International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 11 | Nov - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Deep Learning-Based Analysis of Snow and Ice Dynamics Using Satellite Imagery and Geospatial Data

Manish Kumar (34797)[1] KARTIKEYA LASHKARI (35993)[2] Suhani Shekhawat (33326)[3] Riya kumari (30852)[4] Hans Raj Meena (36279)[5] Dr. Amit Sharma [6]

1 Research Scholar, School of Computer Application and Technology, Vivekananda global university Jaipur

2 Research Scholar, School of Computer Application and Technology, Vivekananda global university Jaipur

3 Research Scholar, school of Agriculture Science Department, Vivekananda global university Jaipur

4 Research Scholar, school of Management Department, Vivekananda global university Jaipur

5 Research Scholar, school of Department of Science, Vivekananda global university Jaipur

6 Associate Professor, Department of Computer Science and Engineering, Vivekananda Global University, Jaipur

Abstract - Satellite imagery may aid in the repeat mapping of snow/glacier surfaces, though the presence of clouds, shadowed areas, pixel mixing and unclear ground reference labels presents challenges in the context of the Himalayan landscape. In this paper, we propose a reproducible deep learning approach for our intended regional investigation at Chhota Shigri Glacier, Himachal Pradesh, India. Our approach incorporates Sentinel-2 L2A reflectance data, glacier inventory context and topographic height data on an aligned 20 m resolution grid. We specify a minimal U-Net architecture that segments visible snow/glaciers and other surface types, while uncertain or obscured pixels are ignored in the process of supervised assessment. We use the Normalized Difference Snow Index as a conventional baseline for the binary task derived from the union of snow and clean-ice surface labels. Geographic blocks and time-series held out satellite acquisitions are employed in order to control for information leakage, in addition to normalization solely based on the training samples. Class-specific segmentation accuracy metrics and area calculation accounting for temporality establish the link between prediction quality and interpretation. The contribution here is in methodological design, while empirical results, satellite-based area measurements and comparisons remain pending until imagery and trained network labels become.

Keywords: Deep learning; snow cover; glacier ice; Sentinel-2 satellite imagery; image segmentation; geospatial analysis.

I. INTRODUCTION

The snow cover and glaciers on the earth's surface constitute interrelated elements of the mountain scenery, but they are still different physical magnitudes. The former one denotes the extent of the territory covered by the snow on the moment of observation. The glacier surface is an observable condition, and the glacier outlines may comprise the parts of ice surface which are covered by the seasonal snow or debris. Studies of operational snow maps and surface classification of the glaciers prove the importance of satellite observations and at the same time show that the chosen target defines what we are able to infer [1], [2].

Repeated spatial observations using satellite imagery can supplement ground-based measurements. However, it should be noted that a change in two classified images could be caused by variations in illumination conditions, acquisition date, presence of clouds, confusion among classes, or genuine change. A transition from snow to ice can happen within a glacier boundary. In turn, disappearance of a snow area outside a glacier does not necessarily mean that a glacier is retreating. Spatial and temporal generalization and prediction uncertainties have recently been studied for glacial mapping [3].

Observations from the Sentinel-2 satellite include visible, near-infrared, and shortwave-infrared bands that have suitable resolutions for regional mapping [4]. The spectral data allow a snow index calculation and a multispectral segmentation algorithm. The deep neural networks can fuse spectral and structural features in images, but an increasingly complex network does not reduce ambiguity in the input data. The U-Net architecture provides an efficient encoder and decoder

with skip layers for semantic segmentation [5]. In the case of regional study in college, it would be easier to train and validate the adapted model.

In this paper, a study case for Chhota Shigri Glacier and surrounding areas in the western Himalayas is defined. The proposed contribution consists of a replicable comparison between image preparation, reference labels review, model validation, and area reporting aware of coverage. The suggested model differentiates snow that can be seen clearly, glacier ice, and other surfaces. A binary classification approach combines the previous two classes to have a simple spectral baseline and the neural model tested against the same target. There is no claim that this is an innovative segmentation architecture or a new application in the Himalayas.

There are three research questions to guide the study. The first question is how effectively does the chosen U-Net classify the three surface types on held-out geographically and temporally observations. The second question is whether or not it outperforms an existing index in respect to the binary classification task of visible snow and clean ice. Finally, the third question is how do areas mapped for the two similar observations differ when both the dates are confined within the same valid footprint?

The current manuscript is a draft of the experiment with all methodologies in place. This manuscript contains all the design of the study, the mathematical formulation, and the model building. It does not contain all the empirical aspects like study scenes obtained, annotation process completed, training done, and the original change measured.

A. Operational meaning of snow and ice dynamics

For the purposes of this paper, dynamics denotes surface class shifts between chosen satellite dates. It is not the same thing as glacier speed or deformation of ice, or a full-fledged mass balance model. This pragmatic definition ties the title of the paper to metrics that are available through the use of visible imagery and a segmentation algorithm. A surface changing from snow to ice is still glaciated in most cases, while off-glacier snow could become rock. Both transitions are legitimate, but they speak to different aspects of the environment.

The study will continue to maintain a geographical summary separately for both the processing window and the inventory context zone. While the former refers to the conditions of visibility in the selected area, the latter refers to the same predictions within an outline of history overlaid with the inventory date. Both do not update the inventory automatically. If there is an overlap of snow or clean ice predictions across an old boundary, this is considered as an error either on the surface, inventory, or registration.

B. Objectives and decision criteria

The first goal is to create a collection of images with traceable reflectance values and identical spatial support. Evidence for this will be a

manifest of scene, aligned stack of bands, and quality masks. The second is to have a comprehensible reference set that identifies exclusions explicitly. Evidence for this will be versioned annotation polygons and class counts. The third goal is to compare a traditional spectral decision against a learning algorithm in the same evaluation conditions. Evidence will be predictions, paired confusion counts, and configuration.

Finally, the goal is to relate these classifications to validated area comparisons. The evidence will take the form of date pairs in tables and maps showing where observations are missing. One goal accomplished does not guarantee the other accomplished: the successful checkpoint does not confirm the annotation, and the visually sensible area curve does not confirm georegistration. These relationships will be tested explicitly. A good study can document an inadequate or negative modeling outcome if its measurements and rationales are clear.

II. LITERATURE REVIEW

A. Spectral approaches to snow mapping

The Normalized Difference Snow Index (NDSI) depends on the differential reflectance of visible-green and shortwave infrared. Procedures for snow mapping using multispectral satellite data were proposed by Hall, Riggs and Salomonson [6]. These procedures provide a clear decision with low computing costs. It is helpful to have such a simple criterion as a baseline, although only its value does not help to determine which cryospheric surfaces one should map and whether a white spot is a glacier. The threshold needs to be determined for a particular sensor and image conditions.

Gascoin et al. provided the description of the Theia Snow dataset and their optical image processing techniques, including clouds treatment and validation [1]. The presented practical experience proves that the snow retrieval process requires preliminary processing and assessment of the quality, besides calculating the index. The described outcomes are not applicable to the suggested Himalayan window due to differences in labels, seasons, and evaluation procedure. Theia dataset is thus considered in this paper as the literature methodological source; the appropriate date matching dataset has not been found.

B. Learned segmentation and glacier surfaces

The U-Net framework was developed for biomedical imaging applications, having an encoder that provides contextual representation and a decoder that predicts spatial predictions through skip connections [5]. This concept when applied to multispectral imagery only modifies the input features and output classes but not the overall architecture itself. The research question in relation to this concept would be its performance when used within a defendable region. It can learn the relationship between local spectral and textural features, which may also include misleading relationships due to improper labeling/splits.

Chu et al. studied attentional enhancement of DeepLab V3+ for glacier extraction using high-resolution optical images [7]. The study addresses challenges relating to snow cover, clouds, and debris-covered glaciers. The study demonstrates sensitivity of learned extraction to both the object of interest and the imaging conditions. The current design does not emulate the images, the architecture used and therefore does not adopt their performance metrics. Their work is only used in motivating the use of explicit surface categories and excluding unsupported claims on completeness of debris-covered glacier boundaries.

GlaViTU was created by Maslov et al., a convolutional transformer architecture for glacier mapping and generalization analysis via open satellite imagery [3]. The research explores spatial, temporal, and cross-sensor aspects as well as uncertainty in mapping. This research has a much wider scope compared to a smaller project for one region. The key point in relation to current research is the fact that deep learning-based glacier mapping already exists. Thus, any contribution should be grounded in regional validation, easy implementation, or some particular behavior, not just the idea that neural networks had not been used before.

The study by Aberle et al. assessed the automated detection of snow cover on glaciers using different remote sensing data with respect to supervised classification and interpretation [2]. This differentiation between snow and other glacier facies is relevant to the creation of the reference mask. It can be problematic to differentiate snow, firn and bare ice based on the imagery. Consequently, the current experiment only assigns labels to surfaces that can be unambiguously identified, while the rest of the pixels will be ignored.

C. Evaluation gap addressed by the proposed study

The interdependence of the neighboring image patches as well as repeated measurement of the same patches will give an impression of making the validation process more robust, but will in reality result in failure of validation. The reason why validation techniques need to consider the temporal and spatial characteristics of the data rather than randomly selected samples is explained by Roberts et al. [8]. In this study, the application will be done in new sites at a later period of time in the regional window.

The selected literature plays a certain role and this is indicated in Table I. It represents a comparative methodology approach rather than a comprehensive systematic review. The study that we have suggested covers an integration problem between the following aspects: target definition and reference independency to like-to-like baseline assessment and qualified temporal area reporting. Whether the application of the region-specific research produces enough empirical data remains an open question.

TABLE I. SELECTED METHODS AND THEIR ROLE

Study	Focus	Role in this study
Gascoin et al. [1]	Operational optical snow maps	Quality masks and validation context
Aberle et al. [2]	Snow on glacier surfaces	Reference class distinctions
Maslov et al. [3]	Generalized glacier mapping	Honest generalization and uncertainty
Ronneberger et al. [5]	U-Net segmentation	Established model architecture
Chu et al. [7]	Attention-based extraction	Optical and debris limitations

D. Reference uncertainty and the comparison target

Paul et al. have analyzed the accuracy of glacier boundary lines produced through remote sensing data and variations resulting from manual interpretation [9]. It relates to the selection of the reference standard since manual drawings of geometrical lines cannot always be accurate. In the proposed research project, disagreements and ambiguous boundaries are thus kept, and no drawing is taken to be infallible. It is a methodological insight from the example of glacier delineation, but not an assumption that identical numerical errors exist for the current surface classes.

Two criteria help compare the literature. An image classifier can estimate the surface facies, whereas the inventory technique will be able to map the whole glacier. Moreover, one research can involve randomly picked image tiles, whereas another one may investigate an entirely new data set or location. Due to these differences, the researches cannot be compared in terms of one score. Therefore, in the current review, the objects of mapping and the methods of their evaluation are considered separately.

Improvement in architecture is also different from improvement in evidence. Adding attention and transformer modules modifies the network architecture, but improved reference independence and blocked inference modify the evidence quality. The proposed research will only focus on the latter pathway while using an existing small architecture. Any further architectural modification at a future time will need to maintain the same test environment and target so that any improvement could be associated with the modification with some level of certainty.

III. METHODOLOGY

A. Study area and scope

Chhota Shigri Glacier is situated in Lahaul Spiti, Himachal Pradesh, at coordinates 32.28°N and 77.58°E, according to a local glaciological study [10]. The suggested processing window spans an area of approximately 30 km x 30 km around this site, thus providing room for including nearby non-glacier terrain and varied snow conditions. The above window is not claimed to represent a hydrological catchment area. The precise polygon and the glacier names covered by the window will be determined after ensuring scene availability and annotation viability.

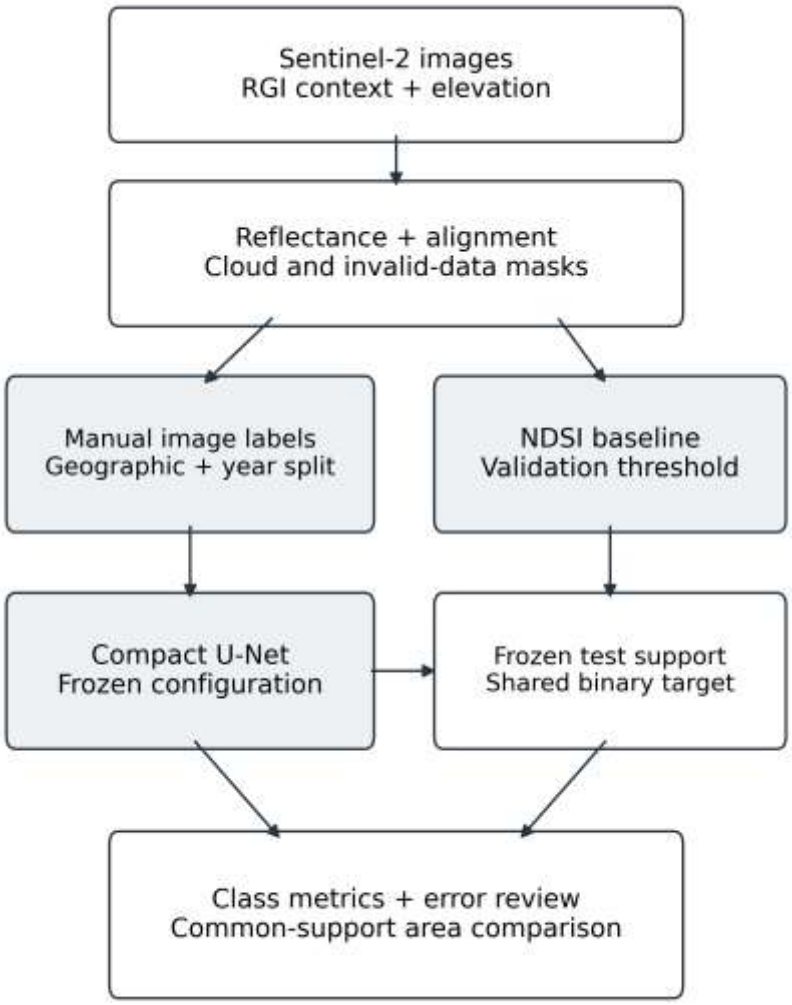
For the first test, the time frame for observations will be from 2021 to 2024. Possible observation periods may be May to June and September to October. This is not yet verified and is suggested as an acquisition window. It is better to have comparisons within the same window than those which may include both springtime and later times of the year. Twelve to eighteen scene counts are aimed at in order to make the first test data set manageable, but no number of usable scenes is yet available.

The geographic coordinate system is kept for queries in the catalogue and annotations on the maps. Raster analysis will be conducted in WGS84/UTM 43N, EPSG:32643, using a static grid of 20m. The final archive should contain the exact area-of-interest polygon, raster starting point, coordinate reference system, and cell size and extent of processing. This reuse of definitions is required for aligned predictions and equal area calculations. Figure 1 illustrates the process flow.

The processing grid will only be constructed one time and stored prior to band stacking. All outputs should have an identical affine transform, pixel orientation, size and coordinate reference system. It is not possible to derive the location of the pixel centers or edges independent of the date using rounded geographic coordinates. A slight shift will cause a

ring of difference to appear around constant surfaces. The test for consistency will, therefore, take into account both equality of metadata and alignment of constant terrain.

The original area-of-interest selection should take into consideration evaluation feasibility as well. There will be no point to use a window where there is only one glacier or one class of surfaces, since it will not be possible to create geographically separate areas for training and testing. The pilot project will verify if such surfaces as snow, clean ice and others are present within the remaining blocks. If it is impossible to use the target through a split, then another



window or another target will be chosen.

Fig. 1. Proposed processing and evaluation workflow. Conceptual design; no experimental imagery or outputs are depicted.

B. Satellite and geospatial data

Sentinel-2 Level-2A surface reflectance data are the main source of images. The official documentation states that bands B02, B03, B04 and B08 have a 10 m resolution, whereas bands B11, B12 and the scene classification band have a 20 m resolution [11]. Bands B02, B03, B04, B08, B11 and B12 constitute the model input, whereas bands B03 and B11 constitute the NDSI input. A common 20 m resolution grid is used to avoid interpolation and representation of shortwave-infrared values as 10 m resolution images.

This version of the Randolph Glacier Inventory version 7 provides context from historical glacier outlines [12]. This inventory's dates and sources should be checked for the subset chosen. It is not an up-to-date snow coverage map; its outlines do not determine changes to glaciers' areas at present. In the current framework, inventory membership is the contextual layer but not the criteria for automatic assignment of all the enclosed pixels as either snow or exposed ice. The attribution of the data set must be preserved.

Copernicus DEM GLO-30 provides elevation information at roughly 30 m resolution, acquired primarily between 2011 and 2015 [13]. This is a digital surface model and is not an elevation measurement that was made simultaneously with

each image date. The use of elevation bands can be used to illustrate where surface variations are mapped. Reprojection to the analysis grid will not improve the true terrain detail. Necessary product notices, release identifier, and cell identifiers are required when producing derived terrain products.

Selected data roles are described in Table II below. Imagery acquisition will make use of the official Copernicus catalog and download services. For each scene, a manifest file will keep the product ID, acquisition time, processing baseline, tile ID, location and checksum. Landsat can be considered as an extension to Copernicus but not as a new data input to the first experimental scenario since it would mean sensor harmonization and more work on validation.

C. Acquisition screening and data provenance

The cloud percentage at the scene level will be used to select candidates, but not as an indicator that the local window is free of clouds. Large tiles may have very low percentages of clouds in total, while the selected glaciers are still occluded by clouds. Candidates will, thus, need to be checked using the actual study polygons. The usable information will then be provided together with the date of acquisition and solar geometry from the product metadata. A visually pleasing image is not chosen only based on future prediction.

If more than one granule covers an acquisition, then the manifest will keep the association between them and the rule used for mosaicking. Duplicates or subsets of the same observation cannot be treated as separate dates. For single date mosaics, all the source granules will be kept to prevent any silent inclusion of consecutive days to create one instantaneous observation. Multi-date composites will only be created when absolutely required, and treated as a new data product.

The database will differentiate between the original files and any derived products. The source images, geographic boundaries, and metadata will remain intact. Aligned stacks of reflectances, masks, and terrain will include the same checksum and processing options applied in generating the product. An incomplete download and/or raster file will be recognized prior to training by checking file integrity and expected band coverage. This documentation allows for recreating the intermediate product and determining if a difference in results is due to different data, preprocessing, or model alterations.

It should also be noted that access paths and product rights should be specified for each downloaded version. Public availability does not make citation or redistribution of someone else's basemap any more acceptable. Map panels in the original research will be created using satellite products properly cited. There will be no accounts and credentials of the providers and data access rights involved. The actual file identifiers and dates of their access will be included after the process of downloading data.

TABLE II. PROPOSED DATA INPUTS

Input	**Native scale**	**Planned use**
S2 B02/B03/B04/B08 [11]	10 m	Visible and near-infrared channels
S2 B11/B12 [11]	20 m	Shortwave-infrared channels; B11 for NDSI
S2 classification [11]	20 m	Quality masks; not reference labels
RGI version 7 [12]	Vector; historical dates	Glacier context; not current surface truth
GLO-30 [13]	About 30 m	Elevation bands; not height change

D. Image preparation and quality masks

All the values from images are going to be scaled to the actual reflectance on the physical surface by either the product metadata or harmonized export. Any changes in the processing baseline for the Sentinel-2 mean that there will be changes in the interpretation of the digital numbers [11]. Therefore, any general division by 10,000 should not be done

without verifying the offsets and export convention. The acquisition manifest needs to show which conversion path was chosen.

Reflectance bands of 10 meters will be combined to the 20 meters analysis grid; continuous bands need to employ an appropriate averaging or interpolation algorithm, while categorical masks will make use of a nearest neighbor technique. Clipping on each date to the same area is required. Registration problems will be evaluated against stable ground features since the offset may be mistaken for a transition of the surface.

The first mask for valid data does not include no data, saturated, and defective pixels; cloud shadow, medium probability, and high probability clouds; and cirrus clouds. These are defined by official classification codes as 0, 1, 3, 8, 9, and 10 [11]. Pixels classified as dark area and low probability clouds need visual identification and not automatic labeling as bare ground. The use of a one-cell dilation from detected clouds is recommended to minimize edge effect. Cost of exclusion in terms of surface is to be noted. Class 11 snow/ice is not an alternative to labels.

Normalization of bands will be done based on the mean and standard deviation of valid training pixels alone. The validation and test scenes will get the same values so that nothing from their distributions affects model building. The equation for standardized channel is presented in Equation (1).

$$z_{i,b} = \frac{x_{i,b} - \mu_b}{\sigma_b + \varepsilon} \quad (1)$$

In the above formula, z(i,b) represents standardized reflectance; x(i,b) represents the reflectance value at pixel i in band b; mu(b) and sigma(b) represent the training only mean and standard deviation respectively; while epsilon is a small positive constant to prevent division by zero. For the NDSI baseline, physical reflectance will be used prior to standardization because of mean subtraction in the index.

E. Quality checks before learning

Spectral validity, image quality, and annotation availability are individual masks. There may be a pixel with spectral validity and no label by humans. Also, there may be pixels with interpretable labels on one date but cloud cover on another. They won't be reduced to the same background value. In order to train the model, it needs supervision which implies a valid image and available label. The model scoring needs reviewed reference. The support for the valid image in a temporal map should be taken into account at both dates regardless of the reference polygons.

In the preprocessing review, channel order, physical units, and percentage of invalid values for each scene will be checked. It is possible to create a plausible array through permutations of channels, but this array will be scientifically incorrect when using it as an input for the model. Also, no-data coded as zero will lead to a biased normalization of training samples in case of ignoring the validity mask. The statistics will be calculated from valid training pixels only.

No terrain correction apart from the given atmospheric correction process is taken into account in the initial pipeline. Rather, the research will focus on analyzing the behavior of errors under hard lighting conditions and determine the scope of the study. A new terrain correction will need documentation of its own process as well as parameter setting and validation. Any such attempt at correction should be documented as an ablation study of the same preparatory process.

F. Reference classes and annotation

The three supervised categories of classification include the other visible surface, visible snow, and exposed clean glacier ice. The last category is the ignore category having the number 255 which indicates cloud cover, bad data, uncertain firn, unresolved dark areas, and boundaries. Rocks without visible ice beneath them do not qualify as exposed clean ice.

This means that the classification should focus on surface conditions observable to the eye and not on all the underlying ice present beneath debris or snow.Reference polygons shall be delineated from date-matched true and false color images using a written interpretative key. The interpreter should not consult the results of the NDSI classification or model predictions when labelling the references since doing so will only serve to test conformity with an algorithm and not with an independent interpretation of the imagery. Where possible, another reviewer will assess a selection of these references and resolve any disputes that arise; this should be documented. Manual interpretation of imagery is still imagery-derived reference data and not field-ground truth.

The limited-scale annotation experiment would determine if the clean ice can be identified consistently at the selected resolution. In case this cannot be done, a combined target of snow and clean ice should be determined prior to the evaluation process, while the finer distinction should be eliminated from the research questions. The ambiguous pixels should not be assigned with values using any of the algorithms under evaluation. The annotation database would keep polygon-based annotations, their dates, reviewer comments and class frequencies.

G. Sampling units and interpretation rules

Reference patches will be selected to contain representative combinations of surfaces and difficult interfaces, and the evaluation population will be specified clearly. Oversampling of rare patches of clean ice is possible during training for purposes of making learning realistic. This oversampling will not mean that there is a large proportion of clean ice in the region. Validation and test statistics must maintain their actual sample and sampling method. In the case where the patches used in testing are nonprobabilistic, area-wide accuracy claims cannot be made.

Polygons will trace the certainty of the interior of a class instead of making a decision on the basis of a mixture of pixels. An uncertainty band can be recorded separately for a problematic interface. It is not possible to selectively remove that band after an analysis of errors in the model. The rule about that band will have to be defined using the annotation guideline, and the percentage of exclusion will be stated by the date and class context.

Consistency of interpretation will be verified before freezing the final masks. The annotator will go back to a small documented sample unseen by him previously, while another reviewer may verify another documented sample. However, disagreement will be tracked not based on a global claim of ground truth, but by surface categories and boundaries. Should there be insufficient evidence to make a reliable distinction between snow and clean ice, the target merged will be used before evaluating the model on the held out dataset.

H. Conventional baseline and comparable target

NDSI thresholding is the standard baseline, based on the spectral theory for snow mapping [6]. Equation (2) shows the definition of the index using green and the first shortwave infrared reflectance bands.

$$NDSI = \frac{\rho_G - \rho_S}{\rho_G + \rho_S} \quad (2)$$

B03 reflectance is indicated by rho(G), and B11 reflectance is indicated by rho(S). Pixels having an unacceptably small denominator are invalid. A base threshold of 0.4 will be considered together with 0.2, 0.3, 0.5, and 0.6 based on the validation dataset. The threshold resulting in the largest value of the validation binary intersection over union will be locked before the test inference step. In case of ties, the larger threshold will win.

NDSI is not a snow/ice discriminator for three classes. In such a case, the comparison test will involve the union of visible snow and clean ice. The output from the U-Net model will also involve the same union, and the two methods will have the same number of valid pixels. Three-class U-Net results will be reported separately. This will avoid an incorrect comparison between binary baseline accuracy and a multiclass model. There will be cases where thresholding might misclassify water or bright rocks.

Binary target has explicit correspondence to the three-class reference target with snow and exposed clean ice constituting the foreground, while all other visible surfaces constitute the background. Ignore mask goes untouched. The corresponding neural decision is found using the argmax function on three-class probabilities with merging of two classes of foreground into one. It results in an identical criterion to that of the reference union. Probability sum criterion would be a different decision criterion and would need independent validation. It will not be substituted afterwards, regardless of the test results.

Choice of baseline thresholds will be recorded as a small controlled search, along with the potential values tested and their validation measures. No threshold will be selected based on minimization of the area change or through correlation with an environmental storyline. The baseline is permitted to outperform the neural network. An optional future reflectance guard or water exclusion rule will form its own baseline variant and will have to be evaluated without removing the initial test.

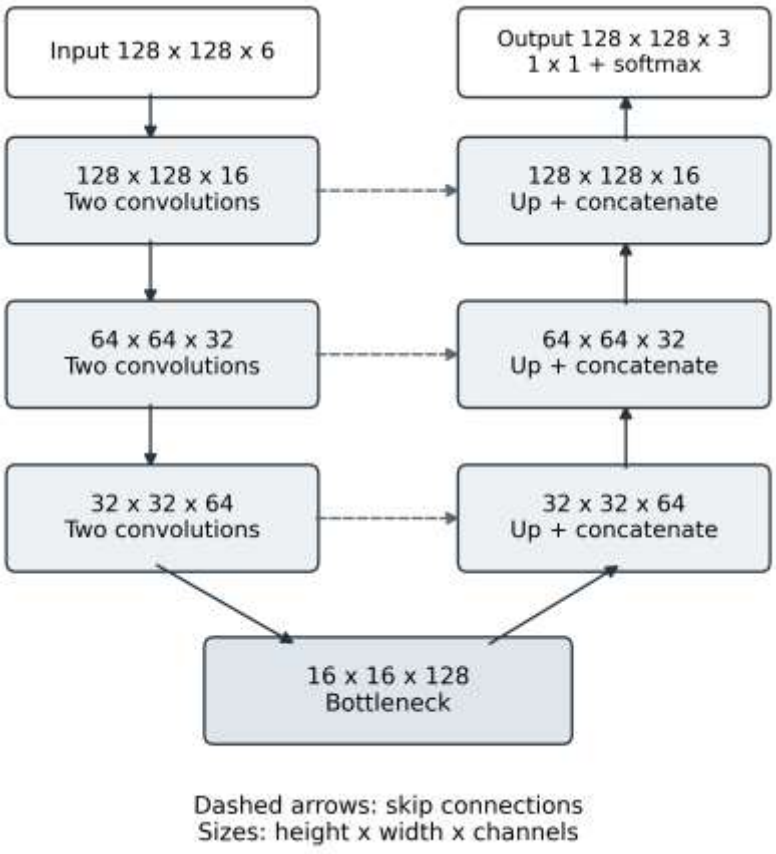
I. Deep learning segmentation model

This network architecture is a small-scale 2D U-Net derived from the well-known encoder-decoder model [5]. This architecture takes inputs of six channels in 128 by 128 pixel patches. The three encoders have channel widths of 16, 32, and 64, and a 128-channel bottleneck comes after the third max-pooling layer. Each of the three encoder layers has two 3 by 3 padded convolutional layers with rectified linear activation function. The decoder expands the features with upsampling and concatenation of the respective encoder features.

See Figure 2 for the sizes of the tensors and skip connections specified. The model is designed compactly to facilitate a low-cost first implementation whose memory requirements and runtime have not yet been evaluated. Dropout, attention layers, pretrained encoders, and extra terrain input channels have all been omitted to ensure that the comparison between models remains clear. Any future modifications would need to be accompanied by justification and new evaluation without retuning on the test set.

B02, B03, B04, B08, B11 and B12 bands shall be maintained for channel arrangement through the entire process of preparation and inference. The above sequence is expected to be preserved in every tile loader as well as the saved model setup. Elevation and glacier inventory are provided for context and stratified analysis purposes but not as hidden classes or channels in the model to begin with.

All convolutional layers are padded to maintain their spatial dimension in that stage. The input layer whose dimension is a multiple of eight allows three stages of pooling. In case of inference, the edge tiles are padded in order to enable compatibility with the neural network and later crop to match their original size. It should be made sure through the padding strategy and valid data mask that the artificial borders of an image do not make the surface



prediction.

Fig. 2. Proposed compact U-Net with tensor sizes and skip connections, adapted from the architectural principle in [5]. All sizes are specifications, not measured results.

Training involves a combination of multiclass cross-entropy loss with a soft Dice loss. Dice loss criteria have previously been utilized in optimization of segmentation overlap [14]; the current multiclass formulation is a deliberate modification and not an attempt at replicating the volumetric formulation. Equation (3) provides the masked cross-entropy loss.

$$L_{CE} = -\frac{\sum_i v_i \ln p_{i,y_i}}{\sum_i v_i} \quad (3)$$

The value of validity indicator of the binary classifier v(i) will have a value of one, only for the case of a supervised training pixel. y(i) is the reference class for such a pixel and p(i,y(i)) is its respective prediction probability. The summation is done over the pixels belonging to a particular batch and only the batch which does not have any valid labels is ignored. y(i,c) is the one-hot indicator of the reference class c.

$$D_c = \frac{2\sum_i v_i p_{i,c} y_{i,c} + \varepsilon}{\sum_i v_i p_{i,c} + \sum_i v_i y_{i,c} + \varepsilon} \quad (4)$$

The totals in Eq. (4) have the same validity mask. The loss function calculates the Dice score for all three categories – the background included. The foreground mean IoU does not consider the background. Eq. (5) is the sum of the losses with equal weights. The smoothing method and no-category policy should stay the same across all runs. In Eqs. (4)-(5), L is the overall training objective, and c = 0, 1, 2 correspond to other surface, snow, and clean ice classes.

The weighting for cross-entropy and Dice loss are predetermined configurations and not empirically optimized values. Cross-entropy measures the probability of the target class being predicted, and overlap loss promotes consistency between the classes through the supervised batches. Both metrics behave differently numerically, meaning that the training plots will be stored separately from the combined objective function. To alter the weights, validation-only selection would be necessary alongside keeping track of candidate configurations.

Losses calculations will be performed on small controlled tensors prior to conducting an experiment with satellites. If the batch is completely ignored by the system, then it needs to be skipped; a perfect categorical prediction will yield expected limits behavior; and the three class probabilities will sum up to 1 in every pixel. All these are just checks on mathematical behavior that cannot be taken for satellite classification ability.

$$L = 0.5L_{CE} + 0.5\left(1 - \frac{1}{3}\sum_{c=0}^{2} D_c\right) \quad (5)$$

J. Experimental setup and leakage control

Implementation is to be carried out in Python using libraries such as PyTorch, NumPy, Pandas, Rasterio, and Matplotlib. Annotation and map visualization is to be handled by QGIS. Environment details will be saved by exporting the list of installed packages and run configuration. Preparing images can be done on the CPU. Training is expected to take place on a GPU when available, even temporarily through Colab GPU. No particular GPU, time taken for training or availability of service is considered. Device details will be provided later.

Training will proceed using Adam [15] with a learning rate of 0.001, a batch size of four, and up to 50 epochs. The early stopping criteria will involve foreground mean IoU for validation with patience of eight epochs, keeping the best validation checkpoint. Horizontal and vertical flipping, and 90° rotation transformations will apply to both input and label pairs only during training patches. The main random seed will be 42, and further sensitivity tests can be carried

out using additional seeds after a valid pipeline has been established. Table III provides the difference between these and actual results.

TABLE III. PROPOSED TRAINING CONFIGURATION

Parameter	**Specified starting value**
Input and patch	6 channels; 128 by 128 pixels
Classes	Other, snow, clean ice; ignore 255
Encoder / bottleneck	16, 32, 64 / 128 channels
Optimizer	Adam; learning rate 0.001
Batch / maximum epochs	4 / 50
Early stopping	Validation foreground mean IoU; patience 8
Loss	0.5 cross-entropy + 0.5 soft Dice loss
Random seed	42; record actual determinism settings

Geographic blocks will be selected prior to extracting patches, and all dates belonging to the same location will be kept in the same block. At least a one-patch-width buffer will keep split patches apart. Potential years for training are 2021-2022, for validation 2023, and for testing 2024, along with disjoint geographic blocks. This is to evaluate joint spatial and temporal transfer without separating their individual contributions. Split feasibility, class balance, and availability of labeled data will need to be verified before freezing the manifest. Overlapping or neighboring patches cannot be randomly allocated in splits [8].

Preprocessing statistics, sampling weights, early stopping, and the NDSI threshold will be chosen without the test set. Predictions for the test set will be made only once the model configuration is fixed. Full scene prediction can take advantage of overlapping windows with averaging to avoid seams; all models will then be evaluated on the same valid test set. If any redesign is done on the basis of evaluation, it should be clearly flagged as such.

K. Prespecified experimental sequence

In the first step of the experiment, a deterministic data loading and masking pipeline will be established, then the regular threshold will be fitted to the validation data, and finally the actual U-Net run will be performed. The source images and labels will be kept constant for both runs. Failures specific to the model under test will be logged, rather than fixing the problem by altering the reference.

The main analysis utilizes the frozen checkpoint and threshold. Other seeds, masking changes or loss removal fall into the second analysis type and reuse the same protected test interpretation. If many exploratory options are considered, a good outcome from among them should not be taken as the unique hypothesis. The count of tried configurations and the

way they were selected will be saved. Figure 3 demonstrates the concept of combined geographical and temporal separation without stating that the real area has already been divided.

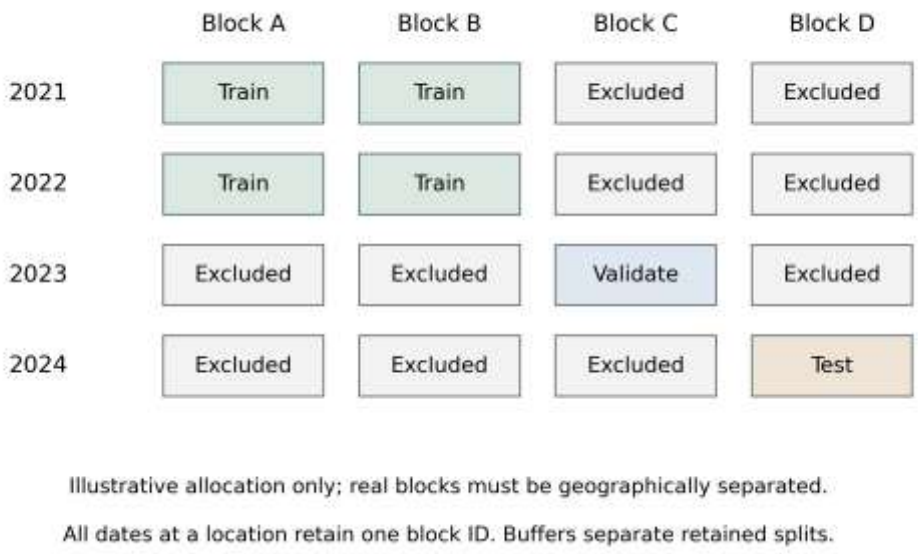


Fig. 3. Conceptual joint spatial and temporal holdout design. Blocks are illustrative labels, not an acquired geographic partition; unselected combinations are excluded from fitting and scoring.

TABLE IV. PRESPECIFIED EXPERIMENTS

Experiment	**Fixed components**	**Question**
Primary binary comparison	Source scenes, reference union, test support	Does the U-Net union improve on NDSI?
Three-class assessment	Reviewed snow and clean-ice labels	Which surface classes are confused?
Mask sensitivity	Predictions and declared reference	How does exclusion of cloud edges affect support?
Seed sensitivity if feasible	Data, split, architecture and settings	How variable is model fitting?
Temporal comparison	Seasonal selection and common footprint	What visible area differences can be supported?

Sensitivity analysis is constrained by its scope. There could be a comparison that changes the cloud-mask dilation value but does not change the model and labels; there could be one that changes the value of the NDSI threshold within the stated range; and there could be one that looks at the effect of training the seed again. Each analysis must specify what observations or parameters were altered. The above examples are used for identifying instabilities; they cannot serve as a substitute for the protected analysis.

L. Segmentation evaluation

For each category, TP, FP, and FN will be tallied based on the labels from the reviewed test images. Formulas (6)-(9) represent IoU, precision, recall, and F1. These formulas apply to the merged foreground as well when comparing against the baseline.

$$IoU_c = \frac{TP_c}{TP_c + FP_c + FN_c} \quad (6)$$

$$Precision_c = \frac{TP_c}{TP_c + FP_c} \quad (7)$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c} \quad (8)$$

$$F1_c = \frac{2TP_c}{2TP_c + FP_c + FN_c} \quad (9)$$

None of the measures excludes ignored pixels. In case the denominator equals to zero, the measure is marked as undefined instead of giving it automatic perfect score. Classes being evaluated and the support of those classes are reported for each aggregated measure. FGM IoU is a mean of defined snow and exposed ice IoUs; the individual results and confusion matrices are needed because one measure can conceal a poor-performing class. Overall accuracy is not considered as a primary measure in case of dominating background.

The principal baseline-model difference is to be estimated on paired test pixels and aggregated across geographical blocks. A block bootstrap with 1,000 bootstrap samples is recommended to estimate intervals in the presence of sufficiently many independent blocks. Block sizes, and not pixels, are the sampling units in the bootstrap. In cases where there are insufficient independent blocks, the research will present variability measures and note that the interval is unstable. There is no intention to combine seed, label and cloud mask uncertainties under one uncertainty measure.

Paired comparisons are needed since the difficulty of held-out surfaces is not stable across blocks. Both the model and the baseline would be compared to a common reference for each eligible block. The counts would initially be kept at the block level before pooling them together. Both the pooled statistics and the distribution of block statistics are interesting measures to consider since the pooled statistics give more weight to blocks with a higher number of labeled pixels than the block averages, which weigh blocks equally.

Resampling with the bootstrap will occur through re-sampling blocks, with replacement, and calculating the paired difference in metrics for each resample. The choice of interval is dependent on the choice of blocks and their capability of representing the intended application. Thousands of pixels in one block do not generate thousands of independent geographic observations. If only a few independent blocks remain after filtering, the focus will be on their results. Absence of an accurate interval is better than the presence of an inaccurate pixel interval.

Error checking will employ a pre-defined list of categories which will be created before choosing representative panels: snow and clean ice confusion, reflective non-glacier regions, shadows misclassification error, mixed boundaries and cloud edge contamination. Illustrations will be selected from the withheld data and will consist of both successful and failed examples. A visually convincing illustration is not an objective assessment of system performance. It is only meant to demonstrate the pattern of the numbers and pinpoint the location of the limiting factors.

M. Geospatial area and temporal comparison

Let V denote the common foot print of the two dates whose clear and valid image support is considered. The mapped area of class c at time t can then be calculated using Equation (10) based on the predicted surface labels. It is not a corrected estimate of the glacier area.

$$A_{c,t|V} = 10^{-6} \sum_{i \in V} a_i \, 1(\hat{y}_{i,t} = c) \quad (10)$$

In this case, a(i) is the area occupied by the pixel in square metres and the indicator function takes value of one in the event that the class predicted is c. The selected metric grid uses a(i) which is around 400 square metres; the conversion to square kilometres is done through division by one million. The same V must be used for both days. Valid-area fraction and excluded area will go together with all pairs.

Absolute area change is shown by equation (11), and percentage change is represented by equation (12) if the first area is positive. If the first area equals zero, the percentage change will be undefined, but the absolute difference will have meaning. The stratified elevation results will use a set distance of 200 m for each band.

$$\Delta A_c = A_{c,t_2|V} - A_{c,t_1|V} \quad (11)$$

$$Change_c(\%) = 100 \, \frac{\Delta A_c}{A_{c,t_1|V}} \quad (12)$$

The difference between mapped area and area estimation aided by an appropriate reference-sampling design has been made by Olofsson et al. [16]. The first experiment will include results of mapped class areas and their constraints regarding classification. In case the representative probability sampling with adequate reference data is available, error adjusted area analysis can be included additionally. Bootstrapped confidence interval around segmentation statistics is not a confidence interval for true snow area.

N. Coverage denominators and change interpretation

For each temporal comparison, there will be three zones considered: the polygon used for the study, the zone covered by an image at a particular date, and the intersection zone for that pair. The percentage of the polygon represented by the intersection zone forms part of the output. In cases where only a small portion or high-elevation portion of the region is retained by a particular date pair, the difference in the areas does not represent the whole time window. One possible eligibility criteria may be 80 percent common image support.

The pairwise comparison sequences may generate an intersection at different points depending on the pairs. These values should not be combined into one curve for the area with variable denominators. In the case of a series of dates that is similar to the one discussed above, one can generate a footprint that applies to all the selected dates, which will indicate the cost of discarding extra areas. The choice will be clearly documented, and the footprint will be depicted in the map or caption.

There are multiple sources of area uncertainty which do not vanish even when pixels are counted precisely. Classification errors modify the surface of interest, registration errors move a boundary, and coverage selection modifies the set of locations being considered. The older inventories and terrain databases have their own contextual constraints. The table below organizes the diagnostics of these sources of uncertainty. It represents a plan of analysis, rather than an estimate of the total uncertainty.

TABLE V. UNCERTAINTY SOURCES AND PLANNED DIAGNOSTICS

Source	Diagnostic	Limit of inference
Reference interpretation	Review disagreements and ignored fraction	Image interpretation is not field truth
Cloud and shadow masks	Coverage and mask sensitivity	Excluded area can be unrepresentative
Spatial registration	Stable-terrain alignment check	A shifted edge can mimic a transition
Model fitting	Protected validation and seed sensitivity	One run does not establish stability
Sampling dependence	Geographic blocks and paired metrics	Pixels within a block are not independent
Inventory and terrain dates	Record source dates and context	No current glacier boundary or thinning estimate

IV. RESULTS

A. Experimental status

Empirical findings have yet to be obtained. None of the following can be provided for this paper: scene shots from experiments, final reference masks, split manifest that has been "frozen", U-Net checkpoint trained and prediction hold-out set. Therefore, no percentages of accuracy, area per class, training loss, nor seasonal difference have been measured.

Tables VI and VII contain definitions of necessary measurements to conduct this experiment. Pending means that there is no information available.

TABLE VI. TEST PERFORMANCE TO BE MEASURED

Method and target	IoU	F1	Precision / recall
NDSI; binary union	Pending	Pending	Pending
U-Net; binary union	Pending	Pending	Pending
U-Net; snow class	Pending	Pending	Pending
U-Net; clean-ice class	Pending	Pending	Pending

The two rows for the binary union in Table VI will use the common visible snow and clean ice target. The rows for the three-class U-Net will be defined by the class-specific rows and a confusion matrix to be developed. The observation that the neural network outperforms the benchmark model should involve a paired difference on the frozen test surface along with block variance. Any equal or worse neural network performance is an acceptable observation and cannot be replaced by any image or seed.

B. Required temporal and visual evidence

TABLE VII. TEMPORAL AREA MEASUREMENTS PENDING

Required quantity	Observed value
First and second acquisition dates	Pending
Common valid area / coverage	Pending
Snow area at each date	Pending
Exposed clean-ice area at each date	Pending
Absolute / percentage change	Pending

Acquisition dates, valid area and mapped class areas need to be provided in place of every pending field in Table VII. Maps shall indicate the date, coordinate reference system, legend, scale, valid footprint and the source. Qualitative evaluation panels need to display the held-out image patch, manual reference, NDSI prediction and U-Net prediction. Examples with terrain shadow and mixed pixels need to be included despite their being challenging. There is no satellite-like example, predicted map or training curve that can take the place of an unexecuted experiment.

The current deliverables are the detailed experimental design and conceptual framework depicted in Figs. 1 and 2. These describe how the experiment will be executed and assessed, but they lack any demonstration of predictive ability or environmental modification. The Results chapter is possible only when the associated scene, annotations, logs, and predictions have been captured.

C. Analysis sequence after evidence is available

The completed section of Results must first state the data that managed to make through the filtering and annotation stage. Scene counts, dates, labeled areas, ignored fractions, and independent block counts will describe the extent of the evaluation even before the scores are analyzed. The next subsection must include the fixed baseline threshold, selected model checkpoint, and accompanying binary metrics. The results from class-specific U-Net must be reported as well along with their supporting evidence to avoid a misleading mean result.

The temporal findings will also have to show the identified season pairs, their common valid region and the surface transitions that are visible. A decrease in snow and an increase in clean ice is not the same as a decrease in their union. This can be stated without attributing it to climate. The conclusion made must rely on the type of transition noted and the quality of images used. Where there is failure to meet the required conditions for a comparison, this becomes a reporting result. An alternative favorable date cannot be picked without reason.

V. DISCUSSION

A. Interpretation supported by the design

The major decision of methodology involves separating the classification of the surface from that of the glacier outline. If a classifier can properly classify the bare ice surfaces, the glacier ice covered with debris is excluded by definition. Similarly, the presence of snow in the landscape could occur on glaciers and non-glaciers. The suggested outputs will be useful in analyzing the surface variability visible, as long as it is clear what areas are the targets and which areas are being excluded. The union map cannot be reclassified as the glacier extent.

Target equivalence is important for a good baseline comparison. NDSI thresholding doesn't give the same set of classes as the network and hence its binary prediction must be compared with an explicitly collapsed neural prediction. This enables the experiment to test the hypothesis that spatial context will help in this task. The three output classes from the neural model cannot be used as proof of better performance. The other experiment involving snow and ice is contingent upon the success of the annotation pilot.

The comparison also has a computational meaning, but performance measures have to be computed instead of being assumed. The threshold-based approach does not require the computation of any weights, but the neural approach will need annotation, optimization, and checkpointing. If additional complexity of the second approach is worthwhile remains to be seen from how much better it performs. A slight increase in accuracy may not matter much if it appears only in surfaces not covered by the application or is less than reference disagreement.

Even a null finding will be beneficial to establish the right approach to use. In case where NDSI gives similar results with the common target under ideal local conditions, the simpler approach may work well for this localized context. Where the neural network brings additional value only at mixed boundaries, this needs to be reported without overstating the superiority. In case where both approaches face challenges in cases of shadows and unclear firn, the finding might mean a limitation of the selected observations and labels.

B. Uncertainty and generalization

Temporal and geographic holdouts are more difficult to implement than random split, yet more realistic when it comes to making predictions about unfamiliar areas or future observations [8]. The score will remain conditional on the regional time window chosen. Since the suggested approach to holdouts affects both the region and time period, it is impossible to determine which of them caused the increase in errors. Thus, two separate experiments should be run – geographic holdout experiment and temporal holdout experiment. A positive outcome of the local experiment would mean its applicability in the region.

The exclusion of uncertain pixels avoids forced labeling but limits the population to be assessed. In case the clear and well-lit pixels are more easily labeled than those in shadow or covered with debris, the resulting metrics will apply only to the easier environment. The proportion excluded and how the exclusion applies should be given along with the table of accuracies. Reference reading may be inaccurate even when independent of the evaluated algorithm. Another reader could reveal disagreement but not generate truth on the ground.

Cloud masking and timing of the image acquisition influence the temporal analysis separate from the process of segmentation. Scenes acquired within one calendar month might have different history of snow cover and different illumination conditions. Limitation of the analysis to a subset of the valid support minimizes one source of the potential bias, but not a proof that this smaller area is representative for the whole processing time period. The limited number of optical images allows characterization of certain visible changes, but not for a long-term trend or its reasons.

C. Practical limitations and reproducibility

The design focuses on a small scale model, a limited number of bands, and a regional collection that is not too large. Its primary task, however, may turn out to be annotation for referencing rather than training. With the use of 20 m grid,

there will be mixed pixels around narrow snow patches and surface edges, and the old model cannot estimate modern thinning. There is also a risk of using a fixed mask which eliminates dark ice along with cloud margins.

Reproducibility goes beyond merely providing the network. The distribution of the code must include the study polygon, the source-product list, the reflectance conversion parameters, the reference-label guide, the versioned masks, the split list, the normalization values, the configuration, the package versions, the random seed, and the checkpoint used. The predicted rasters must retain their spatial metadata. External data that is licensed and/or restricted data should be referenced and not distributed in violation of their licensing conditions, while public datasets must be credited.

The next milestone of evidence is the completion of a feasibility pilot with annotations using a limited number of scenes. Following the freezing of target and splits, baseline and U-Net can be run without any adjustment based on tests. The outcome of the evaluation will confirm whether the suggested approach provides any regional comparison value. At that point, feasibility will become an experimental result rather than a premise.

D. Transferability and extension boundaries

Application to another region within the Himalayas would be more complicated than just using the trained checkpoint. The new region may have different characteristics in terms of ice appearance, background rock, or light acquisition. The original algorithm could be tested in the new location by the way of fixed transfer test, using a separately defined reference sample. Training on the new region would be considered a separate test. Combining both processes would make it hard to understand whether the regional algorithm transferred.

An additional satellite mission would affect the interpretation as well. Band responses, native resolutions and processing traditions would need to be harmonized and the analysis would need to sort out cross-instrument effects from date-to-date differences on the ground. Likewise, measuring the area of debris-covered glaciers would entail an enlarged coverage area along with appropriate evidence. This current design does not address such needs through omission; rather, it identifies them as potential extensions.

VI. CONCLUSION

This manuscript describes a replicable methodology to study visible snow cover dynamics and clean ice exposure through Deep Learning using Sentinel-2 images and geospatial information. This connects a small U-Net architecture with an explicit NDSI benchmark, stand-alone image classification and withheld area calculation based only on feasible observations. Contribution is a precisely described regional analysis including distinct demarcation between surface properties, glacier extent and mass balance.

The framework is still not proven through empirical testing. Superiority of the models, accuracy of segmentation and changes in the temporal area are thus open questions, and there is no inference made about environmental trends. What is left to do is real image acquisition, mask verification, splitting for evaluation and storage of the model predictions. Future work could include more areas, debris covered glaciers, additional terrain features or field observations independent of the first test.

REFERENCES

[1] S. Gascoin, M. Grizonnet, M. Bouchet, G. Salgues, and O. Hagolle, “Theia Snow collection: high-resolution operational snow cover maps from Sentinel-2 and Landsat-8 data,” Earth System Science Data, vol. 11, pp. 493-514, 2019. doi: 10.5194/essd-11-493-2019.

[2] R. Aberle et al., “Automated snow cover detection on mountain glaciers using spaceborne imagery and machine learning,” The Cryosphere, vol. 19, pp. 1675-1693, 2025. doi: 10.5194/tc-19-1675-2025.

[3] K. A. Maslov, C. Persello, T. Schellenberger, and A. Stein, “Globally scalable glacier mapping by deep learning matches expert delineation accuracy,” Nature Communications, vol. 16, Art. no. 43, 2025. doi: 10.1038/s41467-024-54956-x.

[4] M. Drusch et al., “Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational Services,” Remote Sensing of Environment, vol. 120, pp. 25-36, 2012. doi: 10.1016/j.rse.2011.11.026.

[5] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015, LNCS, vol. 9351, pp. 234-241, 2015. doi: 10.1007/978-3-319-24574-4_28.

[6] D. K. Hall, G. A. Riggs, and V. V. Salomonson, "Development of methods for mapping global snow cover using moderate resolution imaging spectroradiometer data," Remote Sensing of Environment, vol. 54, no. 2, pp. 127-140, 1995. doi: 10.1016/0034-4257(95)00137-P.

[7] X. Chu, X. Yao, H. Duan, C. Chen, J. Li, and W. Pang, "Glacier extraction based on high-spatial-resolution remote-sensing images using a deep-learning approach with attention mechanism," The Cryosphere, vol. 16, pp. 4273-4289, 2022. doi: 10.5194/tc-16-4273-2022.

[8] D. R. Roberts et al., "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," Ecography, vol. 40, no. 8, pp. 913-929, 2017. doi: 10.1111/ecog.02881.

[9] F. Paul et al., "On the accuracy of glacier outlines derived from remote-sensing data," Annals of Glaciology, vol. 54, no. 63, pp. 171-182, 2013. doi: 10.3189/2013AoG63A296.

[10] M. F. Azam et al., "Reanalysis of the longest mass balance series in Himalaya using a nonlinear model: Chhota Shigri Glacier (India)," The Cryosphere, vol. 18, pp. 5653-5672, 2024. doi: 10.5194/tc-18-5653-2024.

[11] Copernicus Data Space Ecosystem, "Sentinel-2 L2A," technical documentation. Accessed: Nov. 15, 2026. [Online]. Available: https://documentation.dataspace.copernicus.eu/APIs/SentinelHub/Data/S2L2A.html

[12] RGI 7.0 Consortium, Randolph Glacier Inventory - A Dataset of Global Glacier Outlines, Version 7. NASA NSIDC DAAC, 2023. doi: 10.5067/F6JMOVY5NAVZ. Accessed: Nov. 15, 2026.

[13] European Space Agency, "Copernicus DEM - Global and European Digital Elevation Model," Copernicus Data Space Ecosystem. Accessed: Nov. 15, 2026. [Online]. Available: https://doi.org/10.5270/ESA-c5d3d65

[14] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in Proc. Fourth International Conference on 3D Vision, 2016, pp. 565-571. doi: 10.1109/3DV.2016.79.

[15] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in Proc. International Conference on Learning Representations, 2015. [Online]. Available: https://arxiv.org/abs/1412.6980

[16] P. Olofsson, G. M. Foody, M. Herold, S. V. Stehman, C. E. Woodcock, and M. A. Wulder, "Good practices for estimating area and assessing accuracy of land change," Remote Sensing of Environment, vol. 148, pp. 42-57, 2014. doi: 10.1016/j.rse.2014.02.015.