A data efficient machine learning model for autonomous operational avalanche forecasting

Michaela Wenner, Reviewer comments

May 2021

1 General comments

Chawla et al. present results from an automated avalanche risk assessment algorithm using machine learning (ML). The authors propose random forest (RF) as their ML-Algorithm of choice. RF is an ensemble learning method meanwhile widely used in the field of natural hazards. The authors thoroughly explain the method and benefits thereof. As input features they use intrinsic snow-meteorological features such as temperature, snow height, wind speed and occurring avalanches of the last 24 hours. Additionally, they derive further features from this data, such as the change in snow surface temperature and new snowfall as well as wind speeds of the last couple of days. They process in total seven years of data; three years to train the model and four years to test the model. They define two different classes: no-avalanche day and avalanche day. In order to counteract imbalanced data (more days without avalanches) they applied a filter to only consider days with a snow height larger than 0.5m. After tuning some hyper parameters (tree depth and number of trees) they show the Heidke Skill Score (HSS), False Alarm Rate (FAR), Probability of Detection (POD) and Precision results of the training process (5-fold-cross validation). They end up with a HSS score of max 0.33. When applying the model on the test data set (called generalization in the paper), the performance increases to a max HSS of 0.42. From thereon specific days and predictions are discussed and an example of a decision tree from this day is shown graphically. They conclude that random forest is superior to other ML techniques for avalanche hazard assessment due to its data efficient behavior and comprehensive feature analysis. It can therefore help forecasters to make a decision not only by the probability output of the model but also its "graphic" decision making process.

The study is well within the scope of NHESS and is interesting for the avalanche hazard community. The difficulty of predicting avalanches is a well known problem, therefore the low prediction accuracy does not come as a surprise.

A strong point of the study is surely the discussion on what the visualisation of a single RF tree can tell us about the prediction making process and the decision rules that were applied for each feature. As RF is a relatively simple ML algorithm this might encourage more people in the community to make use of automatic techniques to find decision making rules in a data set. Additionally the paper profits from a clear introduction.

However the paper suffers from several points: (1) Performance analysis: the authors refrain from setting a classification threshold which therefore does not allow to give clear evaluation of a confusion matrix and how well the model actually performs. There are ways to evaluate the performance more clearly, which I will mention further in the specific comments. Additionally, a more thorough feature analysis (over all trees) should be performed to find most important features on a RF level and not just a single decision tree level (2) A direct comparison to other ML algorithms is missing. The authors do compare the algorithm performance to other studies, however, the amount of data they use is very different. A simple performance comparison to another ML technique (such as k-nearest-neighbours) would give a much better insight on how capable RF is. (3) There are

several typos and unclear sentences with sometimes incorrect grammar and often incorrect punctuation/white spaces. Therefore, before publication can be granted, I recommend a thoughtful review of the aspects detailed below.

2 Specific comments

The paper suffers from an unclear structure, especially in the methodology, results and discussion part. I recommend restructuring the paper according to the conventional structure: Start out with explaining every method used in this paper, then show and describe the results obtained and in a last step discuss what this might mean. Specifically for this paper I would recommend moving section 3 to directly after the introduction, then continuing with a methods part which can be subdivided into the description of random forest, then the data pre-processing, performance measures, model training and hyper parameters tuning and model generalization. Then show and describe the results of above mentioned methods. This means to basically describe Figures 4-7. In the discussion part the authors can then mention the model output interpretability and data efficiency and potential for autonomous process.

Another aspect which would greatly improve the paper is a more extensive performance analysis. I suggest using receiver operating characteristics (ROC) and the area under the curve (AUC) to evaluate the model performance additionally. This way, one can directly see the trade-off between the true positive rate and false positive rates. Then I suggest to define a threshold and include a confusion matrix in the performance analysis. I understand that in the end the threshold value is for the "operator" to define, but I think for this study it would benefit to give an example threshold and evaluate the model with a confusion matrix based on this value. I do like the probability presentation in Figure 5, but I think an additionally confusion matrix is needed. Furthermore, instead of only showing single trees, a feature importance analysis over the whole forest would greatly improve the manuscript.

In line 291-304 and table 7 the authors compare their model to other studies using different ML techniques. However, the amount of training data used in other studies is quite different to what has been used in this case. This of course makes it hard to actually compare the performance between two ML algorithms. I think it would be beneficial for this study to compare the results obtained with RF to another ML technique, e.g., support vector machines or k-nearest neighbours. This way, the authors can describe (1) why they prefer RF as techniques - of course also because of its interpretability and (2) how the amount of training data influences the performance - in comparison with the other studies and (3) explain differences to other studies, e.g., in the features that were used.

Line 326-329 the authors state that new features can be added easily to improve the model. I do not understand completely how this is done. Please explain more excessively and give references.

Please remake Figure 2. Instead of only showing the medians I suggest to use boxplots for a clearer picture of climatic conditions at the study area.

Add a few sentences in the discussion on how the model could be generalized for example for other sites. Is it possible to use one and the same model for all sites? Or sites at similar altitudes? Please discuss.

3 Technical corrections

3.1 Abstract

l11 Delete "world over"

3.2 Introduction

l28 "e.g.," comes with a comma behind the second dot. This is forgotten several times throughout the manuscript

l33 Please add white spaces. Also this is encountered quite often

l45 Please rephrase the sentence.

l59-65 Please clarify what is meant here.

l78 If you mention detection of avalanches using seismic data, you might add our recent paper (Wenner et. al 2021) in which we actually also use RF to detect mass movements (also avalanches)

l83 Maybe rephrase to "(...) for binary classification (avalanche day – no-avalanche day)

l87-88 Is the input not the same as for other ML models? Please clarify what you mean with this statement

l90-95 As mentioned in the specific comments, I think a restructuring of the paper would benefit the overall reading experience

3.3 Random Forest Technique

l104 Please write "until" instead of "till" (same for line 107)

l108-111 Please rephrase and clarify

3.4 Study Area and Data Characteristics

l148 In the north-western part of "the" Indian Himalaya

Table 1 "Intrinsic snow-meteorological features used by "the" model" (I suggest to include a "the" before model)

3.5 Data pre-processing

l171 Leave out "only"

Table 3 I would find it more informative to see the total number of avalanches days instead of the mean, and accordingly total number of avalanches that happened with a snow height below 0.5m

Table 4 I like the table, it gives a great overview. Maybe you could just add in the caption what i,j are (true label vs classifier label)

3.6 Performance measures

l186 Here, the confusion matrix is described, but then not used in the paper. I strongly suggest to do that though (as mentioned in specific comments)

3.7 Model training and hyper-parameter tuning

Figure 4 Add standard deviation from 5-fold-cross-validation. Additionally, make sure that the term "5-fold-cross-validation" is equally written in the text and the figure title

l197 Explain quickly what cross-validation is and what the 5-fold means

l198 There are more than two hyper-parameters for RF. Explain why you didn't change those or rather why you set it to this value.

l202 Please rephrase and clarify this statement

l205-209 This should definitely go to the discussion section of the paper

l210-212 Delete "Based on this premise" and rephrase sentence
l226 Yes, but this is intrinsic with the definition of both. How about the the POD? Also, please address Figure 4a and Figure 4b separately and explain what is shown.

l232 What would those complex situations be? Please discuss (in the discussion section)

l233 Consider using normal words instead of the feature names - new snow instead of new_snow (the connection is easy enough)

3.8 Model generalisation

l239-241 Consider rephrasing the sentence to: "After hyper-parameter tuning through cross-validation experiments for optimized performance, we constructed (...)"

3.9 Model Output Interpretability

l254 How can you find unstable slopes with that? Please clarify

l262 Consider rewriting to "(...) high wind speed as the most indicative feature for avalanche hazard"

l280-281 Please clarify the sentence

l283 I suggest to skip the last sentence

l290-304 I like the comparison to other studies, however I think it could be a bit more comprehensive. Maybe rephrase and make sure to clarify your message. Also, this should be in the discussion section.

3.10 Comparison with other models

l309 "The model uses lesser": what do you mean by lesser? In terms of quality or quantity?

3.11 Data Efficiency and Potential for Autonomous Process

l326-328 Please rephrase and clarify

4 References

Wenner, M., Hibert, C., van Herwijnen, A., Meier, L., and Walter, F.: Near-real-time automated classification of seismic signals of slope failures with continuous random forests, Nat. Hazards Earth Syst. Sci., 21, 339–361, https://doi.org/10.5194/nhess-21-339-2021, 2021.