Recognition of Polar Lows in Sentinel-1 SAR Images With Deep Learning

Jakob Grahn and Filippo Maria Bianchi

Abstract—In this article, we explore the possibility of detecting polar lows in C-band synthetic aperture radar (SAR) images by means of deep learning. Specifically, we introduce a novel dataset consisting of Sentinel-1 images divided into two classes, representing the presence and absence of a maritime mesocyclone, respectively. The dataset is constructed using the ECMWF reanalysis version 5 (ERA5) dataset as baseline and it consists of 2004 annotated images. To our knowledge, this is the first dataset of its kind to be publicly released. The dataset is used to train a deep learning model to classify the labeled images. Evaluated on an independent test set, the model yields an F1 score of 0.95, indicating that polar lows can be consistently detected from SAR images. Interpretability techniques applied to the deep learning model reveal that atmospheric fronts and cyclonic eyes are key features in the classification. Moreover, experimental results show that the model is accurate even if: 1) such features are significantly cropped due to the limited swath width of the SAR; 2) the features are partly covered by sea ice; and 3) land is covering significant parts of the images. By evaluating the model performance on multiple input image resolutions (pixel sizes of 500 m, 1 km, and 2 km), it is found that higher resolution yield the best performance. This emphasizes the potential of using high-resolution sensors like SAR for detecting polar lows, as compared to conventionally used sensors such as scatterometers.

Index Terms—Deep learning, mesocyclones, polar lows, synthetic aperture radar (SAR).

I. INTRODUCTION

POLAR lows belong to the class of mesoscale maritime cyclones (from now on referred to as mesocyclones) that form at high latitudes, typically due to cold air outbreaks from sea ice or snow covered regions [1]. They are characterized by rapid development, small scale, strong winds, and heavy snowfall. This makes them both difficult to predict and extremely hazardous for maritime activities such as fishing, shipping, petroleum extraction, and offshore wind power production. When making landfall, polar lows are prone to disrupt land and air traffic, destroy infrastructure, and trigger high snow avalanche activity in mountainous regions.

Due to their unpredictable and destructive nature, reliable and precise methods for early detection and tracking of polar lows are desirable. Meteorologists and scientists largely rely on direct observations in terms of satellite imagery or numerical weather prediction (NWP) models constrained by observations for detecting polar lows [2], [3], [4], [5], [6], [7], [8], [9]. In maritime and polar regions, observations almost exclusively originate from satellites. Conventionally, data from scatterometers, radiometers and optical sensors are assimilated into the NWP models [10], [11]. However, these sensors either rely on sunlight or have a coarse spatial resolution (typically a few to tens of kilometers). Considering that polar lows often occur during the polar night and are small scaled, featuring wind streaks, sharp atmospheric fronts, and precipitation cells, observations at higher resolution regardless of light conditions could be beneficial.

Synthetic aperture radars (SARs) are independent of solar illumination and provide imagery at very high spatial resolution (typically a few to tens of meters). Researchers have already indicated that SAR data adds value to polar low monitoring [12], [13], [14]. Assimilation of SAR data into NWP models is however challenging, since the exact relationships between radar measurement and geophysical parameters are not trivial, especially at high wind speeds [14], [15], [16]. An alternative approach to make use of synthetic aperture radar (SAR) data is to rely on data-driven techniques, such as deep learning.

Deep learning has successfully been applied to several remote sensing applications and achieved state-of-the-art results [17], [18], [19], [20]. Cyclone type phenomena specifically, has been considered in assimilated data [21], [22], [23], passive microwave data [24], thermal infra-red (IR) data [25], [26], [27], [28], [29] and scatterometer data [30]. With the exception of [31], deep learning has, however, been largely overlooked for detecting mesocyclones in SAR data.

This article investigates the possibility of using deep learning for detecting mesocyclones in general, and polar lows in particular, in SAR images. We aim to answer two main questions.

1) Can a deep learning model recognize polar lows in SAR images?
2) What significance does the image resolution have on the performance?
To answer these questions, we first show that a training dataset can be constructed from the Sentinel-1 data archive, which is large enough for a deep neural network to be trained. In order to make the dataset large enough, we relax the definition of a polar low to the broader class of mesocyclones. The constructed dataset contains image samples divided in two classes, representing the presence and the absence of mesocyclones, respectively. In the following, we explain in detail how the dataset is built. To our knowledge, it is the first of its kind to be publicly released.

Then, we show how a deep neural network can be trained on the dataset to perform binary classification with very good performance. The deep learning model and the training procedure is carefully motivated by considering the training dataset size, input image size, and class imbalance. The performance of the model is evaluated for multiple input image resolutions and interpretability techniques are applied on the model to evaluate what image features are most relevant for the classification.

II. SAR DATASET

This section describes the construction of the dataset for classifying mesocyclones, observed by the Sentinel-1 satellites. The dataset is publicly available (https://doi.org/10.18710/FV5T9U) and consists of 2004 images divided in two classes: the positive class (318 images with mesocyclones) and the negative class (1686 images without mesocyclones).

A. Positive Class: Mesocyclone Present

To build the positive class, polar lows monitored by the Sentinel-1 satellites were required. Historic catalogs of polar lows exist [32], [33], based on manual analysis of NWP model data as well as satellite data (thermal IR, passive microwave, and scatterometer data). However, these catalogs are regional and, more importantly, do not cover the time period when the Sentinel-1 satellites were operational. On the other hand, studies like [2], [3], and [34] proposed objective criteria based on meteorological parameters that produce results similar to the manually annotated catalogs. Such objective criteria can be applied on reanalysis data, enabling identification of candidate low pressures that were coincident with the Sentinel-1 satellites.

Although a variety of objective criteria have been proposed, they are typically associated with either:
1) the low-pressure intensity;
2) the presence of a cold air outbreak; or
3) the location of the low pressure in relation to the polar front.

In [2], a combination of such criteria were imposed on the ECMWF reanalysis Interim (ERA-I) dataset and the most effective criteria for detecting polar lows were identified using the manual catalog by Noer et al. [32] as reference. However, events meeting all criteria are infrequent, since polar lows are rare. For reference, in [32], only 12 events per year were recorded over the Nordic seas on average, from year 2000 to 2009. Moreover, considering the limited spatio-temporal coverage of the Sentinel-1 satellites, not all events are imaged, making the number of image candidates even lower. Therefore, to include as many observed events as possible in our dataset, the cold air outbreak and locality type criteria were neglected. By considering only an intensity criteria, mesocyclones that are not necessarily driven by baroclinic instabilities or located in the polar air masses were included. Assuming that such mesocyclones share substantial similarities to polar lows, they can still provide valuable information to train a deep learning model, which motivates their inclusion in the dataset.

The intensity criteria was imposed on the ECMWF reanalysis version 5 (ERA5) dataset. Specifically, it was formulated in terms of the depression in the sea level pressure (SLP) relative to the local mean. This type of criteria was considered by Stoll et al. [2], where different SLP depression thresholds were tested. In our study, the threshold was set at 230 Pa and the local mean was computed within a 9 × 9 grid cell neighborhood (corresponding to 270 km × 270 km at the equator). The spatio-temporal distributions of resulting candidates and the subsequent matched SAR observations are shown in Fig. 1. The highest concentrations of candidates were found in the subtropical regions of the North Pacific and North Atlantic. However, due to higher satellite revisit frequencies at higher latitudes, most SAR observations were found in the extra tropical and polar parts of the North Atlantic.

![Fig. 1. (a) Spatial distribution of ERA5 candidates and Sentinel-1 matches. (b) Corresponding temporal distribution, as counts per month. Background map: © OpenStreetMap contributors/CARTO.](image-url)
3) merging time-adjacent products to a common grid in SAR geometry;
4) multilooking to 500 m resolution in range and azimuth\(^2\);
5) geocoding to a 400 km × 400 km grid (centered at the AOI) in a UTM coordinate system with a 500 m grid spacing;
6) generating red-green-blue (RGB) color composites.

The RGB color composites were generated by first rescaling the radar cross section (assumed in decibel scale) to a value \( x \in [0, 1] \). Specifically, the second and 98th percentiles of each separate image and polarization channel were rescaled to 0 and 1, respectively.\(^3\)

For data with dual polarization channels, the rescaled values were used to make RGB color composites as
\[
R = G = \frac{x_{||} + x_x}{2}, \quad B = x_{||}
\]
where \( x_{||} \) and \( x_x \) corresponds to the co- and cross-polarized\(^4\) channels, respectively. For single polarization data, containing only the co-polarized channel, the color channels were defined as: \( R = G = B = x_{||} \). Both dual and single polarization data were thus considered jointly in the training dataset,\(^5\) however, the dual polarization data constituted the great majority of the samples (see Fig. 6).

3) Manual Validation and Offset Correction: Each RGB color composite was manually validated. Specifically, in each positive image, we asserted the presence of distinctive features (typically an eye or a comma-shaped pattern). In general, these features were not centered in the processed images, since the image grid was centered at the candidate AOI originating from ERA5. Therefore, offsets were corrected for by manually centering the AOIs on the eye or comma-shaped pattern. The samples were then reprocessed with the refined AOIs.

B. Negative Class: Mesocyclone Absent

To obtain samples of the negative class, representing the absence of a cyclone, we considered repeat-pass SAR acquisitions successive to those of the positive image samples (i.e., images acquired at the same relative orbit). The motivation of our choice was twofold.

1) Sentinel-1 repeat-passes are separated by at least six days, which is enough time for the sea state (and thus the image features) to decorrelate.
2) The imaging geometry of repeat-pass acquisitions is nearly identical, such that static/background features appear similar.

The second point is important in order to factor out land features from the dataset. Indeed, if the same land features

\[^{2}\]In terms of number of looks, EW mode products are in total multilooked by 60 × 20 looks in range and azimuth, while interferometric wide-swath (IW) mode products are multilooked by 250 × 50 looks in range and azimuth. Speckle noise is thus significantly suppressed in the processed images.

\[^{3}\]The second percentile was clipped to the range −25 to −15 dB and the 98th percentile was clipped to the range −10 to 0 dB. The clipping values were chosen to harmonize the scaling across image samples. Pixels without data were excluded when computing the percentiles and replaced by zeros.

\[^{4}\]The co-polarized channel can be either HH or VV, and the corresponding cross-polarized channel can be either HV or VH.

\[^{5}\]A dedicated experiment using only the co- or cross-polarized channel separately, can be found in the supplementary material.
Fig. 3. (a) Repeat-pass set is shown, consisting of one positive (central image) and eight negative samples. All samples within a set are processed onto the exact same grid, centered at the low pressure in the positive sample. The polarization channels for this set are HH/HV. (b) Location of the individual products and the grid is displayed. Background map: © OpenStreetMap contributors/CARTO.

As an example, a repeat-pass image set consisting of one positive and eight negative samples is shown in Fig. 3. To the left, the processed RGB composites are shown. The south tip of Svalbard can be seen statically in all images, while ocean features appear dynamically. The positive sample in the center, contains a distinct vortex structure. To the right, a map with the footprints of the individual Sentinel-1 products involved is shown, together with the footprint of the image grid. Typically, due to the limited swath of the SAR, the products do not cover the whole image grid across track, leading to missing data in the RGBs in the cross-track direction. Occasionally, some products are not captured, leading to missing data in the along-track direction as well.

Fig. 4. Distribution of SLP depression for the two classes. The depression is measured as the SLP averaged over the whole image, minus the SLP averaged over the center 100 × 100 pixels.

appear in both the positive and negative class, it is expected that a machine learning model will be able to ignore them in the classification task.

The distribution of SLP depression, defined as the difference in SLP (extracted from ERA5) between the image wide average and the average of the center 100 × 100 pixels, is shown in Fig. 4. The positive samples have a strong depression, while the negative samples exhibit a symmetric distribution centered around 0 Pa. This indicate that the negative class, indeed, represents a sea state not biased toward a centered low pressure. It should however be emphasized that no particular features are excluded from the negative class.

In total, 1686 negative sample were generated from the 318 positive ones, resulting in a total of 2004 samples in the dataset. The number of negatives per positive varied depending on the existence of repeat-pass acquisitions, as shown in Fig. 6. Furthermore, most samples were acquired in the extended wide-swath (EW) mode with two polarisations.

Fig. 5. Overview of the repeat-pass sets, in terms of distributions in size (samples per set), imaging mode, and polarization mode. For each positive sample, we extracted a maximum of ten negative samples.
III. Deep Learning

Three immediate challenges can be identified when choosing and training an appropriate deep learning model to perform classification on the dataset:
1) the input image size is relatively large;
2) the training dataset is relatively small;
3) the classes are imbalanced.

In the following, we discuss how these were dealt with.

A. Deep Learning Architecture

One of the major benefits of using SAR data, compared to, e.g., scatterometer or passive microwave data, is the high image resolution. Although the images in the training dataset were already heavily downsampled from the original resolution of 10–40 m to 500 m, resulting in an image size of 800 × 800 pixels, they are relatively large in the context of many popular deep learning models. These are often designed for images of size 256 × 256 pixels or lower. To preserve details that are specific for the SAR data, such as wind streaks, rain cells, or sharp atmospheric fronts, and to enable us to study the added value of high input image resolution, we wish to avoid further downsampling and rather let the model handle the high input image resolution. Convolutional neural networks (CNNs) used for image classification usually consist of a stack of convolutional layers followed by pooling. Each such processing block sequentially increases the feature dimensionality through convolutions, while reducing spatial resolution through pooling. As such, relevant spatial information will gradually become embedded in the feature space. If the input image is large, the model must either apply an aggressive downsampling in each processing block, or include many blocks and, thus, become very deep. The former can be obtained by large stride in the convolutional and pooling layers, or by using Atrous convolutions [35]. These techniques do, however, come at a cost of discarding spatial information, which we wish to avoid. This leaves us with the option of using a deep architecture, which gradually distill the spatial information and embeds it into the feature space.

Training a very deep network poses two fundamental challenges. Firstly, the gradients of the loss used to update the parameters may gradually vanish as they are backpropagated through the network. Secondly, an architecture with many layers contains many trainable parameters. This makes the model prone to overfitting, unless the training set is exceptionally large, which was not the case in our study.

A solution to address the first problem is to use residual connections, popularized by architectures such as ResNet [36], which facilitate the flow of the gradients during the backpropagation.

Considering the second problem, a ResNet is unfortunately characterized by many trainable parameters. There are, however, more recent deep architectures which include residual connections but have fewer parameters. For example, MobileNet [37] and Xception [38] implement separable 2D convolutions (Sep2DConv), which allows to greatly reduce the number of trainable parameters,

Therefore, we opted for a customized Xception architecture, whose details are depicted in Fig. 7. The entry block consists of a convolutional layer, followed by a batch normalization layer and a ReLU activation function. There are L residual blocks, each one including Sep2DConv layers, batch normalization, ReLU activations, and a max-pooling layer. The max-pooling output is combined with the input of the residual block through a skip connection. The convolutional layer in the middle of the skip connection has no activation function and simply applies a kernel of size 1 with stride 2, to match the shape of the input with the one of the output. A global pooling layer reduces the feature map generated by the last residual layer to a single vector, which is processed by the final classifier consisting of a dropout layer [40], a dense layer, and a softmax activation.

B. Data Augmentation

Augmenting the training data by applying random transformations is a common technique used to prevent overfitting. By exposing the deep learning model to perturbations of original inputs, it is possible to improve the robustness of the model. In addition, data augmentation allows to get rid of some bias in the dataset and increase the generalization performance on new unseen data.

Our dataset has been designed by keeping image augmentation in mind. Each low pressure is centered in the image and has a wide area around that can be partially cropped. Each...
time a batch of images is fetched to our deep learning model, the following random transformations are applied on the fly:

1) horizontal and vertical translation (between 0% and 10% of the image size);
2) horizontal and vertical flip;
3) rotation (0°–40°);
4) zoom (−10% to 10% of the original scale);
5) cropping to the center 512 × 512 pixels.

If after the transformation some points fall outside the boundaries of the original input image, these are filled with zeros. Notably, after data augmentation the low pressures are no longer centered in each image. Fig. 8 shows an example of augmented images randomly generated during training.

C. Class Imbalance

While the number of positive samples were restricted by the number of matches found on the Sentinel-1 archive, multiple negative samples could be generated for each positive sample. This led to a natural skewness in the distribution of classes in the dataset: 84% of the samples belong to the negative class and 16% to the positive. We tested and compared three different approaches to train the deep learning model in the presence of class imbalance.

1) Class-Weighting: We reweighted the loss function according to class frequencies. Denoting the number of samples in the negative and positive classes, respectively, by $n_0$ and $n_1$, the loss for samples in the corresponding classes were weighted by $w_0 = (1/n_0)(n_0 + n_1/2)$ and $w_1 = (1/n_1)(n_0 + n_1/2)$. This means that each error on the positive class affects the optimization of the model weights to a greater extent.

2) Oversampling: We ensured that in each batch there were always the same amount of samples of both classes. Specifically, this was obtained by designating half of the batch to the negative class, and half to the positive. In each epoch, samples from the negative class are seen only once, while samples from the positive class are repeated. We made sure to observe all samples of the negative class at least once during an epoch and to not sample any image twice within a batch.

3) Rejection Sampling: This strategy drops samples until a balanced distribution across the two classes is obtained. Contrarily to oversampling, each sample is seen at most once in each epoch, which makes the overall training faster.

By empirical comparison (see the supplementary material), we found that oversampling yields the best performance and, therefore, was the strategy adopted in our experiments.

D. Hyperparameter Tuning

To find the optimal configuration of the deep learning model, we searched several hyperparameters and selected those giving the best performance on a validation set. As validation set, we used 10% of the training set. The hyperparameter space and the optimal values found after the optimization procedure are reported in Table I. To reduce the hyperparameters space, we only search the number of filters of the first residual blocks and then we double the number in the following blocks.

Since the dataset contains large images and we consider deep models with many parameters, evaluating each hyperparameter configuration is computationally expensive. Therefore, rather than performing an exhaustive search with grid search or evaluating a large number of configurations with a random search, we opted for a more efficient approach. In particular, we used Bayesian hyperparameter optimization [41].

We used a batch size of 16 and the Adam optimizer [42]. During the hyperparameter tuning, we trained the model for 50 epochs. After finding the optimal configuration, we trained the final model for 200 epochs.

![Fig. 7. Examples of random image augmentation. (Top left) Original. The polarization channels for the image are VV/VH.](image-url)
E. Model Interpretability

Due to the presence of many nonlinear transformations, it is difficult to interpret the decision process of a neural network and considerable research effort has been devoted to improve our understandings. Gradient-based approaches try to find which inputs have the most influence on the model scoring function for a given class [43], [44], [45]. This is usually done by taking the gradient of the class activation score with respect to each input feature [46]. A drawback of gradient-based methods is that they give zero contribution to inputs that saturate the ReLU or MaxPool. To capture such shortcomings, a formal notion of explainability was introduced in [47] with the axiom of conservation of total relevance, which states that the sum of relevance of all pixels must match the class score of the model. Specifically, the authors propose to distribute the total relevance of the class score to the input features with layer-wise relevance propagation (LRP). While LRP follows the conservation axiom, it does not specify how to distribute the relevance among the input features. To address this problem DeepLiFT [48] enforces an additional axiom on how to propagate the relevance by following the chain rule.

In this work, we adopt two recent interpretability techniques, that address some of the shortcomings discussed above and are able to provide valuable insights into the decision problem of our model.

1) Integrated Gradients (IGs): IGs [49] has become a popular interpretability technique since it can be applied to any neural network model, is easy to implement, and theoretically grounded. IGs aims to satisfy two additional axioms that are not jointly ensured by other existing attribution schemes:

1) if the input and an uninformative baseline differ in exactly one feature, such a feature should be given nonzero attribution;
2) when two models are functionally equivalent, they must have identical attributions to input features.

Denoting the model scoring function $F$, the attributions given by integrated gradients (IG) are

$$\text{IG}(x) := (x - x') \cdot \int_{a=0}^{1} \frac{\partial F(x' + a \cdot (x - x'))}{\partial x} da$$

(1)

where $x$ is a sample in the dataset, $x'$ is the uninformative baseline, and $\alpha$ is an interpolation constant used to perturb the input features.

In our study, we let $x'$ be a black image (all zeros) as the uninformative baseline. As empirically confirmed in our experiments, such a baseline is classified with high confidence to be negative. Let $X'$ be the set of interpolated images from $x'$ to $x$. The computation of the integral in (1) is approximated with the sum of the partial derivatives of the images in $X'$. Fig. 8 depicts a small interpolation set $X'$ from the mean-baseline to a positive sample and shows how the classification score changes. By summing the gradients ($\partial F(X')/\partial x_i$) one quantifies the relationship between the changes in the input features and the changes in the predictions of the model $F$.

2) Gradient-Weighted Class Activation Mapping: While IG can be used on any neural network model, gradient-weighted class activation mapping (Grad-CAM) is specific for CNNs. It uses the gradients of a given target class flowing into the final convolutional layer to produce a coarse localization map, which highlights the important regions in the image for predicting the class [50]. We summarize at a high level the main steps of the algorithmic procedure and we refer the interested reader to [50] for further details:

1) take a trained model and cut it at the $k$th layer, which is the layer for which we want to create a Grad-CAM heatmap (usually, the activation after the last convolutional layer);
2) feed an input image to the complete model and collect the total loss and the output of layer $k$;
3) compute the gradients of the output of layer $k$ with respect to the loss;
4) take parts of the gradient which contribute to the prediction and use to build a heatmap;
5) resize the heatmap so that it can be overlaid to the original image.

IV. RESULTS

The dataset presented in Section II was used to train the model described in Section III. Specifically, the dataset was partitioned such that 79% of the samples were used for training and validation and 21% for testing. The partitioning was done by randomly assigning complete repeat-pass sets to either the test or the training set. In such a way, positive and negative samples with the same land features cannot appear both in the training and test set. This:

1) encouraged the model to factor land features out as irrelevant to the classification task;
2) allowed us to evaluate the generalization capability of the model by testing on new locations, unseen during training.

The arguably most attractive property of SAR data, as compared to, e.g., scatterometer data, is the high spatial resolution. In order to evaluate the added value of higher spatial resolution, the model accuracy was examined for three different input image resolutions9; 500, 1000, and 2000 m (the latter two obtained by bi-linear down sampling of the first).

9The highest resolution here (500 m) is still considerably lower than the original resolution of the SAR images. However, as discussed in Section III-A, the input image size is limited by the depth of the network architecture in relation to the size of the training dataset. Therefore, we did not considered even higher input image resolutions, even if the original data allowed for it.
Hyperparameter tuning was performed independently for each resolution (see the supplementary material for details), and the classification performance on the test set is shown in Table II. The table displays the mean and standard deviation of true negatives (TNs), false negatives (FNs), false positives (FPs), true positives (TPs) and $F_1$ score obtained from ten independent runs. It is clear that higher image resolution significantly improves the classification results. In fact, for the highest input resolution, the model is misclassifying on average less than eight samples (as FN or FP) out of the 435 samples in the test set, with a mean $F_1$ score of 0.94 (in the supplementary materials, results of the performance for the highest input image resolution using the co- or cross-polarized channels separately are also presented).

A model trained on the 500 m resolution images was further examined using the IG and Grad-CAM techniques presented in Section III-E. The performance of this specific model is shown in Table III and the images it classifies as TPs, FPs, and FNs

---

**TABLE II**

| Pixel size | TN       | FN       | FP       | TP       | $F_1$ score |
|------------|----------|----------|----------|----------|-------------|
| 2 km       | 346.6±1.1 | 6.4±1.1  | 9.8±1.9  | 54.2±1.9 | 0.87±0.01   |
| 1 km       | 364.4±2.1 | 6.6±2.1  | 7.4±2.1  | 56.6±2.1 | 0.89±0.02   |
| 500 m      | 367.8±2.7 | 3.2±2.7  | 4.6±1.7  | 59.4±1.7 | 0.94±0.04   |

**TABLE III**

| Pixel size | TN | FN | FP  | TP  | $F_1$ score |
|------------|----|----|-----|-----|-------------|
| 366        | 2  | 5  | 62  | 0.95 |             |

---

10A detailed comparison based on Grad-CAM between the model trained on 2000 and 500 m resolution is presented in the supplementary materials.
are discussed in the following. The deep learning model used in our experiments and the code to apply the interpretability techniques is available online.\(^\text{11}\)

### A. True Positives

Of the 62 TP samples (i.e., low pressures correctly classified as low pressures), four samples are displayed in Fig. 9. The first column shows the input RGB color composites, the second column shows the IG in green, and the third column shows the Grad-CAM as a heat map. Three out of these four samples are located in polar regions, while the sample on the second row is an extra-tropical cyclone observed off the coast of Japan. The IG and Grad-CAM overlays indicate that the model is focusing on the cyclonic eye features. The IG overlay has a slightly higher emphasis on the wind fronts as compared to the Grad-CAM. Both the IG and Grad-CAM indicate that the model is effectively disregarding land features as well as the sea ice features appearing in the top row. Notably, in the top row, a large part of the cyclonic eye feature is also cropped due to the limited swath width of the SAR. This is the case in multiple samples classified as TP, indicating a certain robustness to image features being cropped or obscured by, e.g., sea ice.

### B. False Positives

The model classified five samples as FP (i.e., absence of low pressures incorrectly classified as low pressures), of which four are shown in Fig. 10. The top two samples are presumably difficult to classify correctly (or the ground truth label could potentially be wrong), as they actually contain some pronounced wind fronts. Considering the IG and Grad-CAM, indeed the model is focusing on these wind features. The sample on the third row also contains a pronounced wind front that the model is focusing on, but the front is not curved.

\(^{11}\)https://github.com/FilippoMB/Recognition-of-polar-lows-in-Sentinel-1-SAR-images-with-deep-learning
Fig. 11. FNs (all samples shown), with classification scores 0.96 and 0.79. The top image is acquired with the HH polarization channel and the bottom image in the HH/HV polarized channels.

The classification score is however only 0.57 for this sample. In the fourth sample, no wind front is visible, but the IG and Grad-CAM reveal that the model focuses on a wind wake (formed behind the Izu peninsula, Japan, located in the image center), which may be misinterpreted as a cyclonic eye.

Finally, we notice that IG and Grad-CAM highlight different areas in the second and third image. Explainability techniques for deep learning are tools meant for diagnostic, which require a certain degree of subjective interpretation. Each technique is based on specific heuristics, which put a bias on what features are considered relevant. Indeed, even for samples classified with high confidence two explainability techniques might focus on different input features [51]. The discrepancy is often exacerbated in samples classified with lower confidence.

C. False Negatives

Only two samples of the positive class were incorrectly classified as negatives, i.e., mistaken as absence of low pressure while being labeled as low pressures. These are shown in Fig. 11. Here, IG are not computed, since a black image cannot be used as a baseline for the negative class. Nevertheless, Grad-CAM can still be computed and is shown in the second column. It can be noted that both images suffer from lacking data due to the limited swath width of the SAR acquisitions. Indeed, the Grad-CAM indicate that the model is not focusing on the darker center features as was the case for the TP samples in Fig. 9. It should however be emphasized that this happens for only two of the 368 negative samples in the test set.

V. CONCLUSION

In this study, we show that SAR images from the Sentinel-1 satellites provide an attractive data source for automatic and accurate detection of maritime mesocyclones, such as polar lows. Specifically, we show that sufficiently many image examples can be found to build a labeled dataset for a deep learning model to be trained. By further comparing our deep learning model when trained on different input image resolutions, we find that higher resolution yields significantly better performance. This highlights the added value of using SAR data, as compared to conventionally used sensors of lower resolution. In particular, at 500-m resolution, we get an F1 score of 0.94, as compared to 0.87 at 2-km resolution (comparable to modern scatterometers).

It should further be noted that the highest resolution tested in this study (500 m) is primarily limited by the size of the training dataset and not the native resolution of the SAR sensor (10–40 m). Thus, even higher input image resolutions could in principle be considered, potentially with even better performance. Larger input image sizes, however, ideally require deeper neural network architectures, with more trainable weights. This in turn require larger training datasets to avoid over-fitting. Even so, with an increasing amount of SAR data being available from new satellites, larger training datasets could be constructed in the future, enabling even better performance.

By design, the training dataset contains spiral-form low pressures in the positive class. By analyzing IG and Grad-CAM on the trained model, we verify that the spiral-shaped atmospheric fronts and the low wind centers yield most of the class attribution. Moreover, we conclude that:

1) these characteristic wind features do not need to be fully covered in the images, but can be substantially cropped due to the limited swath width of the SAR;
2) wind features can be partly covered by sea ice and still be identified by the model;
3) the model is able to ignore land features in the images.

The last point can be verified thanks to the procedure used to obtain the negative samples, i.e., through repeat-pass acquisitions (see Section II-B).

In summary, we conclude that the application of deep learning on SAR images for recognizing maritime mesocyclones is promising. Further evaluation and comparison to detection based on data from other sensors or NWP models is encouraged as a future work direction.

ACKNOWLEDGMENT

The authors would like to thank Patrick Stoll for his valuable feedback. They thank those involved in developing the Generic DAta Raster (GDAR) software used to process synthetic aperture radar (SAR) data, especially Heidi Hindberg, Yngvar Larsen, and Tom Grydeland. They also thank Temesgen Gebrie Yitayew and Hannah Vickers for their help in establishing this project. Finally, they would like to thank the reviewers for their insightful comments.

REFERENCES

[1] E. A. Rasmussen, “Polar lows,” in A Half Century of Progress in Meteorology: A Tribute to Richard Reed. Berlin, Germany: Springer, 2003, pp. 61–78.
2. P. J. Stoll, R. G. Graversen, G. Noer, and K. Hodges, “An objective global climatology of polar lows based on reanalysis data,” Quart. J. Roy. Meteorol. Soc., vol. 144, no. 716, pp. 2099–2117, 2018.

3. W. Yanase et al., “Climatology of polar lows over the sea of Japan using the JRA-55 reanalysis,” J. Climate, vol. 29, no. 2, pp. 419–437, Jan. 2016.

4. A.-M. Blechschmidt, “A 2-year climatology of polar low events over the Nordic seas from satellite remote sensing,” Geophys. Res. Lett., vol. 35, no. 9, pp. 1–5, 2008.

5. K. Wilhelmsen, “Climatological study of gale-producing polar lows near Norway,” Tellus A, Dyn. Meteorol. Oceanogr., vol. 37, no. 5, pp. 451–459, 1985.

6. G. Zappa, L. Shaffrey, and K. Hodges, “Can polar lows be objectively identified and tracked in ERA-MPI operational analysis and the ERA-interim reanalysis?” Monthly Weather Rev., vol. 142, no. 8, pp. 2596–2608, 2014.

7. C. Michel, A. Terpstra, and T. Spengler, “Polar mesoscale cyclone climatology for the Nordic seas based on ERA-interim,” J. Climate, vol. 31, no. 6, pp. 2511–2532, 2018.

8. L. Xia, M. Zahn, K. Hodges, F. Feser, and H. Storch, “A comparison of two identification and tracking methods for polar lows,” Tellus A, Dyn. Meteorol. Oceanogr., vol. 64, no. 1, p. 17196, 2012.

9. P. J. Stoll, T. Spengler, A. Terpstra, and R. G. Graversen, “Polar lows—Moist-baroclinic cyclones developing in four different vertical wind shear environments,” Weather Climate Dyn., vol. 2, no. 1, pp. 19–36, 2021.

10. M. Pulit et al., “AROME-MetCoOp: A Nordic convective-scale operational weather prediction model,” Weather Forecasting, vol. 32, no. 2, pp. 609–627, 2017.

11. H. Hersbach et al., “The ERA5 global reanalysis,” Quart. J. Roy. Meteorol. Soc., vol. 146, no. 730, pp. 1999–2049, Jun. 2020.

12. G. W. K. Moore and P. W. Vachon, “A polar low over the Labrador sea: Interactions with topography and an upper-level potential vorticity anomaly, and an observation by RADARSAT-1 SAR,” Geophys. Res. Lett., vol. 29, no. 16, p. 1773, Aug. 2002.

13. B. R. Furevik, H. Schyberg, G. Noer, F. Tveten, and J. Røhrs, “ASAR and ASCAT polar low situations,” J. Atmos. Ocean. Technol., vol. 32, no. 4, pp. 783–792, 2015.

14. M. Tollinger, R. Graversen, and H. Johnsen, “High-resolution polar low winds obtained from unsupervised SAR wind retrieval,” Remote Sens., vol. 13, no. 22, p. 4655, 2021.

15. B. Chapron, H. Johnsen, and R. Garelo, “Wave and wind retrieval from SAR images of the ocean,” Annales Des Telecommun., vol. 56, no. 11, pp. 682–699, Nov. 2001.

16. A. A. Mouche et al., “On the use of Doppler shift for sea surface wind retrieval from SAR,” IEEE Trans. Geosci. Remote Sens., vol. 50, no. 7, pp. 2901–2909, Jul. 2012.

17. X. X. Zhu et al., “Deep learning in remote sensing: A comprehensive review and list of resources,” IEEE Geosci. Remote Sens. Mag., vol. 5, no. 4, pp. 8–36, Dec. 2018.

18. F. M. Bianchi, M. M. Espeseth, and N. Borche, “Large-scale detection and categorization of oil spills from SAR images with deep learning,” Remote Sens., vol. 12, no. 14, p. 2260, Jul. 2020.

19. F. M. Bianchi, J. Grahn, M. Eckerstorfer, E. Malnes, and H. Vickers, “Snow avalanche segmentation in SAR images with fully convolutional neural networks,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 75–82, 2021.

20. L. T. Luppino et al., “Deep image translation with an affinity-based change prior for unsupervised multimodal change detection,” IEEE Trans. Geosci. Remote Sens., vol. 60, pp. 1–22, 2022.

21. Y. Liu et al., “Application of deep convolutional neural networks for detecting extreme weather in climate datasets,” 2016, arXiv:1605.01156.

22. D. Matsuoka, M. Nakano, D. Sugiyama, and S. Uchida, “Deep learning approach for detecting tropical cyclones and their precursors in the simulation by a cloud-resolution global nonhydrostatic atmospheric model,” Prog. Earth Planet. Sci., vol. 5, no. 1, pp. 1–16, Dec. 2018.

23. S. Giffard-Roisin, R. Yang, G. Charpiat, C. K. Bonfanti, B. Kégl, and C. Monteleoni, “Tropical cyclone track forecasting using fused deep learning from aligned reanalysis data,” Frontiers Big Data, vol. 3, p. 1, Jan. 2020.

24. A. Wimmers, C. Velden, and J. H. Cossuth, “Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery,” Mon. Weather Rev., vol. 147, no. 6, pp. 2261–2282, 2019.

25. P. Gulubkin, J. Sinnotra, and L. Bloybey, “Satellite-derived spatio-temporal distribution and parameters of North Atlantic polar lows for 2015–2017,” Atmosphere, vol. 12, no. 2, p. 224, Feb. 2021.
[50] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 618–626.

[51] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, vol. 11700. Berlin, Germany: Springer, 2019.

[52] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. 3rd Int. Conf. Learn. Represent., Y. Bengio and Y. LeCun, Eds. San Diego, CA, USA, 2015, pp. 1–14.

[53] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[54] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. 9th Int. Conf. Learn. Represent., Virtual Event, Austria, 2021, pp. 1–22. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[55] I. Tolstikhin et al., “MLP-Mixer: An all-MLP architecture for vision,” in Proc. 35th Conf. Neural Inf. Process. Syst., vol. 34, 2021, pp. 24261–24272.

[56] G. Zhang, X. Li, W. Perrie, P. A. Hwang, B. Zhang, and X. Yang, “A hurricane wind speed retrieval model for C-band RADARSAT-2 cross-polarization ScanSAR images,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 8, pp. 4766–4774, Aug. 2017.

Jakob Grahn received the M.Sc. degree in radio and space science from the Chalmers University of Technology, Gothenburg, Sweden, in 2012, and the Ph.D. degree in physics from the University of Tromsø, Tromsø, Norway, in 2018. He is currently a Researcher with the Norwegian Research Center (NORCE), Tromsø, focusing on the use of radar remote sensing for various detection and prediction applications.

Filippo Maria Bianchi is currently an Associate Professor with the Department of Mathematics and Statistics, UiT The Arctic University of Norway, Tromsø, Norway. He also holds a position at the Norwegian Research Centre (NORCE), Tromsø. His research interests lie at the intersection between machine learning, dynamical systems, and complex networks. The main areas where he applies his research are energy analytics and remote sensing.