Review Report

Manuscript: Rapid Landslide Mapping During the 2023 Emilia-Romagna Disaster: Assessing Automated Approaches with Limited Training Data

The study employed deep learning techniques based on convolutional neural network architectures (U-Net and SegFormer) to segment landslides across regions with distinct geological characteristics. The primary objective was to assess the effectiveness and limitations of automated landslide mapping in practical scenarios and to evaluate whether deep-learning approaches can reliably replace manual mapping. The results highlight several important aspects, including a comparison of models trained using different architectures and input layers. The discussion further examines the impact of incorporating a lithological layer into the model and presents an analysis of buildings potentially affected by landslide hazards. Despite the interesting approach to landslide segmentation, the manuscript presents structural and organizational weaknesses that at times hinder readability. A revision is therefore recommended to improve the clarity of the methodological framework and the presentation of results.

Major Review

1. Introduction:

The paragraph in lines 56–64, which reviews studies employing Convolutional Neural Networks (CNNs) and deep learning techniques for landslide mapping, should be strengthened by incorporating a broader range of relevant literature, particularly more recent studies. Expanding this section would provide a more comprehensive and up-to-date overview of current advances in deep learning–based landslide mapping and better contextualize the contribution of the present work.

The paragraph in lines 65–73 correctly notes that previous studies typically adopt large train-to-validation ratios (80:20 or 70:30), but it does not sufficiently clarify why this constitutes a limitation in real-time crisis scenarios. The authors should explicitly explain that, during an emergency, only a very limited portion of the affected area is usually mapped and annotated in the early hours or days, making large training sets unrealistic. Additional constraints, such as delays in acquiring cloud-free post-event imagery, the absence of recent pre-event data, and the operational need for rapid deployment, further restrict the amount and quality of training data available. The paragraph also states that training procedures often fail to represent the diversity of landslide types and geological conditions encountered in practice; however, this point would benefit from brief elaboration.

3. Study Area

The description of the study areas (Casola Valsenio, Brisighella, Modigliana, and Predappio) should be moved before the initial paragraphs of this section. These initial paragraphs discuss geological conditions and terrain variability that are only detailed later in the text, resulting in a forward reference that disrupts the narrative flow.

The section lists the number of landslides, their types, and relative percentages for each study area. A summary table compiling this information for all municipalities would greatly facilitate comparison and comprehension of the landslides among sites.

4. Methods

The paragraphs from lines 298–354 provide a comprehensive overview of the different data-availability scenarios (L1–L7) considered in the study; however, they are rather long and difficult to follow in their current form. The description alternates between data-availability assumptions, sensor characteristics, and computational constraints, which difficulties the understanding of the logical structure of the scenarios summarized in Table 2. Explicitly linking each paragraph to the corresponding cases in Table 2 would significantly improve readability.

Lines 301–303 state that, in Table 2, the availability of a slope map (L7) is assumed based on the free accessibility of global DEMs (e.g., SRTM, ASTER, ALOS, Copernicus DEM). However, earlier in lines 295–297, the slope map is described as being derived from the 1:5000 Regional Technical Map. This creates ambiguity regarding the actual source(s) of the L7 slope product. If global DEMs were also used, alone or in combination with the regional dataset, they should be explicitly described as part of the L7 product. Alternatively, if L7 is derived exclusively from the regional map, this should be clearly stated and the reference to global DEMs clarified accordingly.

4.x Deep Learning Semantic Segmentation Models

The section first introduces the models and only later describes the training areas and data preparation. It would be clearer to reorganize the content to follow a more logical workflow: data preparation, model training, and data evaluation (metrics). In addition, some paragraphs are difficult to follow and would benefit from improved clarity and structure. Including additional figures illustrating the model architectures would also help facilitate understanding.

4.x Model Application

The first two paragraphs discuss aspects related to the test area and would be more appropriately placed in the training–testing split section. Additionally, including a table showing the percentage of each test area within each lithological unit would improve clarity and facilitate understanding of how the models were evaluated.

5 Results

The second paragraph (563-573) should explicitly show the results and instead of using words like “highest” , “achieve results close to those of the top-performing models”..

In general the paragraphs do not explicitly state the values encountered by the models and some paragraphs should be rewrite to be more precise. I recommend reviewing this section completely.

6 Discussion

The section discusses some of the results; however, it does not adequately interpret them in terms of the spectral response of the mapped targets. In addition, certain results, such as the inclusion of a lithological layer and the assessment of buildings potentially at risk, are introduced without having been previously described in the Methods section, which makes the paragraph difficult to follow. To improve clarity and coherence, these methodological aspects should be clearly presented earlier, and portions of the discussion should be relocated to the Results section to enhance overall readability.

Minor Review

Abstract: Well written but slightly long. Can be more objective and tightened by removing methodological repetition.

3. Study Area

136-137: Readability would be improved by explicitly citing the name of the municipalities;
144: Text states that “about 25% of the recorded 80,997 landslides events”, adding the quantity of landslides would also improve the comprehension of the impacted area;
Figure 2 - Increase the legend size;
193: Correct from “sandstone-rice” to “sandstone-rich”;

3. Methods
Review the numbering of the sections
Figure 4: Sentinel images are with low contrast which difficult the visualization;
304-309 - Should also describe the product level that was used in the study.
320-354 - Text refers to cases (e.g case 1, case 2 ..) however this definition is not represented or cited in table 2.

3.7 Evaluation Metrics and Expert Judgment

Lines 498-501 where the F1 metric is described, the text says that the metric is useful for imbalanced problems, but this was not clearly discussed in the methodology, was the data used in the experiment imbalanced?

5 Results
First paragraph (556-562) could be adapted and turned into the legend of table 4, as it does not describe any result, this would improve readability.

Figure 7 - Legend states “case 1 - 7” but in the table it is referred as “U1”, “S1”, “U2”, “S2”, also inputting these names would facilitate the comprehension of the table. The size of the legends should be increased.
Figure 8 and 9 - Changing the colors used to highlight the TP and FP would improve the visualization.
Paragraphs from 628-655 highlight difficulties of the model in the segmentation over different areas, but figure 9 does not highlight those areas making it harder to follow the paragraph.

6 Discussion

Paragraph from 715-726 states that another model was trained to evaluate the impact of including lithology but no explanation was added to describe how this was done. The categorical layer was converted to dummy layers (one for each category) or was added as a categorical layer? Extra information would improve the comprehension

The discussion in lines 734–741 regarding the F1-score relies on comparisons with generic machine-learning benchmarks rather than on metrics reported in landslide-mapping studies. The analysis should instead be grounded in the context of landslide detection and segmentation literature. In particular, the manuscript should address why F1-scores in landslide mapping are often relatively low, discussing contributing factors such as class imbalance between landslide and non-landslide areas, the spatial heterogeneity of landslides, uncertainties in reference inventories, and the influence of image resolution and labeling quality.

7 Conclusion

This section should also address future research directions aimed at overcoming the limitations identified in the study.