Interactive comment on "Auroral Classification Ergonomics and the Implications for Machine Learning" *by* Derek McKay and Andreas Kvammen

Anonymous Referee #1

Received and published: 16 March 2020

This paper addresses the sources of expert-injected bias in the training samples for auroral classification. By developing an auroral training data set, the authors listed the bias to serve as a checklist for improving the training data integrity. It is an interesting point for machine learning researchers. However, the organization of the manuscript needs improvement to get published. My concerns and questions are listed below.

Major issues: This paper reads like neither a behavior study paper nor a data paper because it does not provide much quantitative results about the annotation procedure or the resultant dataset. Providing a checklist of the expert-related biases does not make much contribution since these biases have been reported in psychology and cognition research. I would like to see a data paper like that about ImageNet, PASCAL VOC in the machine learning field.

Minor issues: 1. Section 2, the original images were "cropped to the central 128x128 pixels and data were binned …", was the image cropped to 128x128 first and then binned to 64x64 or less, or the cropped image was smoothed by binning? The description is not clear. In addition, what is the criteria for choosing the parameter 128? 2. Section 2, two physicists processed the images using different software implementations, did they follow the same protocol? Why did not they use a same software to exclude unnecessary discrepancy? 3. Section 2, a brief introduction to the characteristics of each category is expected. Showing example images of each category is helpful to more intuitively explain the problems encountered during classification, such as data contrast bias and environment contrast bias. 4. Section 2, images not suitable for training machine learning algorithm were also removed. What is the standard of suitable for training? 5. Section 3.2, the classifications were done randomly and chronologically, respectively, by the two experts. However, classify chronologically seems more likely to encounter the situation that authors mentioned 'If the classifier has just seen a faint, patchy aurora, then a following faint, patchy aurora is likely to be classified the same'. 6. It is expected to see some experimental results on the labelled dataset using currently available machine learning methods, which can be treated as a baseline for further research. 7. Since the authors provided the data list in the supplement, I tried to download the original images accordingly. Unfortunately, 4767 images in the 13947 list are not exist on the website (http://www2.irf.se/allsky/). The web link provided in the manuscript (http://seid.uit.no/data/) is broken. As a researcher in auroral image analysis, I do hope the authors publicly release the dataset completely (original images + annotations).