***Information****: The following content is written in markdown syntax. The journal requested to wrap everything in a word document and did not accept the plain `.md` files. You might want to move the content into a `.md` file and preview it with your favourite markdown editor.*

*# Reviewer 2 - Iteration 2*

*> The manuscript deals with the optimization of input data for the detection of different damage classes in trees. The different feature selection approaches should simplify the evaluation of hyperspectral data and lead to better results. The topic and the methods are exciting and deal with current interesting problems. The manuscript is well structured and well written even if some minor shortcomings are noticeable.*

*## General comments*

*> The biggest problem I see in the hyperspectral data which were analyzed. The spectral signatures are too noisy and do not correspond to any physical basis. Such ups and downs are not justifiable and therefore the data must be filtered/smoothed beforehand - as usual in other studies.*

*Image preprocessing is an important step that was handled in our case by an experienced external service provider, which applied any necessary radiometric, geometric and atmospheric corrections.*

*Additional preprocessing steps in the spectral and spatial domains may potentially be beneficial ([Vaiphasa 2006](https://www.sciencedirect.com/science/article/pii/S0924271605001012?via%3Dihub)), but smoothing techniques also have the potential to introduce new artifacts, for example by propagating errors from one spectral band into adjacent ones.*

*In this study we opted for a data-driven approach since feature selection techniques are designed to filter spectral characteristics that most reliably predict the response.*

*The chosen machine learning models are furthermore capable of weighting or averaging the numerous features either explicitly or implicitly, depending on model architecture.*

*It would be a topic for an interesting follow-up study to determine whether model-based spectral smoothing or data-driven approaches are more effective at using the available information.*

*We believe that this goes well beyond the scope of our contribution.*

*We would like to note that all code and data related to this publication is made available on Zenodo, which will facilitate future research related to preprocessing steps or other more specific aspects.*

*> I would like to see examples of all damage classes from both the field response and the canopy delineated in the EO data.*

*We attached some pictures taken in the field to our reply.*

*Note that these should not be generalized as every tree could look different, especially with respect to defoliation caused by Diplodia (which causes brown coloring of the shoots).*

*> It is not clear how the hyperspectral information was extracted for each tree. I expect that they had GNSS points for each tree. The approach of using a buffer to avoid the use of non-tree pixels is not correct. If the point is outside the tree crown also the have of the four pixels will be outside. On the other hand, if the pixel is on the tree crown the buffer can lead to include pixels from outside. Therefore, the tree crowns should (manual) delineated, and e.g. the mean of each object should be used!*

*Thanks for discussing the use of the buffer again.*

*The tree positions were referenced via GPS sensors in the field, correct.*

*We believe that the use helps to reduce possible positional errors and disagree that this approach is not correct.*

*Manual delineation is an alternative approach which is fine to use but which we did not do in this study.*

*> The discussion chapter is mainly a description of the results and doesn’t include a real discussion. I have to be revised and some parts should be shifted to the results. I want to see more comparison with other studies, both dealing with feature selection and hyperspectral data but also some tree health studies should be included.*

*Thanks for sharing your thoughts related to the discussion section.*

*We agree with the reviewer that the first part of the discussion deals with the results of this study.*

*However, rather than describing the results as in the "Results" section, results are discussed and interpreted in the "Discussion" section.*

*We believe that this is a good structural division and would like to avoid mixing result description and interpretation.*

*With respect to discussing other studies, there is already a dedicated paragraph titled "Comparison to other studies" which discusses similar studies and includes 18 references.*

*In a paragraph of this section we also note the following:*

*"In summary, no studies which used filter methods for FS or made use of NRI indices in their work and had a relation to tree health could be found.*

*This might relate to the fact that most environmental datasets are not high-dimensional.*

*In fact, many studies use fewer than ten features and issues related to correlations are often solved manually instead of relying on an automated approach."*

*The terminology "real discussion" used by the reviewer furthermore indicates that their view of structuring a discussion might differ from ours and that we most likely cannot match the reviewers desired discussion structure perfectly.*

*However, we believe that the general discussion structure of this manuscript is acceptable.*

*Nevertheless, to showcase our willingness to incorporate constructive suggestions, we added more references to studies which used feature selection in combination with hyperspectral data ([Pal 2010](https://ieeexplore.ieee.org/abstract/document/5419028?casa\_token=MpK3WG8ZqLEAAAAA:E7cICk6rfmT8H8jf4rzyFLPR0ijL5rm\_idUOix\_WvJyOB6uiZqAwanqkrSd8cIpvHT4lHxMMcd\_w), [Keller 2016](https://ieeexplore.ieee.org/abstract/document/8071759)), added two more references showcasing that most studies only describe the loss of defoliation in the field rather than modeling it ([Hlebarska 2018](https://www.cabdirect.org/cabdirect/abstract/20183376613), [Kaya 2019](https://doaj.org/article/3d9b8162586c480caea4dd3e04753868)), and pruned sections "Model differences" and "Performance vs. plot characteristics" to have less potential overlap with the "Results" section.*

*> Abbreviations introduced should be used consistently afterward.*

*We harmonized the use of abbreviations.*

*> The citation style is sometimes strange and I am not sure if it meets the journal's requirements. If there are multiple sources, I would put all numbers in one square bracket. If the source is used in the body text, I think the name of the author should be also given.*

*We've used the LaTeX citation style as recommended by the author instructions of the Journal.*

*We are using the following options*

*```tex*

*\documentclass[final]{IEEEtran}*

*\usepackage[noadjust]{cite}*

*\bibliographystyle{IEEEtran}*

*```*

*which should be accepted by the journal and community judging from [this Texexchange post](https://tex.stackexchange.com/a/44086/104937).*

*> Tables and Figures have to be understandable stand-alone. Therefore, the captions are often not sufficient. Please revise it and provide more details. Also, units are often missing and abbreviations are not explained.*

*We have updated the captions for the following figures and tables and hope they are more descriptive now and understandable standalone:*

*- 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, respectively. Values for Laukiz1, Luiando and Oiartzun were observed in 5\% intervals, for Laukiz2 defoliation was observed at multiple heights and then averaged, leading to smaller defoliation differences than 5\%."*

*- Fig. 2: "Study area maps showing information about location, size and spatial distribution of trees for all plots (Laukiz1, Laukiz2, Luiando, Oiartzun). The background maps used should give a visual impression of the individual plot area but do not necessarily represent the plot state during data acquisition."*

*- Fig. 3: "Predictive performance in RMSE (p.p.) of models across tasks. Different feature sets are shown on the y-axis. Labels show the feature selection method (e.g. NF = no filter, Car = 'Carscore', Info Gain = 'Information Gain', Borda = 'Borda'). The second value of each label shows the RMSE value (p.p.) and the standard error (SE) of the respective setting."*

*- Fig. 4: "Model performances in RMSE across all tasks, split up in facets, when using no filter method (blue dot) compared to any other filter method (red cross) for learners RF, SVM and XGBoost (XG)"*

*- Fig. 5: "Predictive performances in RMSE (p.p.) when using the Borda filter method (blue dot) compared to any other filter (red cross) for each learner across all tasks."*

*- Table II: "List of filter methods used in this work, their categorization and scientific reference."*

*- Table III: "Best ten results among all learner-task-filter combinations, sorted in decreasing order of RMSE (p.p.) and their respective standard error (SE)."*

*- Table IV: "Worst ten results among all learner-task-filter combinations, sorted in decreasing order of RMSE (p.p.) and their respective standard error (SE)."*

*- Table V: "The overall best individual learner performance across any task and filter method for RF, SVM, XGBoost, Lasso and Ridge, sorted ascending by RMSE (p.p.) including the respective standard error (SE) of the cross-validation run."*

*- Table VI: "Test fold performances in RMSE (p.p.) for learner SVM on the HR dataset without using a filter, showcasing performance variance on the fold level. For each row, the model was trained on observations from all others plots but the given one and tested on the observations of the given plot."*

*- Appendix A: "Spearman correlations of NRI feature rankings obtained with different filters. Filter names refer to the nomenclature used by the mlr R package. Underscores in names divide the terminology into their upstream R package and the actual filter name."*

*- Appendix B: "Spearman correlations of rankings obtained with the information gain filter using different $n\_{bins}$ values for discretization of the numeric response. Filter names refer to the nomenclature used by the mlr R package. Underscores in names divide the terminology into their upstream R package and the actual filter name."*

*> The quality of the maps should be improved.*

*We've made some small tweaks to the maps even though we think it was fine before:*

*- Removed red plot indicators on plot maps*

*- Referenced Bing Maps as the image source*

*- Refined legend icon size*

*- Added Burgos district to overview map (Burgos is not part of the Basque Country)*

*> Page 2, Column 2, Line 22: Please use the already introduced abbreviation ML*

*Thanks, we replaced all instances of machine-learning/machine learning with the abbreviation ML.*

*> P2, C2, L25: citation style is strange*

*We've used the LaTeX citation style as recommended by the author instructions of the Journal.*

*We are using the following options*

*```tex*

*\documentclass[final]{IEEEtran}*

*\usepackage[noadjust]{cite}*

*\bibliographystyle{IEEEtran}*

*```*

*which should be accepted by the journal and community judging from [this Texexchange post](https://tex.stackexchange.com/a/44086/104937).*

*> P2, C2, L25: I didn’t get it, please revise the sentence.*

*Thanks, we clarified the sentence as follows:*

*"The ability of predicting into unknown space qualifies ML algorithms as a helpful tool for such environmental analyses."*

*> P2, C2, L46: the index type is well known, but you are right, often not all possible combinations were tested.*

*> P3, C1, L3, L17: ML instead of machine-learning*

*Thanks, we replaced all instances of machine-learning/machine learning with the abbreviation ML.*

*> P3, C1, L44: Are you sure that all these pathogens are invasive?*

*We removed the word "invasive" as it is not relevant for this methodological study.*

*> P3, C1, L54: How was this data obtained? Please add some more details.*

*We have added "by visual inspection of experts in the field" to make the data collection even more clear.*

*> P3, C2, Figure 1, L36: The 5% intervals are clear for three test plots. Four Laukiz 2 the intervals between 0 and 15 are smaller. Please clarify. I am also not sure if you can use these values for a regression approach because they are not really metric data.*

*Thanks for taking such a close look at the data.*

*The observed values for Laukiz2 are indeed smaller than for all other plots.*

*The reason is that for plot Laukiz2 defoliation was observed at three different heights (bottom, middle, top) with the final value being composed out of the mathematical average of these three values.*

*In contrast, the observation data from the other plots consisted of a single value only.*

*This information was missing so far in the manuscript, we have rephrased the respective part as follows:*

*"Defoliation was measured via visual inspection using 5\% intervals with the help of a dedicated score card.*

*For Laukiz2, values at three height levels (bottom, mid, top) were available and averaged into an overall defoliation value, leading to values outside of the 5\% interval of the other three plots (e.g. 8.33 \%)."*

*> P3, C2, L29: Standardize the spelling of the plot names (with or without spaces).*

*We have not found any instances of Laukiz1 or Laukiz2 with spaces between "Laukiz" and its respective count ID, also not in the sentence you referenced.*

*> P4, Figure 2: Please revise the maps. Add a detailed view of the tree crowns. What is the background image? Single coordinates are not meaningful.*

*We do not have custom images showing the plots from a birds-eye perspective; note that the hyperspectral image resolution is 1 m, which does not provide any visual detail at the crown level.*

*The background images shown are "Bing Aerial" maps queried from within QGIS.*

*The purpose of these is not to show a high-resolution bird-eye view of the plots but just to give a rough overview of the plot area and its surroundings.*

*> P4, C1, L44: Spatial resolution instead of geometric resolution?*

*The difference between these terms is not fully clear, often is assumed that they are the same.*

*See also [this discussion](https://gis.stackexchange.com/a/170324/43409).*

*> P5, C1, L15: I traditionally use wrapper methods that obtain sufficient results. Therefore, a comparison of the used filter methods with wrapper methods would be very interesting, both in terms of time and accuracy.*

*We agree that the comparison of both concepts is an interesting topic, though it is out of scope for this study as the focus is on filter methods.*

*A literature research for this dedicated question might give more interesting insights.*

*> P5, C1, L21: I am not a statistician but for me, a filter changes something in the data and does not only rank the original data, but maybe I am wrong.*

*Filters in the sense of machine learning do not change the data, they simply create a ranking of variables based on their heuristic.*

*Such rankings can then be used for further variable selection or combined with rankings from other filters (in an ensemble filter scenario).*

*> P5, C1, L58: Not clear to me. I expect that specific filters work better for specific algorithms. Therefore, I am skeptical that the combination is better if every filter results in different ranks?*

*The effectiveness of filters is a function of the algorithm, dataset and tuning method used.*

*Varying one of these parameters might yield substantially different results.*

*Therefore a benchmarking matrix covering multiple settings and/or combining filter methods is usually used (as also done in this work).*

*This paragraph refers to the weighting of single filter in the ensemble compilations, which is an important precondition for an unbiased application.*

*> P5, C2, L41: use the already introduces abbreviation RF*

*Thanks, we replaced all instances with the acronym.*

*> P6, C1, L6: The last Table was II, what is the reason to place the next table after five other tables?*

*Table VIII is placed in the appendix, hence the numbering. We appended the term "appendix" to make this more clear.*

*> P6, C1, L21: Include Author’s name in the sentence*

*We manually added it even though it seems not officially supported by the citation style.*

*> P6, C1, L33-38: I don’t agree. From my experience, the feature selection should be connected to the following method. Therefore, the importance of the features is not the same for all models. Also, high correlated input data are problematic for some algorithms but not for others.*

*Thank your for this comments, which we believe points to a misunderstanding since this paragraph was not very clear.*

*The only point we make in this paragraph is that features that are obviously completely redundant were removed.*

*No filter and no model will be able to retrieve useful information from features that have a correlation of 1 with another feature.*

*We changed the paragraph to read as follows: 'Some individual features were finally removed from these feature sets if they were numerically equivalent to another feature based on their pairwise correlation being greater than $1 - 10^{-10}$.'"*

*> P6, C2, L47: not clear to me. Please explain.*

*For the permutation-based feature importance on the HR dataset, a learner is required, i.e. a ML model.*

*We used the learner which performed best on this task, i.e. SVM.*

*> P7, C1, L27: What is PCA-based correlation analysis?*

*Thanks for pointing out that the wording is imprecise.*

*We changed the subsection title to 'PCA of feature sets' and shortened the paragraph, which is intended to give a reader a very simplified impression of the complexity of feature space, as requested by a reviewer in the previous round. We changed the text as follows:*

*"PCA was used to assess the complexity of the three feature sets.*

*Depending on the feature set, 95\% of the variance is explained by two (HR), twelve (VI) and 42 (NRI) \ac{PC}s.*

*HR features are therefore highly redundant, while the applied feature transformations enrich the data set, at least from an exploratory linear perspective."*

*> P7, C1, L44: use the introduced abbreviation p.p.*

*Thanks, using the abbreviation now.*

*> P8, C1, Table III, and IV: From my point of view the Tasks with all input data should outperform the others if the feature selections work well. Therefore, I prefer wrapper algorithms and again I think they should be included for comparison (e.g. stepwise feature reduction based on the RF importance measure).*

*Thanks for suggesting the use/comparison of wrapper methods.*

*While feature selection wrappers can provide good results and enhance the predictive performance of a model, they also come with certain costs.*

*They usually take way more time to compute and used with a computational intensive algorithm (e.g. XGBoost or SVM), fitting times become quite large.*

*In contrast, filter methods are more efficient and can be cached after their first calculation.*

*Also we wanted to avoid making a subjective pre-selection of feature selection methods in this study for an unbiased comparison and the evaluation of new ensemble filters (Borda).*

*Comparing different feature selection strategies/categories (i.e. wrapper vs. filter) would be subject of a different study which has access to an unreasonably larger amount computational resources when operating with datasets of 7k or more features.*

*Besides, such studies already exist, even with a subfocus on specific subareas of modeling ([Talavera 2005](https://link.springer.com/chapter/10.1007/11552253\_40)).*

*However, more recent studies try to make use of hybrid ensemble approaches which combine filter and wrapper methods ([Chen et al. 2020](https://onlinelibrary.wiley.com/doi/full/10.1111/exsy.12553)).*

*> The order in Table 3 is not clear, 2 to 10 have the same RMSE. Also, the different SE in 8 is not clear.*

*The results in table VIII are simply the model results.*

*They are highly similar for the SVM learner, which can happen and is not unusual.*

*Learner SVM on the HR-NR-VI dataset has a slightly higher standard error.*

*>Table IV: Why does PCA not work and again it is not clear why PCA based on VI only outperforms PCA based on HR-NRI-VI. Also, the better results (compared to PCA) without filters are not clear.*

*Applying a filter or using more variables does not guarantee to result in a better performance.*

*Exploring these differences was one motivation for the extensive benchmark matrix of this study.*

*Often enough, model results do not follow a pattern one might have from previous modeling experiences.*

*Explaining why model X is better than model Y can often not be fully determined.*

*The case mentioned by you here might be one case for which this applies but there might be more to find in this extensive benchmark matrix, with the potential increasing when adding more settings to the matrix.*

*Judging from the results, it looks like PCA does not go very well with XGBoost - in our view this is everything one can draw from these non-optimal results in Table IV.*

*> P8, C1, Table VI: Why did you present these results based on the models without filters? I also want to see the best results using filters.*

*The purpose of this table is only to show the variability of results between the folds without making a deliberate choice which filter to use to estimate such.*

*We use this table in the first paragraph of Section IV. B. where we want to emphasize that performance differences between the plots are large.*

*At this point there is no added value to us in showing this information for all learner/filter/dataset combinations, given that it would require a table with four rows for each.*

*> P8, C1, L50: p.p. instead of percentage points*

*Thanks, using the abbreviation now.*

*> P8, C2 Table VII: Please explain Features (%) and values of 0.999, Values are sorted by test plots not ascending RMSE; # means the number of input data? RMSE Values are strange because higher as without filter and also the completely different results for the test plots are not clear to me: Luiando achieved the best result without filter, Laukiz 2 (the worst without filter) the best result with RF, Oiartzun the best with XGB and SVM*

*Thanks for discussing the results of Table VII.*

*`#` denotes the absolute number of features used of the respective training set. We've updated the caption.*

*`features (%)` refer to the relative percentage of features selected during tuning judging from the total amount of features of the respective training set which is the sum of all plots but the test set.*

*However values were not in percent, we've updated them now. Thanks.*

*Values were sorted by test plot name which did not match the table caption.*

*Now the table is sorted by learner and test plot.*

*Thanks.*

*In addition we discovered that some results were assigned to the wrong plots names due to an issue in the code generating the table.*

*We've updated the table.*

*The results shown here refers only the HR-NRI-VI task and we used the best learner combinations for this feature set.*

*We assume that by saying "without filter" you mean that almost all features were used for the specific plot and learner combination as there are no "no filter" combinations present in this table.*

*This table has the sole purpose of showcasing the variance within plots with respect to selected features during tuning and how these might differ among learners.*

*> P8, C2, Figure 3, include a unit for RMSE, change order of the learners (and also colors) Lasso and Ridge should be next to each other. Why did you include MBO in the name?*

*Thanks for suggesting possible improvements to Figure 3.*

*We omitted the unit for RMSE in all other tables/figures.*

*Adding it only here would make things inconsistent.*

*We believe that it should be clear to readers of the study that the unit is percentage points as it is mentioned in various places.*

*Lasso and Ridge were appended with a "MBO" suffix to highlight the use of MBO for internal optimization in contrast to using their respective integrated tuning algorithms.*

*The added value might not be fully clear, hence we removed the suffix.*

*We have changed the color palette to "Viridis", a palette with support for colorblind people.*

*> P9, Figure 4: The filters improved the results mainly for the task HR. For all others, the no Filter results are within the best results. Please explain how can a filter worsen the performance?*

*Thanks for the detailed look on the results.*

*There is no guarantee that the use of filters always improves the result, even though the selection of features is optimized via tuning.*

*As mentioned and discussed in the study, the outcome is a combination of learner+dataset+filter and using a subset of features might not always result in the best performance.*

*Also the ranking of certain filters might simply not have a positive effect and result in features being removed even though they would have been helpful to the model in the end.*

*> P9, Figure 5: Why did you highlight Borda Filter?*

*We highlighted the Borda filter because the use of ensemble filters (i.e. the Borda filter) is a central evaluation point of this work.*

*> P10, Figure 6: As I mentioned in the general comments you have to apply smoothing before analyzing the spectral signatures. The peak around 950nm is also very strange.*

*Please see our previous comment on this topic.*

*> What does scaled Reflectance or in the caption normalized reflectance mean?*

*Normalized reflectance between [0, 1] as shown on the axis title.*

*> P11, C1, L47: not instead of no*

*Thanks.*

*> P11, C1, L56-58: PCA obtained the worst results!*

*Thanks for double-checking the results.*

*Indeed PCA scored the worst results and this is what we wrote in l.52 - 56.*

*It also reduced model fitting times, which is fair to mention.*

*> P11, C2, L10-11: Instead of “not surprisingly” I would prefer citing other studies which obtained similar results.*

*Using references to support statements is usually a good idea.*

*In this case however, we believe that this statement does not need to backed by additional references.*

*> P11, C2, L36-37: Sentence is not necessary.*

*Thanks, we removed it.*

*> P11, C2, L48: Why only in almost all cases. I don’t see an option to get less or more pixels. Anyway, this approach is not a solution for your mentioned problem of pixels outside the tree. On the contrary, the method leads to the fact that neighboring pixels are included.*

*We agree that there are no possibilities to get a different number of pixels than exactly four and removed "almost" from the sentence.*

*As mentioned in previous comments, we still believe that this approach is a good compromise for the issue at hand.*

*> P13, C1, 56: Please avoid citations in the conclusion.*

*We tried to avoid or at least minimize the number of citations in the Conclusion, but we find this one reference particularly pertinent.*

*> The appendix is often not mentioned in the manuscript except Table VIII.*

*That's why the resources are listed in the appendix.*

*> Why do some of the Figures and Tables have regular numbering and some not?*

*As mentioned we are using the IEEE Tran Latex class and numbering is determined by such.*