AI4EO - Artificial Intelligence for Earth Observation

Final Report

Baseline study on supervised snow classification with Sentinel-2 L2A imagery

Ahmed Habib & Lucas Velciov
https://github.com/bigwahbs/ai4eo-snow-segmentation
6. October 2025

Introduction

Seasonal snow in the Alpine region plays a pivotal role in hydrology, water supply, ecology, and natural hazard management. Snowmelt from the Alps feeds river basins across Central Europe, supports hydroelectric production, and regulates downstream flood timing. But when snow cover changes unpredictably due to climate shifts or local anomalies, operations such as avalanche forecasting, infrastructure maintenance (roads, ski lifts), and water resource planning face increased uncertainty. Satellite remote sensing offers a unique view into snow dynamics over large, inaccessible mountain terrain where in situ measurements are sparse. Yet optical algorithms like NDSI struggle in forested areas, mixed pixels, and shadows while microwave or SAR methods face other trade-offs. Recent reviews call the spatial and temporal variability of snow cover in mountainous regions “a clear knowledge gap” in remote sensing science. Frontiers In the Alps, even small misestimates of snow extent or melt timing can cascade into operational risks like avalanche release zones, unplanned water release, or misaligned maintenance schedules. Remote-based ice/snow monitoring therefore isn’t just academic, but it actually underpins real decision systems in alpine environments.

The Normalized Difference Snow Index (NDSI) is the classical approach for optical snow detection and has been widely used for over two decades. It exploits the high reflectance of snow in the visible green band (B03) and its strong absorption in the shortwave infrared (SWIR, B11), furthermore it`s defined as:

$$NDSI = \frac{B03 - B11}{B03 + B11}$$

Pixels with NDSI values above a threshold (typically 0.4–0.5) are classified as snow. NDSI performs well over bright, homogeneous snow but deteriorates under forest canopy, terrain shadows, and mixed pixels containing vegetation or soil. In such cases, snow’s spectral signature becomes ambiguous, and thresholding alone fails to capture partial cover or cloud–snow confusion. Moreover, fixed thresholds are not universally transferable and therefore illumination, sensor geometry, and atmospheric effects vary regionally and seasonally. Studies over alpine regions show that while NDSI achieves overall accuracies above 90 % in open terrain, errors can exceed 25 % in complex topography (Hu & Sean, 2022). These limitations motivate data driven approaches that learn from multiple spectral bands and contextual information rather than relying on a single fixed ratio.

Recent studies have demonstrated the potential of machine learning for snow cover mapping using multispectral satellite data. Random Forest remains a strong baseline due to its simplicity and robustness, while deep networks such as U-Net generally achieve higher accuracies when sufficient labelled data are available. Wang et al. (2022) showed that U-Net outperformed Random Forest on Sentinel-2 imagery, but that well-chosen band subsets, especially combinations of visible, near-infrared, and shortwave-infrared channels, narrowed the gap significantly. Similarly, Hu and Shean (2022) reported improved performance in alpine regions when including SWIR and NIR bands instead of relying solely on RGB.

These findings emphasize that spectral diversity and spatial evaluation strategies are more critical than model complexity alone. Geographic cross validation, now common in recent works, provides a fairer measure of generalization than random sampling. Moreover, several studies use Sentinel-2 Scene Classification Layer (SCL) maps as pseudo-labels when ground truth is unavailable. Following this evidence, the present project adopts a Random Forest baseline with selected multispectral bands, SCL-derived snow labels, and geographic cross-validation to establish a reference for future deep-learning extensions.

Dataset

The dataset used in this study is a Sentinel-2 Level-2A product acquired over the Alpine region (tile T32TPS, January 2024). The L2A product provides atmospherically corrected surface reflectance and an accompanying Scene Classification Layer (SCL). All bands were resampled to a common 20 m grid to ensure spatial alignment. Six spectral bands were selected based on prior work and physical relevance: B02 (blue), B03 (green), B04 (red), B8A (narrow NIR), B11, and B12 (short-wave infrared). These channels capture the strong spectral contrast between snow, vegetation, and soil, while maintaining manageable data size.

The SCL layer was used as pseudo-ground truth for supervised learning. Its “snow/ice” class, originally produced by the Sen2Cor processor at 60 m resolution, was up-sampled to match the 20 m grid. Although SCL provides consistent, globally available labels, it is known to misclassify bright clouds, thin cirrus, or shaded snow. Such inaccuracies impose an upper bound on achievable model accuracy, as the model can only reproduce the quality of its reference data. Nevertheless, this approach allows reproducible training without external labelling campaigns and is widely used in similar remote-sensing studies such as the ones mentioned in this report`s References section.

To evaluate generalization, the dataset was split geographically rather than randomly. The image tile was divided into left and right halves, alternately used for training and testing. This geographic cross validation simulates out-of-distribution conditions, where topography, illumination, and land cover differ spatially. Such evaluation is increasingly recognized as a more realistic indicator of performance in heterogeneous mountain environments than traditional random splits.

Methods

The study first established an analytical baseline using the NDSI previously introduced. Thresholding was applied at 0.4 to derive a binary snow mask, which served as a simple, reproducible reference. This method provided a fast way to visualize snow distribution but was known to struggle under challenging conditions such as forested areas, shadowed slopes, and cloud- or snow confusion. These limitations motivated the use of a learning-based approach capable of integrating additional spectral information and spatial variability.

Building on this, a Random Forest classifier was trained using the selected Sentinel-2 bands together with the NDSI as input features and the SCL snow/ice class as target labels. The model aggregated 200 decision trees to capture nonlinear relationships between spectral responses. Hyperparameters were tuned empirically for stable results while keeping runtime low. Training and testing were performed under geographic cross validation to assess the model's ability to generalize across distinct terrain regions. Model performance was evaluated using precision, recall, F1-score, and Intersection-over-Union (IoU) metrics for the snow class. This approach provided a more flexible and spatially aware baseline compared to the threshold-based NDSI.

Results

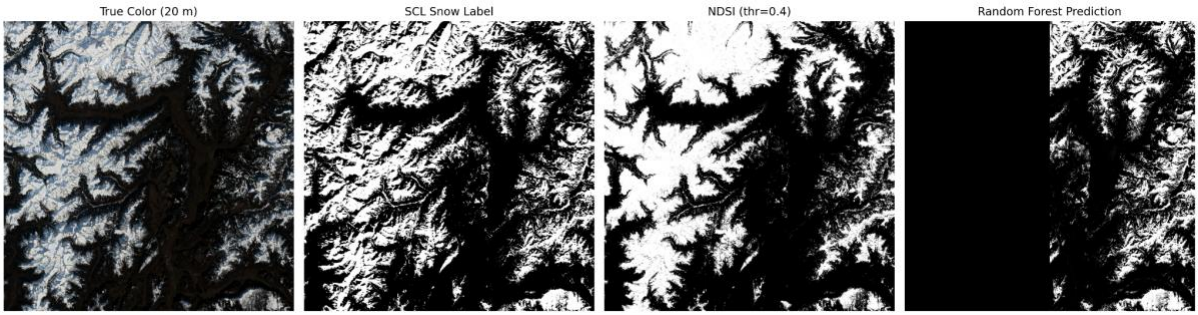

Figure 1 illustrates the visual comparison between the true-colour Sentinel-2 image, the SCL snow label, the classical NDSI mask, and the Random Forest prediction. The reference scene covers the central Alpine region near Innsbruck on 28 January 2024. The NDSI map successfully captures most high-altitude snow cover but tends to overestimate snow in shaded valleys and partially cloud-covered zones. The SCL label shows a generally consistent snow pattern but includes some cloud contamination, especially over bright slopes. The Random Forest output provides a visually smoother and more continuous snow mask, reducing false detections in forested and shadowed areas while maintaining good coverage over open terrain. It shows valid results only for one half of the scene because the other black half corresponds to the withheld training area, which was intentionally masked to prevent information leakage.

Quantitative evaluation is summarized in Table 1. Under geographic cross validation, the model achieved an overall accuracy of 0.936 in both splits. When trained on the western (left) half and tested on the eastern (right) half, the snow class reached a precision of 0.898, recall 0.863, F1-score 0.880, and Intersection-over-Union (IoU) 0.786. Reversing the split improved the snow IoU to 0.869, indicating that the model generalized well across different

mountain sectors. The small performance difference between the two directions suggests consistent predictive behaviour and limited overfitting to local conditions.

Split	Precision	Recall	F1	IoU	Overall Acc
Left->Right	0.898	0.863	0.880	0.786	93.6 %
Right->Left	0.943	0.917	0.930	0.869	93.6 %

The main failure cases occur in deep shadows, forested north-facing slopes, and along cloud edges where the spectral response of snow overlaps with bright soil or cirrus. These errors reflect both sensor limitations and the noise in the SCL pseudo-labels, which occasionally misclassify snow-free bright regions. Calibration using more precise reference data or incorporating topographic and texture features could reduce these effects. Overall, the results confirm that the Random Forest baseline offers a clear improvement over simple NDSI thresholding and provides a solid reference for future deep-learning experiments.

Discussion and Limitations

The main limitation of this study lies in the quality of the reference data. The SCL snow/ice class used as pseudo-ground truth introduces a bias ceiling, since any errors in the Sen2Cor classification are directly learned by the model. As a result, some of the remaining misclassifications and especially bright soil or thin cloud confusion reflect labelling noise rather than model weakness. Independent validation with externally derived snow maps, such as MODI, would allow more objective accuracy assessment and calibration.

The geographic cross validation approach demonstrated that the Random Forest model can generalize across spatially distinct alpine regions, yet the experiment was limited to a single Sentinel-2 tile and acquisition date. Broader temporal sampling would be necessary to evaluate robustness under varying illumination, seasonal conditions, and snow textures. Likewise, a deep-learning baseline such as U-Net, as proposed in recent studies (Wang et al. 2022; Hu and Shean, 2022), could better capture spatial context and subtle transitions between snow and no-snow surfaces. Multi-sensor approaches combining Sentinel-2 with MODIS or SAR data have also shown strong performance in the literature and could be explored in future work.

From an operational perspective, automated snow detection must be applied carefully, especially in alpine risk management and water operations. Inaccurate snow masks can misinform avalanche warnings, hydropower scheduling, or road maintenance planning. Therefore, models should always be complemented by human validation or independent data sources before deployment. Transparent reporting of uncertainty and performance limits is essential to ensure that such methods support, rather than mislead, operational decision-making.

Conclusion

This project explored snow cover mapping in alpine terrain using Sentinel-2 Level-2A data and compared a traditional index-based approach with a supervised machine-learning model. The Normalized Difference Snow Index (NDSI) served as a physically interpretable baseline, while a Random Forest classifier was trained on selected multispectral bands and the NDSI itself using the Sentinel-2 Scene Classification Layer (SCL) as pseudo-ground truth. The introduction of geographic cross validation enabled an unbiased assessment of spatial generalization across distinct terrain sectors.

The Random Forest model achieved consistent performance, with snow-class F1-scores between 0.88 and 0.93 and IoU values up to 0.87, indicating clear improvements over simple thresholding, particularly in shaded and forested areas. These results demonstrate that even a relatively lightweight ensemble method can leverage multispectral information to outperform classical band ratio techniques in complex alpine topography.

However, several limitations remain. The reliance on SCL as training data introduces label noise that constrains achievable accuracy. The study was limited to a single Sentinel-2 tile and acquisition date, which restricts conclusions on temporal robustness. Incorporating additional scenes, topographic variables, or snow depth observations would allow a more comprehensive analysis of seasonal and illumination effects. Similarly, deep-learning architectures such as U-Net or transformer-based models could capture finer spatial context and improve discrimination under cloud or canopy cover, as shown in recent literature.

The project was completed under significant time constraints due to concurrent coursework and examinations, which limited the opportunity for deeper experimentation and model tuning. Nevertheless, the implemented workflow, comprising reproducible preprocessing, pseudo-label handling, and geographic cross validation, forms a solid, extensible baseline. It demonstrates a clear methodological path for future studies aimed at operational, large-scale snow monitoring across alpine and other mountainous regions.

References

Hall, D. K., Riggs, G. A., & Salomonson, V. V. (1995). Development of methods for mapping global snow cover using moderate resolution imaging spectroradiometer (MODIS) data. *Remote Sensing of Environment, 54*(2)

Hu, F. S., & Shean, D. E. (2022). Machine-learning snow classification in complex alpine terrain using Sentinel-2 multispectral data. *Remote Sensing, 14*(17)

Wang, W., Huang, X., Li, Z., Liu, J., & Liang, T. (2022). Comparison of random forest and deep learning methods for snow cover mapping from multispectral imagery. *Remote Sensing, 14*(3)

Frei, M., Kääb, A., & Huggel, C. (2025). Advances in multi-sensor snow and glacier mapping for alpine hazard assessment. *The Cryosphere, 19*(5)

Liu, Y., Chen, X., & Zhao, Y. (2017). Regional snow cover mapping using Sentinel-2 and Landsat data with support vector machines and random forests. *IEEE Transactions on Geoscience and Remote Sensing, 55*(9)

Outline of own grading thoughts

Breadth: The project combined concepts from several course topics, including supervised learning, evaluation metrics, feature selection and processing sentinel 2 Images. It demonstrated that we have a practical understanding of remote sensing workflows.

Depth: Additional focus was given to multispectral feature design and geographic cross validation, testing how models generalize spatially rather than through random sampling.

Results: The Random Forest classifier improved over the classical NDSI baseline, achieving snow class F! scores up to 0.93 and producing spatially consistent predictions aligned with alpine topography.

Novelty: while the scope was modest, the use of SCL-based pseudo-labels with geographic cross validation provides a transparent, reproducible baseline for potential future work.