

Ice-Core Micro-CT Image Segmentation With Deep Learning and Gaussian Mixture Model

Faramarz Bagherzadeh^{1b}, Johannes Freitag, Udo Frese, and Frank Wilhelms

Abstract—Ice cores of polar regions (ice sheets) are one of the most prominent natural archives that can reveal essential historical information from the past environment of our planet. The ice-core microstructure is a key feature in determining the principal properties of ice such as pore close-off, albedo, and melt events. Microcomputer tomography (CT) scans can provide valuable information about the microstructure of materials, although achieving a high-quality automated segmentation of porous materials, especially with phase/density changes is still a challenge. This work proposes a new method for improving the segmentation of porous microstructures where a weak segmentation [Gaussian mixture model (GMM)] on high-resolution (30 μm) data is used as ground truth to train a deep-learning model (U-net) for segmentation of low-resolution (60 μm) data. This approach has reached high segmentation accuracy in terms of quantitative metrics having the F1-score of 92.5% and an intersection over the union (IoU) of 91%, with a considerable improvement compared to thresholding and unsupervised methods. Also, the segmentation results of U-net are closer to the real weight, density, and specific surface area (SSA) of the specimen.

Index Terms—Deep learning, ice core, microcomputer tomography (CT), segmentation, unsupervised learning.

I. INTRODUCTION

OVER time, snow accumulates in polar ice sheets (Antarctica and Greenland) and undergoes a process of compaction due to the weight of the overlying snow layers. This compaction causes the snow grains to become more tightly packed, resulting in the formation of a more dense structure ranging from sintered snow to firn and ice.

As snow accumulates, the underlying firn becomes more compact, and it can eventually transform into solid glacial ice. This transformation can take many years or even centuries,

Manuscript received 22 March 2023; revised 8 June 2023 and 20 September 2023; accepted 12 November 2023. Date of publication 20 November 2023; date of current version 4 December 2023. This work was supported by the Open Access Publication Funds of Alfred Wegener Institute of Polar and Marine Science. The work of Faramarz Bagherzadeh was supported by the Helmholtz School for Marine Data Science (MarDATA) under Grant HIDSS-0005. (Corresponding author: Faramarz Bagherzadeh.)

Faramarz Bagherzadeh is with the Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, 27568 Bremerhaven, Germany, and also with the Faculty of Mathematics and Computer Science, University of Bremen, 28359 Bremen, Germany (e-mail: Faramarz.bagherzadeh@awi.de).

Johannes Freitag is with the Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, 27568 Bremerhaven, Germany.

Udo Frese is with the Faculty of Mathematics and Computer Science, University of Bremen, 28359 Bremen, Germany.

Frank Wilhelms is with the Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, 27568 Bremerhaven, Germany, and also with the Geoscience Centre, University of Göttingen, 37077 Göttingen, Germany. Digital Object Identifier 10.1109/TGRS.2023.3334867

depending on factors such as temperature, precipitation, and other environmental conditions [1]. Air that was once in the open pore structure of snow gets preserved in the form of air bubbles surrounded by ice. These air bubbles are snapshots of the atmosphere of the past [2].

Polar ice cores are an excellent repository of historical climate, due to their ability to preserve a wide range of proxies, including greenhouse gas concentrations [3], aerosol-related atmospheric impurity records [4], on timescales ranging from decades to hundreds of millennia, and they could be utilized as a tool to track the history of global temperature [5], [6].

One of the main features of the ice core is microstructure. Ice-core microstructure contains invaluable data about melting events (melt layers) [7], optical properties, and global warming information [8], [9], [10].

The evolution from snow to ice occurs typically over the top 120 m. Hence, to understand the driving processes, it is necessary to investigate this depth range. The structure of the snow/firn column plays a critical role in determining the pore close-off point (age gap between ice and gas) as well as ice-core dating [11], [12]. This section of the ice core is the part that this study will focus on.

To interpret the microstructure of snow/ice cores, there are different techniques such as fabric analysis, Raman spectroscopy, large-area scanning microscope (LASM), optical line scanner, scanning electron microscope (SEM), and computer tomography (CT). All these methods produce image data that could benefit computer vision techniques to process [13], [14], [15], [16], [17], [18].

One of the most promising methods of ice-core microstructure analysis is X-ray imaging, specifically CT. The CT scan data is a sequence of 2-D images being horizontal slices through the core and in this study, each 2-D image is considered individually. Ice-core micro-CT images can be taken in various types, modalities, resolutions, and ice characteristics. This leads to the creation of CT images of a porous material (e.g., ice-air mixture) that has a diverse density across the top 120 m of ice core starting with low-density snow (0.08 g/cm^3) and then the middle part firn (0.35 g/cm^3) and high-density Polar ice (0.91 g/cm^3). This is a major challenge to develop a robust image segmentation tool for such a range of densities.

To perform image segmentation, the current state-of-the-art algorithms are divided into two main groups. The first group is traditional algorithmic-based models such as the region growing [19], random walk [20], active contour models [21], and graph cut models [22]. These models perform many

computations while producing a segmented mask and some of them are prone to changes in image histogram and range of intensity values. Also, they require user intervention (manual inputs) to obtain an accurate segmentation. For example, the user needs to provide seeds for region growing, graph construction parameters, adjacency matrix parameters, energy function parameters, and so on.

The second group is based on artificial intelligence (A.I.) models that rely on convolutional neural networks (CNNs) which can get the segmentation results with a significant decrease in inference time and minimal need for human intervention. Among CNN models, the U-shaped encoder–decoder architecture (U-Net) is well-known for image segmentation and used by researchers in different fields [23], [24], [25], [26]. Despite the high accuracy of these models, they are computationally expensive during the training phase, and they mostly require plenty of labeled data (supervised learning). Also, these models are not fully interpretable, and there is a possibility of creating unexplained features while performing segmentation [27], [28].

By utilizing the above-mentioned segmentation methods, different research groups tried to perform segmentation on micro-CT images of porous materials in different fields of study [29], [30]. Some researchers have used aggregated physical microstructure parameters to validate the segmentation results, that is, density, porosity, specific surface area (SSA), trapped air bubbles, and casting, while others with ground-truth images used pixelwise metrics such as intersection over union (IoU), Dice, and accuracy.

Researchers with Otsu's thresholding and triangle algorithm of ImageJ (a Java-based image-processing program) segmented ice-core X-rays with an empirical approach. Their algorithm was able to segment young sea ice through X-ray microtomographics with an absolute uncertainty of 0.5%–1% brine volume [31].

Another research group has shown the capabilities of methods such as sequential filtering (denoising, thresholding, and postprocessing) as well as energy-based segmentation on CT scan volumes of snow. Then, the microstructure parameters such as SSA were computed from the segmented images. It was shown that both of the methods were prone to hardware setup, but the energy-based (graph cut) method is not subjective to the choice of a human user and it benefits from local spatial information, with less affection of beam hardening [32], [33], [34].

Similarly, researchers implemented energy-based methods on super-voxels with a majority voting approach (Q-Cuts-3-D) to segment a porous media with various particle shapes. After comparing the model outputs with the given ground truth, they reported reaching an average IoU of 0.88, which was higher than previous studies [35].

On the other hand, with a deep-learning (3-D-Unet) approach, researchers were able to segment lithium-ion battery microstructures with an averaged Dice score of 0.58 via training their model on synthetic data. They have also shown that the accuracy of the 3-D-Unet is higher than that of the random-walk and k-means clustering methods [36]. In a similar study, to facilitate the process of obtaining labeled data

for CNN models, a multi simple linear iterative clustering (MultiSLIC) algorithm was developed to help human users in pixelwise labeling of rock images [37].

To the best of our knowledge, **U-net architecture** was never employed for ice-core microstructure segmentation in CT images. Overall, generating ice-core structures with enough details was always a challenge for researchers in the field of glaciology due to several significant obstacles such as the following.

- 1) Complexity of the shape of bubbles and inner structures which varies with the depth of the ice core (porosity range of 0.05–0.8).
- 2) Considering the shape complexity of the microstructure of Polar ice cores, low-resolution scanning leads to the creation of many mixed pixels (i.e., partially air and partially ice).
- 3) Taking a high-resolution (30 μm) CT scan is tedious and time-consuming which makes it impossible to scan the whole ice-core column (120 m) with high resolution.
- 4) Lack of ground truth leads to rule-based or unsupervised methods which demand operator intervention all the time.
- 5) Manual thresholding, rule-based algorithms, and unsupervised methods have low accuracy while segmenting low-resolution scans.
- 6) Ice is a crystal material, so scattering increases with the increase of ice density in the specimen. This leads to higher intensity values of air pixels in dense ice. Thus, no range of intensity can define the air pixels in all specimens in different depths.
- 7) As the X-ray source is broadband (cone type), and the X-ray filament life decays, various types of other noise and artifacts might occur as well.

For the first time by utilizing unsupervised and supervised machine-learning models along with a unique CT scan system at AWI CT lab, it is possible to perform a full segmentation and study the polar ice microstructure on many various specimens. Here, we propose a combined approach of supervised and unsupervised models to compensate for the lack of ground truth. The low-resolution scans (60 μm) are considered input data for the deep-learning model. The Gaussian mixture model (GMM) receives the high-resolution (30 μm) inputs and generates the segmentation. Then, these segmentations are downsized to the same size as low-resolution scans to be used as ground truth for training the deep neural network. This approach can solve the above-mentioned issues and offers considerable improvements in ice-core segmentation. This article describes the methodology and validation of the proposed model.

II. DATASET GENERATION

For this study, three ice cores (specimens) were taken from Alfred Wegener Institute ice-core storage and transferred to the AWI-Ice-CT lab considering frozen transport protocols ($< -20^\circ\text{C}$). The lab's CT scan machine can make microscale CT images from ice cores with a max length of 1 m and diameter of 15 cm, and the machine source of X-ray is producing cone shape beams [38].

TABLE I
ICE-CORE SPECIMEN META-DATA

N	Name	Characteristic	Date	L(long:lat)	Region	d(m)	H(mm)	D(mm)	W(g)	HP(μm)	LP(μm)
2620	KohnenQK_D38_KF540	Sintered snow	12/01/2018	-0.791, -75.128	Antarctica	0.48	150	100	397.1	29.105	58.21
2848	TEDRIST_bag43	Compacted firn	01/01/2016	-8.267, -70.667	Antarctica	42.83	156	97.2	837.1	29.123	58.24
2324	ExS5-1_bag102	Bubbly ice	07/01/2022	-35.568, +75.568	Greenland	101.94	123	97.5	806.6	29.123	58.24

N: number of images L: location, d: Depth of sample, H: Height, D: Diameter, W: Weight, HP: High-resolution pixel size, LP: Low-resolution pixel size

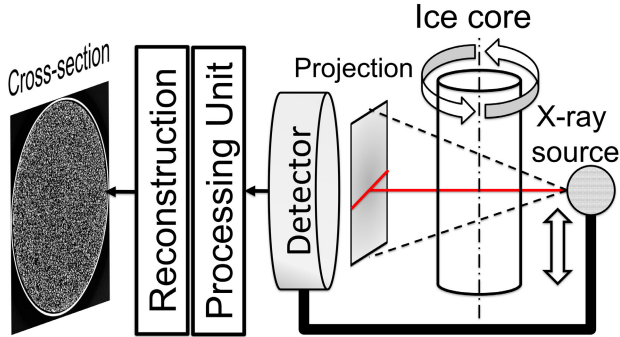


Fig. 1. Micro-CT imaging procedure.

To increase the accuracy of the measurement, the imaging method was set to helical mode. In this mode, the projections are improved by having the signals that are collected from the center-line beams which were perpendicular to the axis of rotation having the width of a pixel. This method increases unification with less distortion and artifacts. As shown in Fig. 1, the specimen is put inside a carbon fiber tube and then fixed to a gripper on the machine table. The X-ray source and the detector are synchronized and move vertically while the specimen rotates around its axis. To minimize the noise, in each imaging angle (0.09° for $30\text{-}\mu\text{m}$ resolution), the CT machine takes an image averaging over eight scans. This forms a helical movement that results in the highest machine accuracy. In the end, all the images are processed in the reconstruction stage to make cross sections of the specimen (see Fig. 2) [39].

To provide images for the current study, the specimens were scanned with $30\text{-}\mu\text{m}$ resolution resulting in cross-section images of 4096×4096 pixels. The beam power was set to 28 W with an applied voltage of 140 KV, a target current of $200 \mu\text{A}$, an exposure time of 1000 ms, and sampling with 4000 images per 360° rotation. For each specimen, the scanning time was around 35 h with another 18 h for the reconstruction process. The reconstruction was performed via nine graphic cards (NVIDIA GeForce GTX 1080 Ti) at AWI ice-core CT lab.

To train the neural network, input images and ground truth should lay on each other perfectly, and any misalignment in the dataset can ruin the training process. To have this alignment between two different resolutions, both of them were reconstructed from the high-resolution projections. To elaborate, the high-resolution images were reconstructed from the $30\text{-}\mu\text{m}$ projections directly, and the low-resolution images were reconstructed from the downsized projection (downsizing from 30 to $60 \mu\text{m}$ with linear averaging). Consequently, two sets of images with different resolutions were reconstructed that could perfectly lay on top of each other after resizing.

The three specimens (see Table I) are selected from three different depths (snow, firn, and bubbly ice) with a center depth of 0.4 m to more than 100 m, two samples from Antarctica and one sample from Greenland to make sure the final developed model is capable of performing the segmentation of different ice structures of the snow/firn column from necessary regions that are the main concerns of polar researchers (see Table I). Among samples of this study, the bubbly ice is the most sensitive one since it has the highest ice density which leads to higher scattering and artifact issues. Also, this specimen is collected from a marginal location that shows the Glacier shear stress effects on bubbles. Therefore, preserving the shape of bubbles during the segmentation is extremely important. Specimens have a circular cross section with a diameter of around 10 cm and a length of 15 cm in cylindrical form. The difference between these specimens is shown in Fig. 2, where the percentage of ice in sintered snow [see Fig. 2(a)] is less than air while this proportion changes drastically in bubbly ice [see Fig. 2(c)]. In this study, each 2-D image (cross section of the specimen) is considered individually.

The weight of each specimen (see Table I) was measured with ($\pm 0.1\text{-g}$ accuracy). The weights are used for the quantitative validation of segmentations. As we know the voxel size (resolution), ice density, and ice temperature, we can calculate the weight of specimens from the segmentation outputs, and later compare it with the actual physical measured weights.

Overall, after preprocessing and removing extra images, the final dataset consisted of 7792 2-D images (2620 snow, 2848 firn, and 2324 bubbly ice) in “.PNG” format of which 6195 were used for training and 1597 were employed for testing purposes. The test images were selected from the lower 20% of each sample (proportion according to Table II). As the testing volume is separated from the training volume, it is easier to identify overfitted models. An overfitted model will have poor performance on a test set located outside of the training volume.

III. METHODOLOGY

A. Flow of the Study

The flow of the study is shown in Fig. 3. High-resolution scanning is very time-consuming, so segmenting with no or few high-resolution images is the goal of this study. The first step is to enhance the image brightness, remove the salt pepper noise, and crop the outer ring (carbon fiber casing) to have only ice particles in the image. Next, the GMM model is used as a weak learner for making ground truth from high-resolution images ($30 \mu\text{m}$). The GMM was fit on the image pixels to classify them by considering a Gaussian distribution of intensities for both ice and air. The GMM

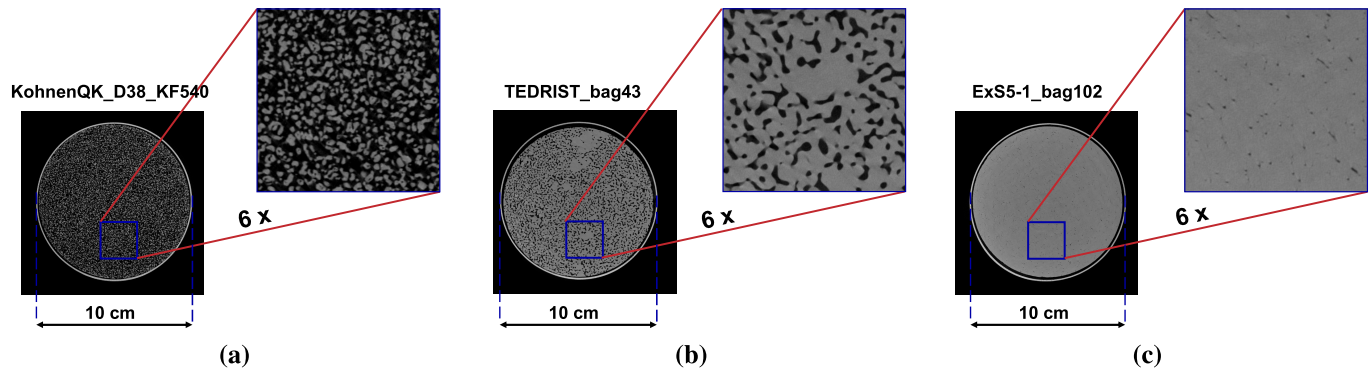


Fig. 2. Dataset sample of 2-D images from (a) sintered snow, (b) firn, and (c) bubbly ice.

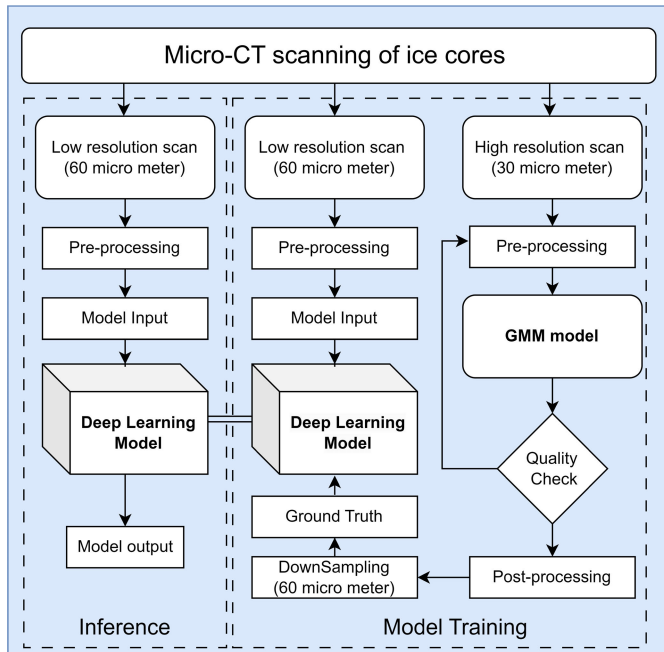


Fig. 3. Flow of the study.

output was a high-resolution binary mask. In the next step, every two binary masks were combined and downsampled (averaged) to one low-resolution mask. Next, with thresholding of the subsampled masks, a binary mask was made. Then, these low-resolution binary masks were used as ground truth. The U-net is trained using the low-resolution images ($60 \mu\text{m}$) and the downsampled binary masks produced from the GMM model as ground truth. Finally, the model weights are saved and transferred to the inference phase, and during the inference workflow, low-resolution images are given to the model for testing and utilization on later scans.

As shown in Fig. 4, the image that is resized to a lower size from high-resolution scanning is much sharper than the low-resolution scans. Low-resolution scans are closer to the CT machine images than the resized images and contain all the noises. Therefore, in this study, low-resolution scans were used as the input for training the U-net model. Consequently, the U-net model is more reliable, as it was trained with all the blurriness, noises, and other issues within the low-resolution scans.

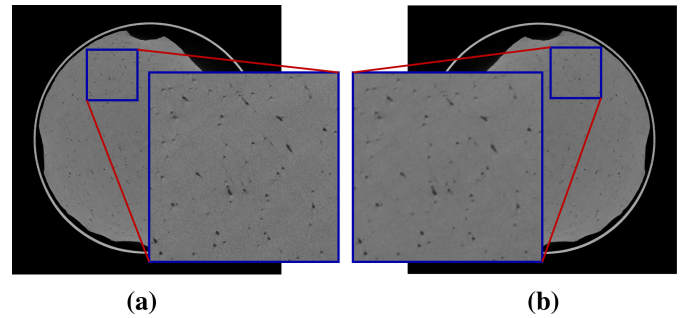


Fig. 4. (a) Resized from high-resolution and (b) low-resolution scans.

B. Producing Ground Truth

Utilizing high-resolution images to provide the ground truth and later merging high-resolution masks to build a downsampled segmentation increases the accuracy of the ground truth. To do so, automatic segmentation of high-resolution scans with nonlearning methods is needed. Among nonlearning methods, the algorithms with minimal need for user manual input (i.e., seed points, energy thresholds, and tuning parameters) are needed to have a robust segmentation pipeline.

As is common in the field of machine vision, the primary step for segmentation is trying basic image-processing techniques that might be able to segment the dataset with minimal user intervention. Thus, global thresholding, as usual, comes first for producing a binary mask. Utilizing this method demands careful investigation of the image histograms along with frequent manual checks. One method (weight-based threshold) is to determine a threshold for a specimen (e.g., 1-m ice core) in a way that the physical weight of the specimen is almost equal to the calculated weight from the generated binary mask. The disadvantage of this method is the need for operator intervention in every meter of ice. Also, the Z-direction noise (noise from one layer to another). As the ice density and image noises differ in the dataset, selecting a specific threshold for the whole specimen increases the uncertainty of segmentation, and therefore, global thresholding is not a feasible way for segmenting ice-core CT images. Also, it is not practical to provide manual thresholds for each image in the dataset. Additionally, due to the crystal shape of ice particles, there is a considerable range of scattering which changes with different ice densities. Consequently, the intensity values of air in bubbly ice are around 15% higher

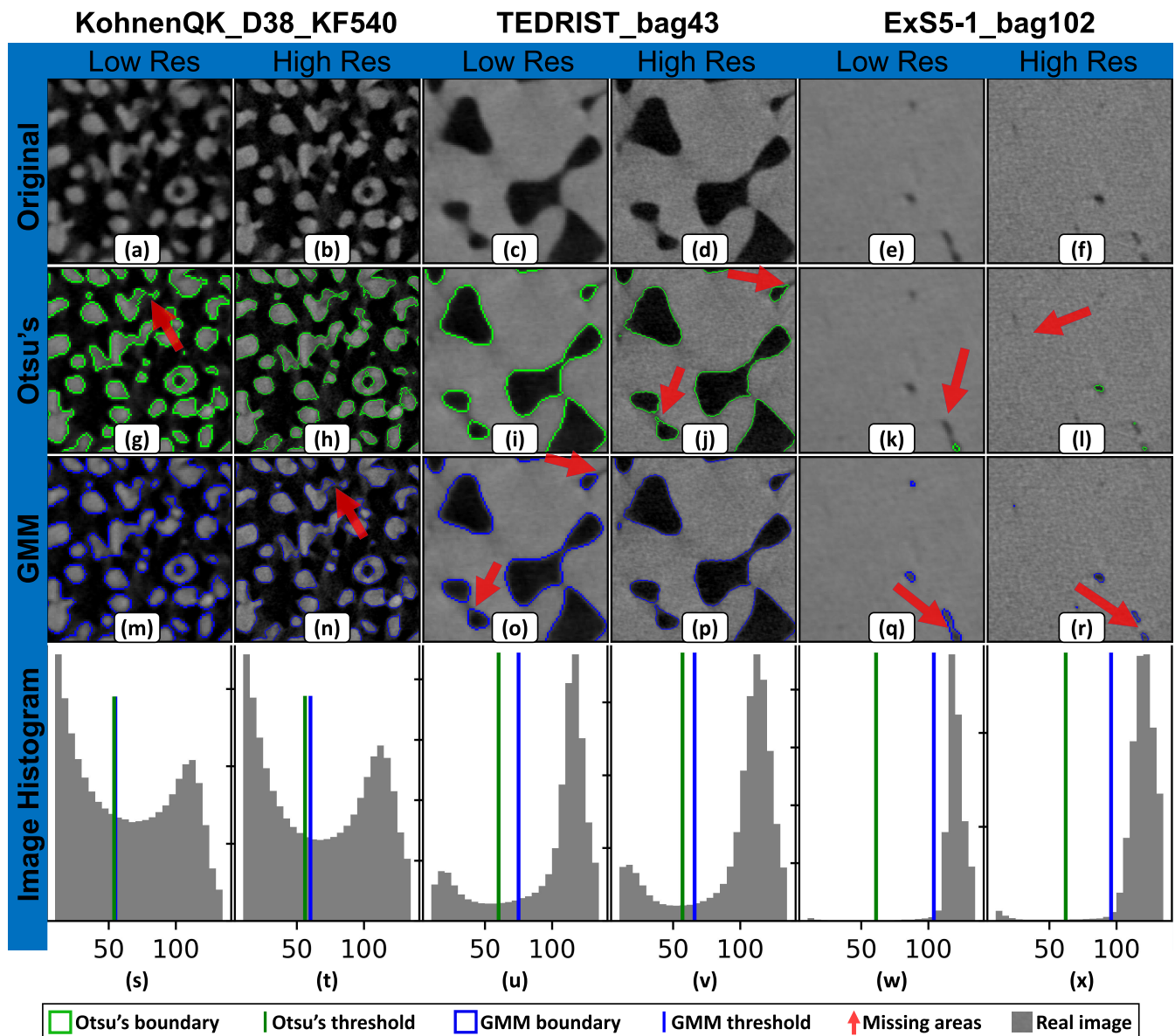


Fig. 5. Different segmentation methods on low-/high-resolution images of three specimens showing missing features and histograms. (a) Original low-res snow. (b) Original high-res snow. (c) Original low-res firn. (d) Original high-res firn. (e) Original low-res bubbly ice. (f) Original high-res bubbly ice. (g) Otsu's method on low-res snow. (h) Otsu's method on high-res snow. (i) Otsu's method on low-res firn. (j) Otsu's method on high-res firn. (k) Otsu's method on low-res bubbly ice. (l) Otsu's method on high-res bubbly ice. (m) GMM on low-res snow. (n) GMM on high-res snow. (o) GMM on low-res firn. (p) GMM on high-res firn. (q) GMM on low-res bubbly ice. (r) GMM on high-res bubbly ice. (s) histogram low-res snow. (t) histogram high-res snow. (u) Histogram low-res firn. (v) Histogram high-res firn. (w) Histogram low-res bubbly ice. (x) Histogram high-res bubbly ice.

than the air pixel intensities in snow. On the other hand, there are automated thresholding methods such as Otsu's method and GMM that can automatically threshold each image individually to tackle these issues.

Otsu's method determines an optimal threshold value from the image histogram automatically and without user intervention for each image. Although Otsu's algorithm automatically performs the segmentation, the segmentation quality is considerably low in some cases. When one cluster possesses more data points than the other one, Otsu always tends to stand in the middle of these two clusters. Thus, in high-density ice specimens, it tends to overestimate the ice (shrinking the holes) or even not be able to identify some air bubbles as shown in Fig. 5(i)–(k).

On the other hand, unsupervised methods such as the GMM demonstrate high-quality segmentation. The GMM algorithm automatically adapts to the image histogram on every image in the dataset and computes the natural distribution of the intensities of the pixels which leads to higher performance independent of the depth of the ice-core specimen and it can identify more bubbles than Otsu [see Fig. 5(p) and (q)].

As shown in Fig. 5(s)–(x) with increasing density of ice (moving from snow to bubbly ice), the shape of the histogram differs considerably, so, the thresholding method should adapt accordingly. While it is observed that Otsu's method remains always in the middle of two clusters, the GMM model can get closer to the ice cluster, for example, detect more bubbles in Fig. 5(q) and (r). It is also observed that small channels,

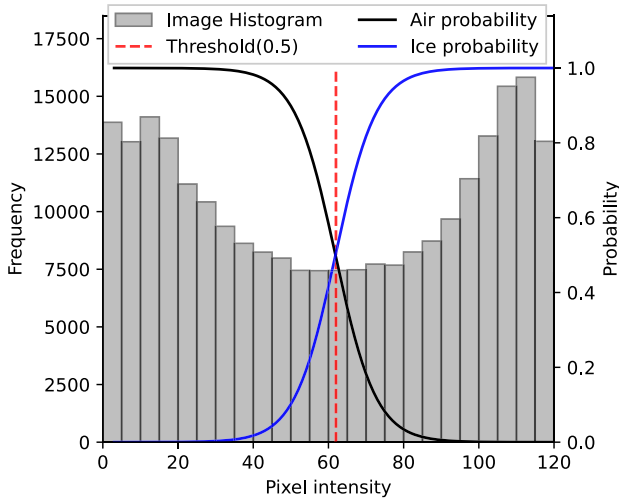


Fig. 6. GMM probability predictions on image histogram of low-resolution sintered snow.

thin bridges, and elongated bubbles are sometimes ignored in low-resolution segmentation, although they are visible with a shadowy appearance [see Fig. 5(n) and (r)].

Overall, the GMM model had better performance both quantitatively [see Fig. 5(v) and (w)] and qualitatively [see Fig. 5(k) and (q)] in comparison to other Otsu's method.

C. Machine-Learning Models

A GMM that is fit on a distribution of pixels is developing a Gaussian probability distribution for each class with an assigned weight to keep the probability between zero and one. Often the Gaussian mixture components correspond to different “types.” Here, the assumption is made, that one component is snow and another one is air. The probability distribution is shown in Fig. 6 with μ as the mean value and σ as the variance.

GMM is a probabilistic model that assumes that the observed data is generated by a mixture of several Gaussian distributions [40], [41]. The basic equation for the likelihood function of a GMM is

$$P(x|\theta) = \sum_{k=1}^K \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \quad (1)$$

where x is the observed data, k is the number of Gaussian components, θ is the vector of parameters of the distribution of observation associated with cluster k including variance and mean, π_k is the mixing coefficient of the k th distributions, μ_k is the mean of the k th component, and Σ_k is the covariance matrix of the k th component. In this study, there are two clusters ice and air, and the observations are the intensity values, that is, scalar, and therefore Σ_k is a 1×1 matrix.

The basic equations for the expectation step (E-step) and the maximization step (M-step) of the expectation-maximization (EM) algorithm for estimating the parameters of a GMM are as follows.

E-Step: This step calculates the expectation of the component C_k for each data point in x , given the model parameters are π_k , Σ_k , and μ_k with initial step

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}. \quad (2)$$

The M-step is then using the calculated expectation γ_{nk} to improve the model as follows:

$$\pi_k = \frac{1}{N} \sum_{n=1}^N \gamma_{nk} \quad (3)$$

$$\mu_k = \frac{1}{\sum_{n=1}^N \gamma_{nk}} \sum_{n=1}^N \gamma_{nk} x_n \quad (4)$$

$$\Sigma_k = \frac{1}{\sum_{n=1}^N \gamma_{nk}} \sum_{n=1}^N \gamma_{nk} (x_n - \mu_k)(x_n - \mu_k)^T. \quad (5)$$

The implementation of GMM was made with Scikit-learn1.2.0. In this study, the covariance matrix type was set to “tied” (i.e., all components share the same general covariance matrix), and K-means clustering was used for the initialization step [42].

The second model that is used in this study is a deep neural network. Deep neural networks are the most typical artificial intelligence (AI) models for processing image data. These models usually consist of convolution layers, max-pooling, upsampling, batch normalization, and skip connections. U-Net uses a series of downsampling and upsampling layers. The downsampling layers are used to extract features from the input image, while the upsampling layers are used to reconstruct the output image [43].

As shown in Fig. 7, the U-Net architecture consists of two main parts: the encoder part is a sequence of convolutional and max-pooling layers that reduce the spatial resolution of the input image. This is used to extract features from the image.

The decoder part is a sequence of convolution and upsampling layers or transposed convolution layers. It is a combination of these layers that increases the spatial resolution of the feature maps. This is used to reconstruct the output image [44].

The U-Net architecture also features skip connections that concatenate feature maps from the encoder part with the corresponding feature maps from the decoder part. This allows the network to combine high-level semantic information with low-level spatial information, which improves the performance of the network. U-Net architecture is widely used for image segmentation and generation tasks in fields of medical imaging, remote sensing, engineering, and so on [45], [46], [47].

With a heuristic approach, a U-net model was developed and tested, and finally, the following hyperparameters were considered an optimal solution. To tune the U-net architecture for the current binary segmentation task, the number of features was reduced compared to standard U-net, to decrease model complexity which led to a decrease in overfitting and computational time. The developed U-net model (see Fig. 7) is made of five stages both in the encoder and the decoder. In the encoder part, the image passes through double 2-D convolution layers with the $[3 \times 3]$ kernel, the stride of one, and padding same from 32 to 128 features. Next, the data is passed through the nonlinear activation function of the rectified linear unit (ReLU). Finally, the data is passed to a max-pooling layer to perform downsampling to reduce the input size. In the decoder part, with five steps of 2-D convolution and upsampling (scale factor of 2 and mode of nearest), the number of features is reduced to reach the input size. The final layer is a Sigmoid function that returns a binary image as the model output.

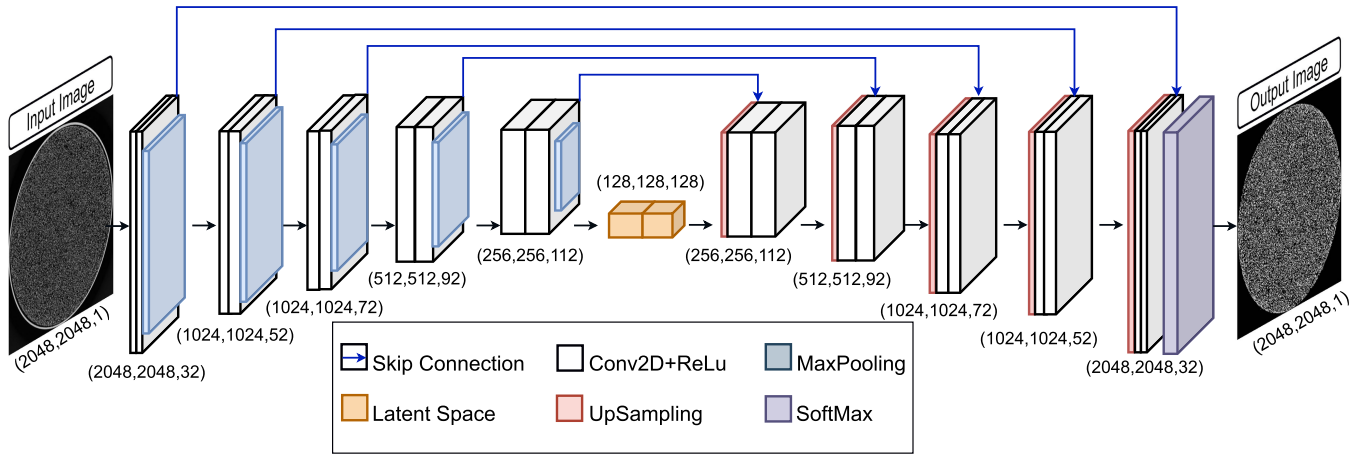


Fig. 7. U-Net architecture (shown in straight shape) with 2-D image input and output.

During the training process, the loss function was calculated via binary cross-entropy, and Adam was selected as the optimizer [48]. The model (1.5M trainable parameters) is developed with the PyTorch package and trained on GPU NVIDIA Quadro RTX 4000 (GPU Memory 8 GB) first with 30 epochs to find the best number of epochs and later was trained again with 15 epochs, and each epoch takes around 30 min.

D. Model Metrics

Segmentation models are used to divide an image into multiple segments or regions, each of which corresponds to a specific class or object. Evaluating the segmentation model performance is often done using specific metrics that are designed to measure the quality of the segmentation results. In this study, ice particles (white pixels) are considered “positive.” To evaluate the performance of the developed deep-learning model, the following metrics are utilized.

Pixel Accuracy: It is a measure of the proportion of pixels that are correctly classified in the predicted segmentation compared to the ground truth segmentation [41]

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (6)$$

where TP is true positives, the number of pixels that are correctly labeled as the target class. FP is false positives, the number of pixels that are incorrectly labeled as the target class. FN is false negatives, the number of pixels that should be labeled as the target class but are not.

The accuracy as a metric might be misleading when the proportion of pixels in each class is considerably different (e.g., many black pixels out of the interest region). The F1-score is a measure of the balance between precision and recall for a binary segmentation problem. It is the harmonic mean of precision and recall

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \quad (7)$$

The F1-score is mostly averaging the performance, and we have the inference from several specimens. Thus, for a better comparison, it needed to use a metric that emphasizes the worst predictions

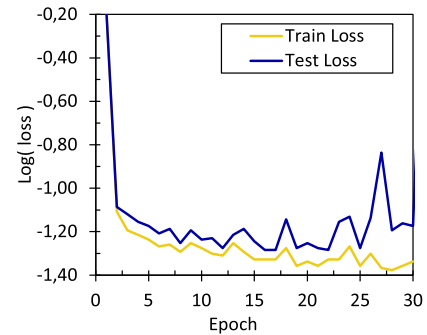


Fig. 8. Prediction error per epoch (one cycle of passing all training data to the U-net model).

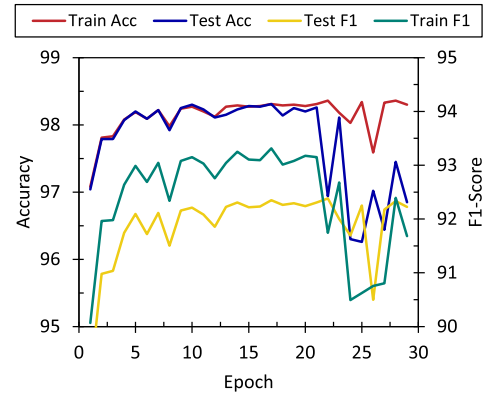


Fig. 9. Prediction accuracy and F1-score per epoch (one cycle of passing all training data to the U-net model).

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \quad (8)$$

IoU: Also known as the Jaccard index, it is a measure of the overlap between the predicted segmentation and the ground-truth segmentation. It is defined as the ratio of the intersection of the two segments to the union of the two segments [49].

IV. RESULT AND DISCUSSION

During the training process, the developed U-net model is trained until it reaches the maximum possible accuracy on the test set. Each training epoch took almost 30 min considering the above-mentioned system characteristics. As the size of the dataset is considerably large, the majority of the parameters

TABLE II
MODEL METRICS PER ICE-CORE SAMPLE

N	Name	N 2D-img	Characteristic	WE_{Gmm} (%)	WE_{DI} (%)	F1-score(%)	Accuracy(%)	IoU(%)
1	KohnenQK_D38_KF540	500	Sintered snow	+ 2.2	+5.1	95.78	98.36	91.92
2	TEDRIST_bag43	600	Compacted firm	- 1.1	-0.1	97.72	98.21	95.52
3	ExSS-1_bag102	500	Bubbly ice	+ 0.5	-0.8	83.22	97.01	81.56

N 2D-img: number of 2D images given as test set for a particular ice core, WE_{Gmm} : GMM model (ground truth) weight error, WE_{DI} : Deep learning model weight error

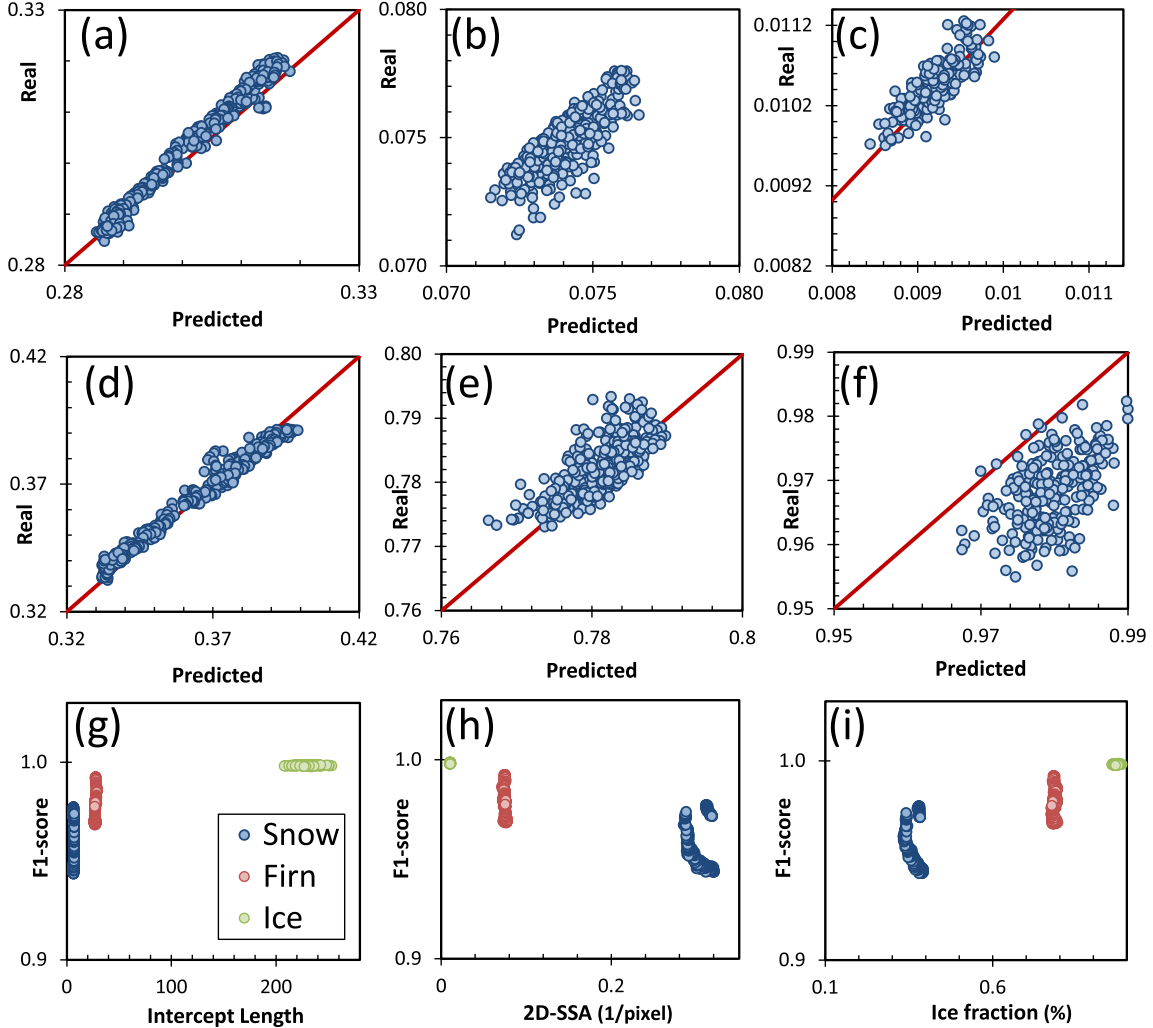


Fig. 10. Microstructure parameters. (a) 2-D-SSA snow. (b) 2-D-SSA firm. (c) 2-D-SSA ice. (d) Density snow. (e) Density firm. (f) Density ice. (g) Ground-truth intercept length. (h) Ground-truth 2-D-SSA. (i) Ground-truth density.

are tuned well during the first epoch of training. The loss that is calculated between the U-net output and the ground truth (GMM output from high-resolution images) is shown in Fig. 8. It is observed that the loss is decreased considerably after the second epoch and training over the 15th epoch pushes the test loss and train loss to diverge (overfitting issue). The training process was first set with 30 epochs (see Fig. 8), to find the minimum generalization gap and an optimal number of epochs. To avoid the overfitting issue, later on, the model was trained again with the optimal number of epochs (15 epochs).

In Fig. 9, U-net model metrics are shown per epoch for all three specimens combined. The accuracy and F1-score of the

test set are moving in the same pattern which shows that the model becomes better in determining the ice particles. The accuracy and F1-score increase until epoch 15 and they drop with training more than 20 epochs, and later on, the model was retrained with only 15 epochs.

The accuracy and F1-score of the test set do not increase considerably after 20 epochs of training, and the F1-score is mostly fluctuating around 91.5% until epoch 30.

To evaluate the model performance on each ice core separately, the model outputs are given in Table II individually. In terms of accuracy, the first core with sintered snow had the highest accuracy, but considering a high number of black

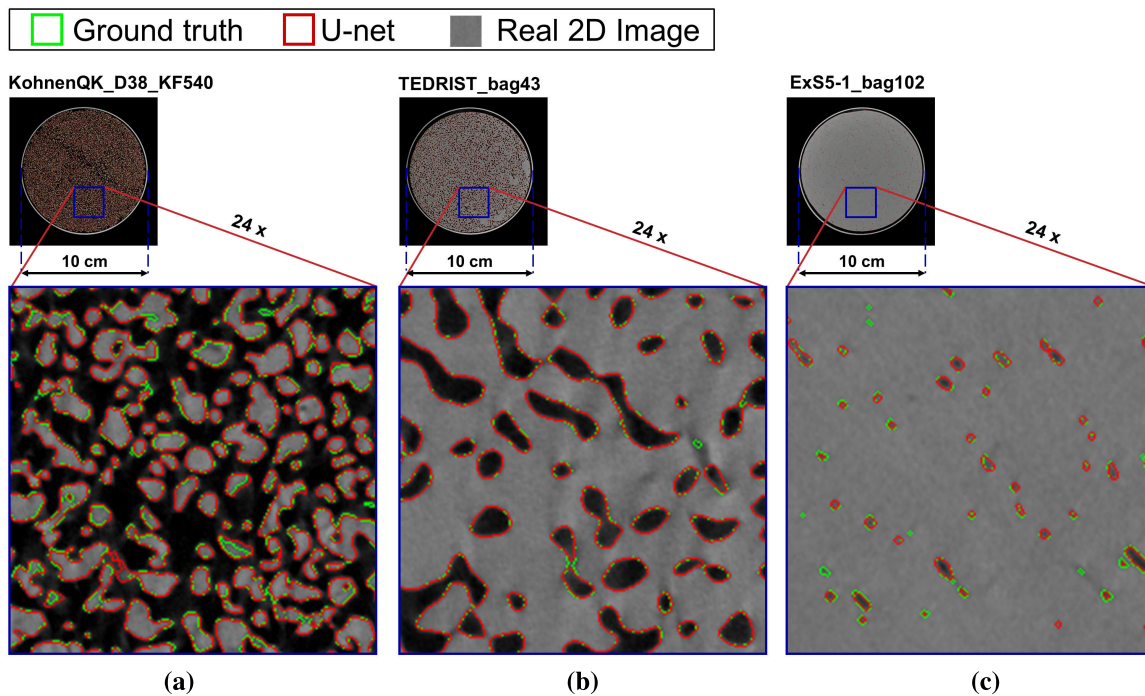


Fig. 11. Detected boundaries by U-net (red) versus the ground truth (GMM from high-resolution, green). The red boundary is drawn on top of the green one, so where no green is visible, both coincide. (a) Sintered snow. (b) Firn. (c) Bubbly ice.

pixels out of the ice-core contour, accuracy might not be a proper metric for evaluation. The F1-score considers the achievements of both classes simultaneously, thus the outer black part of the image affects this metric less. The F1-score of the snow and firn is higher than the bubbly ice, and this difference also exists in IoU. Bubbly ice has a considerably lower ice–air interface, and therefore, the weight error is less affected by segmentation issues.

The relation between several basic microstructure parameters and the U-net prediction for a few images of the test set (not including broken ice samples) is shown in Fig. 10. According to Fig. 10(a)–(f), the predicted 2-D-SSA and ice fraction are relatively accurate and the majority of the points are laying on the 45° bisector (shown with red color). Although the ice fraction is very high and the ice 2-D-SSA is very low, the predicted values are slightly biased [see Fig. 10(c) and (f)]. Due to decreasing the interface between the two clusters going from snow to ice, it is expected that the F1-score will increase and the segmentation task becomes easier by considering F1-score numbers Fig. 10(g)–(i). However, with a quality check (see Fig. 11), it is obvious that detecting small bubbles in ice becomes more difficult.

Despite having very close metrics for cores 1 and 2, the snow core has a lower percentage of IoU. This is an indication that the ground truth is considering a wider border for ice particles than the deep-learning model, which can also be related to the ground-truth weight error (+2.2%). The DL model is considering even tighter borders leading to (+5.1%) which is also visible in Fig. 11(a).

The ground truth for the deepest ice core (bubbly ice) had the lowest quality, as the amount of ice is proportionally much higher than the air and the air bubbles are very small and

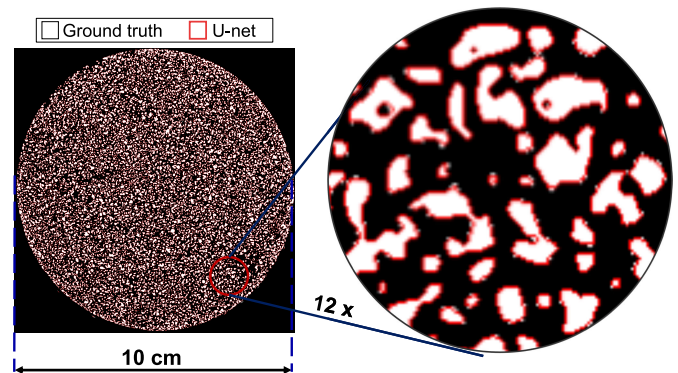


Fig. 12. Detected boundaries by U-net versus ground-truth mask (GMM on high-resolution) for sintered snow.

might be lost among the noise. Also, this specimen is collected from the high-stress areas of the glacier, and bubbles and air channels are pulled horizontally, and therefore, many air channels lose their vertical shape and have a tangential position against the image plane (CT cross-section images). Thus, the slightly low model metrics of this ice core are to be expected.

Low-resolution images (inputs of U-net) are illustrated in Fig. 11 along with the ground-truth boundary and the U-net model output boundary for three given specimens. In terms of qualitative analysis, the firn [see Fig. 11(b)] had the highest alignment between predictions and ground truth. Similarly, sintered snow [see Fig. 11(a)] was segmented with high accuracy, despite the weight error that comes with imaging uncertainty. In contrast, the performance of the model on segmenting the bubbly ice [see Fig. 11(c)] was less successful than the others, taking into account that the provided ground truth by the GMM model also had less quality than other ice cores.

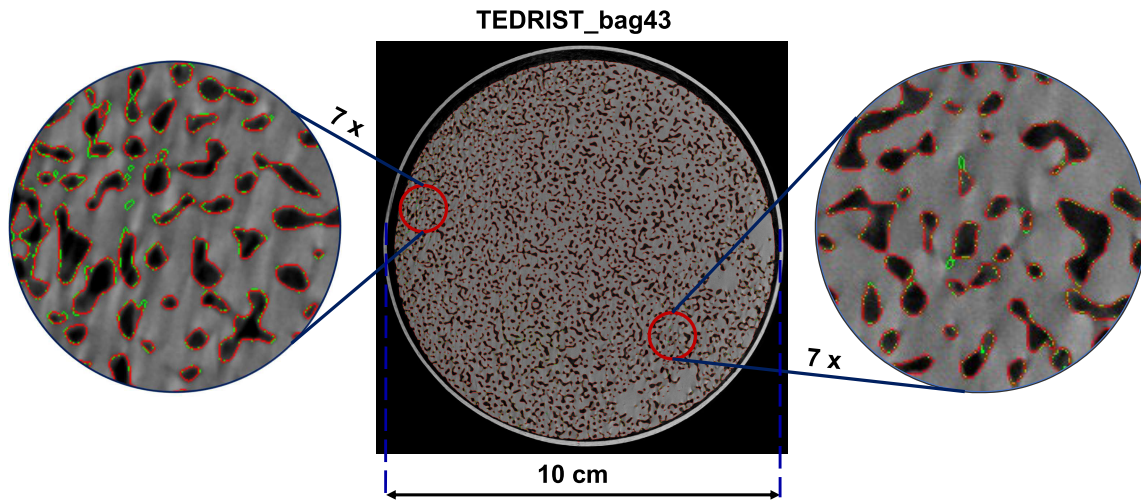


Fig. 13. U-net (red) and ground truth (green) for TEDRIST_bag43.

According to Table II, the higher weight error was produced by the model while segmenting the sintered snow (KohnenQK_D38_KF540). This error (5.5%) is visible in Fig. 12 where the ground truth is filled with white color and the U-net output is shown with red borders. It is observed that overall, the deep-learning model has tighter borders around ice particles compared to the ground truth. From Fig. 12, we also observed that some small ice particles are available in the ground truth but the deep-learning model misses them. The sintered snow has a large weight error adds a comparatively more significant uncertainty in the ground-truth segmentation which plays an important role in deep-learning model training.

On the other hand, in the second core (TEDRIST_bag43), the deep-learning model is improved both in terms of weight error and segmentation quality. The weight error was decreased from -1.1 to -0.1 as shown in Table II. Also, in this ice core, the deep-learning model is correcting some parts of the ground-truth mistakes. There are several small air bubbles in the ground truth caused by noise and shadows, which we know from the field knowledge is incorrect, and the deep-learning model is ignoring them properly (see Fig. 13).

V. CONCLUSION

Ice-core microstructure investigation begins with having an accurate segmentation of ice and air. Deep-learning algorithms trained with the help of weak learners can generalize the image patterns to have a wide range of abilities for segmentation. Therefore, utilizing the deep-learning models can bring the following advantages.

- 1) Quality of the segmentation is increased considerably, and one model can segment ice cores gathered from different depths and regions without needing user intervention.
- 2) It is not needed to scan the cores with high resolution anymore, so it saves time and effort.
- 3) The validation process specifically with the weight of the specimen is a steady approximation of the overall error of the model, and the deep-learning model demonstrated considerable improvement and versatility.

- 4) The deep-learning performance on segmenting low-resolution images is better than unsupervised methods (GMM) or algorithmic approaches such as Otsu's method.

Despite these advantages, the deep-learning model relies on the produced ground truth that makes it vulnerable in case the weak learner produces a wide range of errors. In this study, the last core (bubbly ice) ground truth had several issues, leading to lower final accuracy of the U-net model in determining some bubbles. Also, the 2-D U-net is sensitive to image misalignment, and before passing images to the U-net we need to be sure the low-resolution input image has the exact corresponding downsampled high-resolution ground truth. This prevents the current approach from reaching further resolution differences.

In future work, the study will continue. In this study, we have used a limited number of samples, and yet we have to discover the performance of the model on a larger domain of specimens, so in the future study, we shall expand the number of samples with more variations. Also, a 3-D deep-learning model will be developed to increase the spatial information given to the model to reduce errors and help the model interpret the objects (ice/air) in a 3-D format with a more significant difference in resolution. The developed models in this study and future studies are going to be used for segmenting the Polar ice cores available in the AWI archive.

DATA AND CODE AVAILABILITY

The dataset produced for this study is available on <https://doi.org/10.5281/zenodo.8325366>. Also, the code is available on the GitHub link https://github.com/Faramarz-bagherzadeh/Icecore_GMM_U-net. The raw data contains 16k images with sizes 4096×4096 and 8k images with sizes 2048×2048 . The total size of the raw dataset is around 120 GB.

REFERENCES

- [1] D. I. Benn and D. J. Evans, *Glaciers & Glaciation*. Evanston, IL, USA: Routledge, 2014.

- [2] H. Friedli, H. Löttscher, H. Oeschger, U. Siegenthaler, and B. Stauffer, "Ice core record of the 13C/12C ratio of atmospheric CO₂ in the past two centuries," *Nature*, vol. 324, no. 6094, pp. 237–238, Nov. 1986.
- [3] C. Lorius, J. Jouzel, D. Raynaud, J. Hansen, and H. L. Treut, "The ice-core record: Climate sensitivity and future greenhouse warming," *Nature*, vol. 347, no. 6289, pp. 139–145, Sep. 1990.
- [4] D. Wagenbach, P. Bohleber, and S. Preunkert, "Cold, Alpine ice bodies revisited: What may we learn from their impurity and isotope content?" *Geografiska Annaler, A, Phys. Geography*, vol. 94, no. 2, pp. 245–263, Jun. 2012.
- [5] M. Hörhold et al., "Modern temperatures in central–north Greenland warmest in past millennium," *Nature*, vol. 613, no. 7944, pp. 503–507, Jan. 2023.
- [6] K. Hoffmann-Abdi et al., "Short-term meteorological and environmental signals recorded in a firn core from a high-accumulation site on Plateau Laclavere, Antarctic Peninsula," *Geosciences*, vol. 11, no. 10, p. 428, Oct. 2021.
- [7] S. H. Faria, J. Freitag, and S. Kipfstuhl, "Polar ice structure and the integrity of ice-core paleoclimate records," *Quaternary Sci. Rev.*, vol. 29, nos. 1–2, pp. 338–351, Jan. 2010.
- [8] M. Yu et al., "Modelled variations of the inherent optical properties of summer Arctic ice and their effects on the radiation budget: A case based on ice cores from CHINARE 2008–2016," *EGU sphere*, vol. 2022, pp. 1–21, Oct. 2022.
- [9] A. R. Macfarlane et al., "Evolution of the microstructure and reflectance of the surface scattering layer on melting, level Arctic sea ice," *Elementa, Sci. Anthropocene*, vol. 11, no. 1, p. 103, 2023.
- [10] E. R. Thomas et al., "Physical properties of shallow ice cores from Antarctic and sub-Antarctic islands," *Cryosphere*, vol. 15, no. 2, pp. 1173–1186, Mar. 2021.
- [11] J. M. Fegyveresi et al., "Five millennia of surface temperatures and ice core bubble characteristics from the WAIS divide deep core, West Antarctica," *Paleoceanography*, vol. 31, no. 3, pp. 416–433, Mar. 2016.
- [12] D. E. Moser, M. Hörhold, S. Kipfstuhl, and J. Freitag, "Microstructure of snow and its link to trace elements and isotopic composition at Kohlen station, Dronning Maud land, Antarctica," *Frontiers Earth Sci.*, vol. 8, p. 23, Feb. 2020.
- [13] P. Bohleber, M. Roman, C. Barbante, S. Vascon, K. Siddiqi, and M. Pelillo, "Ice core science meets computer vision: Challenges and perspectives," *Frontiers Comput. Sci.*, vol. 3, Jun. 2021, Art. no. 690276.
- [14] R. Drews et al., "Layer disturbances and the radio-echo free zone in ice sheets," *Cryosphere*, vol. 3, no. 2, pp. 195–203, Aug. 2009.
- [15] A. Krischke, U. Oechsner, and S. Kipfstuhl, "Rapid microstructure analysis of polar ice cores: Analyzing past climates using the large area scan microscope," *Optik Photonik*, vol. 10, no. 2, pp. 32–35, Apr. 2015.
- [16] J. Eichler et al., "Location and distribution of micro-inclusions in the EDML and NEEM ice cores using optical microscopy and in situ Raman spectroscopy," *Cryosphere*, vol. 11, no. 3, pp. 1075–1090, May 2017.
- [17] P. R. F. Barnes, R. Mulvaney, E. W. Wolff, and K. Robinson, "A technique for the examination of polar ice using the scanning electron microscope," *J. Microsc.*, vol. 205, no. 2, pp. 118–124, Feb. 2002.
- [18] I. Weikusat, S. Kipfstuhl, S. H. Faria, N. Azuma, and A. Miyamoto, "Subgrain boundaries and related microstructural features in EDML (Antarctica) deep ice core," *J. Glaciol.*, vol. 55, no. 191, pp. 461–472, 2009.
- [19] P. Neubert and P. Protzel, "Compact watershed and preemptive SLIC: On improving trade-offs of superpixel segmentation algorithms," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 996–1001.
- [20] L. Grady, "Random walks for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 11, pp. 1768–1783, Nov. 2006.
- [21] Y. Yang, X. Hou, and H. Ren, "Efficient active contour model for medical image segmentation and correction based on edge and region information," *Expert Syst. Appl.*, vol. 194, May 2022, Art. no. 116436.
- [22] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.
- [23] X. Yang, X. Li, Y. Ye, R. Y. K. Lau, X. Zhang, and X. Huang, "Road detection and centerline extraction via deep recurrent convolutional neural network U-Net," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7209–7220, Sep. 2019.
- [24] N. Maffezzoli et al., "Detection of ice core particles via deep neural networks," *Cryosphere*, vol. 17, no. 2, pp. 539–565, Feb. 2023.
- [25] Z. Li, N. Sun, H. Gao, N. Qin, and Z. Li, "Adaptive subtraction based on U-Net for removing seismic multiples," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9796–9812, Nov. 2021.
- [26] Y. Chen, L. Bruzzone, L. Jiang, and Q. Sun, "ARU-Net: Reduction of atmospheric phase screen in SAR interferometry using attention-based deep residual U-Net," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5780–5793, Jul. 2021.
- [27] M. Egmont-Petersen, D. de Ridder, and H. Handels, "Image processing with neural networks—A review," *Pattern Recognit.*, vol. 35, no. 10, pp. 2279–2301, 2002.
- [28] P. Angelov and E. Soares, "Towards explainable deep neural networks (xDNN)," *Neural Netw.*, vol. 130, pp. 185–194, Oct. 2020.
- [29] Q. Zhong et al., "Filamentous target segmentation of weft micro-CT image based on U-Net," *Micron*, vol. 146, Jul. 2021, Art. no. 102923.
- [30] T. R. Ferreira, F. A. Cássaro, H. Zhou, and L. F. Pires, "X-ray computed tomography image processing & segmentation: A case study applying machine learning and deep learning-based strategies," in *X-Ray Imaging of the Soil Porous Architecture*. Cham, Switzerland: Springer, 2022, pp. 57–80.
- [31] S. Maus, M. Schneebeli, and A. Wiegmann, "An X-ray microtomographic study of the pore space, permeability and percolation threshold of Young sea ice," *Cryosphere*, vol. 15, no. 8, pp. 4047–4072, Aug. 2021.
- [32] P. Hagenmuller, M. Matzl, G. Chambon, and M. Schneebeli, "Sensitivity of snow density and specific surface area measured by microtomography to different image processing algorithms," *Cryosphere*, vol. 10, no. 3, pp. 1039–1054, May 2016.
- [33] P. Hagenmuller, G. Chambon, B. Lesaffre, F. Flin, and M. Naaim, "Energy-based binary segmentation of snow microtomographic images," *J. Glaciol.*, vol. 59, no. 217, pp. 859–873, 2013.
- [34] M. Lombardo, M. Schneebeli, and H. Löwe, "A casting method using contrast-enhanced diethylphthalate for micro-computed tomography of snow," *J. Glaciol.*, vol. 67, no. 265, pp. 847–861, Oct. 2021.
- [35] J. Malik et al., "3D quantum cuts for automatic segmentation of porous media in tomography images," *Comput. Geosci.*, vol. 159, Feb. 2022, Art. no. 105017.
- [36] S. Müller et al., "Deep learning-based segmentation of lithium-ion battery microstructures enhanced by artificially generated electrodes," *Nature Commun.*, vol. 12, no. 1, p. 6205, Oct. 2021.
- [37] J. Yu et al., "Superpixel segmentations for thin sections: Evaluation of methods to enable the generation of machine learning training data sets," *Comput. Geosci.*, vol. 170, Jan. 2023, Art. no. 105232.
- [38] J. Freitag, S. Kipfstuhl, and T. Laepple, "Core-scale radiosopic imaging: A new method reveals density–calcium link in Antarctic firn," *J. Glaciol.*, vol. 59, no. 218, pp. 1009–1014, 2013.
- [39] C. F. Schaller, J. Freitag, S. Kipfstuhl, T. Laepple, H. C. Steen-Larsen, and O. Eisen, "A representative density profile of the North Greenland snowpack," *Cryosphere*, vol. 10, no. 5, pp. 1991–2002, Sep. 2016.
- [40] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc., B, Methodolog.*, vol. 39, no. 1, pp. 1–22, Sep. 1977.
- [41] R. Szeliski, *Computer Vision: Algorithms and Applications*. Cham, Switzerland: Springer, 2022.
- [42] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2012.
- [43] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
- [44] R. Gupta, A. K. Rana, S. Dhawan, and K. Cengiz, *Advanced Sensing in Image Processing and IoT*. Boca Raton, FL, USA: CRC Press, 2022.
- [45] Y. Chen, L. Li, X. Liu, and X. Su, "A multi-task framework for infrared small target detection and segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5003109.
- [46] B. Li, J. Gao, S. Chen, S. Lim, and H. Jiang, "POI detection of high-rise buildings using remote sensing images: A semantic segmentation method based on multitask attention Res-U-Net," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4706916.
- [47] T. Falk et al., "U-Net: Deep learning for cell counting, detection, and morphometry," *Nature Methods*, vol. 16, no. 1, pp. 67–70, Jan. 2019.
- [48] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
- [49] A. Thakur, *Approaching (Almost) any Machine Learning Problem*. India: Pothi, 2020.