Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P0089-cd

Uncertainties in Iceberg Detection from Satellite Data: Error-Modelling for the Quantification of Total Uncertainty in Image Segmentation

Peter Kuhn

Fraunhofer Institute for High-Speed Dynamics, Ernst-Mach-Institut, EMI, Germany.
E-mail: peter.kuhn@emi.fraunhofer.de

Daniel Schweizer

Fraunhofer Institute for High-Speed Dynamics, Ernst-Mach-Institut, EMI, Germany.
E-mail: daniel.schweizer@emi.fraunhofer.de

Christoph Brockt-Haßauer

Fraunhofer Institute for High-Speed Dynamics, Ernst-Mach-Institut, EMI, Germany.
E-mail: christoph.brockt-hassauer@emi.fraunhofer.de

Keywords: Predictive uncertainty, Monte Carlo dropout, error-modelling.

1. Introduction

In many real-life contexts, particularly those where the impact of potential errors is high, making maximum-likelihood predictions is not sufficient. Rather, one is interested in whether a prediction is certain or uncertain, that is how large one should expect the error to be. This paper focuses on a use case from the EU HORIZON Project AI-ARC (AI-ARC, 2024), which involves AI-based detection of icebergs from satellite images for naval navigation. The high-risk task carries significant consequences for incorrect predictions. Thus, an automated analysis of the uncertainty of iceberg predictions is essential.

A relevant restriction on the uncertainty quantification task was the fact that the model under evaluation, what we will call the *first-order model*, was provided by external partners and thus had to be treated as a black box. Standard techniques of uncertainty quantification, like ensembling or Monte Carlo dropout (Gal and Ghahramani, 2016), require in depth tinkering with the model architecture or at least various retrainings of the same model. This paper applies the simple but under-explored approach of training a second-order machine learning model to predict the error of the first-order model. The results are interpreted as follows: High predicted error corresponds to uncertainty, low predicted error corresponds to certainty in the first-order model.

The contribution of this paper consist in a presentation of error-modelling as a simple and effective way of quantifying the uncertainty of machine learning models. The approach has the advantage of not relying on changing or even knowing the architecture of the first-order model we are trying to evaluate. We furthermore compare our approach to uncertainty estimates produced using Monte Carlo dropout qualitatively and quantitatively. While Monte Carlo dropout outperforms

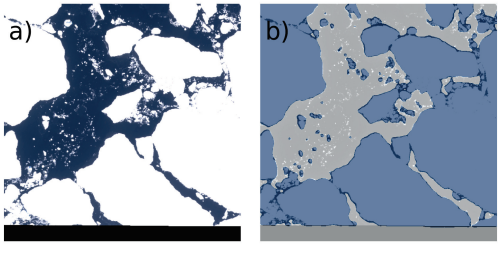

Fig. 1. Example of the iceberg detection results for which we want to quantify uncertainty. (a) shows a satellite image, (b) shows iceberg predictions generated by an FPN architecture (Lin et al., 2017) in blue. The black bar is the result of image padding.

error-modelling, the results they produce are very similar. Finally, we investigate the impact of having an additional dataset reserved to train the error-model versus using the original dataset that has been seen by the first-order model before. We show that error-modelling does not necessarily require an independent dataset. Thus, error-modelling is an effective alternative to standard uncertainty quantification techniques in black-box settings.

2. Related Work

One might assume that the numeric output of a classification model encapsulates all uncertainty information. By calculating the entropy of this output, one can derive uncertainties usually referred to as *predictive entropies*. However, the numeric result, for instance as produced by a sigmoid layer, does not properly represent whether a new sample is close to the training distribution (Gal and Ghahramani, 2016). More sophisticated solutions are called for, even though predictive entropy can sometimes perform surprisingly well (Weiss and Tonella, 2022).

Monte Carlo dropout (MCD) has for some time been considered the standard technique for the quantification of predictive uncertainty in machine learning. In MCD one randomly deactivates nodes during inference, given the model was trained with dropout, to obtain a distribution over predictions instead of a singular point prediction. The technique goes back to a paper by Gal that demonstrated that MCD approximates the predicted variance of a Bayesian neural network, a neural network where every node is associated with a second value representing uncertainty or variance (Gal and Ghahramani, 2016).

It is well established now that MCD is typically outperformed by using different kinds of ensembling techniques (Rahaman and thiery, 2021), either by re-initializing the same model with different weights (Lakshminarayanan et al., 2017), or by retraining a model with different subsets of the training data (Rahaman and thiery, 2021). The variance of the predictions at an inference point can be used as a powerful uncertainty measure. However, the construction of ensembles of this kind is computationally expensive. Furthermore, just as MCD, ensembling techniques are inapplicable where the model under evaluation is a black box because model retraining requires access to the model and to the relevant training data.

The approach of modelling uncertainty by directly building a second-order model has not gotten much traction in the machine learning literature. The approach can be used in combination with other approaches like ensembles (Rahaman and thiery, 2021; Hu et al., 2020) and there have been some attempts to capture uncertainty in this manner in the context of predicting material properties (Tavazza et al., 2021). Here, the approach was limited to a regression task and the authors did not systematically investigated the relation of error-modelling as a standalone technique to other state of the art uncertainty quantification approaches.

3. Methods

3.1. *Evaluating Uncertainty Quantification Methods*

A key challenge in uncertainty quantification is the limited access to ground truth data, meaning true uncertainty values. Consequently, the reliability of uncertainty results must be evaluated based on their correlation with error in the test set. Although there is no consensus on evaluation methods in the light of these difficulties, we will adopt those proposed in Mobiny (2021). We aim at classify predictions into certain and uncertain ones, which facilitates efficient decision making for the end users. Thus, the evaluation methods use the binary (thresholded) output of the first-order model. If we then also threshold the output of the uncertainty quantification method (assuming it gives float results) we have a binary classification into certain and uncertain results. We can treat results that are both certain and correct as true positives (TP), ones that are certain and incorrect as false positives (FP), ones that are uncertain and correct as false negatives (FN), and finally, ones that are uncertain and incorrect as true negatives (TN). A good uncertainty metric will then be one where the following probabilities are high. First the *correct-certain ratio*:

$$\begin{aligned}\text{CCR} =& P(\text{correct}|\text{certain}) \\ =& \frac{P(\text{certain, correct})}{P(\text{certain})} = \frac{\text{TP}}{\text{TP+FP}}\end{aligned} \quad (1)$$

Secondly, the *incorrect-uncertain ratio*:

$$\begin{aligned}\text{IUR} =& P(\text{uncertain}|\text{incorrect}) \\ =& \frac{P(\text{uncertain, incorrect})}{P(\text{incorrect})} = \frac{\text{TN}}{\text{TN+FP}}\end{aligned} \quad (2)$$

And finally, the *uncertainty-accuracy*:

$$\text{UA} = \frac{\text{TP+TN}}{\text{TP+TN+FN+FP}} \quad (3)$$

Here we run into the problem that the scores are defined with respect to a classification problem. For continuous values this problem can be alleviated by considering the area under curve between the minimum and maximum of the uncertainty predictions on the relevant dataset. Note that the resulting scores are relative to the minimum and maximum of the uncertainty metric on the relevant dataset. Our application of a minmax scaler on the uncertainty results will thus not impact scores.

3.2. *Error-Modelling*

Uncertainty can be defined as the expected error in some inference task (Hüllermeier and Waegeman, 2021). The idea of building a model that predicts the error of the first-order model directly from the data is thus straightforward. Assume a dataset $D = \{(x_1, y_1), (x_2, y_2)...\}$ and a first-order model f_{FOM} approximating the functional relationship $f(x_i) = y_i$, that generates the ground truth, by $f_{\text{FOM}}(x_i) = \hat{y}_i$. Typically, the first-order model will be trained using some subset $D_{\text{train}}^{FOM} \subset D$ by minimizing a loss function $L_{\text{FOM}}(\mathbf{y}, \hat{\mathbf{y}})$, with $\mathbf{y} = \{f(x)|x \in D_{\text{train}}^{\text{FOM}}\}$ and $\hat{\mathbf{y}} = \{f_{\text{FOM}}(x)|x \in D_{\text{train}}^{\text{FOM}}\}$.

We define a second-order dataset for some dataset D, given some loss function L, which will typically be the mean squared error MSE, as:

$$D^* = \{(x_1, L(y_1, \hat{y}_1)), (x_2, L(y_2, \hat{y}_2))...\} \quad (4)$$

A (second-order) error model will then be a model $f_{\text{SOM}}(x_i) = \hat{L}_{\text{FOM}}(y_i, \hat{y}_i)$ approximating the functional relation obtaining in the second-order dataset by minimizing a second-order loss $L_{\text{SOM}}(L_{\text{FOM}}(\mathbf{y}, \hat{\mathbf{y}}), \hat{L}_{\text{FOM}}(\mathbf{y}, \hat{\mathbf{y}}))$.

Intuitively, the effective training of an error model requires that we reserve a subset of the available data $D_{\text{train}}^{\text{SOM}} \subset D$ that can be used for the generation of the second-order dataset $D_{\text{train}}^{\text{SOM}*}$ after the first-order model has been trained (Tavazza et al., 2021). However, below we will also consider the case where first- and second-order model are trained based on the same dataset, i.e. we will assume that $D_{\text{train}}^{\text{SOM}} = D_{\text{train}}^{\text{FOM}}$ in these cases.

An error model captures the *total uncertainty* involved in an inference, i.e. it does not differentiate between *aleatoric uncertainty* due to the inherently in-deterministic nature of the data and *epistemic uncertainty*, i.e. uncertainty due to limitations of our modelling capacities, may they be due to lack of data or due to the model architecture we employ.

4. Experimental Setup

We use two datasets. The primary use case consisted of 481 satellite images from the SENTINEL satellite together with iceberg- and land-masks. The iceberg dataset is quite small, considering that there is a serious under-representation of iceberg images as opposed to clouds and open water. This imbalance results in poor performance both for the first-order model and for the uncertainty evaluation techniques. As the primary purpose of this paper is the demonstration of the feasibility of the error modelling approach, we also tested our pipeline on the public UAV landing site dataset which consists of 1359 aerial imagery captured by unmanned aerial vehicles (UAVs). The target is to identify potential landing sites. It includes various environmental conditions and terrains, providing a richer resource evaluating the performance of our approach.

To address the limited and imbalanced nature of these datasets, we employed data augmentation techniques using the Albumentations library (A. Buslaev and Kalinin, 2018). For the

ice dataset, spatial augmentations—including horizontal and vertical flips—were utilized to increase orientation variability. In contrast, the drone dataset underwent comprehensive pixel-level spatial augmentations, where multiple transformations such as random cropping, brightness and contrast adjustments, geometric distortions, rotations, and resizing were applied simultaneously to create diverse imaging conditions. Multiple augmented versions were generated for each original image to ensure sufficient variability. These augmentation methods not only expanded the dataset size but also improved the overall quality and diversity of the otherwise scarce data.

The first-order model used in both cases was an FPN (Lin et al., 2017) segmentation model from Iakubovskii (2019) with a resnet34 encoder and imagenet pretrained weights. A simply sigmoid was chosen as activation of the output layer. The loss-function was DICE to counteract target imbalances (Milletarì et al., 2016). The network was further trained with decoder dropout with a dropout rate of 0.3 for the later application of MCD. Our MCD variances are calculated from a drawing of 30 samples. The only difference between models in the ice vs. the drone case is that the ice model has an additional input channel for land-masks (ice on land does not count as an iceberg).

Our second-order model is a UNET model (Ronneberger et al., 2015), with a resnet34 encoder but featuring no pre-trained weights and no dropout. We employed Smooth L1 loss to effectively address the imbalance between predominant zero values and the non-zero target pixels in the image error prediction task. The loss function was chosen because an error prediction task is not a segmentation task, where DICE would have been preferable. Rather, we are interested in pixel-wise uncertainty estimates (i.e. estimates with maximal log likelihood). Furthermore, we employed an ADAM and ADAGRAD optimizer with weight decay of 0.0005 and a learning rate scheduler in the first-order model and the second-order model, respectively. To augment the input, we include a fourth channel that incorporates the prediction from the first-order model, which have been scaled to match the RGB channel range of 0 to 255.

Where we create wholly distinct datasets, we do this by diving the dataset in halve. We then use a train, test and validation split of 0.7, 0.2 and 0.1, both for the first-order dataset the second-order dataset (which may or may not overlap).

5. Results

The central result of our experiments is that neither dropout, nor error-modeling gives superior uncertainty measurements across the board. While dropout produces tighter uncertainty bounds, the error model is better at detecting some sources of uncertainty like clouds in the ice-dataset or different anomalies in the UAV dataset. Quantitatively, with respect to the more robust UAV datset, the error-model outperforms both reference metrics, as is visible in table 1. Here, we will first discuss the visual, qualitative results and then validate our observations using the quantitative analysis.

Generally, the error-models generates more equally distributed uncertainty estimates on the iceberg dataset. Our model predicts more or less continuous higher errors for white surfaces on the ice dataset. Predicted peaks in error are sparse. What might seem like a mistake first is a natural and desirable result. The main obstacle to straightforward iceberg predictions are clouds. All white surfaces might potentially turn out to be clouds. The central distinguishing factors are the tight borders of icebergs. Thus, the further one moves from the borders of an iceberg the higher one should expect the error to be. In this way the error-model captures what we would intuitively call uncertainty better than the MCD variance. Of course, this does not explain the systematic spatial distortions visible in figure 2. These come about largely due to the inherent limitations of the dataset mentioned above. As these constrain the kinds of inference that can be drawn from the results, we will now focus on result obtained using larger and more evenly distributed UAV dataset.

Qualitatively, the results on the UAV dataset, illustrated by figure 3, are more favourable to the error-modelling approach. While MCD mainly picks up on the the boundaries of segmented areas,

the error-model is more sensitive to anomalous and hard to identify features.

The quantitative analysis on the test sets, shown in table 1, supports the above conclusions. Dropout consistently outperforms the error

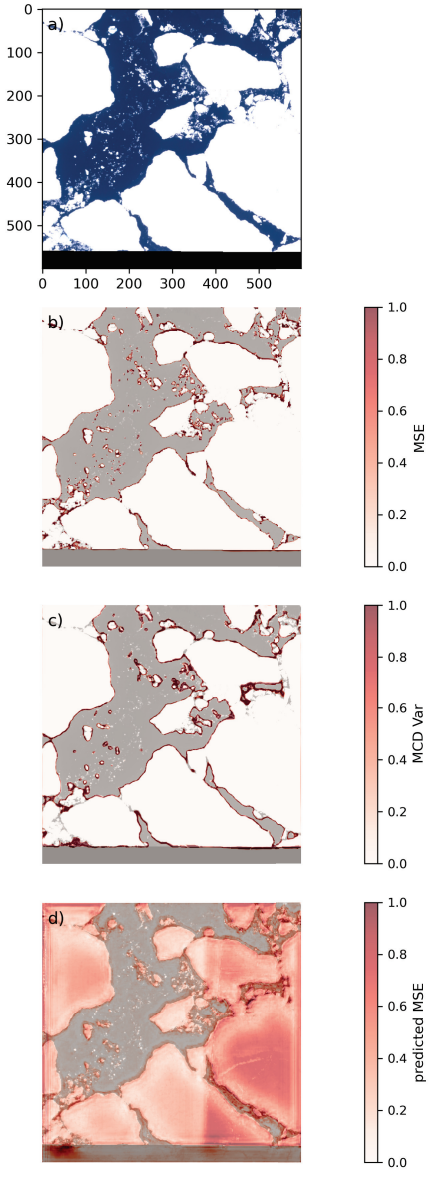

Fig. 2. Example results from the iceberg dataset. (a) shows the original image, (b) shows the error, (c) depicts the normalized variance of MCD samples and (d) shows the error predicted by the second-order model. (All results are scaled using a minmax scaler across the test set.)

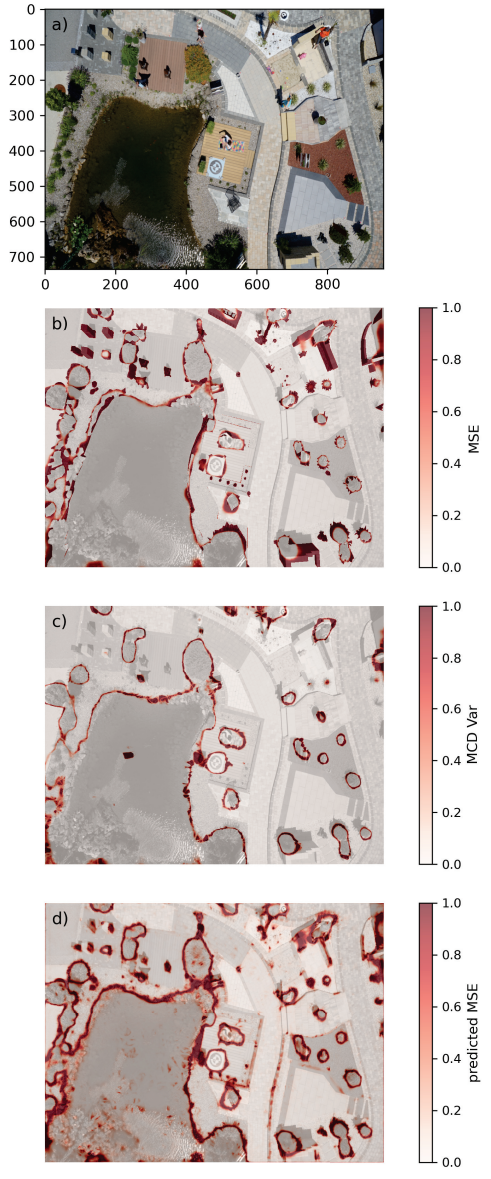

Fig. 3. Example results from the UAV dataset, using the same data for the error-model as for training the first-order model. (a) shows the original image, (b) shows the error, (c) depicts the normalized variance of MCD samples and (d) shows the error predicted by the second-order model. (All results are scaled using a minmax scaler across the test set.)

model on the iceberg dataset. As the black box approach is here set against the white box approach of dropout, the results are still promising. The main disadvantage of the error model is its lower incorrect-uncertain ratio, i.e. there are more errors that are not correctly captured as uncertain. This might seem surprising at first, given higher propensity of the error model to predict uncertainty, relative to MCD. However, given that the relevant scores are calculated across thresholds, the low incorrect-uncertain ration can be explained by the reluctance of the model to predict an error of 1.

The additional splitting of the dataset into an independent first-order and second-order dataset systematically decreases the performance of *all* uncertainty metrics. This might seem surprising, but it is readily explainable. The scores we calculate quantify the performance of the whole setup - first order model plus uncertainty quantification technique - to generate predictions with an uncertainty estimate. The better the first-order model the easier the task. Thus, at least as long as the first-order model is not over-fitted, using the same data for both models is a reasonable approach. Note however that this assessment is relative to the model used, so the validity of the approach should be assessed for every new model architecture.

On the more robust UAV dataset the error model outperforms both reference metrics. Using the same data for both models the uncertainty-accuracy and correct-certain ratio are close, while the incorrect-uncertain ratio is considerably higher. Thus, the approach produced much less instances where one is certain, even though one is incorrect, a desirable result where the task is to minimize risks.

Table 1. A summary of the scores for the two datasets.

Uncertainty metric	CCR [%]	IUR [%]	UA [%]
Icebergs			
predictive entropy	91.0	16.3	90.1
MCD	90.6	24.8	89.3
error-model	91.0	23.0	81.5
UAV (same data)			
predictive entropy	94.2	27.6	92.0
MCD	94.1	27.3	91.4
error-model	96.0	55.0	89.6
UAV (independent)			
predictive entropy	92.4	25.2	89.8
MCD	92.3	25.1	88.9
error-model	95.0	55.7	85.1

6. Conclusion

We have shown that error modeling is an effective black box approach to quantifying the uncertainty of a machine learning model. The method produces results comparable to state-of-the-art techniques like MCD and predictive entropy, but yields the advantage of being easy to implement without the need to change the prediction model. We have addressed the intrinsically hard task to measure the performance of uncertainty quantification methods by adapting established metrics for binary classification. No single method outperformed the others on all metrics consistently over all data sets, but our results indicate, that the error modeling is especially sensible to cases with incorrect prediction, making it a valid candidate for uncertainty quantification in high risk tasks.

We compared error modelling and MCD, which capture total uncertainty and epistemic uncertainty (Kendall and Gal, 2017), respectively. Even under ideal conditions MCD would not yield perfect scores because it would not represent aleatoric uncertainty. Comparing those is not straightforward. In principle, an error model with uncertainty disentanglement, i.e. one differentiating between aleatoric and epistemic uncertainty, would be necessary for a fair comparison. Such a model might be based on an additional variance layer, as proposed by Nix and Weigend (1994). On the other hand, for practical purposes, the quantification of total uncertainty is usually most relevant.

While in our analysis we saw no negative impact of using an independent dataset for the training of the error-model, there is an obvious path to making our proposed method more data efficient: Data augmentation on the training dataset, i.e. utilizing assumed symmetries of the ground truth

data to add synthetic images, one could effectively enlarge the dataset. As we are not modelling a generative process outside our control but a *model*, we can effectively probe its capacities by feeding it such additional augmented datapoints. This should be considered where the method is implemented in a context where data is sparse or where one fears the first-order model to be overfitted.

Looking ahead further, the choice of UNET as the error model may not be ideal for this task. While it effectively captures uncertainty at the edges, it sometimes struggles to identify uncertain shapes more broadly. Future work could explore the use of GANs, particularly through their ability to generate high-quality representations, in combination with perceptual loss as a potential alternative to enhance uncertainty quantification in segmentation tasks. Specifically, allowing GANs to generate multiple plausible outputs could provide valuable insights into the range of possible outcomes in uncertain regions, further improving the robustness of uncertainty estimation.

7. Aknowledgements

This work is part of the AI-ARC project funded by the European Union's Horizon 2020 research and innovation program. We would like to acknowledge Telespazio-France, specifically Guillaume Cavaro and Elodie Guasch, for providing the iceberg data.

References

A. Buslaev, A. Parinov, E. K. V. I. I. and A. A. Kalinin (2018). Albumentations: fast and flexible image augmentations. *ArXiv e-prints*.

AI-ARC (2021-2024). project webpage. `https://ai-arc.eu/`.

Gal, Y. and Z. Ghahramani (2016, 20–22 Jun). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In M. F. Balcan and K. Q. Weinberger (Eds.), *Proceedings of The 33rd International Conference on Machine Learning*, Volume 48 of *Proceedings of Machine Learning Research*, New York, New York, USA, pp. 1050–1059. PMLR.

Hu, S., N. Pezzotti, D. Mavroeidis, and M. Welling (2020). Simple and accurate uncertainty quantification from bias-variance decomposition. *ArXiv abs/2002.05582*.

Hüllermeier, E. and W. Waegeman (2021, March). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning 110*(3), 457–506.

Iakubovskii, P. (2019). Segmentation models pytorch. `https://github.com/qubvel/segmentation_models.pytorch`.

Kendall, A. and Y. Gal (2017). What uncertainties do we need in bayesian deep learning for computer vision?

Lakshminarayanan, B., A. Pritzel, and C. Blundell (2017). Simple and scalable predictive uncertainty estimation using deep ensembles.

Lin, T.-Y., P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie (2017, July). Feature Pyramid Networks for Object Detection . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, pp. 936–944. IEEE Computer Society.

Milletarì, F., N. Navab, and S.-A. Ahmadi (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 Fourth International Conference on 3D Vision (3DV)*, 565–571.

Mobiny, A., Y. P. M. S. e. a. (2021). Dropconnect is effective in modeling uncertainty of bayesian deep networks. *Sci Rep 11*(5458).

Nix, D. and A. Weigend (1994). Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, Volume 1, pp. 55–60 vol.1.

Rahaman, R. and a. thiery (2021). Uncertainty quantification and deep ensembles. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems*, Volume 34, pp. 20063–20075. Curran Associates, Inc.

Ronneberger, O., P. Fischer, and T. Brox (2015). U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi (Eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Cham, pp. 234–

241. Springer International Publishing.

Tavazza, F., B. DeCost, and K. Choudhary (2021). Uncertainty prediction for machine learning models of material properties. *ACS Omega 6*(48), 32431–32440.

Weiss, M. and P. Tonella (2022). Uncertainty quantification for deep neural networks: An empirical comparison and usage guidelines.