

[52] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[53] M. B. Kursa, *Praznik: Collection of Information-Based Feature Selection Filters*, 2018.

[54] Z. Zawadzki and M. Kosinski, *FSelectorRcpp: 'Rcpp' Implementation of 'FSelector' Entropy-Based Feature Selection Algorithms with a Sparse Matrix Support*, 2019.

[55] B. Bischl, M. Lang, L. Kotthoff, J. Schiffner, J. Richter, E. Studerus, G. Casalicchio, and Z. M. Jones, "mlr: Machine learning in R," *Journal of Machine Learning Research*, vol. 17, no. 170, pp. 1–5, 2016.

[56] W. M. Landau, "The drake R package: A pipeline toolkit for reproducibility and high-performance computing," *Journal of Open Source Software*, vol. 3, no. 21, 2018.

[57] M. Ghosh, S. Adhikary, K. K. Ghosh, A. Sardar, S. Begum, and R. Sarkar, "Genetic algorithm based cancerous gene identification from microarray data using ensemble of filter methods," *Medical & Biological Engineering & Computing*, vol. 57, no. 1, pp. 159–176, Jan. 2019.

[58] L. W. Lehnert, H. Meyer, and J. Bendix, *Hsdar: Manage, Analyse and Simulate Hyperspectral Data in R*, 2018.

[59] D. N. H. Horler, M. Dockray, and J. Barber, "The red edge of plant leaf reflectance," *International Journal of Remote Sensing*, vol. 4, no. 2, pp. 273–288, Jan. 1983.

[60] K. M. de Beurs and P. A. Townsend, "Estimating the effect of gypsy moth defoliation using MODIS," *Remote Sensing of Environment*, vol. 112, no. 10, pp. 3983–3990, Oct. 2008.

[61] R. Rengarajan and J. R. Schott, "Modeling forest defoliation using simulated BRDF and assessing its effect on reflectance and sensor reaching radiance," in *Remote Sensing and Modeling of Ecosystems for Sustainability XIII*, vol. 9975. International Society for Optics and Photonics, Sep. 2016, p. 997503.

[62] R. Meng, P. E. Dennison, F. Zhao, I. Shendryk, A. Rickert, R. P. Hanavan, B. D. Cook, and S. P. Serbin, "Mapping canopy defoliation by herbivorous insects at the individual tree level using bi-temporal airborne imaging spectroscopy and LiDAR measurements," *Remote Sensing of Environment*, vol. 215, pp. 170–183, Sep. 2018.

[63] U. Kälin, N. Lang, C. Hug, A. Gessler, and J. D. Wegner, "Defoliation estimation of forest trees from ground-level images," *Remote Sensing of Environment*, vol. 223, pp. 143–153, Mar. 2019.

[64] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, Apr. 2016.

[65] J. Xia, W. Liao, J. Chanussot, P. Du, G. Song, and W. Philips, "Improving Random Forest With Ensemble of Features and Semisupervised Feature Extraction," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 7, pp. 1471–1475, Jul. 2015.

[66] F. E. Fassnacht, C. Neumann, M. Förster, H. Buddenbaum, A. Ghosh, A. Clasen, P. K. Joshi, and B. Koch, "Comparison of Feature Reduction Algorithms for Classifying Tree Species With Hyperspectral Data on Three Central European Test Sites," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2547–2561, Jun. 2014.

[67] J. Feng, L. Jiao, F. Liu, T. Sun, and X. Zhang, "Unsupervised feature selection based on maximum information and minimum redundancy for hyperspectral images," *Pattern Recognition*, vol. 51, pp. 295–309, Mar. 2016.

[68] S. Georganos, T. Grippa, S. Vanhuysse, M. Lennert, M. Shimoni, S. Kalogirou, and E. Wolff, "Less is more: Optimizing classification performance through feature selection in a very-high-resolution remote sensing object-based urban application," *GIScience & Remote Sensing*, vol. 55, no. 2, pp. 221–242, Mar. 2018.

[69] J. F. R. Rochac and N. Zhang, "Feature extraction in hyperspectral imaging using adaptive feature selection approach," in *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)*, Feb. 2016, pp. 36–40.

[70] S. Xu, Q. Zhao, K. Yin, F. Zhang, D. Liu, and G. Yang, "Combining random forest and support vector machines for object-based rural-land-cover classification using high spatial resolution imagery," *Journal of Applied Remote Sensing*, vol. 13, no. 1, p. 014521, Feb. 2019.

[71] I. Shendryk, M. Broich, M. G. Tulbure, A. McGrath, D. Keith, and S. V. Alexandrov, "Mapping individual tree health using full-waveform airborne laser scans and imaging spectroscopy: A case study for a floodplain eu-

calypt forest," *Remote Sensing of Environment*, vol. 187, pp. 202–217, Dec. 2016.

[72] M. Ludwig, T. Morgenthal, F. Detsch, T. P. Higginbottom, M. Lezama Valdes, T. Nauß, and H. Meyer, "Machine learning and multi-sensor based modelling of woody vegetation in the Molopo Area, South Africa," *Remote Sensing of Environment*, vol. 222, pp. 195–203, Mar. 2019.

[73] H. Meyer, C. Reudenbach, T. Hengl, M. Katurji, and T. Nauss, "Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation," *Environmental Modelling & Software*, vol. 101, pp. 1–9, Mar. 2018.

[74] Y. Guo, F.-L. Chung, G. Li, and L. Zhang, "Multi-Label Bioinformatics Data Classification With Ensemble Embedded Feature Selection," *IEEE Access*, vol. 7, pp. 103 863–103 875, 2019.

[75] M. Radovic, M. Ghalwash, N. Filipovic, and Z. Obradovic, "Minimum redundancy maximum relevance feature selection approach for temporal gene expression data," *BMC Bioinformatics*, vol. 18, no. 1, p. 9, Jan. 2017.

[76] J. E. Vogelmann, B. N. Rock, and D. M. Moss, "Red edge spectral measurements from sugar maple leaves," *International Journal of Remote Sensing*, vol. 14, no. 8, pp. 1563–1575, May 1993.

References