Review: A comparative analysis of deep learning models for classifying shallow mesoscale cloud patterns in satellite images

By: Anna Granberg, Vilma Lundholm, Pouria Khalaj, Manu Anna Thomas, Yifan Ding, Daniel Jönsson, and Abhay Devasthale

Anonymous Reviewer #2

This manuscript compares CNN, EuroSat transfer-learning ResNet50, supervised DINOv2+MLP, and DINOv2 feature-based clustering approaches for classifying shallow mesoscale cloud patterns in SEVIRI imagery. The research question is clear, the dataset and benchmark design are useful, and the results consistently suggest that DINOv2 features are more robust under domain shift. The manuscript is generally well structured and the figures and tables are comprehensive. However, several aspects require clarification, especially data splitting, annotation reliability, model comparability, and reproducibility. I recommend minor revision.

We thank the reviewer for their thoughtful and detailed assessment of our manuscript. We value your suggestions that will help to bring more clarity to the manuscript and we have tried to address your concerns. Please find below the point-by-point response to your comments.

Major Minor-Revision Comments

Clarify whether all models were evaluated on exactly the same test set.

Please see the response below.

Sections 4.2 and 6.1 state that all models use a unified 15% hold-out test set, but the class counts in Tables 5/6 differ from those in Table 7. For example, Closed Cell has a count of 71 for CNN/ResNet50 but 98 for DINOv2+MLP. This affects direct comparison of Top-1/Top-2 and per-class metrics. The authors should explain this discrepancy or re-report all supervised model results on an identical test set.

Thank you for this point. We agree that for an ideal model-to-model comparison, all models should have been evaluated on the same test images. We further acknowledge that our preference for an improved training efficiency in the representation learning stage has unavoidably led to slightly different per-class counts which could be viewed as a methodological limitation. However, our main intention of comparison was not to claim a definitive ranking but rather to explore and demonstrate that the DINOv2+MLP pipeline can achieve improved performance compared to supervised CNN/ResNet50 baselines. To clarify this point, we have added the following to the caption of Table 7:
“Note that, for all models, the sample size of the test dataset is 534 (15% of 3554 according to Table 1). The test dataset of CNN and ResNet50 are exactly the same (tables 5 and 6). However, they differ from that of DINOv2+MLP, hence the difference in per-class counts.”

Provide more detail on the manual annotation process and label consistency.

We have now added more information to sections 2.1, 2.2 and 4.1 (Section 3.4.1 in the revised manuscript) to further clarify this.

The dataset was annotated by a meteorologist, and the Discussion notes that some annotations are incomplete or inconsistent. The authors should describe the annotation criteria, how ambiguous cases were handled, whether any quality control or second review was performed, and whether inter-annotator agreement was assessed. If only one annotator was used, this should be explicitly acknowledged as a limitation.

This same concern was raised by the Anonymous Referee # 1.
The labeled dataset was created by a single expert annotator with a background in meteorology. A preliminary attempt involving non-expert annotators resulted in misclassification, primarily due to the subtle and often ambiguous distinctions between cloud pattern types. Previous studies (e.g., Stevens et al., 2020) used annotators with expertise. Using a single expert annotator can ensure higher degree of consistency although we acknowledge, it may introduce subjective bias. The annotator focused on selecting clear, representative examples for each cloud pattern from the SEVIRI images, There are cases where some mesoscale structures within an image may not have been annotated, for example when very small, ambiguous, or partially obscured features were deliberately left unlabelled to avoid over-interpretation.

Due to the use of a single annotator, the variability in the labeling when one uses multiple annotators can not be evaluated and no independent second review was carried out. To compensate for the limited size of the labelled data set and to introduce more diversity, data augmentation techniques such as geometric transformations (e.g., rotations, flips) and intensity variations designed to preserve the physical characteristics of cloud structures were applied. This approach helps the model generalize better and reduces overfitting, thereby mitigating some of the limitations associated with a single-annotator dataset.

We agree that the use of multiple annotators would strengthen the dataset and we consider this an important step for future work.

In Section 7 line 450:
The dataset used in this study is relatively small, comprising of 3,554 cropped images across eight classes.

In Section 7, lines 459-466 (in the revised manuscript), the following paragraph is added:

It is important to point out that the labelled dataset was produced by a single expert annotator with a background in meteorology, who focused on selecting clear and representative examples of each mesoscale cloud pattern from the SEVIRI imagery. Consequently, some mesoscale structures within an image were intentionally left unannotated, particularly when features are very small, ambiguous, or partially obscured, in order to avoid over-interpretation and maintain

annotation quality. To address the limited size of the labelled dataset and increase sample diversity, data augmentation techniques are applied, including geometric transformations (e.g., rotations and flips) and intensity variations that preserve the physical characteristics of the cloud structures. These augmentations improve the model's ability to generalize and reduce overfitting, thereby partially mitigating the limitations associated with the relatively small, single-annotator dataset.

Describe the benchmark domain shift more concretely.

Regarding the domain shift, it arises from the fact that the original dataset and the benchmark dataset were created separately using different tools, and approximately six months apart by the human annotator. In particular, they differ in the size of the bounding boxes as also evident in table 1. We have clarified this in the revised manuscript. Regarding the analysis of the domain shift, Section 6.4 in the revised manuscript elaborates upon this.
The following lines are added to the beginning of Section 6.4 in the revised manuscript: “The domain shift arises from the fact that the original dataset and the benchmark dataset were created separately using different tools, and approximately six months apart by the same human annotator. In particular, they differ in the size of the bounding boxes as also evident in Table 1.”

The manuscript states that benchmark images are larger and differ in annotation conditions, causing domain shift. Please provide more information on the benchmark sampling strategy, date/month distribution, bounding-box size distribution, class distribution, and how these differ from the original dataset. If possible, a simple stratified analysis by bounding-box size or month would strengthen the claim.

We have elaborated upon this in Section 2.2, Table 1, and also revised the beginning of Section 6.4 in the revised manuscript.

Add uncertainty estimates or significance testing for model comparisons.

We appreciate the reviewer’s comment. We have now added Wilson 95% confidence intervals to tables 5, 6, and 7 for Top-1 accuracies.

On the benchmark set, DINOv2+MLP achieves a Top-1 accuracy of 0.61, compared with 0.56 for both CNN and ResNet50. The direction is clear, but the margin is modest. Bootstrap confidence intervals, McNemar tests, or at least macro/weighted averages would make the comparison more convincing.

We appreciate the reviewer’s comment. We have now added macro-F scores and balanced accuracies to Table 13 to add further to the comparison metrics for benchmark data.

Improve reproducibility details for DINOv2 and clustering.

Please see the response below.

Please specify the exact DINOv2 version/model size, embedding dimension, input normalization, k-means initialization settings, number of runs, random seeds, and whether PCA or feature standardization was used.

Thank you for the suggestion. We elaborate upon the model and size of the DINOv2 in Section 3.3 as follows: "... For the teacher model, the DINOv2-ViT-S14 variant (vision transformer with 21 million parameters) is used, which processes input images into fixed-size 14×14..." Regarding other information, we have now added three more tables, namely Table A3, A4, and A5 to add more information about the models, parameters and also the reproducibility. In addition, The relevant information regarding the performance and details of the DINOv2, model is given in the reference which we already cite, i.e. Oquab et al. 2023. Moreover, we have added two new footnotes (6, 10) to include the URL of the paper on Hugging Face and also the model on GitHub.

Also, Tables 11/12 report cluster-label prediction performance, which is not equivalent to semantic cloud-class accuracy; this distinction should be made clearer.

Thank you for pointing this out. This is a valid point and we totally agree that results of our k-means analysis are not directly comparable to that of visual inspections by the human annotator. Our purpose by including the k-means clustering was to merely examine the clustering of DINOv2 embeddings as an exploratory approach, and not treating it as a fourth classifier per se. This has been further clarified in the Discussion and the caption of Table 12.

The first paragraph of the Discussion has been revised as: "In this study three supervised deep learning based models — CNN, ResNet50, and DINOv2 — are developed to classify shallow oceanic clouds using satellite image data. Furthermore, an unsupervised clustering of frozen DINOv2 embeddings was performed using the k-means algorithm. This was to examine the extent to which the model can successfully recover the features of the cloud patterns (identified visually) in the latent space. Being merely an exploratory approach, the results of our k-means analysis are not directly comparable to those of visual inspections by the human annotator. In other words, our k-means analysis is not meant to be interpreted as a fourth model to classify cloud patterns."

The caption for Table 12 is re-written as:
"Table 12. Classification performance per predicted cluster (six clusters). Note that these results merely correspond to cluster memberships according to the k-means algorithm, and therefore cannot be directly compared with the performance metrics of the three supervised models trained on the annotated dataset (identified visually) provided by the human expert."

Strengthen the code and data availability statement.

This section is rephrased as: “The ML tools developed in this study will be made public after the completion of the AI4PEX project (by April 2028). Meanwhile, the code can be provided upon request. The annotated dataset is currently being used in ongoing research for a second manuscript and, therefore, cannot be made publicly available at this time. The dataset will be considered for release upon completion of the related work. The training and hyper-parameter configurations of the different architectures are provided in Tables A1, A2, A3, A4 and A5 in the Appendix, as well as in Tables 2 and 3, and Figure 3.”

Minor Comments

The abstract would benefit from including the main benchmark result, such as the Top-1/Top-2 accuracy of DINOv2+MLP.

We have added to the end of Abstract:
“In particular, Top-1 and Top-2 accuracies of the DINOv2 model when tested on an independent benchmark dataset, was 61% and 84%, respectively.”

Tables 13/14 should also report macro-F1 or balanced accuracy, since per-class performance varies substantially.

Thank you for your suggestion, we have added macro-F1 and balanced accuracies to Table 13.

Figure 10 is dense, and some Wasserstein/Silhouette labels are difficult to read. Please enlarge the font or move some panels to supplementary material.

Thank you for this comment. We added a new table (Table B1) to include Wasserstein distances and Silhouette values in a table format which is more readable and readily accessible. We have added the following to the caption of Figure 10:
“Refer to Table B1 to see the Wasserstein and Silhouette values in a table format.”

Please standardize terminology, such as “data set” vs. “dataset,” “DINOv2+MLP” vs. “supervised DINOv2+MLP,” and “Custom CNN” vs. “CNN.”

Thank you for pointing this out. We have now used “dataset”, replaced “custom CNN” with “CNN” and “supervised DINOv2+MLP” with “DINOv2+MLP”, everywhere.

Several minor language issues should be corrected, for example “Each metric capture” should be “Each metric captures,” and “class class” is repeated in the caption of Table 14.

This is corrected in the revised manuscript.

The Discussion would benefit from a clearer distinction between transfer learning from general remote-sensing imagery and transfer learning from cloud/atmospheric imagery.

Among the architectures evaluated in this study, only ResNet-50 employed a transfer learning strategy, with weights pretrained on the EuroSAT dataset. EuroSAT consists of Sentinel-2 satellite images and is primarily designed for land-use and land-cover classification. Consequently, the pretrained model had learned representations relevant to land surface characteristics rather than cloud structures or atmospheric processes. While these pretrained features can improve convergence and provide robust low-level representations (e.g., edges, textures, and spatial patterns), they are not specifically optimized for cloud morphology.

Here, in this study, the main aim has been to explore different ML tools to find out the one that could best suit our purpose.

Please define how Top-2 accuracy is computed, especially in cases where class probabilities are close or tied.

We have now added the following to the end of the first paragraph of Section 7.2:
“Top-1 and Top-2 accuracies have been calculated using the topk function from PyTorch, by setting k=1 and k=2, respectively.”

The conclusion that cloud patterns “can be effectively classified” should be softened, since the best benchmark Top-1 accuracy is only 0.61. A more cautious phrasing would be “moderately classified under domain shift, with DINOv2 showing the strongest robustness.”

Thank you for pointing this out. We have revised this in the new version of the manuscript. In particular, we have rephrased the first sentence of the second paragraph of Conclusions, according to your suggestion.