# Reducing Manual Workload in SAR-Based Oil Spill Detection Through Uncertainty-Aware Deep Learning

Anonymous Full Paper
Submission 26

## Abstract

Constant monitoring of the oceans is required to detect oil spills and reduce environmental damage associated with spills. Synthetic Aperture Radar (SAR) imaging is a critical tool for oil spill detection, but is complex and requires time-consuming manual labor for analysis. Deep learning has shown encouraging performance in automatic classification of oil spills on these images, but the performance is not yet sufficient for a deep learning classifier to act autonomously, making manual assessment essential. However, if only a reduced subset of uncertain samples had to be analyzed by human experts while the remaining samples could be automatically classified, it could greatly reduce the manual workload. In this study, we investigate if uncertainty estimates can identify which samples should be prioritized for manual inspection. Specifically, we propose a pipeline of defining a user-specified error tolerance and identifying an uncertainty threshold that filters out samples for automatic/manual thresholding. We evaluate the proposed pipeline on challenging real-world data. The results show that our proposed uncertainty-based ranking technique can reduce the manual workload by 41%, paving the way for new and more efficient ways to detect marine oil spills.

## 1 Introduction

Marine oil spills are common, with several thousand spills occurring each year in the United States [1]. It can have major environmental impact due to the damage on the marine ecosystem and the wildlife at both the sea and shore [1]. Therefore, the ocean needs to be monitored to minimize the damage by removing them quickly after release. There are multiple ways of large-scale monitoring of the ocean, but Synthetic Aperture Radar (SAR) imaging is often a preferred choice [2]. SAR is independent of daylight and cloud coverage while covering large areas, which is a big advantage compared to other options such as optical imagery [2].

Despite all the advantages of SAR imaging, the resulting images are complex and cover large areas, making the manual image analysis time-consuming. Automatic systems based on deep learning models have shown encouraging performance for this specific application [2, 3] and could potentially reduce the manual workload. Even though the DL-based systems achieve high accuracy [3–5], it is challenging to put them into operational use-cases because of the harsh consequences of a false negative. However, the vast majority of images are still correctly classified. Therefore, if there was a process that could reliably identify samples that required human evaluation and let the remaining samples be automatically classified, it could significantly reduce the manual workload. The bottom row of Figure 1 illustrates how such a pipeline could be constructed.

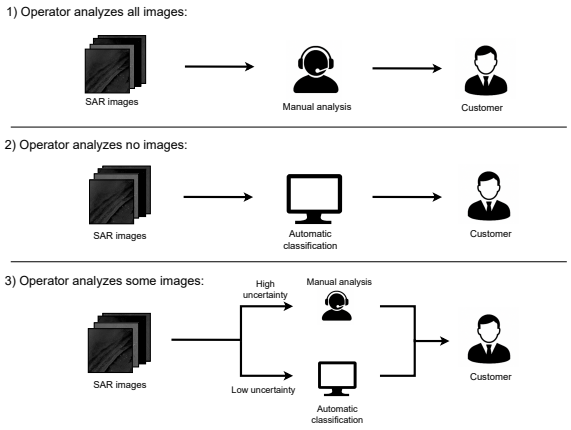


**Figure 1.** Three scenarios for SAR-based oil spill detection: (1) human analysis with low error rate but high workload, (2) automatic analysis with high error rate but low workload, and (3) combined human and machine analysis for low error and low workload. Images in figure are artificially enhanced SAR images. © Copernicus Sentinel data, processed by KSAT.

In this work, we propose an uncertainty-guided selection process for identifying SAR images of the ocean with potential oil spills detected by a deep learning model that must be processed by human evaluators to ensure sufficient performance. It is well-known that relying on the softmax output of deep learning models is ill-advised due to their high degree of overconfidence [6], and that more sophisticated uncertainty estimation techniques can more accurately identify samples that are likely to be misclassified [7]. The key idea in our proposed selection process is to automatically identify an uncertainty threshold that ensures a certain performance, where all samples above the threshold are sent for human evaluation. This approach allows for utilizing all

the benefits of DL models without having to achieve perfect performance, while minimizing the risk of false negatives. Additionally, the automated process has the added benefit of minimizing inconsistent analysis due to human error in simple cases. Our contributions are:

- A new procedure in automated deep learning-based classification for uncertainty-based selection of samples for human evaluation.
- Identification of domain-relevant augmentation strategies to allow for uncertainty estimation at test-time.
- An in-depth analysis of the proposed pipeline on real-world challenging data. Our results show that using uncertainty-based selection for human evaluation can significantly reduce the manual workload of operators.

## 2 Related Work

Automatic detection of oil spills from SAR images has been extensively studied [2]. These approaches often revolve around sophisticated feature extraction techniques in combination with classical classification algorithms [8, 9]. More recently, such methods have been outperformed by deep learning-based approaches [2]. Bianchi et al. [3] proposed a deep learning model based on convolutional neural networks with encouraging performance. More recently, Trujillo-Acatitla, Tuxpan-Vargas, et al. [10] also demonstrated the high potential for deep learning-based oil spill detection but across a wider range of deep learning architectures. While all of these works demonstrate that automatic systems have great potential for alleviating the manual workload associated with oil spill detection, we are not aware of any works that have considered how uncertainty estimation could be practically integrated into the oil spill detection pipeline.

## 3 Reducing Manual Workload With Uncertainty-Filtering

Here, we present our proposed approach for filtering out samples that requires human evaluation.

### 3.1 Uncertainty-Filtering With Test-Time Augmentation

In this work, we focus on test-time augmentation (TTA) [11] to model uncertainty in the prediction $\hat{y}$ of a deep learning model $f$. This choice is motivated by the flexibility of TTA, as it requires no modifications to the model like Monte Carlo Dropout [12] or SWAG [13], nor does it require storing multiple models like ensemble approaches [14]. This is highly beneficial in industrial applications, where a working model might already be in place and it is undesirable to alter the existing pipeline solely for the uncertainty estimation. TTA works by generating augmented views of an input and aggregating predictions across all augmented views, and has demonstrated impressive performance across a wide range of applications [11, 15]. This can be mathematically described as taking an input $\mathbf{x}$ and transforming it using a stochastic augmentation procedure $\mathcal{T}$ that produces augmented versions $\tilde{\mathbf{x}}$. Assuming $M$ augmentations are generated, a set of $M$ predictions $\{\hat{y}_1, \cdots, \hat{y}_M\}$ are made. The uncertainty associated with $f$'s prediction of $\mathbf{x}$ is calculated as:

$$\sigma^{(tta)} = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\hat{y}_m - \bar{y})}, \qquad (1)$$

where $\bar{y}$ is the mean of the $M$ predictions.

**Choosing augmentations in TTA** A key aspect of TTA is choosing a suitable data augmentation procedure that fits the data and task at hand. A common choice is to apply dropout [16] on the input to generate augmented samples [11], which is motivated by its flexibility, speed, and good performance in many cases [11]. However, in the case of oil spill detection from SAR images, we hypothesize that dropout is not the most effective choice for augmentation. Oil spills typically appear as dark areas in SAR images, and dropping these pixels would do little to alter their appearance, which would therefore induce less informative uncertainty estimates. Also, a high dropout rate could create areas that appear similar to oil spills. Therefore, we instead propose to use augmentations that are more suitable for the particular task and data at hand.

The pixel-value shift (PVS) method shifts each pixel based on the average pixel intensity in the training set. PVS has demonstrated encouraging results for hyperspectral imaging [17], which is often attributed to its ability of generating in-distribution samples. Mathematically, PVS generates samples by

$$\mathbf{x}^{(pvs)} = \mathbf{x} \pm \gamma \cdot \boldsymbol{\mu}^{(tr)}, \qquad (2)$$

where $\boldsymbol{\mu}^{(tr)}$ is the average pixel intensity estimated across the training set and $\gamma$ is a coefficient that controls the strength of the shift. This coefficient is randomly sampled for each generated sample to induce different shifts for the same sample, and the sampling procedure for $\gamma$ is an important hyperparameters for this type of augmentation.

Elastic transformations are transformations that alter the geometry of an image, which is accomplished by generating displacement vectors for all

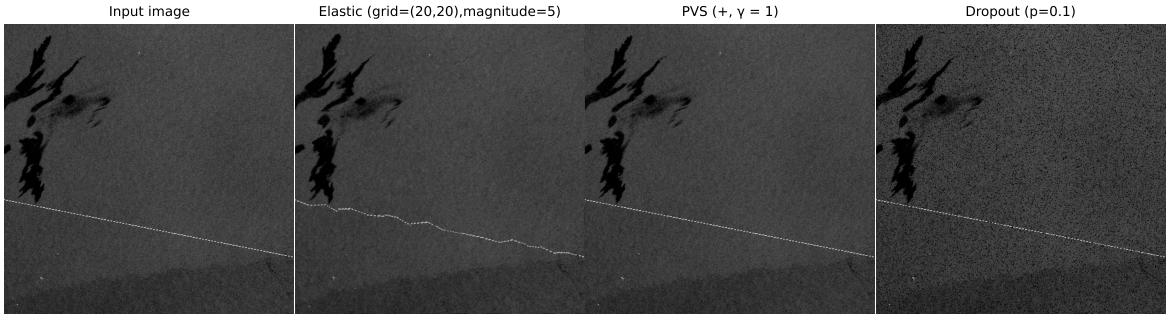


**Figure 2.** An illustration of three types of data augmentation when applied to a SAR image with oil spills.

pixels based on random offsets. The displacement vectors are added to the image through an identity grid that performs the transformation. Our choice of the elastic transformation in the context of oil spills is motivated by the fact that oil spills often exhibit a curved shape, and the elastic transformation generates variation of such curved shapes in a realistic looking manner.

Figure 2 shows each of the augmentations applied to a SAR image containing oil spills. The elastic transformation slightly distorts the content of the image, which is apparent both on the oil spills in the top left corner and for the white line (interference from ground radar) that crosses the image. The difference induced by PVS is more difficult to see, since the change is on the pixel values. As for dropout, the image appears more noisy with more dark spots across the entire image.

## 3.2 Uncertainty-Sorting Procedure

Our core idea is to send all samples that exceed a threshold of uncertainty to manual analysis, assuming that the human analysis will result in only correct predictions. The threshold can be adjusted in such a way that the overall system can obtain a desirable trade-off between the error rate and the reduction in manual workload. First, we assume an independent validation set $X^{(val)} = \{\mathbf{x}_i\}_{i=1}^{N_{val}}$ with $N_{val}$ samples, and that this validation set is sorted from least to most uncertain according to a corresponding set of uncertainty estimates $U^{(val)} = \{\sigma_i\}_{i=1}^{N_{val}}$.

Next, we assume a function that measures the error rate of the classifier for a given set of samples, in this case $\text{Er} = 1 - \text{accuracy}$. As part of the procedure, a user specified error rate $e$ must be provided that indicates an acceptable error for the classifier $f$. The complete procedure is described in Listing 1, and the output of the algorithm is the uncertainty threshold $\tau_e$ that can be used to sort future samples into either automatic or human evaluation. A key assumption here is human evaluators correctly classifies all samples they are provided.

**Listing 1.** Python-like pseudocode for proposed uncertainty-sorting procedure

```
# X_val - Sorted validation set
# U_val - Sorted uncertainties
# f     - Classifier
# Er    - Error rate function
# e     - User specified error rate
# tau   - Uncertainty threshold

while Er(f, X_val) >= e:

    X_val.pop()

tau = U_val[len(X_val)]

return tau
```

For the process outlined in Listing 1 to be useful, it is critical that the uncertainty estimates $\sigma_i$ are informative and highlight samples where there is a high likelihood of error. There are numerous ways to estimate uncertainty, and below we describe three ways to estimate the uncertainties in $U^{(val)}$.

**Probability-based ranking** While often criticized for being overconfident [6], the softmax probabilities of deep learning-based classification models can be used as confidence scores to filter out uncertain predictions, where scores closer to 0.5 are the most uncertain. Since the probabilities for oil being present or not can be confident in both ends (close to 1 or close to 0), we make the following transformation such that the confidence scores can be ranked:

$$\sigma_i^{(p)} = |0.5 - \hat{y}_i|, \tag{3}$$

where the $p$ indicates that the uncertainty comes from the output probabilities of the model.

**Uncertainty-based filtering** The standard deviation of the softmax output over the TTA samples described in Equation 1 can be used as an uncertainty estimation to identify uncertain predictions. The samples are ranked such that the images with the highest standard deviation are considered the most uncertain.

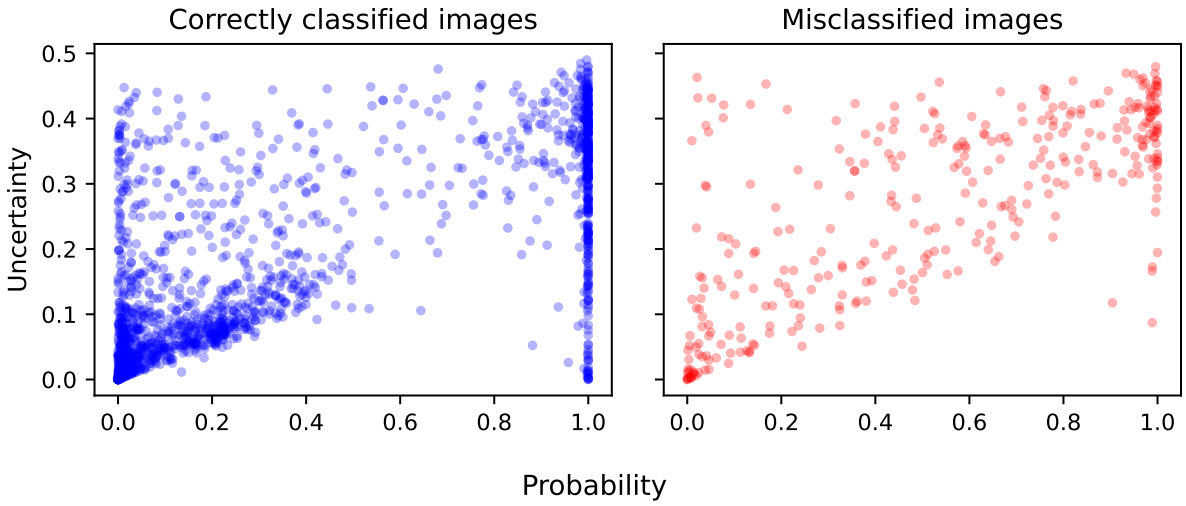


**Figure 3.** The standard deviation from the uncertainty estimation using PVS ($+/-, \gamma < 1$) and the sigmoid values from the network for the testing set. Image on the left is for the correctly classified images and the misclassified images are on the right.

**Probability and uncertainty-based filtering** An alternative approach is to combine both the output probabilities and the uncertainty estimates to perform the filtering. Figure 3 shows the probabilities plotted versus the uncertainty estimates for correctly and incorrectly classified samples. A key observation here is that many misclassified samples have a probability of approximately 1, which means they will be among the last to be filtered out for human analysis. However, when taking uncertainty into account, many of the misclassified samples with an output of almost 1 also have high uncertainty, which means that they would be picked up in the filtering process. The complementary information in the probabilities and uncertainties could therefore provide complementary information for the ranking.

# 4 Experimental Setup

Here, we describe the data used in this study, the hyperparameters for each augmentation scheme, the model used to perform the oil spill detection, and how the model was trained.

**Data** The dataset consists of 313 SAR images taken from the Satellite Sentinel-1A. The images were preprocessed by downsampling to a $480 \times 480$ pixels corresponding to a resolution of 60m and cropped into 10317 patches of which 6960 patches fell into a chosen area of interest which were used for this study. The images were segmented into 7 classes: (i) Background: All pixels not explicitly as one of the following labels, present in 763 samples; (ii) Undefined: Low confidence oil spill, present in 1223 samples; (iii) Possible spill: Medium confidence oil spills, present in 497 samples; (iv) Probable spill: High confidence oil spills, present in 497 samples; (v) Seep: Oil seeping naturally from reservoirs at known seep locations, present in 65 samples; (vi) Produced water: Liquid byproduct of oil production, contain some oil as well as wastewater, present in 30 samples; (vii) Ignore: Boundaries of all oil spills and also missing data. These segmentation labels were compiled into binary image-level labels based on the occurrence of any of the oil spill classes (ii)-(vi).

The dataset is divided into 56% images for training (3926 images), 30% for testing (2050 images), and 14% for validation (984 images). The division is provided by Kongsberg Satellite Services (KSAT) and based on the acquisition days to ensure no data leakage between the splits as some images overlap and there are potentially several images over the same area each day.

**Data augmentation hyperparameters** For dropout, we investigate a range of hyperparameter ranging from a low dropout rate (0.01) to a high dropout rate (0.5). For PVS, we consider sampling $\gamma$ in Equation 2 from either $U(-a, b)$, $U(0, b)$, $U(-a, 0)$, with $a = b = 1$ or $a = b = 2$. Due to the computational load of the elastic transformation, we qualitatively identified a set of hyperparameters that induced some distortion and used those hyperparameters throughout all experiments. We use a grid size of $20 \times 20$ and a distortion magnitude of 5.

**Model and training** A ResNet-50 model [18] was trained for binary classification for 50 epochs using the Adam optimizer [19], binary cross-entropy loss, learning rate of 0.00001, weight decay of 0.0001 and batch size 16. As a stopping criteria, the highest AUC (area-under-the-curve) for the validation set

**Table 1.** Percentage of testing dataset needed to be manually analyzed to get a human error rate of 5%. Bold numbers indicate improved performance compared to the regular model.

| Method | Sigmoid | Uncertainty (std) | Sigmoid + Uncertainty (std) |
|---|---|---|---|
| Regular model | 51 ± 10 % | - | - |
| PVS (+/-, $\gamma < 1$) | **42** ± 1 % | **45** ± 4 % | **42** ± 1 % |
| Dropout (p=0.01) | 54 ± 11 % | **50** ± 9 % | 54 ± 11 % |
| Elastic transform | **44** ± 4 % | **41** ± 2 % | **45** ± 4 % |

is used. The network is initialized with pretrained weights using contrastive learning for Sentinel-1 and Sentinel-2 data [20, 21]. The weights are obtained from TorchGeo [22].

# 5 Results

Here, we present the evaluation of the proposed uncertainty-filtering approach to reduce manual workload in classification of oil spills in SAR images. We first evaluate the performance, before we present the quantitative evaluation for the proposed uncertainty-filtering procedure. Afterwards, we present an investigation of the effect of oil spill type and size on the proecdure before an in-depth analysis of the hyperparameters used in TTA.

**Model performance** A ResNet-50 is trained and classification performance is evaluated based on the experimental setup described in Section 4. The performance of the model was evaluated w.r.t. accuracy, F1 score, and AUC score. On the independent test set, the model achieves an accuracy of 83.6%, an AUC of 87.8%, and an F1 score of 75.9%. This model forms the basis for the following experiments in this section.

**Uncertainty-filtering reduces workload** We evaluate the probability-based, uncertainty-based, and the combined approach for uncertainty filtering with dropout, PVS, and elastic transform as the data augmentation. Table 1 shows the results from the best performing setup across all hyperparameters settings (see Table 2 for evaluation of hyperparameters), with the acceptable error rate set to 5 %. First, note that using the sigmoid output for the probability-based filtering ("regular model"-row) can already provide a reduction in the workload of the human evaluator. However, the greatest reduction occurse when uncertainty is taken into account through TTA, with a reduction of up to ≈ 40%. Concretely, the test dataset in this manuscript has 2050 images. Using the probability-based filtering reduces the amount of images need to be manually analyzed from 2050 to ≈ 1000 images, and taking uncertainty into account reduces the number further down to ≈ 850. This is a significant reduction in workload with great potential for practical gains. Interestingly, using TTA with dropout rarely leads to noteworthy improvements, which we attribute to its poor fit with the oil spill detection task.

Figure 4 shows the error on as a function of the percentage of samples sent to the operator. The plots show how TTA with PVS and elastic transformations requires less samples to be analyzed by the operator, which corroborates Table 1.

**Alignment between human and machine uncertainty of oil spill identification** The labels used for classification are either 0 (no oil) or 1 (oil), but more fine-grained labels in form of segmentation masks are available, as explained in Section 4. Figure 5 shows the uncertainty of images containing pixels of the respective segmentation classes. The boxplots are calculated using all images in the testing dataset that contain the specified class. The background class is in all images.

The uncertainty profiles of TTA based on PVS and elastic transformations differ notably: PVS yields generally high uncertainty (up to 0.5) with few outliers, while elastic transformations result in mostly low uncertainty (maximum 0.35) but with many outliers. This contrast likely arises from the nature of each augmentation. PVS uniformly shifts pixel values, altering image intensity and making samples less familiar to the model, whereas elastic transformation introduces local deformations but preserves overall intensity, resulting in lower uncertainty. Both methods show similar uncertainty for ignore and background classes, with PVS having slightly higher uncertainty for images containing the ignore class, likely due to its proximity to oil spills. Operator uncertainty for oil spill categories (undefined, possible, probable) aligns with average uncertainty values, and the seep class consistently shows the lowest uncertainty. Produced water exhibits the lowest uncertainty for PVS but nearly the highest for elastic transformation, though its limited representation makes conclusions difficult.

**On the effect of oil spill size** Another important factor that characterizes an oil spill is the size, which refers to the surface area the oil spill covers. Figure 6 shows the standard deviation from using PVS $(+/-, \gamma < 1)$ and elastic transformation as augmen-

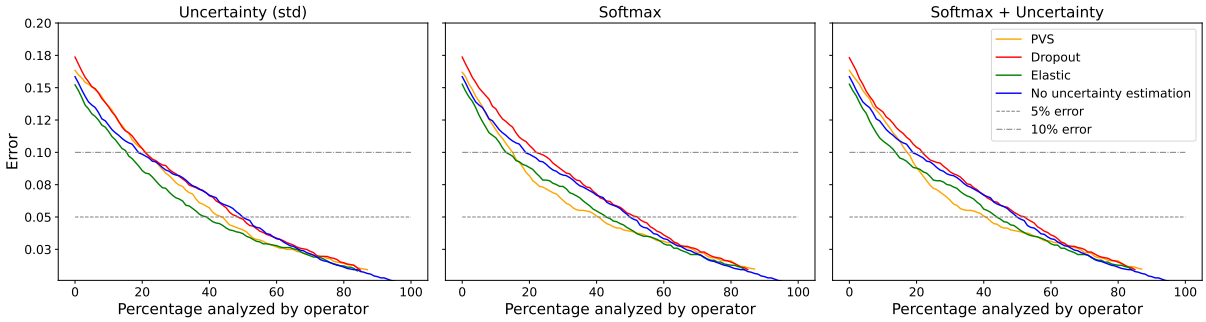


**Figure 4.** Plots showing the classification error with respect to the percentage of images that have to be analyzed by an operator. The three error thresholds are in grey. The three curves are from the methods with best performance, which are dropout with probability 0.01, PVS $(+/-, \gamma < 1)$ and elastic transformation. The first plot is the performance when sending the images with highest standard deviation. The second for sending the ones with most uncertain sigmoid values, which is the mean from the predictions using TTA. The last plot combines both methods.

**Table 2.** Percentage of testing dataset needed to be manually analyzed to get a human error rate of 5% across different hyperparameters for each augmentation technique.

| Method | Sigmoid | Uncertainty (std) | Sigmoid + Uncertainty |
|---|---|---|---|
| Regular model | 51 ± 10 % | - | - |
| PVS $(+/-, \gamma < 1)$ | **42** ± 1 % | **45** ± 4 % | **42** ± 1 % |
| $PVS(+, \gamma < 1)$ | 60 ± 7 % | 58 ± 5 % | 60 ± 7 % |
| PVS $(-, \gamma < 1)$ | 54 ± 4 % | **46** ± 6 % | 54 ± 4 % |
| PVS $(+/-, \gamma < 2)$ | **49** ± 7 % | 52 ± 4 % | **49** ± 7 % |
| PVS $(+, \gamma < 2)$ | 62 ± 5 % | 56 ± 5 % | 62 ± 5 % |
| PVS $(-, \gamma < 2)$ | 59 ± 6 % | 52 ± 7 % | 59 ± 6 % |
| Dropout (p=0.01) | 54 ± 11 % | **50** ± 9 % | 54 ± 11 % |
| Dropout (p=0.1) | 70 ± 4 % | 66 ± 5 % | 70 ± 4 % |
| Dropout (p=0.25) | 79 ± 4 % | 77 ± 4 % | 79 ± 4 % |
| Dropout (p=0.5) | 84 ± 1 % | 82 ± 3 % | 84 ± 1 % |

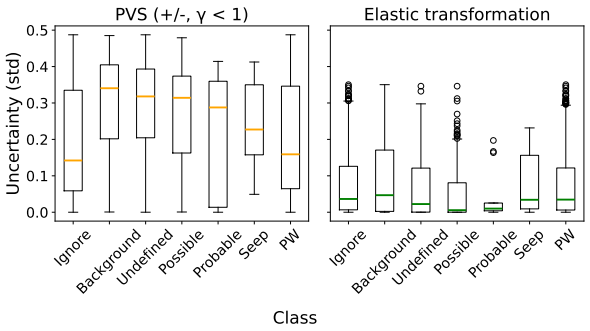


**Figure 5.** Boxplot of the standard deviation for the different classes in the pixel annotations. Plots are made using PVS $(+/-, \gamma < 1)$ as augmentation (left) and elastic transformation (right) as data augmentation.

tation for the different oil spill sizes in the testing dataset. Each data point in the plot corresponds to the uncertainty and oil spill size for a testing image containing oil. It shows that images with small oil spills generally have high uncertainty for using PVS $(+/-, \gamma < 1)$. However, it is opposite for the elastic transformation. For this method, most images with small oil spills have low uncertainty.

Images with small oil spills tend to show high uncertainty when using the PVS $(+/-, \gamma < 1)$ method, likely because such spills are difficult to distinguish from natural features or may be partially visible near image borders, making classification challenging. In contrast, small oil spills generally exhibit low uncertainty with the elastic transformation method, possibly because most images have low uncertainty for this approach and the grid size used in the transformation means small spills may remain largely unaffected.

**Investigating TTA hyperparameters** A critical component of TTA is applying a suitable strength of augmentation, for example the amount of shift in PVS. In Table 2, we evaluate hyperparameters associated with PVS and Dropout. Due to the computational demand, we do not investigate the hyperparameters associated with the elastic transformation, as described in Section 4. For dropout, we see that the minimal amount of dropout noise gives the best performance, while only slightly increasing

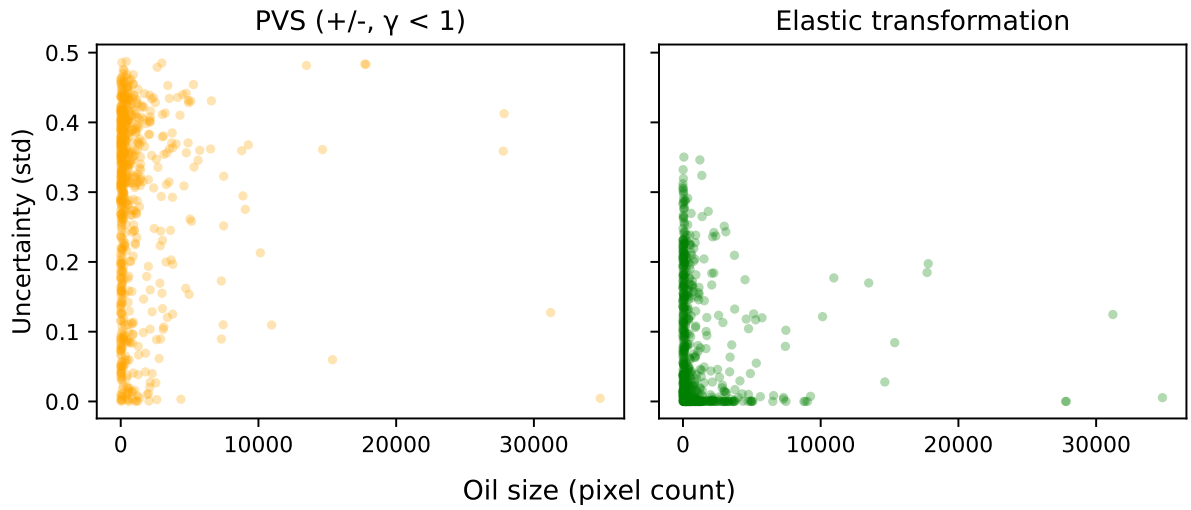


**Figure 6.** Plot of the standard deviation for each of the oil spill sizes in the testing dataset. Plots are made using PVS $(+/-, \gamma < 1)$ (left) and elastic transformation (right) as data augmentation.

the dropout rate leads to much worse performance. We attribute this to the previously discussed hypothesis that dropout is not a suitable augmentation for oil spill detection due to its potential similarity with oil spills. For PVS, it is evident that pixel values should be shifted both in the positive and negative direction. Also, a weaker shift seems to be beneficial.

# 6 Discussion

**On the compositions of augmentations** In our main results, we have only considered applying one type of augmentation in the TTA procedure. However, it has been shown that combing augmentations can lead to improved performance [15]. In Appendix A, we show some initial results where PVS and elastic transformations are combined, with encouraging results compared to the results in Table 1. However, the combination of augmentations introduces an additional element of complexity, as both the strength of the augmentations and the order in which they applied mush be investigated. We believe that future works could more thoroughly investigate this aspect of TTA in the context of uncertainty filtering.

**On the calibration of the classifier** We trained our model following standard procedures in deep learning (see Section 4). This shows encouraging results, but it is well known that the output probabilities might not be well-calibrated [6]. This certainly affects the output probabilities (see Figure 3), but could potentially also affect the uncertainty estimates. We investigated standard methods for improving the calibration of the model (see Appendix B), but saw little difference between the models with different calibration. An interesting line of future research could be to incorporate more sophisticated robustness strategies [23], to see if more calibrated classifiers could improve the filtering further.

# 7 Conclusion

We proposed an uncertainty-guided approach to reduce manual labor in oil spill detection from SAR imagery of the ocean. Given a user specified acceptable error threshold and an independent validation set of data, we automatically tune an uncertainty threshold to achieve a desirable trade-off between performance and efficiency. Our extensive evaluation on challenging real-world data shows that our proposed filtering approach can significantly reduce the manual workload associated with SAR-based oil spill detection. We believe that our proposed uncertainty-filtering also has potential outside this particular application, and that it can play an important part in automated deep learning systems for industrial applications in years to come.

# References

[1] N. Oceanic and A. Administration. *Oil Spills*. Accessed: 21 May 2025. 2025. URL: https://www.noaa.gov/education/resource-collections/ocean-coasts/oil-spills.

[2] R. Al-Ruzouq, M. B. A. Gibril, A. Shanableh, et al. "Sensors, Features, and Machine Learning for Oil Spill Detection and Monitoring: A Review". In: *Remote Sensing* (2020). DOI: 10.3390/rs12203338.

[3] F. M. Bianchi, M. M. Espeseth, and N. Borch. "Large-Scale Detection and Categorization of Oil Spills from SAR Images with Deep Learning". In: *Remote Sensing* (2020). DOI: 10.3390/rs12142260.

[4] O. Garcia-Pineda, G. Staples, C. E. Jones, et al. "Classification of oil spill by thicknesses using multiple remote sensors". In: *Remote Sensing of Environment* (2020). DOI: 10.1016/j.rse.2019.111421.

[5] N. A. Bui, Y. Oh, and I. Lee. "Oil spill detection and classification through deep learning and tailored data augmentation". In: *International Journal of Applied Earth Observation and Geoinformation* (2024). DOI: 10.1016/j.jag.2024.103845.

[6] C. Guo, G. Pleiss, Y. Sun, et al. "On Calibration of Modern Neural Networks". In: *ICML*. 2017.

[7] "A survey of uncertainty in deep neural networks". In: *Artificial Intelligence Review* (2023). DOI: 10.1007/s10462-023-10562-9.

[8] C. Brekke and A. H. S. Solberg. "Feature Extraction for Oil Spill Detection Based on SAR Images". In: *Image Analysis*. Springer Berlin Heidelberg, 2005, pp. 75–84. DOI: 10.1007/11499145_9.

[9] A. H. S. Solberg. "Remote Sensing of Ocean Oil-Spill Pollution". In: *Proceedings of the IEEE* 100 (2012). DOI: 10.1109/jproc.2012.2196250.

[10] R. Trujillo-Acatitla, J. Tuxpan-Vargas, et al. "Marine oil spill detection and segmentation in SAR data with two steps Deep Learning framework". In: *Marine Pollution Bulletin* (2024). DOI: 10.1016/j.marpolbul.2024.116549.

[11] G. Wang, W. Li, M. Aertsen, et al. "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks". In: *Neurocomputing* (2019). DOI: 10.1016/j.neucom.2019.01.103.

[12] Y. Gal and Z. Ghahramani. "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *ICML*. URL: https://proceedings.mlr.press/v48/gal16.html.

[13] W. J. Maddox, T. Garipov, P. Izmailov, et al. "A simple baseline for Bayesian uncertainty in deep learning". In: *NeurIPS*. 2019.

[14] B. Lakshminarayanan, A. Pritzel, and C. Blundell. "Simple and scalable predictive uncertainty estimation using deep ensembles". In: *NeurIPS*. 2017.

[15] I. Martinsen, S. A. Sørensen, S. Ortega, et al. "Quantifying uncertainty in foraminifera classification: How deep learning methods compare to human experts". In: *Artificial Intelligence in Geosciences* (2025). DOI: 10.1016/j.aiig.2025.100145.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* (2014). URL: http://jmlr.org/papers/v15/srivastava14a.html.

[17] J. Nalepa, M. Myller, and M. Kawulok. "Training- and Test-Time Data Augmentation for Hyperspectral Image Segmentation". In: *IEEE Geoscience and Remote Sensing Letters* (2020). DOI: 10.1109/LGRS.2019.2921011.

[18] K. He, X. Zhang, S. Ren, et al. "Deep Residual Learning for Image Recognition". In: *CVPR*. 2016. DOI: 10.1109/CVPR.2016.90.

[19] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *ICLR*. 2014.

[20] K. He, H. Fan, Y. Wu, et al. "Momentum Contrast for Unsupervised Visual Representation Learning". In: *CVPR*. 2020. DOI: 10.1109/CVPR42600.2020.00975.

[21] Y. Wang, N. A. A. Braham, Z. Xiong, et al. "SSL4EO-S12: A large-scale multimodal, multitemporal dataset for self-supervised learning in Earth observation [Software and Data Sets]". In: *IEEE Geoscience and Remote Sensing Magazine* (). DOI: 10.1109/MGRS.2023.3281651.

[22] A. J. Stewart, C. Robinson, and I. A. a. Corley. "TorchGeo: Deep Learning With Geospatial Data". In: *ACM Transactions on Spatial Algorithms and Systems* (2024). DOI: 10.1145/3707459.

[23] M. Heinonen, B.-H. Tran, M. Kampffmeyer, et al. "Robust Classification by Coupling Data Mollification with Label Smoothing". In: *AISTATS*. 2025. URL: https://proceedings.mlr.press/v258/heinonen25a.html.

# A Combining augmentations

Using first PVS(+/-, $\gamma$ ¡ 1) and then elastic transformation as data augmentation, results in performance similar, or slightly better than the two methods alone. However, the combination increases the computational complexity, making the slight increase in performance a relatively small gain. To get even better performance using the combination, the parameters for the two methods should be tuned together. Using the combination in opposite order

gives drastically worse results, seen in Table A.1. This might relate to the PVS being based on the data distribution of the training images. The elastic transformation shifts the values and possibly resulting in another distribution. The usage of PVS on the elastic transformed images might then not be appropriate, as the parameters estimated in PVS are not fitting anymore.

**Table A.1.** Percentage of testing dataset needed to be manually analyzed to get a human error rate of 5% for the PVS (+/-, $\gamma < 1$), elastic transformation and using both.

| Method | Sigmoid | Uncertainty (std) | Sigmoid + Uncertainty (std) |
|---|---|---|---|
| PVS (+/-, $\gamma$ <1) | 42 ± 1 % | 45 ± 4 % | 42 ± 1 % |
| Elastic transform | 44 ± 4 % | 41 ± 2 % | 45 ± 4 % |
| PVS + Elastic | 40 ± 7 % | 43 ± 6 % | 40 ± 8 % |
| Elastic + PVS | 77 ± 4 % | 79 ± 6 % | 77 ± 4 % |

# B Calibration

Temperature scaling is a standard way to prevent overconfidence by calibrating the network [6]. This is when the output from the network is divided by the temperature value of the network. The temperature is found using an optimization algorithm which minimizes the loss. The benefits of using temperature scaling on the model are explored. Results show that temperature scaling improves calibration, as shown in Table B.1. However, we found little improvement in terms of the uncertainty filtering, as seen in Figure B.1. We believe that this is because a single scaling factor is applied to the entire model, which might have less of an impact of the resulting ordering of samples.

**Table B.1.** The Negative Log Likelihood (NLL) and Expected Calibration Error (ECE) of the model using the validation set before and after temperature scaling. The temperature used for the network is 1.159

| | Before temperature scaling | After temperature scaling |
|---|---|---|
| NLL | 0.573 | 0.574 |
| ECE | 0.466 | 0.439 |

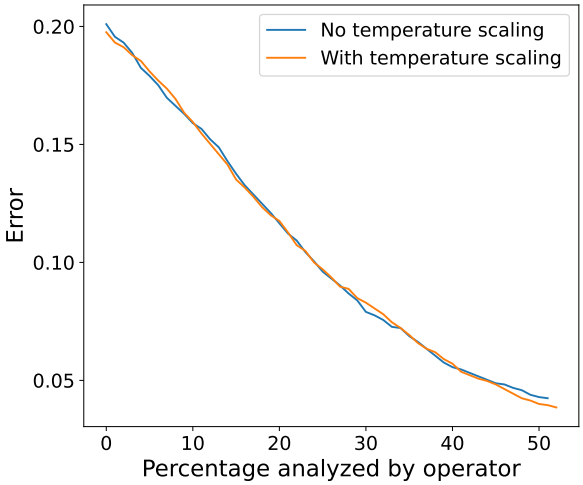


**Figure B.1.** The figure shows the error for the different percentages of images being sent to the operator, for both with and without temperature scaling in the ResNet-50 network. The images being sent to the operator are the ones with highest standard deviation using PVS (-, $\gamma$ <1)