Review of “Leveraging Machine Learning techniques and SEVIRI data to detect volcanic clouds composed of ash, ice, and SO_2” by Naranjo et al.

Reviewed by Andrew T. Prata (andrew.prata@bom.gov.au)

General comments

In this paper, Naranjo et al. propose a new volcanic cloud detection algorithm using a neural network (NN) applied to eruptions from Mt Etna. The main issue they tackle is volcanic cloud detection when the volcanic clouds do not exhibit a ‘traditional’ volcanic ash signal (i.e. a negative 11-12 BTD). They show that this issue is common for a series of Etna eruptions from 2020-2022 and 2024. The authors rely on manual labelling, primarily informed by Ash RGB composites, to derive training data and build a balanced, 50/50 Volcanic Cloud (VC) and non-Volcanic Cloud (NVC), labelled dataset for training and testing. The authors found very good results when validating their algorithm against three Mt Etna case studies that the algorithm had not seen in the training phase (accuracy=99%, precision > 89% and recall >74%). The authors also find that performance degrades over time as the VCs become more dilute and detection sensitivity becomes challenging. Interestingly, the authors find that the false positive rate is lowest for the most complex case which they argue might be because the human operator (manual labeller) tends to be more conservative in these situations (drawing a smaller volcanic cloud boundary) and therefore dilute pixels (where the NN may struggle) are less likely to be labelled as VC. Neural networks are becoming increasingly useful in feature detection in satellite imagery and I am pleased to see this method being applied to the complex task of classifying volcanic clouds composed of ash, ice and SO2. I certainly find the contribution novel and that it addresses a longstanding and complex problem - detection of hazardous volcanic clouds that are not dominated by an ash particle spectral signature. The theme of this review is essentially to ask for further details. In that respect my comments are mostly minor, depending on how the authors choose to address them. I don’t believe the analysis needs to be redone - just more details on how it was performed are required. I would also comment that providing the manually labelled dataset to the research community would be very beneficial. This would allow researchers to design and compare new algorithms based on this valuable training dataset that has been put together.

Specific comments

Training dataset

It would be useful to have some more detail on how the training dataset was constructed. I'm also interested in understanding how the data were randomised. Can the authors provide the spatial region that was used for the analysis (i.e. latitude/longitude box) and how many pixels are contained within this region? Was there any consideration of balancing the non-VC part of the dataset by defining other non-VC classes? I'm wondering if the non-VC pixels were dominated by some other non-VC class (e.g. is it mostly clear-sky pixels?). Similarly, in time, were the non-VC pixels only sampled during eruptions of Mt Etna? These questions have relevance for understanding any bias that might be in the non-VC labelling e.g. If you only sample non-VC pixels in the vicinity of Etna, then the technique might not generalise to other regions of the world (Indonesia, Iceland etc). Similarly, if pixels are only selected during Etna eruptions then it is possible that you may have a temporal bias in your non-VC class but only if the Etna eruptions don't adequately cover all seasons.

Selection of BT and BTD features

I'm interested in understanding how these combinations were chosen. Some are clear (e.g. 11-12 BTD for ash presence) but others are new and interesting (e.g. 11 - 9.7). One could construct several more BTD combinations (e.g. 7.3 - 6.2 may contain a UTLS SO2 signal), but I'm wondering if any statistical technique was used to choose the BTD combinations to begin with? Or was it simply the approach of just pinning all the differences to the 11 μm channel?

Clarification around ML methods

Some choices are simply stated without much motivation or reasoning. I wonder if this is because the choices are standard practice in ML research (which I am not an expert in)? In my line-by-line comments I have raised questions where I don't fully understand what was done. Any attempt to address these questions will help improve the clarity of the paper.

Plume tracking

The plume tracking algorithm looks really effective. However, I'm unsure how the analysis presented in Fig. 4 is related to the analyses presented later in Figs. 5, 6 and 7. For example, how is the "Enclosing" purple line on Figs. 5, 6, 7 related to the expanding circular radius in Fig. 4? Also, how do you prevent the circular region in the plume tracking from growing very large? If the circular region is expanded at every time step, then for long lasting eruptions (>12 h) like the Fig. 6 (4 March 2021) for example, what prevents the region from continuously growing (thus potentially introducing more false positives)? Further, if this region is not limited, then could this be partly the reason for your increase in false positives with time? I.e. the plume tracking radius is becoming very large and is introducing noise from other regions far away from the volcanic cloud?

Evaluation metrics

I know the metrics reported are standard in ML literature but I think it would be beneficial to include their mathematical definitions when describing how these metrics were implemented as there are several pre- and post-processing steps. I also think the confusion matrix results should potentially be reported in the abstract or at least be more upfront alongside the accuracy, precision and recall. The analysis presented in Fig. 8 and the discussion around it are very useful and insightful. Here we see how the algorithm's performance degrades over time due to the volcanic cloud signal becoming "diluted" as the authors hypothesise. But I also find it useful to know that the model only classifies a VC *when it actually is a VC* 66%, 49% and 84% of the time for the three validation events (assuming I'm interpreting the confusion matrix plots correctly). I think these numbers are more insightful than, say, the accuracy metric (e.g. 99% accuracy sounds like you have a 'perfect' model).

Mt Etna VC compositions and mass loading retrievals

The mass loading comparison is useful as it demonstrates that while the NN may miss the diluted plume pixels, it is not too serious in the context of total mass loading estimates because the diluted pixels do not contribute much to the total mass per unit time. However, I'm not sure I understand how the individual mass loading time-series were calculated for the NN masks if the algorithm simply classifies pixels as Volcanic Cloud (VC) or non-VC. If the pixel is "VC", how do you separate the mass loading into ice, ash and SO2? I think it would be useful to define what you mean by VC. E.g. Are volcanic clouds = ash + SO2 + ice, or could they be either ash, SO2 or ice or something else? Also, a component that hasn't been mentioned is sulphate. Have the authors considered the influence of sulphate aerosols on the RGB signals and BTD combinations they're using? E.g. Sellitto and Legras (2016): https://amt.copernicus.org/articles/9/115/2016/

Conclusions/future work

Based on results and analysis in the paper, it appears to me that the authors have a very good algorithm for detecting ash+ice+SO2 clouds from Mt Etna. This kind of analysis would really benefit VAACs that deal with these types of clouds routinely e.g. the Darwin VAAC routinely see volcanic clouds rich with ice and SO2. I think an area for future work could be to see whether or not the algorithm can characterise ash/ice/SO2 clouds from other eruption case studies (in different atmospheres e.g. Russian/Alaskan/Icelandic volcanoes at high latitudes or Indonesian volcanoes in the tropics). I agree that addressing the diluted parts of the plume is an area for improvement. However, I think it would also be valuable to understand if this method/algorithm would generalise to other geostationary sensors and/or other eruption case studies. Or would a new algorithm (maybe with the same network architecture and input features) need to be re-trained on the relevant dataset (e.g. Himawari/GOES)? Some comment or recommendation on this aspect of the research would be helpful.

Line-by-line comments

Title: I don't think you need to capitalise "Machine Learning". Just use "machine learning".

L12: "Etna volcano" -> "Mount Etna volcano (Italy)".

L14: "BTD's" -> "BTDs"

L30-31: I understand the concept but for readers not familiar with the research area I suggest being more explicit by revising to: "Specifically, ash absorbs more energy around the 11 μm band than the 12 μm band, whereas water and ice absorb more energy around the 12 μm band than the 11 μm band. "

L57: Again does "Machine Learning" need to be capitalised? Why not just "machine learning"?

L83: "SEVIRI instrument" -> "The SEVIRI instrument"

L86: "All the SEVIRI data used in this work was…" -> All the SEVIRI data used in this work were…"

L91: When describing the natural color composite can you provide which SEVIRI channels (wavelengths) are assigned to R, G, B? Same goes for the Ash RGB. I know it can be found in the EUMETSAT guide but it would be useful to have them at hand when reading the paper and it's easy to insert this information.

L93-95: Ash RGB - can you state how it's constructed (see comment above)? I know it's not simply single channel BTs assigned to R, G, B.

L120-123: Pre-processing of BTs and BTDs before input into the model - I see later on you state: "Additionally, the data were standardized by removing the mean and scaling to unit variance". I think it would be useful to have this information at this point in the paper. Also, how is this actually done in practice? How is the mean computed? Is it the mean of all pixels considered (VC+nonVC)? Or do you need to group by category and then take the mean?

L140: Parameter optimisation: Is this what is typically referred to in the literature as "hyper-parameter tuning"? If so, I wonder if similar language should be used. If not, how is this method different from hyper-parameter tuning?

L144: Training/test split - Can the authors be more explicit here e.g. "Moreover, the balanced dataset (50% VC, 50% NVC, totalling 2.2 million pixels) was split…". Also was the train/test split randomised? I see in the validation section you mention that the model hadn't seen some eruption cases. It would be useful to state this at this stage of the paper. It would be interesting to know if the training pixels were sampled evenly across the different Etna eruptions. I've seen 80/20 used many times but wonder if there is any previous literature that indicates the 80/20 rule is best practice for satellite pixel classification?

L145: "both accuracy and the average time required for the NN" - It would be useful to have some numbers to support this statement. For example, were there any settings combinations that had a high accuracy but weren't used because the run took too long? And how long is "too long"?

L155: Formatting issue with this line.

L156: Activation function and optimisation algorithm - are these standard choices? Again, is there any literature you can refer to justify their use over other potential options?

L158: Can you expand upon what the "early stopping strategy" is?

L159-160: "the NN model was calibrated so that the predicted class probabilities correspond to the distribution of the observed pixels in the dataset." - Can you expand on this. What is done mathematically to "calibrate" the NN model? Are you applying transformations to the input BT

and BTD data? More detail is needed here to understand the method and how Fig. 3 is generated.
L181-182: The authors state that all cases here used to validate the model were challenging because they were composed of ash, ice and SO2. I'm wondering if the case studies included in the training data were also similarly complex (i.e. ice, SO2 dominated)? Or were some of the case studies included in the training ash-rich eruption clouds?
L190: It might be more informative to include a real image (alongside the schematic referred to in Fig. 2.6), to see the impact of the non-local mean filter i.e. before/after the NL mean filter is applied. The plume tracking images (Figure 4) are great.
L210: "values are based on typical wind speeds of 10-16 m/s" - At what height? I assume there should be a height dependence for this calculation as well.
L239: Evaluation metrics - I know these are standard metrics but it would be useful to have these mathematically defined in the paper, with attention to exactly how they were implemented for this study and the datasets used. E.g. were they before or after the non-local mean spatial filter? Were they only calculated for pixels falling inside the plume tracking region?
L262: "volcanic cloud evolution is quite similar" - Do you mean the volcanic cloud composition is quite similar? The transport and dispersion is markedly different as you point out in the next sentence.
L270: I suggest being consistent in reporting the metrics e.g. here I would include the precision alongside the F1-score and Recall metrics to be consistent with the previous Event description. Accuracy doesn't need to be reported here as it's pretty much the same for all events as you state earlier.
L297: "4.4 Metrics analyse" -> "Metrics analysis"
L333-334: False positive rates - How is the false positive rate defined and calculated here?
L340: "consistent with expectations" - Describe what these expectations are.
L350: "where traditional methods failed to detect volcanic clouds" - I would be a bit more specific here. Perhaps say "where traditional methods (i.e. a negative 11-12 BTD) cannot be used to detect volcanic clouds". I think the authors need to be clear and specific here as some of the papers cited e.g. Prata et al. (2020) deliberately used a positive BTD signature to detect volcanic clouds.
L354: VPR retrievals - I understand how this is done using the VPR method (e.g. Guerrieri et al., 2023). However, for the neural network, how are these retrievals applied to ash, ice and SO2 individually if the neural network algorithm simply classifies pixels as VC? I'm not sure I follow how this analysis was conducted.
L378: "Figure 6 presents … SHAP values" - I think you mean Figure 10? It would also benefit the reader if the authors could add a brief explanation on how to interpret the SHAP figure. E.g. What do negative vs. positive SHAP values mean? How are "Feature Values" interpreted? e.g. is feature value normalised from 0-1? Some initial text on this would then make the physical interpretation of the channels easier to follow (e.g. L393-396).
L390-392: Commentary around the BTD[10.8-12.0] - Again, I would rephrase this to say something like: the result is expected given that a negative BTD[10.8-12.0] is most sensitive to semi-transparent ash-rich volcanic clouds and the volcanic clouds considered in the present study were either opaque and/or ice/SO2-rich. I would also add the BTD[10.8-8.7] combination is used for semi-transparent ice clouds e.g. see Wang et al. (2011)

(https://doi.org/10.1175/JAMC-D-11-067.1) and it appears as an important feature here because of the presence of ice in some of these volcanic clouds.
L395: Presence of SO2 in the 7.3 μm band - I agree this is probably the main reason but I would add that cold opaque clouds in general will be picked up by the 7.3 μm band. So regardless of the SO2 signal, if the plume is opaque and cold the 7.3 μm would be a good discriminator from the clear land/sea surface (you actually can see this in your Figure 11).
L410-412: “BTD values around 0.0. This is expected, given that the presence of ice in volcanic clouds can compromise the performance of the BTD method.” - While I agree that ice in volcanic clouds can impact the ‘ash signal’, a BTD of 0 K between the 11-12 μm channels indicates an opaque cloud, meaning that the compositional information cannot be inferred. In other words, a BTD=0 K could be a dense ash cloud or ice cloud or liquid water cloud - we just don’t know until it becomes semi-transparent.
L418: If it is straightforward (within scope of the study), it would be interesting to understand (quantify) how the plume tracking improved the NN model’s performance.
L428-429: “middle and final parts of the image sequences.” -> I suggest adding further context to this statement e.g. “middle and final parts of the image sequences for the three Mt Etna eruption case studies on 22 Feb 2021, 4 March 2021 and 4 August 2024 that were analysed here” or similar.

Figures
Figure 1: I wonder if it would help adding coastlines on these figures to help the reader see where the land masses are in the Ash RGB image?
Figure 2: Should the panels be labelled (a), (b), (c) etc? Instead of 1, 2, 3.. That way in the text you would refer to “Figure 2(a)” which is more typical of the AMT journal style.
Figure 3: The figure caption is incomplete. Can you describe both Fig. 3(a) and (b)?
Figure 5: How is the purple enclosed circle calculated and how is it related to the plume tracking radius model described earlier?
Figures 5, 6, 7: I suggest moving the “(a)”, “(b)” panels to the top right outside of the image axis bounds to improve readability. I didn’t notice where these labels were at first.

Supplementary material
The animations are really useful to have alongside the paper. Thank you for providing them.

References
Check the DOI links. I noticed for Guerrieri et al. (2023) for example, this link didn’t work: https://doi.org/10.3390/RS15082055/S1