> This manuscript deals with analysis of hyper-spectral imagery with the aim to predict tree defoliation. It is not a typical research paper – it resembles more a technical/methodological document. As the authors state: “This study aims to show how high-dimensional datasets can be handled effectively with ML methods while still being able to interpret the fitted models”. This is a typical challenge for a researcher who applies ML, successfully resolved many times over. Therefore, the level of originality is low and inflated in places (NRI, line 49, is not a “less known index type”). Nevertheless the manuscript contains a significant amount of worthy material, despite suffering from being based on a relatively poor dataset. I would like to suggest that after a major revision the manuscript could be published in Remote Sensing.

Thanks for reviewing this manuscript!  
This manuscript has a strong focus on methodology as it compares multiple feature selection methods, algorithms and datasets.  
The definition of a "typical" research paper is quite subjective and hard to define, yet, in our view this article fulfils the criteria of a common research paper in the applied spatial modelling field.

> As the authors state: “This study aims to show how high-dimensional datasets can be handled effectively with ML methods while still being able to interpret the fitted models”. This is a typical challenge for a researcher who applies ML, successfully resolved many times over.

Every modeling case is somewhat unique and while certainly specific tasks seem similar across studies, we could not find a study which used filter-based feature selection on a highly-correlated hyperspectral dataset.  
In addition our approach shows one way how to solve this challenge while there of course other legit ways to tackle this issue.  
Hence we think that this work is beneficial to the community and worth publishing.

We are sorry that you find our classification of NRI indices "inflating".  
The interpretation whether NRI is a well-known index or not certainly varies within the community and target audience.  
Yet, we still believe that this attribution holds true and that NRI is a "less known index type" (e.g. when comparing with vegetation indices) as it is rarely used in the literature and also sparsely known within our local remote sensing community.

We agree that the dataset in this work has some issues which are extensively discussed in the manuscript (see especially section 4.5).

Please find our line-by-line replies below and thanks again for reviewing our manuscript!

> Title

> I find the title misleading, as the paper compares more variants to the ML application than just feature selection. Perhaps the sentence from Methods “The development of robust methods which enable an unbiased estimation of feature importance for highly correlated variables are subject to current research” is good clue about what the authors are after, and in such case it is placed in a wrong place.

Thanks for the feedback.  
The main focus of the methodology is the comparison of different filter methods (including the ensemble filter method "Borda").  
Filters are a sub-category of feature selection methods.  
We included several ML techniques in order to be able to determine whether the superiority of specific filter techniques - or lack of superiority - is valid across a wide range of ML models or specific to some types.  
The original title reflects this perspective, and that the manuscript's storyline is aligned with this focus as well.  
The issue of model interpretability is certainly an important discussion item, which is why it was addressed in the Discussion section; however, overall the paper focuses on measuring performance improvements, which is reflected by the title as well as the clear focus of the Results section on performance measures.  
We therefore prefer not to change the title but appreciate your remark to reconsider it.

> '...which is subject to current research':

We have added a reference (https://arxiv.org/abs/2104.04295) to this sentence in the manuscript which should strengthen this statement.

> Introduction is short in comparison to Methods and Discussion. The latter section seems to contain some paragraphs which point out to important relevant research and which would be better placed in the Introduction. Some parts of the Methods contain general remarks that seem to be also suited better for Introduction (such as the Filter Methods section).

Thanks for the feedback.  
In our view, having more content in the "Methods" and "Discussion" parts than in the "Introduction" is favorable and not a shortcoming.  
Linking to relevant research can be done both in the "Introduction" and "Discussion"  sections and, in our view, it is common to reference research which links to the discussed topics in their respective paragraphs.  
If you can provide us specific pointers to paragraphs which might potentially be misplaced, we are happy to take a closer look.

With respect to the mentioned paragraph on the introduction of the filter methods, we believe that this should live in the "Methods" section as it describes parts of the used toolstack.  
We are aware that some scientists prefer to introduce their methods in the introduction and solely focus on a high-level workflow description in the "Methods" section, however we believe that explaining/introducing the methodology in the "Methods" section is appropriate.

> Moreover, Introduction does not narrow down, or explain why the scope includes filter methods and not wrapper methods. Only in the Methods (too late) the authors state: “Due to the focus on filter methods in this manuscript, only this sub-group of feature selection methods will be introduced in greater detail in the following sections.”

Thanks for the feedback.  
The understanding of why filter methods are used in this manuscript is important and we are sorry if this was not mentioned clearly enough in the manuscript.

In l.61, which already introduces the use and focus of filter methods, we rephrased the text to

"Instead of using wrapper feature selection methods, which add a substantial overhead to a nested model optimization approach, especially for datasets with many features, this study focuses on the use of (ensemble) filter methods which can be directly integrated into the hyperparameter optimization step during model creation."

In l.151 we expanded the already existing explanation to

"In addition, filters are less known than wrapper methods and in recent years ensemble filters were introduced which have shown promising results compared to single filter algorithms \cite{drotar2017}.  
These two points mainly led to the decision to focus on filter methods for this work and evaluate their effectiveness on highly-correlated, high-dimensional datasets.  
Due to this focus, only this sub-group of feature selection methods will be introduced in greater detail in the following sections."

> The objectives are not clearly stated in the Introduction. Some objectives are stated as late as line 398 (Discussion): “One objective of this study was whether expert-based or data-driven feature engineering has a positive influence on model performance.”

Thank you for the feedback.  
In the objectives in the Introduction we state

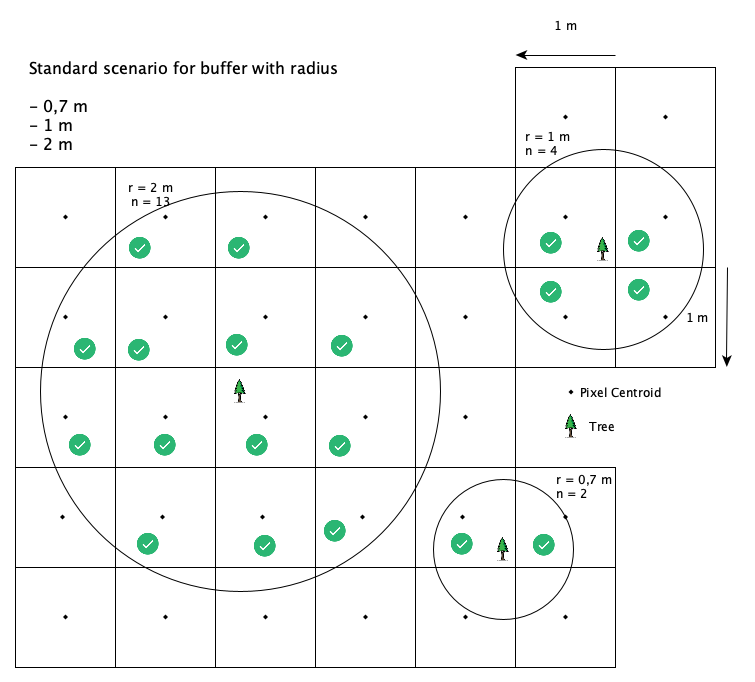
"Do different (environmental) feature sets show differences in performance when modeling defoliation of trees?"

Which is a rephrased version of the question referred to in l.398.

> Design:  
> Here I would like to flag three significant problems:

> Sampling of the hyperspectral data is based on points with 1m buffer (to account for a positioning error), which in practice translates to 4 pixels per tree, in most cases. I would suggest that this means that each tree might be very poorly represented. I assume that these trees have crowns wider than 1 -2 m, and that ~4 m2 (possibly with some spectral information from outside of the tree) might represent only a small fraction of the crown. I recommend revising this approach.

Thanks for asking for clarification on this matter.  
A buffer of 1 m radius always results in four contributing pixels for each tree.  
The authors discussed this issue very carefully and came to the conclusion that using a buffer of 1 m is the best compromise between ensuring that the data actually corresponds to the respective tree, and covering the tree's camopy as completely as possible.  
There has to be a trade-off because there will be no perfect solution; either we decide to not account for possible positioning errors which would increase the risk of using a pixel value which is not representing the tree at all (and possibly even bare ground, especially in the corners of the plot) or, by using a too large buffer, average out the actual tree reflectances by including too many surrounding pixels (and again possibly including bare ground).





In our view, a buffer of 1 m radius actually matches crown sizes of most trees quite well (please see the attached images for a better illustration).  
Actual tree crown sizes depend on the age of the tree and will vary across plots.  
These sample trees will also be included in the graphical abstract of this work.

> The VI feature set is of completely different nature than HR and NRI, because the VIs only represent some portion of the analysed spectra, which is equivalent to elimination of a lot of features a priori, whereas in the case of the other sets this elimination occurs later.

Thanks for mentioning this.  
We agree that the VI dataset only covers parts of the spectral region and that this might be a potential disadvantage compared to the other feature sets.  
However, this is also part of the research question of this work, whether the different feature sets will show substantial differences, with the advantages and disadvantages of each taken into account.  
This is already briefly mentioned in section 4.4 in the discussion, however we also think it is worth highlighting this more explicitly when introducing the feature sets.  
We have added the following note to section 2.4.2 in the manuscript:

"Each feature set has distinct capabilities which differentiates it from the others.  
This can have both a positive and negative effect on the resulting performance which is one issue this study aims to explore.  
For example, feature set VI misses certain parts of the spectral range as the chosen indices only use specific spectral bands.  
Feature set NRI will introduce highly correlated features, for which some algorithms may be more suitable than others."

> Moreover it is not clear how many VIs were used, as it is only mentioned that the 400-1000 nm allowed to calculate only some of them.

This is already mentioned in l.126 in which we state

"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."

> Computational requirements are not included in the comparison of variants, while constraints in this regards seem to have affected the results, as it was decided that 70 iterations for the XGBoost will be executed.

Thanks for asking about computational resources.  
Computational requirements for executing computational studies always depend on the hardware resources at hand.  
These will always vary (in CPU time, cores and memory) and hence there is usually not much added value in describing this matter in greater detail.  
In our case, the analysis was run in parallel on a High-Performance Cluster with 6 nodes.  
There are different resources per node (from 32 cores, 128 GB to 48 cores, 256 GB).  
The runtime to execute all 156 settings was several days, however the resources of the cluster were shared with other scientists and the actual runtime at full capacity is therefore hard to estimate.

There were no restrictions with respect to runtime for the model optimization.  
One always needs to decide on tuning set ranges before optimizing and we ensured that our boundaries were not set in a way that the models are limited by these, i.e. in this particular case we have seen the hyperparameter `nrounds` to level off at earlier values and there was no justification to increase the right border of this hyperparameter.  
We agree that in other scenarios a higher value of this hyperparameter could have been helpful for the model.  
However, algorithms always rely on a combination of hyperparameters and their optimal configuration always differs depending on the dataset at hand.

> Line 231 – State by which % the number of features has been reduced through the analysis of pairwise correlation. Also, it would be worth discussing why do authors think that allowing most of the features to pass this test is better than eliminating them in subsequent, more computationally intensive, stage of the analysis.

Thanks for requesting details for this part.  
Due to an update in the code we had to rerun some benchmarks which also had some effect on the pairwise-correlation preprocessing step.  
We made it more clear how many VIs were removed by the pairwise correlation check and why we chose this limit:

"This preprocessing step reduced the number of covariates for feature set VI to 86 (from 89).  
This decision was made to prevent issues for the subsequent tuning, filtering and model fitting steps which might occur when passing features with a pairwise correlation of (almost) one.  
Additional variable selection was solely devoted to the respective feature selection methods in this study."

> Table 3. Please revise the results. RMSE and SE values are the same everywhere. Line 320 cites different RMSE value than what is in the table.

Thanks for taking a close look on the results.  
In fact, the results for the best ten models are equal when rounding to two decimal places.  
We've increased the table to use three decimal places now to showcase the small differences.  
The wrong value in l.320 is an oversight on our side for which we apologize.  
We have corrected the value.

> Figure 3. SE values not presented, but mentioned in the caption. Were RMSE the same for all SVM cases?

Thanks for checking Figure 3 in detail.  
In fact, when rounding to two decimal places, results for SVM were identical for certain feature sets and learners.  
We have also increased the number of digits to three to highlight small differences here.  
With respect to the missing SE values: these were available in the plot in a previous version of the figure, however we decided to remove the information again as it made the figure overloaded in our view.  
We now removed the reference in the caption as well, thanks for catching this!