Interactive comment on “Estimation of effective calibration sample size using visible near infrared spectroscopy: deep learning vs machine learning” by Wartini Ng et al.

Anonymous Referee #3

The paper discusses the performance of convolutional neural networks (CNN) compared to traditional machine learning techniques in function of the number of calibration samples. The first reviewer of the previous version misunderstood the objectives and therefore had serious doubts on the novelty and the strategies. The authors have changed the title and clarified the objectives in order to deal withthis misunderstanding.

Thank you for taking the time to review our manuscript. We shall address the comments and revise the paper accordingly. Our detailed responses are as follow:

The second reviewer mentions a lack of a discussion section. This has now been added.

Specific comments

Line 59, 85-86, 130, 139 …. (please check throughout the anuscript) …spectral data… or ‘spectra’ on its own.

Response: We have checked the consistency through-out the text.

Line 59-60 …predict clay content..

Response: We have corrected this.

Line 64 …soil spectral libraries…

Response: We have corrected this.

Line 68 What are ‘increased calibration samples’? Is there a word missing?

Response: We have corrected this. We are refererring to calibration sample size.

Line 68 Split up the sentence: …performance. However, there ….

Response: We have split the sentence.

Line 70 A strategy …

Response: We have added the article “A” as suggested.

Line 73 …. How many samples….

Response: We have corrected the word ‘much’ to ‘many’

Line 75 CNN will outperform traditional…

Response: We have corrected this

Line 78 …CNN model to outperform machine….

Response: We have corrected this

Lines 81, 82 Delete either ‘specifically’ or ‘specific’

Response: We have removed one of the word

Line 83 ..models will reach..

Response: We have corrected this

Line 87 There seems to be a word missing : …achieved wen the number…

Response: We have added a word

Line 92 Delete ‘of’

Response: We have deleted the word.

Line 95 …and sedimentary rocks..

Response: We have corrected the word

Line 95 Delete ‘samples’

Response: We have deleted this

Line 102 In general it is better to work as much as possible with the original data. There is no need to multiply all organic carbon contents with 1.724. I would prefer if you use soil organic carbon (SOC) for all prediction models. After the Van Bemmelen factor is an empirical one and adds unnecessary noise to the data.

Response: We reported the value in OC as suggested

Line 103 .. to extract exchangeable aluminium, calcium and magnesium…

Response: We have corrected this

Line 117 …spectral measurement…

Response: We used the word spectra through-out

Line 152 …as one-dimensional data…

Response: We have corrected this

Line 156 Word missing : …was trained using a batch…

Response: We have added the word

Line 200 ….each filter of…

Response: We have corrected this

Line 260 I do not understand what you want to illustrate in Fig. 7. The sensitivity analysis is shown in Fig.8.

Response: We have corrected the reference for the figure

Line 333 Delete ‘in this paper’

Response: We have removed the word.

Anonymous Referee #1

Thank you for taking the time to review our manuscript. Kindly find our responses below:

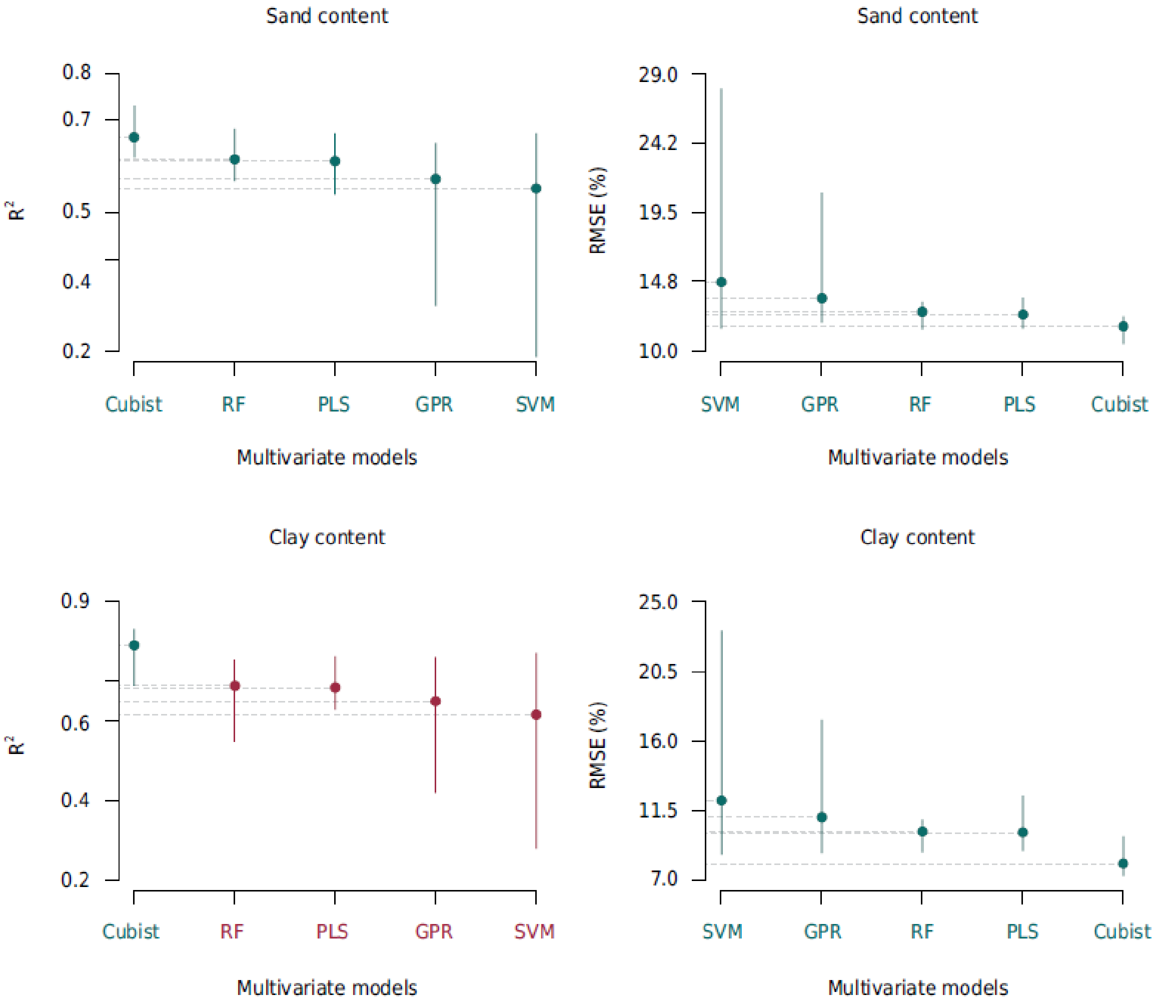
- I still think comparison with RF and SVM needs to be conducted as many studies have shown the superiority of these algorithms rather than other machine learing techniques. Moreover, as many researchers in soil community use these machine learning techniques, they need to know their performance compared to deep learning.

Response:

The aim of this paper is not to compare various machine learning algorithms. There are many papers who have done that already. The aim of this paper is to assess the effect of training sample size on the accuracy of deep learning and compare it with some machine learning models as benchmark.

Here we included findings from other researchers that conducted the study as requested by the reviewer for various soil properties below.

Silva, E. B., Giasson, É., Dotto, A. C., Caten, A. t., Demattê, J. A. M., Bacic, I. L. Z., and Veiga, M. d. 2019. A Regional Legacy Soil Dataset for Prediction of Sand and Clay Content with Vis-Nir-Swir, in Southern Brazil. Revista Brasileira De Ciencia Do Solo, 43.



Sorenson, P. T., Small, C., Tappert, M. C., Quideau, S. A., Drozdowski, B., Underwood, A., and Janz, A. 2017. Monitoring organic carbon, total nitrogen, and pH for reclaimed soils using field reflectance spectroscopy. Canadian Journal of Soil Science, 97(2), 241-248.

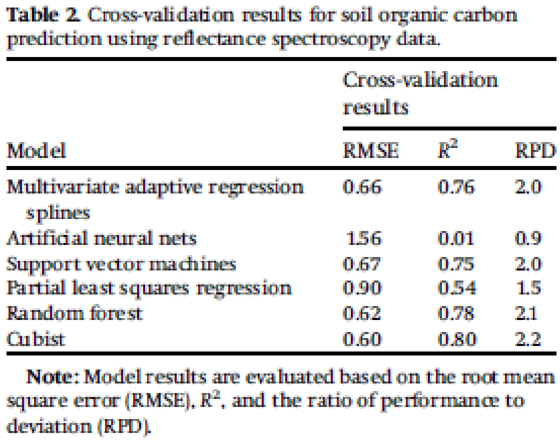


Table 2. Cross-validation results for soil organic carbon prediction using reflectance spectroscopy data.

	Cross-validation results		
Model	RMSE	R^2	RPD
Multivariate adaptive regression splines	0.66	0.76	2.0
Artificial neural nets	1.56	0.01	0.9
Support vector machines	0.67	0.75	2.0
Partial least squares regression	0.90	0.54	1.5
Random forest	0.62	0.78	2.1
Cubist	0.60	0.80	2.2

Note: Model results are evaluated based on the root mean square error (RMSE), R^2, and the ratio of performance to deviation (RPD).

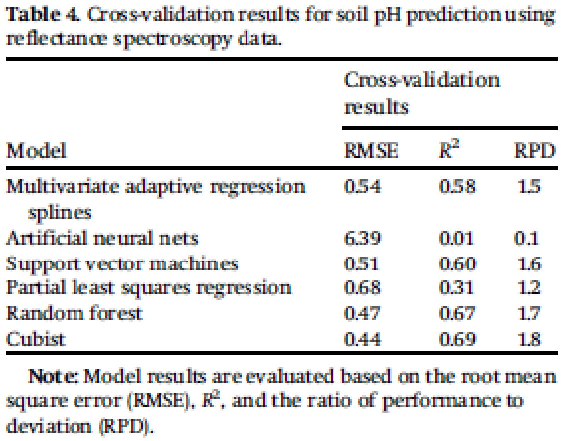


Table 4. Cross-validation results for soil pH prediction using reflectance spectroscopy data.

	Cross-validation results		
Model	RMSE	R^2	RPD
Multivariate adaptive regression splines	0.54	0.58	1.5
Artificial neural nets	6.39	0.01	0.1
Support vector machines	0.51	0.60	1.6
Partial least squares regression	0.68	0.31	1.2
Random forest	0.47	0.67	1.7
Cubist	0.44	0.69	1.8

Note: Model results are evaluated based on the root mean square error (RMSE), R^2, and the ratio of performance to deviation (RPD).

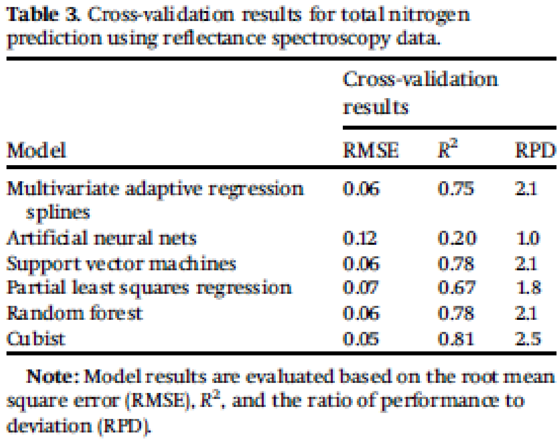


Table 3. Cross-validation results for total nitrogen prediction using reflectance spectroscopy data.

	Cross-validation results		
Model	RMSE	R^2	RPD
Multivariate adaptive regression splines	0.06	0.75	2.1
Artificial neural nets	0.12	0.20	1.0
Support vector machines	0.06	0.78	2.1
Partial least squares regression	0.07	0.67	1.8
Random forest	0.06	0.78	2.1
Cubist	0.05	0.81	2.5

Note: Model results are evaluated based on the root mean square error (RMSE), R^2, and the ratio of performance to deviation (RPD).

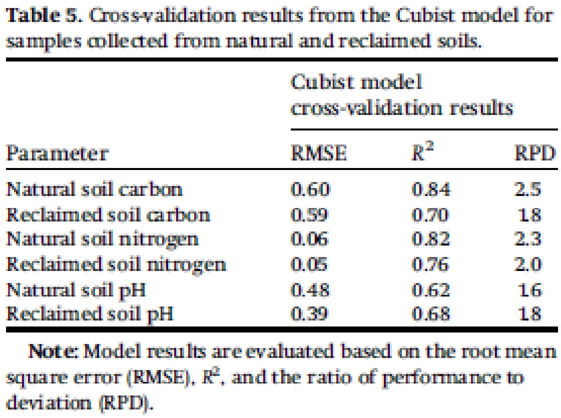


Table 5. Cross-validation results from the Cubist model for samples collected from natural and reclaimed soils.

	Cubist model cross-validation results		
Parameter	RMSE	R^2	RPD
Natural soil carbon	0.60	0.84	2.5
Reclaimed soil carbon	0.59	0.70	1.8
Natural soil nitrogen	0.06	0.82	2.3
Reclaimed soil nitrogen	0.05	0.76	2.0
Natural soil pH	0.48	0.62	1.6
Reclaimed soil pH	0.39	0.68	1.8

Note: Model results are evaluated based on the root mean square error (RMSE), R^2, and the ratio of performance to deviation (RPD).

From both studies, Cubist > RF and SVM.

Several studies have also shown the superiority of CNN to SVM and other models for large dataset, for example, Tsakiridis et al. showed that CNN performs better than PLS, Cubist, SVM, and SBL algorithms. SVM has the same performance as Cubist.

Tsakiridis, N.L., Keramaris, K.D., Theocharis, J.B. and Zalidis, G.C., 2020. Simultaneous prediction of soil properties from VNIR-SWIR spectra using a localized multi-channel 1-D convolutional neural network. Geoderma, 367, p.114208.

Hence, we did not see the usefulness of repeating another comparison, which is not the aim of the paper. There are a lot of machine learning algorithms out there; including RT, RF, SVM,NN, ANN, MBL, etc. It is endless if we keep making comparison with every single model.

- In my opinion the section "sensitivity analysis" in mathodology does not add much to manuscript and in addition it is confusion and not clear, it needs to be removed or at least revised or shortend.

Response: As we mentioned in the manuscript, CNN is known as black box. As this is a soil science journal, we believe it is important to be able to interpret a complex model to understand how the decision was made by the model.

- The Conclusion is vague and requires revision and clarification.

Response:

We have clearly outlined our conclusion which answer the objectives of our paper. For sample size < 2000, the performance of CNN (with its current architecture) is not better than PLSR and Cubist. The increase of performance can only be seen when sample size > 2000. Thus we can recommend to other researchers that it would not be useful to try CNN unless you have enough data.