Uncertainty Informed Anomaly Scores with Deep Learning: Robust Fault Detection with Limited Data

Jannik Zgraggen[1], Gianmarco Pizza[2], and Lilach Goren Huber[3]

[1,3] *Zurich University of Applied Sciences, Technikumstrasse 9, Winterthur, 8400 Switzerland*
jannik.zgraggen@zhaw.ch
lilach.gorenhuber@zhaw.ch

[2] *Nispera AG, Hornbachstrasse 50, CH-8008 Zurich, Switzerland*
gianmarco.pizza@nispera.com

ABSTRACT

Quantifying the predictive uncertainty of a model is an important ingredient in data-driven decision making. Uncertainty quantification has been gaining interest especially for deep learning models, which are often hard to justify or explain. Various techniques for deep learning based uncertainty estimates have been developed primarily for image classification and segmentation, but also for regression and forecasting tasks. Uncertainty quantification for anomaly detection tasks is still rather limited for image data and has not yet been demonstrated for machine fault detection in PHM applications.

In this paper we suggest an approach to derive an uncertainty-informed anomaly score for regression models trained with normal data only. The score is derived using a deep ensemble of probabilistic neural networks for uncertainty quantification. Using an example of wind-turbine fault detection, we demonstrate the superiority of the uncertainty-informed anomaly score over the conventional score. The advantage is particularly clear in an "out-of-distribution" scenario, in which the model is trained with limited data which does not represent all *normal* regimes that are observed during model deployment.

1. INTRODUCTION

Assessing the predictive uncertainty of machine learning (ML) and deep learning (DL) algorithms is essential for any decision taken on the basis of such algorithms. Some popular examples for taking decisions under uncertainty include image classification for autonomous-driving (Kraus & Dietmayer, 2019; He, Zhu, Wang, Savvides, & Zhang, 2019; Miller, Dayoub, Milford, & Sünderhauf, 2019) or for medical purposes (Leibig, Allken, Ayhan, Berens, & Wahl, 2017; Herzog, Murina, Dürr, Wegener, & Sick, 2020) as well as time series forecasting models (Laptev, Yosinski, Li, & Smyl, 2017).

The applications of uncertainty quantification (UQ) to machine learning anomaly detection are still rare, and these focus mostly on anomaly detection in images (Seeböck et al., 2019; Cai, Lu, & Sato, 2020; Sato, Hama, Matsubara, & Uehara, 2019). In time series data, and in particular for machine sensor data, DL based UQ has been primarily used for prognostics models aimed at the estimation of remaining useful life (Biggio, Wieland, Chao, Kastanis, & Fink, 2021). Combining uncertainty estimates in the most fundamental (and application relevant) step of machine fault detection is still missing. As condition-based maintenance often relies on the output of anomaly detection algorithms, uncertainty of such algorithms is necessarily propagated onto uncertainty in maintenance decisions.

In this paper we introduce a method to incorporate the uncertainty quantification of a DL model into an anomaly score. In particular, we suggest to use a regression-based anomaly detection model, in which a model is trained with normal data exclusively and anomalies are detected in the test data based on the deviations (residuals) of the true measurements from the model predictions. Using such a deep regression-based anomaly detection model, the UQ is carried out similarly to a standard regression task, independent of the anomaly detection step. In a subsequent step we derive an anomaly score that combines information about the prediction error together with the prediction uncertainty. We show that the uncertainty-informed anomaly score outperforms the conventional uncertainty agnostic score especially under difficult training conditions, when the training data is not representative for all testing conditions. This scenario is very common for PHM applications, in which machine data is often collected over a

limited period of time prior to commercial deployment (Fink et al., 2020). Despite this, new operating conditions should in general not be detected as anomalies, but as a healthy "out-of-distribution" behaviour. In this sense, exploiting uncertainty information in anomaly detection is more challenging than in classification or forecasting tasks. We show that uncertainty-informed anomaly scores can distinguish between true anomalies and unknown but healthy conditions. An important advantage of the uncertainty-informed score, is that there is no need to use an uncertainty-based filter of the predicted outputs, in order to disqualify or discard the most uncertain predictions, as commonly done in classification or segmentation tasks (Abdar et al., 2021; Schwaiger, Sinhamahapatra, Gansloser, & Roscher, 2020). Instead, each and every prediction obtains an anomaly score and its health condition is assessed given a detection threshold.

There are various approaches for UQ with DL models (Gawlikowski et al., 2021; Abdar et al., 2021). Some methods are based on training an ensemble of networks and using the variance of predictions as a measure for uncertainty (Lakshminarayanan, Pritzel, & Blundell, 2017), other focus on variational inference using MC-Dropout as an estimate for model uncertainty (Gal & Ghahramani, 2016). In this paper we choose to focus on deep ensemble methods. However, since our neural network includes dropout layers for regularization, we in fact combine MC-dropout with ensembling.

In the first part of the paper we focus on the selection of a useful UQ method for our problem. A useful uncertainty measure is on one hand sharp enough to be informative and on the other hand does not suffer from over-confidence, i.e is well calibrated (Kuleshov, Fenner, & Ermon, 2018). The calibration of an uncertainty estimate can be quantified (Kuleshov et al., 2018; Levi, Gispan, Giladi, & Fetaya, 2019; Tran et al., 2020), and the model can be recalibrated in various ways if needed (Kuleshov et al., 2018; Levi et al., 2019). In order to select a properly calibrated model, we contrast the performance of two models: an ensemble of CNN models trained with a Mean Squared Error (MSE) loss is compared with an ensemble of probabilistic CNN models trained with a Negative Log Likelihood (NLL) loss. In the latter case the output of the network includes a mean and a variance of the conditional distribution function (Dürr, Sick, & Murina, 2020). Using an example aimed at wind-turbine fault detection, we demonstrate that the NLL-based ensemble provides a well calibrated uncertainty estimate, as opposed to the MSE-based ensemble. This conclusion is similar to the one in (Lakshminarayanan et al., 2017), however we quantify it here in several different ways.

The second part of the paper is dedicated to using the uncertainty informed regression model for an anomaly detection task. After selecting a reliable uncertainty measure we use it for the derivation of an uncertainty-informed anomaly score. We show that such a score can improve the fault detection performance compared to standard uncertainty-agnostic scores, particularly when the healthy training data is limited and does not cover all possible (healthy) operational conditions observed during testing. This approach to anomaly detection is the main contribution of the paper and has a potential impact beyond the specific application to wind turbine condition-based maintenance that we provide here as an example.

2. Introduction to the Use-Case: Wind Turbine Fault Detection

We demonstrate the usefulness of the uncertainty-informed anomaly score on 4 years of real operational data from the Supervisory Control and Data Acquisition (SCADA) system of a wind turbine. The data contains time series with 10-minute averaged values of environmental and operational variables. The fault detection task is aimed at detecting anomalous patterns in the temperature measurements of various turbine components (Tautz-Weinert & Watson, 2016), focusing primarily on heating rather than cooling effects (one-sided deviations of the temperature). This is achieved by using the component temperature at time t as a target variable y_t in a regression setup with the wind speed, ambient temperature, output power and rotational speed as model inputs. Training the model with data from healthy conditions exclusively, we expect large regression residuals (prediction errors) to be correlated with anomalous behavior. In a previous publication we showed the advantage of using a Convolutional Neural Network (CNN) for this task, and specified our selected architecture (Ulmer, Jarlskog, Pizza, Manninen, & Goren Huber, 2020). Here we repeat only details that are necessary for the performance evaluation of the uncertainty-informed anomaly score.

In the example shown throughout the paper we select the gearbox bearing temperature of the wind turbine as the target variable y_t, in which anomalies are to be detected. The predicting variables are the four mentioned above. The regression CNNs are aimed at providing an uncertainty quantification along with every prediction $\hat{y}_t$ of the bearing temperature at time step t.

A standard approach to anomaly detection based on normal state modeling is to assign anomaly scores to each prediction and set a threshold, above which a prediction is considered anomalous. The conventional anomaly scores are based on the magnitude of the prediction residuals. For example, the anomaly score of a test point at time t can be related to the Cumulative Distribution Function (CDF) of the training residuals, evaluated at the residual $r_t = y_t - \hat{y}_t$ of point t (Clifton et al., 2008):

$$S_t^{(0)} = F(r_t; \mu^{(tr)}, \sigma^{(tr)}) \tag{1}$$

where the mean $\mu^{(tr)}$ and standard deviation $\sigma^{(tr)}$ are esti-

mated from the distribution of the residuals of the entire training data set. The Gaussian CDF is defined as

$$F(y;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx. \tag{2}$$

In this way, a test point whose residual strongly exceeds the typical training residuals will be detected as an anomaly based on its dissimilarity with the training data. Naturally, this approach is bound to perform less well in case the test data is not well represented in the training set. This applies also when the test data is healthy, i.e with no anomalies. The result in this case may be frequent false positives, leading to unnecessary alarms. In this context it is important to distinguish between such "out of distribution (OoD)" normal data in contrast to true anomalies (e.g machine faults). The main purpose of the uncertainty-informed anomaly score we introduce here is to be able to distinguish between the two, thereby detecting the true anomalies and minimizing the false alarms due to "normal" OoD data.

3. Useful Uncertainty Quantification

The anomaly detection task essentially decomposes into two sequential steps: (i) a supervised prediction model trained with normal data only (ii) a clustering task of the mixed (normal and abnormal) data, based on anomaly scores assigned to each prediction.

In decision making problems it is beneficial to quantify the uncertainty inherent to the prediction step (i). There are two main sources for uncertainty; aleatoric and epistemic uncertainty (Dürr et al., 2020). Aleatoric uncertainty is also known as data uncertainty and refers to the inherent ambiguity present in the data. Epistemic uncertainty, on the other hand, is known as model uncertainty and is caused by a lack of knowledge of our model.

By including an uncertainty quantification, the prediction model provides not only a single predicted value $\hat{y}_t$, but an effective predictive distribution, $f(\tilde{\mu}_t, \tilde{\sigma}_t)$, where $\tilde{\mu}_t$ provides an estimate for the predicted value and $\tilde{\sigma}_t$) and estimate for the prediction uncertainty at step t. Since the prediction model is trained with normal data, we expect the predictive distribution of a regression model not to depend on the true value y_t at test time, that is to be independent of whether the ground truth is normal or abnormal. This observation allows us to use for step (i) standard frameworks for UQ commonly used for regression models, ignoring at this point the fact that our ultimate goal is to use this UQ for the anomaly detection task.

In the following we compare different UQ methods in order to select the most useful one. A useful UQ is capable of providing reliable uncertainty estimates for the model predicted output, which is on one hand sharp enough and on the other hand does not suffer from over-confidence (Kuleshov et al., 2018). Selecting a reliable (calibrated) uncertainty quantification is relevant for any prediction model, independent of the anomaly detection task following the prediction step.

Similarly to other regression tasks, the purpose here is to identify the most reliable uncertainty measure amongst possible candidates. In this paper we focus on ensemble-based methods for uncertainty estimates. As ensemble members we select CNNs that have been proven to perform well on the anomaly detection task for wind turbines in our previous work (Ulmer et al., 2020). These CNNs already include dropout layers for regularization, which we retain also here. This implies that our UQ is based on deep ensembles with dropout, which is turned on also at prediction time. We thus generate an ensemble of different dropout configurations, where each member of the ensemble is initialized and trained individually. We compare the uncertainty quantifications of two types of CNN ensembles:

MSE ensemble. We train an ensemble of $M = 30$ CNNs by minimizing the prediction MSE. We denote the weights of the m^{th} trained model with θ_m and the predicted value at step t with $\hat{y}_{t,\theta_m}$. For every time step we use the ensemble mean as the prediction and the variance over the ensemble as the uncertainty measure:

$$\begin{aligned} \tilde{\mu}_t &= \frac{1}{M}\sum_{m=1}^{M} \hat{y}_{t,\theta_m} \\ \tilde{\sigma}_t^2 &= \frac{1}{M-1}\sum_{m=1}^{M} (\hat{y}_{t,\theta_m} - \tilde{\mu}_t)^2 \end{aligned} \tag{3}$$

NLL ensemble. We train an ensemble of $M = 30$ CNNs by minimizing the prediction NLL. Each member m of the ensemble outputs a predictive distribution $N(\hat{\mu}_{t,\theta_m}, \hat{\sigma}_{t,\theta_m})$. In order to combine the predictive distributions of the NLL-ensemble members we sample a value $\hat{s}_{t,\theta_m}$ from the predicted distribution for each step t and each ensemble member m. The estimated mean and uncertainty of the prediction are then defined as:

$$\begin{aligned} \tilde{\mu}_t &= \frac{1}{M}\sum_{m=1}^{M} \hat{s}_{t,\theta_m} \\ \tilde{\sigma}_t^2 &= \frac{1}{M-1}\sum_{m=1}^{M} (\hat{s}_{t,\theta_m} - \tilde{\mu}_t)^2 \end{aligned} \tag{4}$$

Note that the variance $\tilde{\sigma}_t^2$ of the sampled values is necessarily larger than the variance of the mean predictions of the same ensemble.

The MSE-ensemble uses the empirical variance of non probabilistic predictions of the CNNs as a measure of uncertainty. This is done differently in the NLL-ensemble. Here each member of the NLL-ensemble models the inherent ambiguity

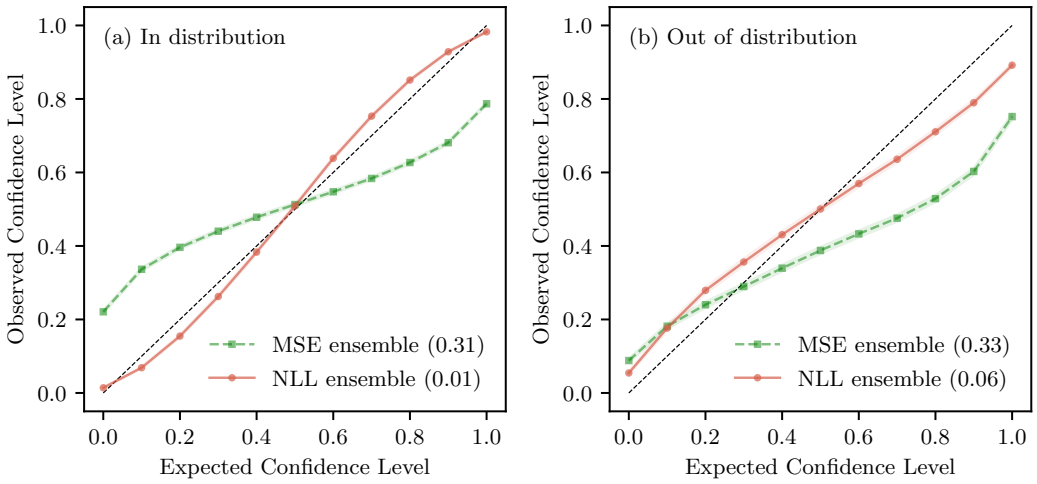

Figure 1. Uncertainty calibration curves. Two uncertainty quantification methods, MSE-ensemble and NLL-ensemble, are compared on two test sets for the prediction of the gearbox bearing temperature: (a) in distribution: a full year healthy test set with both models trained with a full year of healthy data (b) out of distribution: a winter healthy test set with both models trained using healthy summer data. The numbers in brackets are the calculated calibration errors ϵ_{cal}. In both cases the NLL-ensemble model is better calibrated than the MSE-ensemble, and achieves a very low calibration error in distribution.

present in the data (aleatoric uncertainty) and the ensembling over these probabilistic predictions approximates the model uncertainty (epistemic uncertainty). We choose to contrast these two approaches for UQ, despite the inherent difference between them, as the former has been widely used and even claimed in the past to outperform other UQ methods for various applications (Lakshminarayanan et al., 2017).

3.1. Uncertainty Calibration Curves

To assess the calibration level of an uncertainty quantification method we use calibration curves (Kuleshov et al., 2018; Tran et al., 2020). A calibration curve compares the true fraction of points in a given confidence interval with the predicted fraction of points in that interval. Following (Kuleshov et al., 2018), for a given test data set $t = 1...T$ we choose n confidence levels $0 \leq p_1 < p_2 < ... < p_n \leq 1$ and calculate for each threshold p_j the empirical fraction of true values below it,

$$\hat{p}_j = \frac{\sum_{t=1}^{T} \mathbb{I}\left\{y_t \leq F_t^{-1}(p)\right\}}{T}. \tag{5}$$

The calibration curve is composed of the pairs $\{(p_j, \hat{p}_j)\}_{j=1}^{n}$. To further quantify the comparison we calculate the calibration error (Kuleshov et al., 2018)

$$\epsilon_{\text{cal}} = \sum_{j=1}^{n} (p_j - \hat{p}_j)^2. \tag{6}$$

Figure 1 shows the calibration curves (including Wilson confidence intervals) for the gearbox bearing temperature predictions of a wind turbine during periods of healthy condition (no faults). The calibration levels of the two UQ methods, MSE- and NLL-ensemble, are compared, with the calibration errors ϵ_{cal} given in brackets in the legend. In panel (a) the models were trained with data from a full year and the curves were calculated for a time period of one full year. The results demonstrate the clear advantage of UQ using the NLL-ensemble approach which seems to be very well calibrated, with a calibration error of 0.01 (compared to 0.31 for the MSE-ensemble). Note that the shape of the MSE-ensemble curve indicates that this quantification tends to be over-confident, for which the true y_t often falls outside the expected confidence band. The NLL-ensemble method, on the other hand, tends towards a slight under-confidence.

Figure 1(b) repeats the comparison in an Out-of-Distribution (OoD) scenario. It is important to clarify the meaning of OoD in the context of our fault detection task. A common scenario in fault detection applications is that not all healthy (normal) operating conditions are observed during training. As a result, some of these conditions may be detected as anomalies during deployment, only because they are out of the training distribution. Here we use the term OoD to describe these normal operating conditions that *have not been observed during training but should not be detected as anomalies*. In order to emulate such a scenario, we intentionally remove part of the operational conditions from our training set, and introduce these conditions only at testing. Thus, in Figure 1(b) the models are trained with only 3 months of summer data and the calibration curves are calculated on 3 months in winter, where both periods are known to be normal with no anomalies. Since the test data here is clearly OoD (this will be

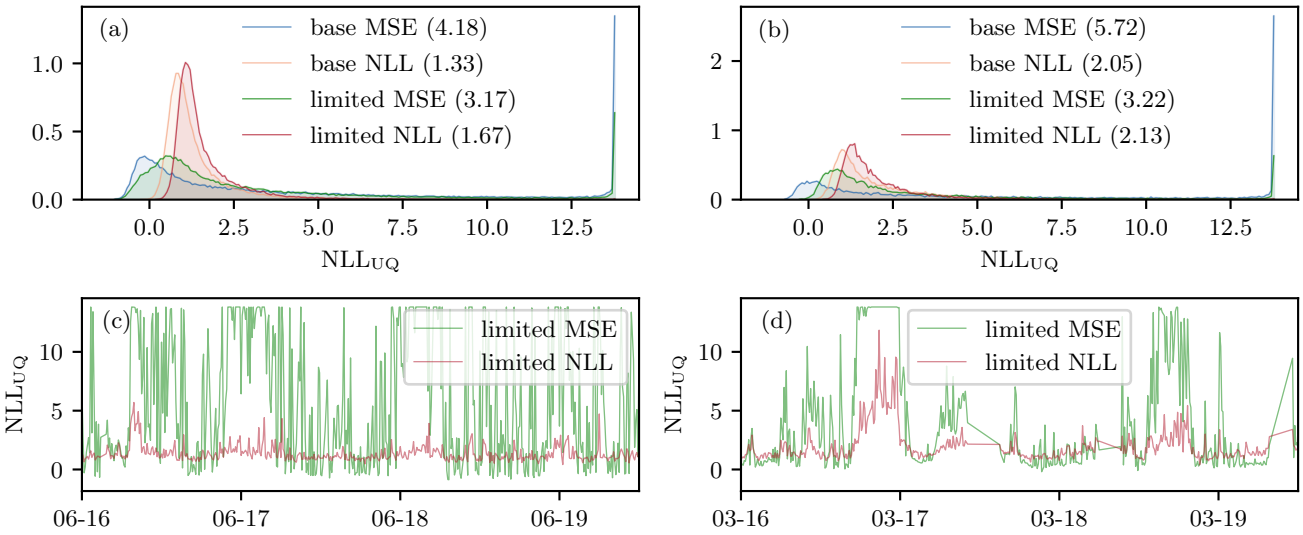

Figure 2. Comparison of the calculated Negative Log Likelihood $\mathrm{NLL_{UQ}}$ of the uncertainty quantification for the gearbox bearing temperature prediction. (a) Distributions of the $\mathrm{NLL_{UQ}}$ score during one year testing for 4 different models: base MSE is the CNN ensemble with the MSE loss trained with a full year data; base NLL uses the NLL loss with 1 year training data; limited MSE uses the MSE loss trained with 3 months of summer data; limited NLL uses the NLL loss with 3 months summer data for training. The number in brackets is the mean $\mathrm{NLL_{UQ}}$ over the entire test period. (b) The same as (a) just for 3 months of test data. The test data is healthy winter data, that is out of the training distribution (OoD) for "limited" training case. (c) An example of the $\mathrm{NLL_{UQ}}$ vs. time during summer (in distribution). (d) An example of the $\mathrm{NLL_{UQ}}$ vs. time during winter (OoD). The NLL-ensemble generally obtains lower values than the MSE-ensemble when both are trained with limited summer data.

demonstrated again below through increased OoD residuals in Figure 4), the UQ of both models is less well calibrated and both suffer from over-confidence. However, the NLL-ensemble is clearly better calibrated than the MSE-ensemble even in the OoD case.

In the examples throughout the paper we chose to train models on summer data and test them on winter data. The opposite case was observed by us to be less interesting since the domain shift seemed to affect the prediction results in a milder manner, such that the OoD effect was less pronounced.

3.2. Likelihood-based UQ Assessment

Another way to assess the usefulness of the different UQ methods is using an NLL-like score (Tran et al., 2020) on a test set. Every UQ method is used to estimate a probability distribution at each point t, which we approximate to be Gaussian and denote with $N(\tilde{\mu}_t, \tilde{\sigma}_t)$. We emphasize the distinction of the estimated mean $\tilde{\mu}_t$ and uncertainty $\tilde{\sigma}_t$ from the mean $\hat{\mu}_t$ and variance $\hat{\sigma}_t^2$ predicted directly by a probabilistic model as the two network outputs (in case of an NLL loss function). Whereas the latter are used to define the conventional NLL loss, we use the former in order to define an NLL-like score that quantifies the usefulness of the uncertainty measure of the method and denote it by $\mathrm{NLL_{UQ}}$:

$$\mathrm{NLL_{UQ}}(t) = -\log P(y_t | N(\tilde{\mu}_t, \tilde{\sigma}_t^2)) \tag{7}$$

For each UQ method we plug in the definitions of $\tilde{\mu}_t$ and $\tilde{\sigma}_t^2$, either from Eqn. 3 or from Eqn. 4.

The $\mathrm{NLL_{UQ}}$ measure is influenced by the predictive accuracy as well as the quality of its UQ (Tran et al., 2020). For a given test set, a lower $\mathrm{NLL_{UQ}}$ value indicates a better combination of prediction accuracy and reliable uncertainty quantification.

Figure 2 compares the uncertainty measures in terms of their $\mathrm{NLL_{UQ}}$ score on test data. Panels (a) and (b) display the empirical distribution of the scores over the test period. In panel (a) the test period is a full year of 10-minute resolution SCADA data from the wind turbine, whereas in panel (b) the test data are 3 months of winter data. In all regimes we selected for this evaluation training and test data without anomalies (healthy data). In each of these panels 4 distributions are displayed: for each of the UQ methods, MSE-ensemble and NLL-ensemble, we train the CNN with a full year data (base) or with summer data (limited) and plot the resulting 4 distributions of the $\mathrm{NLL_{UQ}}$ scores for the test set. The numbers in brackets are the mean $\mathrm{NLL_{UQ}}$ over the test set.

The most pronounced observation from the empirical distributions is the high score peaks of the MSE-ensemble model in all 4 cases (base and limited training with both test sets). The high negative log likelihood scores are indicative of test points with true values which lie at the extreme tail of the

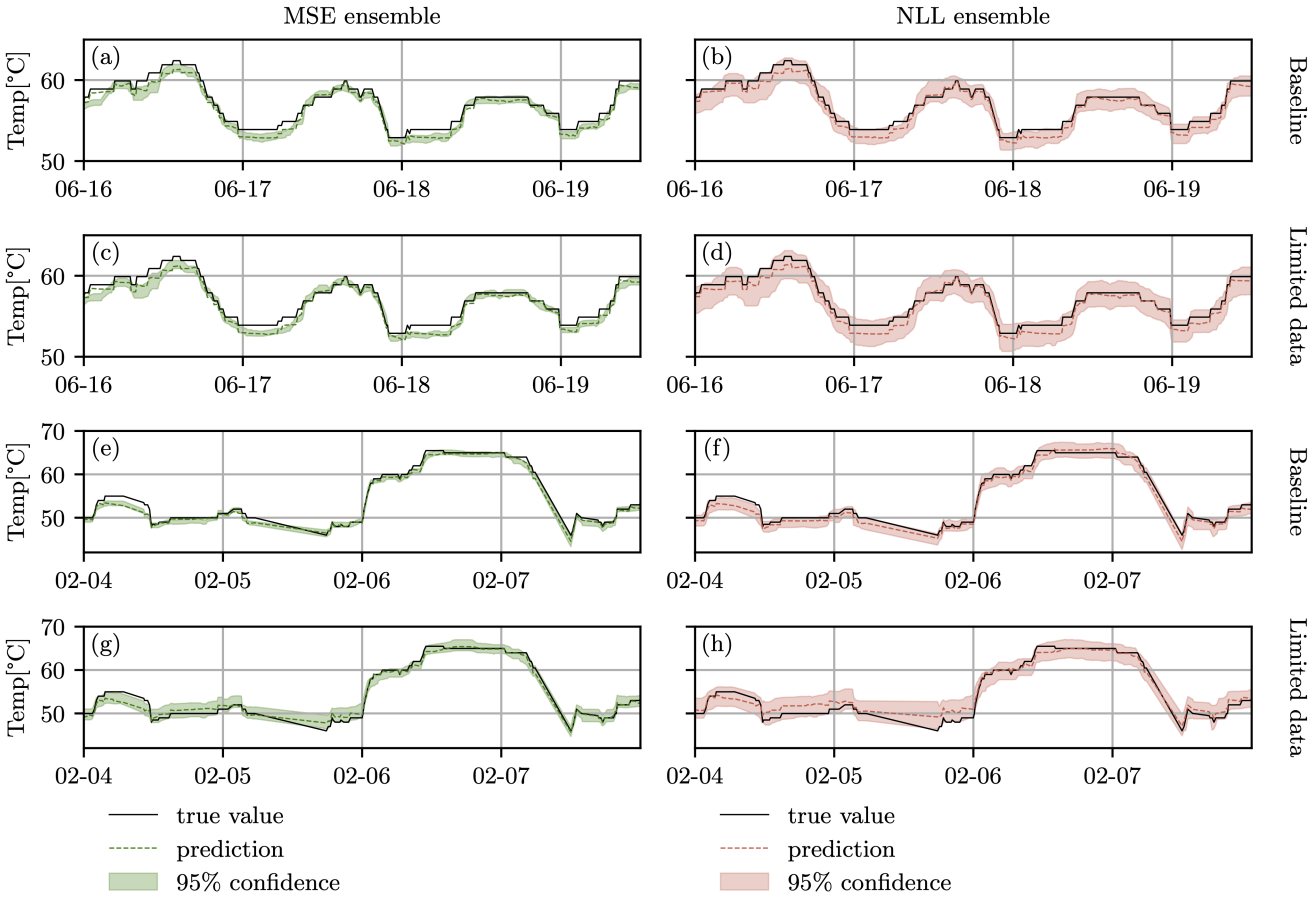

Figure 3. Confidence bands based on different uncertainty quantification (UQ) methods. Three days of the CNN ensemble predictions of the wind turbine gearbox bearing temperature (dashed) together with 95% confidence bands are compared for two different UQ methods: deep MSE ensemble (left), and deep NLL ensemble (right). The true values y_t are shown in solid lines, and the predicted mean $\tilde{\mu}_t$ as dashed lines. Whereas for the MSE ensemble, true values often lie outside the predicted confidence band, the NLL ensemble clearly provides more reliable UQ, for which around 95% of the measured values lie within the predicted 95% confidence band. This improved calibration of the NLL ensemble is seen also for the case of limited training data (2nd and 4th row) and not only with a baseline of full year training data (1st and 3rd row). The two upper rows (a)-(d) show an example from summer (similar to the training conditions) whereas the two lower rows (e)-(h) show an example from winter (OoD). In both cases we chose time slots of normal (healthy) turbine conditions.

predicted uncertainty distribution $N(\tilde{\mu}_t, \tilde{\sigma}_t)$. In these cases the uncertainty estimated by $\tilde{\sigma}_t$ is too small to explain the measured value y_t given the estimated predicted value $\tilde{\mu}_t$. In other words, the peaks result from test points with strongly over-confident predictions. The over-confident predictions are characteristic of the MSE-ensemble based UQ in an "in distribution" scenario, depicted in blue (base MSE) in panel (a) (that is, when both the training and the test data cover a full year). However, this is also the case "out of distribution", when the model is trained with summer data and tested in winter (green distribution in panel (b)).

As opposed to the MSE ensemble, the UQ based on the NLL-ensemble does not suffer from strongly over-confident predictions that lead to the high $\mathrm{NLL}_{\mathrm{UQ}}$ peaks. The advantage of the NLL ensemble for UQ is also evident through the lower mean $\mathrm{NLL}_{\mathrm{UQ}}$ values (in brackets), both in distribution (panel (a) base models) and out of distribution (panel (b) limited models). We stress again that the term "out of distribution" is used here to describe normal (not anomalous) regimes which are not observed during training.

A direct comparison between the two UQ models is demonstrated in Figure 2(c) and (d). The $\mathrm{NLL}_{\mathrm{UQ}}$ score is plotted against time for a period of 3 days using the MSE and NLL models trained with 3 months of summer data. We note that the capped values in the plots results from a regularization constant to avoid exploding logarithms. The NLL ensemble model reaches considerably lower scores as it does not suffer from over confident predictions.

This fact is visualized clearly in Figure 3. Here the 95% confidence bands around the predicted values (dashed lines) are contrasted with the true values (solid lines) of the gearbox bearing temperature of the turbine. The left and right columns of plots display the results using the MSE-ensemble and NLL-ensemble based UQ respectively, with panels (a)-(d) showing summer test data and panels (e)-(h) showing winter test data for both baseline (1 year) and limited data (3 summer months) training. Here it is clearly seen that the MSE-ensemble is strongly over confident in all regimes except OoD (panel (g)), where it is only lightly over-confident. Over-confident behaviour is easy to identify whenever the true values lie considerably outside the 95% confidence band. In a calibrated model this is expected to happen approximately 95% of the time. However, the MSE-ensemble model suffers from this considerably more often. In contrast to this, the NLL-ensemble method (right column) demonstrates almost no cases of true values well outside the confidence band, which is consistent with our observation that this model is well calibrated. The sharpness of the UQ of this model can also be observed here: the predicted uncertainty is repeatedly higher in periods of high prediction errors and lower in periods of low prediction errors.

After having demonstrated the high calibration level of the NLL-ensemble UQ, in the next section we use this UQ method to derive an uncertainty-informed anomaly score for the fault detection task.

4. Uncertainty Informed Anomaly Score

In order to benefit from UQ for more accurate and robust anomaly detection, we suggest to incorporate the uncertainty information inside the anomaly score assigned to every prediction. In this way, the anomaly score is not based on the prediction residual alone, but takes into account the confidence (or uncertainty) of the prediction when assigning an anomaly score to a point. As a natural extension of the conventional anomaly score we described in Section 2, we define the uncertainty-informed (UI) score at step t to be the predicted CDF, evaluated at the true value y_t,

$$S_t^{(UI)} = F(y_t; \tilde{\mu}_t, \tilde{\sigma}_t). \tag{8}$$

where $\tilde{\mu}_t$ and $\tilde{\sigma}_t$ depend on the selected UQ method. In this case, as shown in Section 3, the NLL ensemble model provides a calibrated UQ. We thus use Eqns. 4 to calculate $\tilde{\mu}_t$ and $\tilde{\sigma}_t$ for the anomaly score $S_t^{(UI)}$. Here, as well, the score is bounded between 0 and 1, and the higher it is, the more likely it is to indicate an anomaly. The threshold can be set similarly to the conventional score, in terms of the parameter α, where $S_t^{(UI)} > 1 - \alpha$ is detected as an anomaly (a fault).

In the following we compare the performance of two anomaly scores; the conventional score of Eqn. 1 and the uncertainty informed anomaly score of Eqn. 8.

Figure 4 compares the different scores as function of time for fault detection in the gearbox bearing temperature of a wind turbine. To elucidate the effect of UQ, we compare the performance using three anomaly scores: (i) the standard score $S_t^{(0)}$ of Eqn. 1 using an ensemble mean predictions of MSE-based CNNs (ii) a score based on the aleatoric uncertainty:

$$S_t^{(alea)} = F(y_t; \hat{\mu}_t, \hat{\sigma}_t). \tag{9}$$

(iii) the uncertainty-informed score $S_t^{(UI)}$ of Eqn. 8.

Every panel in Figure 4 displays the prediction residuals of the gearbox bearing temperature as function of time. Each point is colored according to its anomaly score, where blue indicates normal and red faulty given a detection threshold. In this example the significance threshold for fault detection was set on $\alpha = 0.0001$ for all methods. Panels (a)-(c) show the results achieved with a training set of a full year (marked in green shade). Panels (d)-(f) were trained with summer data only. As a result we observe a strong seasonality of the residuals, which tend to be considerably higher in winter, that is OoD, even in the absence of true faults. Panels (a) and (d) display the results of the MSE ensemble using the standard anomaly score $S_t^{(0)}$ derived using the training distribution of

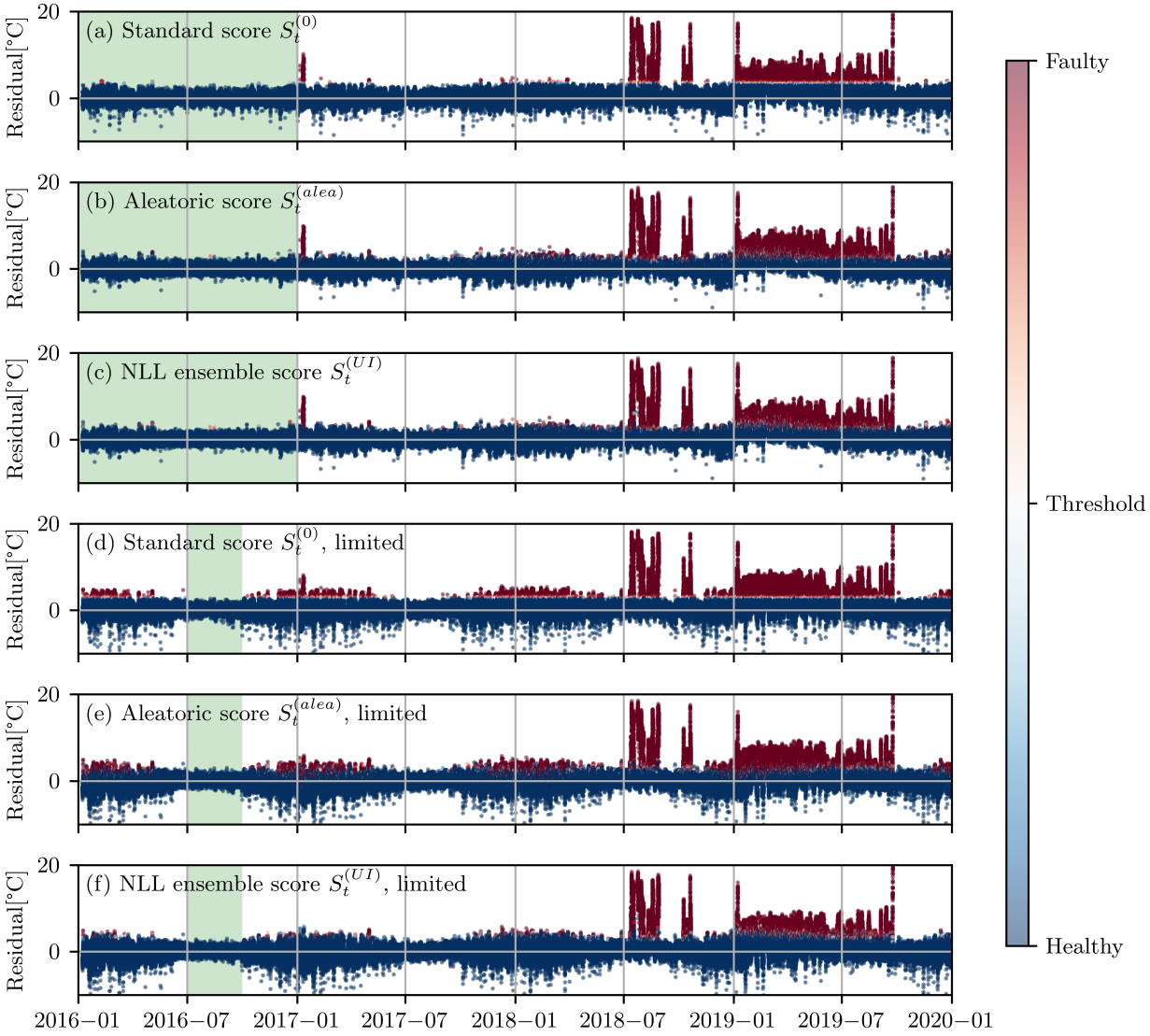

Figure 4. Comparing anomaly scores with and without uncertainty information. The prediction residuals of the gearbox bearing temperature are plotted during 4 years. In panels (a)-(c) the first year was used to train the CNN, whereas in panels (d)-(f) only three summer months were used for training, leading to strongly periodic residuals. The training period is shaded green. Three types of anomaly scores are compared: the standard score (a and d), the aleatoric score (b and e) and the NLL ensemble score (c and f). The same significance threshold $\alpha = 0.0001$ was used for all plots.

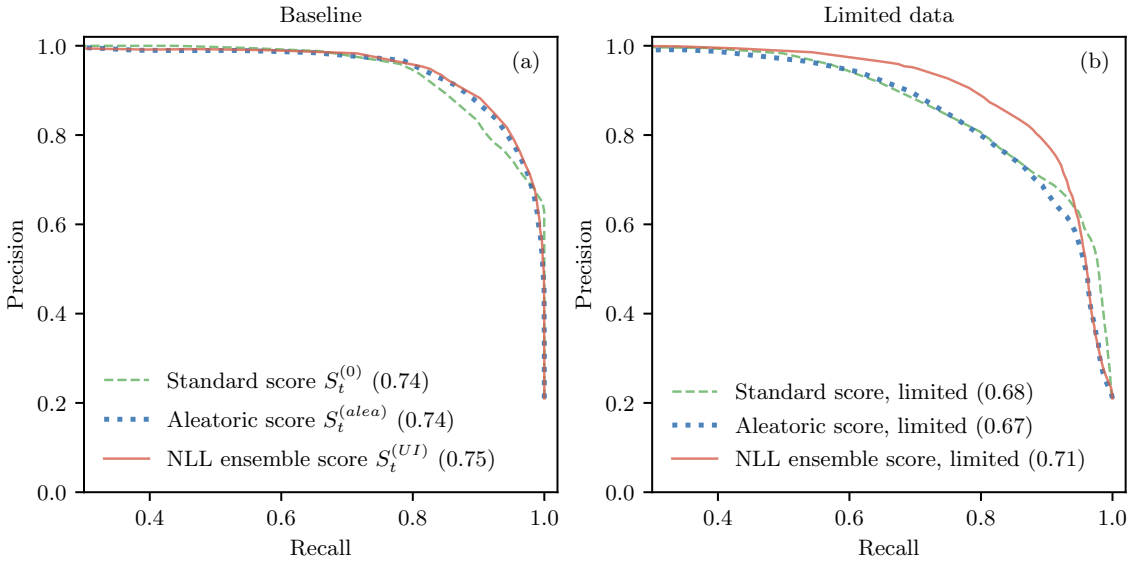

Figure 5. Fault detection performance of anomaly scores with and without uncertainty information. Three different anomaly scores are compared for two training sets: (a) one year training data (b) training data limited to 3 months from summer only. In the latter case, the test data includes winter data, which is outside the training distribution, both in healthy and in faulty conditions.

the ensemble mean residuals (see Eqn. 1). In panels (b) and (e) the residuals are the NLL model residuals based on a single realization from the ensemble. The anomaly score is the aleatoric score $S_t^{(alea)}$. In panels (c) and (f) the residuals are the ensemble mean prediction errors of the NLL model and the anomaly score is the uncertainty informed score $S_t^{(UI)}$ of Eqn.8.

It is evident that in the baseline scenrio of panels (a)-(c), where the models were trained with a full year data, representing all operational conditions, the differences in performance of the three anomaly scores are small. However, when we test the scores under the OoD scenario, where only partial data was used for training we realize the need to compensate for the biased model residuals during normal periods out of distribution (i.e not during summer). The models are prone to high residuals during winter which often lead to false positives (red points) under normal conditions. Such false alarms should be avoided, as they can lead to high costs related to unnecessary downtimes and maintenance. The majority of false positives OoD is indeed avoided when the uncertainty-informed anomaly score $S_t^{(UI)}$ is used (panel (f)). OoD predictions are typically characterized by a high prediction uncertainty, and thus a wide predictive distribution. They are detected as anomalies only if their residual is large enough to reach the tail of the distribution. In this way, most of the false positives of the standard anomaly score (panel (d)) are avoided if we use the uncertainty informed score of panel (f). As expected, the score $S_t^{(alea)}$, based on the aleatoric part of the uncertainty only, does not assess the epistemic uncertainty, and thus provides over-confident predictions OoD whose distribution is not wide enough to avoid the false positives in winter.

In order to quantify the performance of the different anomaly scores irrespective of a specific threshold, we plot their precision recall curves in Figure 5. In the absence of true normal/ abnormal labels we use the baseline MSE-ensemble model as a reference, and assign the label "faulty" to predictions with an ensemble mean residual above the 95%. We observe that even with this bias in favour of the MSE-ensemble model, the NLL-ensemble score outperforms it in its fault detection fidelity.

Figure 5(a) displays the precision-recall curves of the three anomaly scores for the "in distribution" scenario, in which training data from a full year was used. In this case the performance of all scores is similar, with a slight advantage for the uncertainty-informed methods. As expected, in distribution the aleatoric score $S_t^{(alea)}$ and the fully informed score $S_t^{(UI)}$ are very similar, with similar average precision (AP) shown in brackets for each score type. On the other hand, panel (b) represents the performance OoD, since the training data is limited to normal summer data only whereas the test data includes data from the entire year. In this case, there is a clear advantage to the NLL-ensemble score (solid red), with $AP = 0.71$ vs. $AP = 0.68$ of the standard score (dashed green). The aleatoric uncertainty informed score (blue dotted) is clearly under-performing OoD, as the epistemic part

of the uncertainty is crucial in this regime.

In summary, uncertainty quantification of the DL prediction model enables us to derive an uncertainty-informed anomaly score and assign it to each new prediction. The new score outperforms the conventional anomaly score based on the entire training distribution. The advantage is more pronounced for unknown test data, which lies outside the training distribution. In case the test data is normal, the uncertainty informed anomaly score accounts for the high uncertainty in this regime and thus avoids assigning false positives, as opposed to the conventional anomaly score.

5. Conclusion

In this paper we introduced an uncertainty informed anomaly score, which combines the information about the prediction residual together with the prediction uncertainty into a single scalar score assigned to each prediction. The uncertainty quantification is derived using a deep ensemble of probabilistic CNNs. We demonstrated the usefulness of the uncertainty-informed score for time series anomaly detection for wind-turbine condition-based maintenance, and showed its high performance compared to conventional uncertainty-agnostic anomaly scores. The advantage is particularly clear under a distribution shift of the healthy test data with respect to the healthy training set. This situation is quite common in PHM applications, where the training data often covers only part of the normal operating conditions expected during deployment. Thus, an approach that can reduce the false alarm rate in these cases is of high relevance. However, since our approach is generic, it can be applied to anomaly detection models trained with healthy (normal) data in any application field and is not limited to time series nor to PHM applications. A central advantage of the uncertainty-informed score is that a health index can be assigned at each and every time step. This is in contrast to the common uncertainty-aware classification methods that suggest to discard high-uncertainty predictions altogether. We believe that this approach provides a systematic and transparent way to include uncertainty in deep learning algorithms for anomaly detection, increasing their reliability in practical applications.

Acknowledgment

The authors would like to thank Beate Sick for her useful advice. The research was funded by Innosuisse - Swiss Innovation Agency under grant No. 32513.1 IP-ICT.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., ... others (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, *76*, 243–297.

Biggio, L., Wieland, A., Chao, M. A., Kastanis, I., & Fink, O. (2021). Uncertainty-aware prognosis via deep gaussian process. *IEEE Access*, *9*, 123517–123527.

Cai, M., Lu, F., & Sato, Y. (2020). Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 14392–14401).

Clifton, D. A., Tarassenko, L., McGrogan, N., King, D., King, S., & Anuzis, P. (2008). Bayesian extreme value statistics for novelty detection in gas-turbine engines. In *2008 ieee aerospace conference* (pp. 1–11).

Dürr, O., Sick, B., & Murina, E. (2020). *Probabilistic deep learning: With python, keras and tensorflow probability*. Manning Publications.

Fink, O., Wang, Q., Svensen, M., Dersin, P., Lee, W.-J., & Ducoffe, M. (2020). Potential, challenges and future directions for deep learning in prognostics and health management applications. *Engineering Applications of Artificial Intelligence*, *92*, 103678.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050–1059).

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., ... others (2021). A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*.

He, Y., Zhu, C., Wang, J., Savvides, M., & Zhang, X. (2019). Bounding box regression with uncertainty for accurate object detection. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 2888–2897).

Herzog, L., Murina, E., Dürr, O., Wegener, S., & Sick, B. (2020). Integrating uncertainty in deep neural networks for mri based stroke analysis. *Medical Image Analysis*, *65*, 101790.

Kraus, F., & Dietmayer, K. (2019). Uncertainty estimation in one-stage object detection. In *2019 ieee intelligent transportation systems conference (itsc)* (pp. 53–60).

Kuleshov, V., Fenner, N., & Ermon, S. (2018). Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning* (pp. 2796–2804).

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, *30*.

Laptev, N., Yosinski, J., Li, L. E., & Smyl, S. (2017). Time-series extreme event forecasting with neural networks at uber. In *International conference on machine learning* (Vol. 34, pp. 1–5).

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., & Wahl,

S. (2017). Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, *7*(1), 1–14.

Levi, D., Gispan, L., Giladi, N., & Fetaya, E. (2019). Evaluating and calibrating uncertainty prediction in regression tasks. *arXiv preprint arXiv:1905.11659*.

Miller, D., Dayoub, F., Milford, M., & Sünderhauf, N. (2019). Evaluating merging strategies for sampling-based uncertainty techniques in object detection. In *2019 international conference on robotics and automation (icra)* (pp. 2348–2354).

Sato, K., Hama, K., Matsubara, T., & Uehara, K. (2019). Predictable uncertainty-aware unsupervised deep anomaly segmentation. In *2019 international joint conference on neural networks (ijcnn)* (pp. 1–7).

Schwaiger, A., Sinhamahapatra, P., Gansloser, J., & Roscher, K. (2020). Is uncertainty quantification in deep learning sufficient for out-of-distribution detection? In *Aisafety@ ijcai*.

Seeböck, P., Orlando, J. I., Schlegl, T., Waldstein, S. M., Bogunović, H., Klimscha, S., ... Schmidt-Erfurth, U. (2019). Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal oct. *IEEE transactions on medical imaging*, *39*(1), 87–98.

Tautz-Weinert, J., & Watson, S. (2016). Using scada data for wind turbine condition monitoring–a review. *IET Renewable Power Generation*, *11*(4), 382–394.

Tran, K., Neiswanger, W., Yoon, J., Zhang, Q., Xing, E., & Ulissi, Z. W. (2020). Methods for comparing uncertainty quantifications for material property predictions. *Machine Learning: Science and Technology*, *1*(2), 025006.

Ulmer, M., Jarlskog, E., Pizza, G., Manninen, J., & Goren Huber, L. (2020). Early fault detection based on wind turbine scada data using convolutional neural networks. In *Proceedings of the european conference of the phm society* (Vol. 5).