Review of *Automating tephra fall building damage assessment using deep learning* by Tennant et al.

Please find below my review of the manuscript *Automating tephra fall building damage assessment using deep learning* by Tennant et al. submitted for publication in NHESS. The manuscript presents a computer vision/machine learning pipeline to automatically assess the physical damage to buildings in volcanic contexts. Most of the paper presents the method, followed by a short application to the 2021 eruption La Soufrière volcano, St Vincent and the Grenadines and a discussion that includes recommendation to conduct such future surveys.

The manuscript is of great interest for the topic of post-event impact assessment in volcanic contexts and lays the path for future use of UAV-derived impact surveys. The methodology is sound and comprehensively described (although it could benefit from more direct writing and clarity - see comments below). I therefore recommend publication after minor corrections.

Best regards,

Sébastien Biass

General comments

My main impression from reading this manuscript is the focus on the method. The method is long, and although I understand the complexity of finding a balance between conciseness and thoroughness in describing the development of such a pipeline, some parts are hard to follow, and some aspects remain somehow obscure after multiple reads. Everything is documented below, but two aspects that remain puzzling at this point are:

1. The introduction to the two classifier tasks - which are mentioned early in the method section with reference to a more detailed description that, unless I am mistaken, never really comes
2. The generation and use of the orthomosaic, which raises several question (i.e., georeferencing of some of the datasets, whether the training and further predictions are performed on the orthomosaic or individual images)

The second aspect is the balance between methodology and application. Although the case-study serves as a basis for the development of a method, its application in section 4 is less than 20 lines, which felt anti-climatic! Perhaps is there a political context that prevents further analyses as it is often the case following recent eruptions, and I don't expect the authors to remodel the manuscript around a more detailed analysis. However, even without a direct application to la Soufrière volcano, some aspects could be discussed in the context of physical impacts to buildings to widen the very method-oriented message to a broader audience. One component I felt was lacking is the analysis of how the defined damage states fit in a wider damage classification scheme. The only mention to this aspect is found in Section 5.5. However, discussion points

raised only seem to focus on the number of classes on the scale, but miss a more systematic comparison with other schemes as well as a critical interpretation of the damages captured by the present methodology. As a result, the reader is left with a long list of computed parameters to assess the quality of the impact assessment, but with no concrete link to reality. Would it be possible to add:

- Images of a limited number of buildings illustrating each damage?
- A description of what these damages capture? → i.e. is the difference between moderate and heavy a structural component? Roof collapse?

In any case, I suggest adding - where possible - a couple of bridges that help interpreting the model results beyond the simple application of the method, and broadening findings to the actual literature on impact assessments on buildings (note: the discussion could also be more supported by references!).

Specific comments

- **Line 84-89**: I suggest rephrasing this sentence as it is both long and in which parts in brackets could be better integrated. Maybe something along the lines of:

> *To our knowledge, only one study attempts automating the assessment of building damage for volcanic hazards (Wang et al., 2024). In contrast, attention has been given to more commonly occurring hazards such as earthquakes and hurricanes, with the development of both mono-temporal (post-event imagery only) and multi-temporal (uses pre- and post-event imagery) approaches (Table 1).*

- In addition, some "multi-temporal" studies might also differ in the use of either a "before-after" approach (i.e., the comparison of two images) vs time-series approach. Maybe worth specifying if applicable to your problem.
- **Line 101-102**: I like the example! I might use a closer analogy to the problem tackled here, but I leave that to you.
- **Line 135**: Not wanting to open a can of worms - and totally aware of the use of post-disaster "opportunities", I find the use of the term "opportunity" perhaps a bit misplaced (i.e., using impacts on people's homes as the basis for research). I know this is not the case, but perhaps a more neutral phrasing would be more appropriate? (something along the lines of "prospect" - though I leave the selection of the most appropriate word to the native English-speaker authors).
- **Line 205**: Was the dataset manually geo-referenced? If yes - how?
- **Line 213**: This dataset does not contain any information regarding the flight path, which also raises the question regarding why buildings are captured with a lower resolution.
- **Line 240**: The reason behind this is not 100% clear
- **Line 242**: For off-nadir or very-off nadir images, how did you ensure that single bounding boxes did not overlap over buildings in the background?
- **Line 254**: That is not clearly intuitive given your description of the datasets. Did you filter out off-nadir images?
- **Line 282/Section 2.3**: This section is not the easiest to follow. For instance, from the first paragraph you mention splitting the classification task in two, a theme that you refer to in almost every paragraph, but without stating how or why. Can't this introduced and adequately presented in Section

2.3.3? Also, this section keeps on referring to sections ahead. Isn't it possible to optimise the writing to better integrate the development of the pipeline with its model components?

- **Line 292**: Datasets?
- **Line 294**: In general, I personally recommend a more direct writing style, for instance changing: > we split the building damage assessment task into two subtasks, training and evaluating models for building localisation, which consists of identifying building bounding boxes within the images and building damage classification separately
    - to:

> *we split the building damage assessment task into two subtasks that include i) building localisation (i.e. identification of building bounding boxes within the images) and ii) building damage classification.*

- **Line 299**: This is true of most ML algorithms, not only deep learning
- **Line 301**: "experiment with different hyperparameter settings" or "optimise hyperparameters"?
- **Line 304**: I don't think you need to state "(localisation, classification 1, classification 2)" at all in this section, especially if just added in brackets. That makes reading heavy. (Same for line 309).
- **Line 309**: Rephrase:

> *Once we identified the best performing experimental setup for each task (building localisation, classification 1, classification 2), we combined the training and validation datasets and conducted K-fold cross-validation using the experimental setup and optimal hyperparameters that were identified (Cross validation: Section 3.1.3, Section 3.2.2)*

    - To:

> *After hyperparameter tuning, model accuracy was assessed using K-fold cross-validation (Cross validation: Section 3.1.3, Section 3.2.2)*

- **Line 316**: In general, "data" is used in a very loose way. Would it work to change:

> *have data from more than one dataset*

    - to:

> *contains images from more than one dataset*

- **Line 328**: This statement is a bit out of place in a methodology section (plus it is somehow true for all contexts!)
- **Line 331**: Two comments here:
    1. Following the comment on line 205, there seems to be georeferencing. This should be explained
    2. Following the comment on line 242, the definition of the bounding boxes seems to be done on the orthomosaic? If yes, if I understand well, i) impact state is inferred from individual images ii) this is added as a label to the bounding boxes defined on the orthomosaic? This needs more clarity. I don't see any reference to the generation of the orthomosaic on Fig 3. The input to the model pipeline should probably be stated earlier.
- **Line 349**: Support opensource by citing the software used to produce the figure!

- **Line 362**: Here again, unclear if "image feature" refers to individual images or the orthomosaic
- **Line 377**: Here - and anywhere else where you describe these "experiments", can you please specified how they were performed? Was it manually? In which case, is there any guidance on how you chose the ranges of each parameter? Or did you use optimisation algorithms?
- **Line 381**: This heading is inadequate (i.e., 2.3.1 is "Building localisation", why does 2.3.2. - and 2.3.3 too - need a "building"?). In addition, why not keeping these heading conceptual - e.g., "Building localisation" and "Building classification"? It seems to me that the sieve network is part of the building localisation task.
- **Line 383**: Define "small" or remove
- **Line 386**: Rephrase:

> *The dataset used for training and evaluating the sieve network consists of randomly cropped background samples from full sized images in the training and validation sets*

- **Line 397**: It seems that up to this point, the purposes of classifiers 1 and 2 have not been defined (unless we count Figure 3 as doing so). I might be mistaken, but I think this highlights the need of rethinking a bit the structure of Section 2.3.
- **Line 419**: "False positive" has been used in line 385 but not defined
- **Line 444**:

> *The five experiments with the highest average precision*

- **Line 455/Table 3**:

> *Hyperparameters for the 5 experiments with highest average precision conducted for building...*

  - By this point, the use of blocks vs boxes etc gets confusing for the reader. Maybe a conceptual sketch could help? Specifically, I don't think "block resizing" has been described in the text. I understand that a lot of the method is described in the SM, but the main m/s should be self-sufficient, therefore any concept shown in table/figures should be described in the text.
  - I suggest renaming the column "All training/ UWITV& GOV/ SRC" to "training dataset", assigning a letter to each dataset and defining it on a table footnote
- **Line 464**: I think you are citing Table 6 before 5
- **Line 519**: Same as 444:

> *The five experiments with the highest macro F1 score*

- **Line 537**: What do you mean by:

> *to understand the potential for our model to generalize to a new dataset*

- **Section 3**: This section is very technical. Since the manuscript is rather oriented towards an impact/operational rather than a computer science audience, would it be possible to attempt better extracting what the raw values of model validation imply for a further application of the model? Discard this comment if you don't believe this is applicable.
- **Line 598**: I would rephrase to:

> *In order to optimise the application of separate models for building localisation and two stages of damage classification for operational contexts, we have integrated a damage assessment pipeline.*

- **Line 601**: Again, here the use of orthomosaics is unclear. Also:
    - Do you need to state these softwares? Isn't "computer vision" or "structure-from-motion" sufficient?
    - Do you need to state "shapefile" - which is a proprietary file format? What about "georeferenced vector dataset"?
- **Line 637-638**: I don't understand this statement. What do you mean by distributions? Datasets? Building typology? If datasets, does it mean that your model is not generalisable?
    - Note: I see that some precisions are provided later. I still believe this should be clear from the beginning.