

# Investigation of the Calibration Robustness of Deep Neural Networks for Optical Quality Monitoring

Untersuchung der Kalibrierungsrobustheit von tiefen  
neuronalen Netzen für die optische  
Qualitätsüberwachung

## Interdisciplinary Project

at the TUM School of Engineering and Design of the Technical University of Munich

<b>Supervised by</b>	Prof. Dr.-Ing. Rüdiger Daub Chair for Production Technology and Energy Storage Systems
<b>Submitted by</b>	Chen Danqing Hararestraße 4 81735 München
<b>Submitted on</b>	April 6, 2024 in Garching



## Scope of Work

**Title of the Interdisciplinary Project:***Investigation of the Calibration Robustness of Deep Neural Networks for Optical Quality Monitoring***Inv.- Nr.:** 2023/0000**Author:** **Danqing Chen****Issued:** 15.09.2023**Supervisor:****Johannes Bauer****Submitted:**

06.04.2024

**Motivation:**

Deep Neural Networks (DNNs) show high potential for automated quality monitoring of various manufacturing processes (e. g., defect classification on weld seam images). However, the limited availability of data in the manufacturing domain and data drift phenomena make it difficult to achieve and maintain the required prediction performance. Data drift occurs when the distribution of incoming data changes, compared to the distribution of the data the DNN model was originally trained on. This results in decreased performance of the model, jeopardizing the reliability of predictions. One way to automatically assess the expected reliability of such predictions are calibration methods. Calibration empowers DNN models to not only make predictions, but also to assign confidence scores that reflect their anticipated precision. Reliable calibration therefore allows to avoid faulty predictions by only accepting predictions with a sufficiently high confidence score. Additionally, by consistently tracking changes in confidence scores over time, the model becomes an effective tool for detecting shifts in data patterns and concept drift scenarios. However, this would require the calibration of a model to be robust to concept drift. On the one hand, a decline in model performance is usually also associated with weakened calibration accuracy. On the other hand, contemporary research shows that Transformer-based models exhibit a certain resilience to data drift, providing an encouraging direction for further investigation.

**Objective:**

The goal of this project is to investigate the robustness of calibration methods, with a focus on quality monitoring tasks. The project aims to evaluate different combinations of DNN architectures and calibration methods across various datasets, encompassing both public benchmarks as well as datasets from the manufacturing context. Identification of common patterns may allow it to identify methods that show good robustness across different datasets and therefore leverage the potential of a well-calibrated model. At first, relevant datasets,

models, and calibration approaches will be identified in current scientific literature. Afterwards selected ones will be implemented, systematically evaluated, and discussed.

### **Procedure and Working Methodology:**

- Conduction of a semi-structured literature review on the following topics
  - State-of-the-art model architectures for quality monitoring in manufacturing
  - Calibration methods for DNN
  - Popular benchmark datasets for model robustness assessment
- Selection of methods
  - Selection of datasets
  - Selection of model architecture
  - Selection of calibration methods
- Implementation of selected model architectures and calibration methods
- Evaluation of robustness to concept drift on the selected datasets
- Development of a method for data drift detection

### **Agreement:**

Through the supervision of B. Sc. Danqing Chen intellectual property of the *iwb* is incorporated in this work. Publication of the work or transfer to third parties requires the permission of the chair holder. I agree to the archiving of the work in the library of the *iwb*, which is only accessible to *iwb* staff, as inventory and in the digital student research project database of the *iwb* as a PDF document.

Garching, 08.09.2023

Prof. Dr.-Ing.  
Rüdiger Daub

B. Sc.  
Chen Danqing

## **Abstract**

In the field of manufacturing, flexibility is getting increasingly important and products and production process frequently change. This increases the need for quality monitoring approaches. Deep Neural Network (DNN) has demonstrated significant potential in this domain, e.g, in the automated detection of defects in weld seam images. Despite their promise, DNNs confront challenges associated with limited data availability and the phenomenon of data drift—a shift in the distribution of process data over time—which can undermine model accuracy and the reliability of quality assessments. This project, conducted at the Institute for Machine Tools and Industrial Management (iwb), aims to investigate the robustness of DNN calibration methods against such data drifts, particularly in the context of optical quality monitoring in manufacturing. Calibration methods enable DNNs to provide confidence scores alongside predictions, reflecting the expected reliability of a prediction. Ensuring that these methods remain reliable, even when the underlying data patterns shift, is essential to maintain high-quality standards and avoid costly errors. The project will critically evaluate variety of DNN architectures and calibration techniques across multiple datasets, including public benchmarks and manufacturing-specific data. Through reviews of current scientific literature and practical experimentation, the project seeks to identify robust calibration methods that consistently predict with high confidence. This work therefore contributes to a more robust application of DNN with a focus on manufacturing environments.



# Contents

<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>x</b>
<b>List of Abbreviations</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Fundamentals and State of the Art</b>	<b>2</b>
2.1 Calibration Methods . . . . .	2
2.2 Post-Training vs. During-Training Calibration Methods . . . . .	2
2.3 Presentation of Relevant Calibration Approaches . . . . .	3
2.4 Calibration in Manufacturing Applications . . . . .	7
2.5 Summary and Need for Research . . . . .	7
<b>3 Datasets</b>	<b>9</b>
3.1 Friction Stir Welding (FSW) Dataset . . . . .	9
3.2 CIFAR-10 and CIFAR-10-C Dataset . . . . .	10
<b>4 Investigation of Robustness of Calibration Techniques</b>	<b>12</b>
4.1 Experiment with FSW Dataset . . . . .	12
4.1.1 Methodologies . . . . .	12
4.1.2 Results and Discussion . . . . .	13
4.2 Benchmarking with Public Dataset CIFAR-10 . . . . .	13
4.2.1 Methodologies . . . . .	14
4.2.2 Results and Discussion . . . . .	14
4.3 Combination of Calibration Methods with Swin Transformer on CIFAR-10 . . . .	16
<b>5 Summary and Outlook</b>	<b>19</b>
<b>A Digital Appendix</b>	<b>20</b>
<b>Bibliography</b>	<b>22</b>





## List of Figures

2.1	Density-Aware Calibration (DAC) combined with existing post-hoc methods from TOMANI et al. 2023 . . . . .	4
2.2	Illustration of mixup ZHANG et al. 2018 . . . . .	5
2.3	The illustration of training a DNN with online label smoothing method. ZHANG et al. 2021 . . . . .	6
2.4	Illustration of focal loss LIN et al. 2017 . . . . .	6
3.1	Illustration of friction stir welding MISHRA and MA 2005 . . . . .	9
3.2	Cutout with toe flash . . . . .	10
3.3	Cutout without toe flash . . . . .	10
4.1	Cifar dataset split . . . . .	14

# List of Tables

4.1	Calibration results for FSW dataset . . . . .	13
4.2	Calibration results for CIFAR dataset . . . . .	15
4.3	Calibration results for CIFAR and FSW datasets . . . . .	15
4.4	Calibration results for combined methods with Swin . . . . .	17

## List of Abbreviations

DAC	Density-Aware Calibration
DNN	Deep Neural Network
ECE	Expected Calibration Error
FSW	Friction Stir Welding
ID	In-distribution
iwb	Institute for Machine Tools and Industrial Management
OD	Out-of-distribution



# Chapter 1

## Introduction

In the current manufacturing landscape, manufacturers are increasingly confronted with the complexities of modern production processes and fluctuating market demands, highlighting the need for advanced quality monitoring solutions. DNN can be a powerful tool in quality monitoring EL BILALI et al. 2022, particularly in manufacturing. The use of DNNs in manufacturing is not new, for example: HARTL et al. 2019, in their work, DNN for optical quality monitoring is highlighted. In the work of EL BILALI et al. 2022, DNNs are used for monitoring water quality. In the work of TANAKA et al. 2022, they described a method to realize optical signal monitoring by training a DNN. The common methods of quality monitoring in industry heavily depend on manual inspection, which are well-known to be slow and error-prone ASHFAHANI et al. 2021. The integration of DNN can offer advanced analytic tools for analyzing manufacturing data WANG et al. 2018.

Data sampled from different distributions can lead to concept drift or "non-stationary learning" problems HOENS et al. 2012. This issue is particularly relevant in manufacturing, where process variations and new techniques can alter data distributions. As NANNAPANENI et al. 2016 suggests, the presence of uncertainty in different processes may lead to significant uncertainty in the overall performance prediction of DNNs, which could disrupt the manufacturing process, and future profits may not be realized SEIFFER et al. 2021.

Calibration techniques are key to enhancing the trustworthiness of DNNs. They offer a measure of confidence in the network's predictions, ensuring that the predicted probabilities match the actual likelihood of an outcome. A well-calibrated DNN remains accurate and reliable even when faced with new and changing data patterns, effectively handling data drift. This study investigates the applications and challenges of deploying DNNs in settings where the accuracy of the models is critical. It specifically focuses on the ability of DNNs to maintain calibration robustness in the manufacturing application domain when faced with data drift, as well as their performance on public benchmark datasets.

This project is conducted within the Institute for Machine Tools and Industrial Management (iwb), which focuses on research of advanced production technology, including friction stir welding. These processes are prone to quality variations and could benefit from enhanced monitoring techniques.

The project begins with a comprehensive review of existing literature on DNN architectures, calibration techniques, and model robustness. This review will guide the selection of methods and datasets for implementation and evaluation. The research will assess the robustness of various DNN architectures and calibration methods using datasets from both public benchmarks and specific manufacturing contexts. The aim is to identify calibration methods that demonstrate strong robustness and applicability.

## Chapter 2

# Fundamentals and State of the Art

In this chapter, we will delve into the various calibration techniques employed in the experiments. Each method will be introduced with a brief overview, accompanied by references to the relevant literature from which these techniques are derived. This comprehensive examination aims to provide a foundational understanding of the calibration methods under investigation.

### 2.1 Calibration Methods

In this project, several calibration techniques are adopted to enhance the reliability and interpretability of deep learning models. Calibration, in deep learning, is a critical process that adjusts a model's predicted probabilities to reflect observed outcome frequencies more accurately. The accuracy of these probabilities is essential for dependable decision-making. The project utilizes a range of calibration techniques, each designed to specifically target various aspects of model uncertainty and prediction accuracy. The following sections detail these methods and their respective roles within the project's scope.

### 2.2 Post-Training vs. During-Training Calibration Methods

Calibration methods for machine learning models can be broadly categorized into two groups: post-training, e.g: TOMANI et al. 2021 and during-training e.g: HEBBALAGUPPE et al. 2022. The primary difference between these two approaches lies in the timing and integration of the calibration process within the model's lifecycle.

Post-training calibration methods are applied after a model has been fully trained. These techniques adjust the model's output without altering its internal parameters. An example of post-training calibration is Temperature Scaling GUO et al. 2017 , which applies a single scalar to modify the logits of a model's outputs to better align predicted probabilities with actual outcomes.

During-training calibration, on the other hand, integrates calibration directly into the training process. Methods such as Label Smoothing SZEGEDY et al. 2016 and Focal Loss LIN et al. 2017 modify the loss function itself to encourage the model to produce more calibrated probabilities. These methods can help the model naturally adjust its confidence levels as it

learns. However, they may require careful hyperparameter tuning to ensure they enhance the model's performance without detriment. Recent deep neural networks largely rely on a wide range of hyperparameter choices about the neural network's architecture and regularization FEURER and HUTTER 2019.

Both approaches aim to mitigate the disconnection between a model's confidence in its predictions and the actual likelihood of those predictions being correct. The choice between post-training and during-training calibration methods depends on the specific requirements of the task, the computational resources available, and the flexibility of the training pipeline.

## 2.3 Presentation of Relevant Calibration Approaches

The DAC method, a post-training calibration method, as outlined by TOMANI et al. 2023, integrates density estimates into the calibration framework. This approach allows a model to modulate its confidence estimations relative to the density of data points within the feature space. The DAC method employs temperature scaling GUO et al. 2017 to fine-tune the softmax outputs of neural networks, seeking to preserve the model's accuracy while enhancing calibration.

Temperature scaling is utilized to adjust the classifier's logits  $\hat{Q}$ , defined by:

$$\hat{Q} = \sigma_{SM} \left( \frac{z^L}{T} \right) \quad (2.1)$$

where  $\sigma_{SM}$  is the softmax function,  $z^L$  represents the logits, and  $T$  is the temperature parameter.

For Density-Aware Calibration (DAC), logits are rescaled using a sample-dependent parameter  $S(x, w)$ :

$$\hat{Q}(x, w) = \sigma_{SM} \left( \frac{z^L}{S(x, w)} \right) \quad (2.2)$$

where  $S(x, w)$  is determined by the density of the data points in feature space, with  $x$  as input data.

The density-aware scaling factor  $S(x, w)$  is computed as a linear combination of density estimates:

$$S(x, w) = \sum_{l=1}^L w_l s_l + w_0 \quad (2.3)$$

where  $w_l$  are the weights for every layer  $l$  and  $w_0$  is a bias term, with the constraint that  $w_l \geq 0$ .

The density estimate  $s_l$  for every feature layer  $l$  is calculated using the k-nearest-neighbor distance in the feature space, an illustration of this can be found in (2.1).

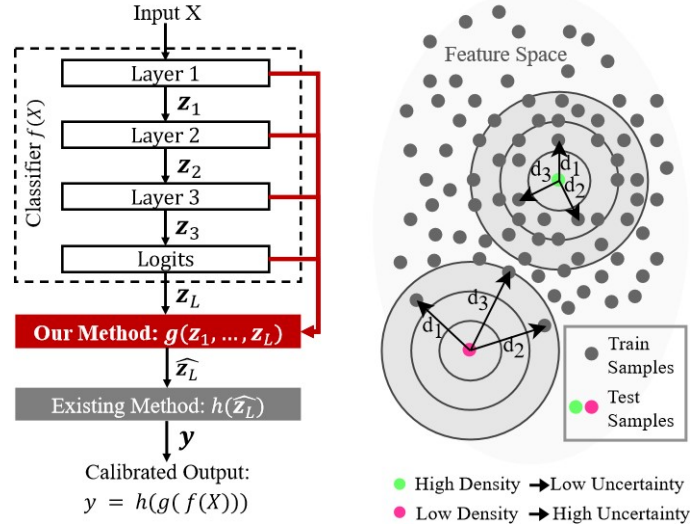
Gal and Ghahramani's interpretation of dropout as a Bayesian Approximation GAL and GHAHRAMANI 2016 introduces a methodological shift in its application.

The process can be mathematically represented as follows:

Given a neural network with dropout, let  $f^W(x)$  denote the output of the network for input  $x$  with a specific realization of weights  $W$ . For a dropout rate  $p$ , the predictive distribution is obtained by averaging over multiple stochastic forward passes:

$$p(y|x, \mathcal{D}) \approx \frac{1}{T} \sum_{t=1}^T f^{W_t}(x) \quad (2.4)$$

where  $\mathcal{D}$  is the training data,  $T$  is the number of stochastic forward passes, and  $W_t$  represents the weights in the  $t$ -th forward pass, with dropout applied. This procedure effectively simulates drawing from the posterior distribution of the weights, providing an approximation to Bayesian inference.



**Figure 2.1:** DAC combined with existing post-hoc methods from TOMANI et al. 2023

The mixup technique, as discussed by ZHANG et al. 2018, employs the combination of input pairs and their labels to foster model generalization. This technique encourages linear interpolations between training instances, which has implications for improved calibration and uncertainty estimation across the model's input space.

**Lambda Distribution:** The mixup parameter  $\lambda \in [0, 1]$ .

**Virtual Feature-Target Vector Creation:** For two randomly selected feature-target vectors  $(x_i, y_i)$  and  $(x_j, y_j)$ , the virtual feature-target vectors  $(\tilde{x}, \tilde{y})$  are created as follows:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j \quad (2.5)$$

$$\tilde{y} = \lambda y_i + (1 - \lambda) y_j \quad (2.6)$$

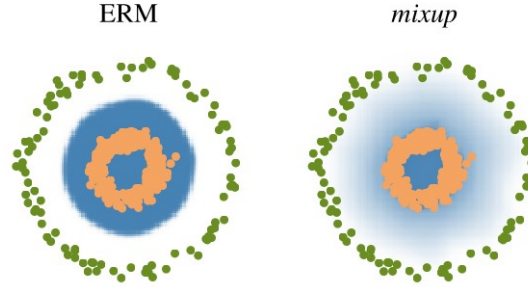
The core idea revolves around the concept of "mixup", where the features and targets of two randomly chosen data points are linearly combined.

Below is an illustration of Effect of mixup ( $\alpha = 1$ ) on a toy problem (2.2). Green: Class 0. Orange: Class 1. Blue shading indicates  $p(y = 1|x)$  ZHANG et al. 2018. The concentric circles (the blue) represent the decision boundaries or the areas where the model is predicting the likelihood of a point belonging to a particular class.

The orange inner circle can be seen as the area where the model is confident that the points belong to Class 1. The blue shading around the orange circle indicates the transitional area where the model's confidence in its prediction decreases from Class 1 to Class 0. The "mixup" technique is shown on the right side of the figure, and it demonstrates how the decision boundary becomes smoother and less distinct compared to traditional methods (shown as



the "ERM(Empirical Risk Minimization)" ). The result of using mixup is that the model learns from blended data, the model becomes less certain about the hard boundaries between classes and more adaptable to variations in the data. This approach helps the model not to be overly confident about its predictions, which is particularly useful when it encounters new data that may not be as clearly separated as the training data.



**Figure 2.2:** Illustration of mixup ZHANG et al. 2018

Investigations into label smoothing SZEGEDY et al. 2016, have revealed its utility as a calibrative tool. The process involves adjusting the training targets and is mathematically encapsulated as follows:

$$\text{LSR}(p(k|x), q'(k|x), \varepsilon) = - \sum_{k=1}^K [(1 - \varepsilon)\delta_{k,y} + \varepsilon u(k)] \log \left( \frac{\exp(z_k)}{\sum_{i=1}^K \exp(z_i)} \right) \quad (2.7)$$

Here,  $p(k|x)$  represents the probability of each label  $k$  for a given input  $x$ , computed using the softmax function on logits  $z_i$ . The ground-truth distribution  $q(k|x)$  is adjusted to  $q'(k|x)$  via label smoothing, with  $\varepsilon$  as the smoothing parameter,  $\delta_{k,y}$  as the Dirac delta function, and  $u(k)$  as a fixed distribution. This adjusted ground-truth distribution is then used in the cross entropy loss function, moderating the model's certainty and aiding in better delineation of classes.

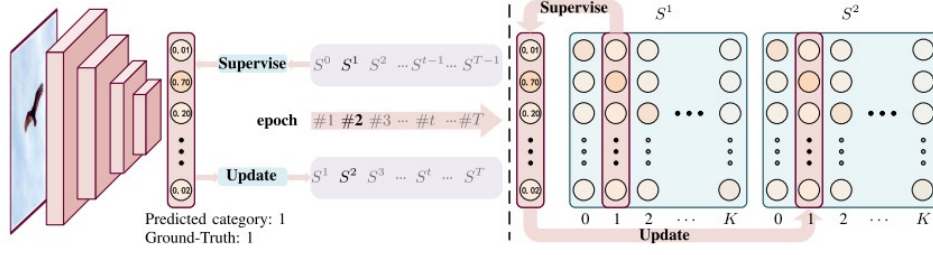
On the left side of the graph (2.3), you see a stack of pink layers with numbers. Each layer represents the computer's knowledge at each practice round. As it practices more (going from  $S^0$  to  $S^{(T-1)}$ ), it updates its guesses. On the right side, there are several columns (0 to  $K$ ), each representing a category (like 'cat', 'dog', etc.). As the computer practices, the colors in these columns change, showing how its guesses get updated. The darker the color, the more confident the guess.

The idea is that as the computer practices, it doesn't just learn from hard right or wrong answers but from these soft, uncertain labels which tell it how close its guess was to being right. This method helps the computer to become better at recognizing animals in pictures, even when new or unclear images are shown.

Focal loss, presented by LIN et al. 2017, addresses class imbalance by adapting the loss function to emphasize challenging, misclassified samples. In the graph (2.4) we can see that when the computer's prediction is very wrong, both of the operands yield very high value, however, as the computer starts guessing closer to the correct answer, the focal loss gets less compared to Cross Entropy, this encourages the computer to focus more on the difficult examples, where it is wrong.

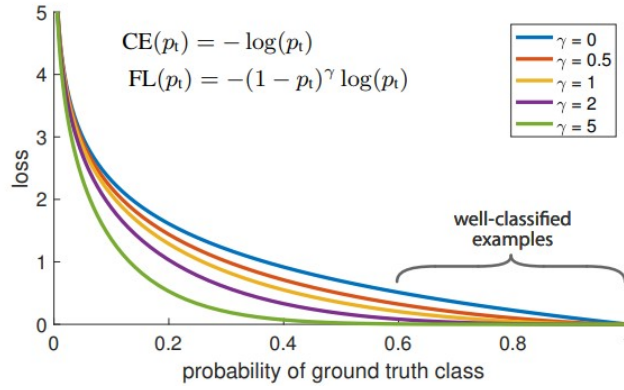
The focal loss is mathematically defined as:

$$\text{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (2.8)$$



**Figure 2.3:** The illustration of training a DNN with online label smoothing method. ZHANG et al. 2021

where  $p_t$  is the model's estimated probability for the class with the true label,  $\alpha_t$  is a weighting factor to balance the importance of different classes, and  $\gamma$  is the focusing parameter that smoothly adjusts the rate at which easy examples are down-weighted, allowing the model to focus more on hard examples.



**Figure 2.4:** Illustration of focal loss LIN et al. 2017

Lastly, the Correctness Ranking Loss (CRL) method, as introduced by MOON et al. 2020, incorporates confidence estimation into the training regimen. The model looks at pairs of predictions it has made. For each pair, it compares the confidence levels. If one prediction in the pair is correct more often than the other, the model should reflect this by having a higher confidence level for that prediction. The CRL calculates a penalty (or loss) based on these comparisons. If the model isn't ranking its confidence correctly (for example, it's more confident about a less accurate prediction), it gets penalized. This penalty encourages the model to align its confidence levels with its actual performance. The Correctness Ranking Loss (CRL) is designed to enforce the desirable ordinal ranking of confidence estimates. For a pair of samples  $x_i$  and  $x_j$ , the CRL is defined as:

$$L_{CR}(x_i, x_j) = \max(0, -g(c_i, c_j)(\kappa_i - \kappa_j) + |c_i - c_j|) \quad (2.9)$$

where  $c_i$  is the proportion of correct prediction events for a sample  $x_i$  over the total number of examinations, The term  $\kappa_i$  represents  $\kappa(p_i|x_i, w)$ , a function or value derived from the model (one can think of it as the model's measure of how much it should pay attention to a particular sample). The function  $g(c_i, c_j)$  is critical in defining the relationship between the proportions of correct predictions for pairs of samples  $x_i$  and  $x_j$ . It is defined as follows:

$$g(c_i, c_j) = \begin{cases} 1 & \text{if } c_i > c_j \\ 0 & \text{if } c_i = c_j \\ -1 & \text{otherwise} \end{cases} \quad (2.10)$$

In summary, the aforementioned methods contribute distinct perspectives and techniques to the calibration of predictive models. Each method offers a unique mechanism by which a model's confidence in its predictions can be adjusted to reflect the underlying data more accurately, thereby enhancing the overall reliability and interpretability of its probabilistic outputs.

## 2.4 Calibration in Manufacturing Applications

The study by ROŽANEC et al. 2023 emphasizes the integration of advanced machine learning calibration techniques in quality control processes, in their paper, they mentioned that no model is perfect, therefore how to calibrate model and make the prediction score more reliable, is still a challenge ROŽANEC et al. 2023.

In the area of high-tech manufacturing products, even slight change of the product state during production can results in costly and time-consuming rework WUEST et al. 2023. Wuest and colleagues' research emphasized the way to increase process and product quality in manufacturing through supervised machine learning.

For applications in engineering design, understanding not only model error but model calibration is very important. In the computational mechanics community, there is already a lot of research done in studying uncertainty quantification and model calibration (ARENDRT et al. 2012, PSAROS et al. 2023, WANG et al. 2020). However, for deep learning models utilized for problems in mechanics in particular, which tend to have low model error but no guarantee of the model being well calibrated (GUO et al. 2017), this is a current research gap. MOHAMMADZADEH et al. 2023

## 2.5 Summary and Need for Research

Many models, especially complex ones like DNNs, can output over-confident or under-confident probabilities, when the input sample is out-of-distribution of the training data, or corrupted by noise LOQUERCIO et al. 2020. This problem arises because the model's primary goal during training is usually to maximize accuracy, not to ensure the predicted probabilities are statistically reliable and reflect true probabilities.

In this chapter, calibration methods in deep learning were examined to understand their practical implications, especially in manufacturing. Calibration plays a crucial role in ensuring machine learning models produce reliable predictions by aligning their confidence with actual outcomes. This alignment is of paramount importance in situations where prediction accuracy is critical.

Model calibration techniques can be broadly categorized into post-training and during-training methods. Post-training methods like Temperature Scaling adjust the model's confidence levels without modifying its internal workings. On the other hand, during-training methods, such as Label Smoothing, Focal Loss, and Correctness Ranking Loss (CRL), incorporate calibration directly into the training cycle, which can inherently improve the model's confidence in its predictions, but they do require careful adjustment of their parameters. Various innovative strategies like Density-Aware Calibration (DAC), Bayesian interpretations of dropout, and mixup further diversify the calibration toolkit, each with its unique way of bolstering a

model's certainty in its outputs. These strategies are exemplified in the literature by TOMANI et al. 2023 for DAC, GAL and GHARAMANI 2016 for Bayesian interpretations of dropout, ZHANG et al. 2018 for mixup, SZEGEDY et al. 2016 for label smoothing, LIN et al. 2017 for focal loss, and MOON et al. 2020 for CRL.

The variety of model calibration methods presents a challenge in reaching a consensus on which is best suited for a given scenario, especially when considering the prevalence of concept drift that can affect model performance over time. Deep neural network is a powerful tool for quality monitoring in manufacturing, but their effectiveness can be compromised by concept drift, therefore, this project embarks on a benchmark using two distinct datasets to scrutinize the robustness of calibration methods, bridging the gap between advanced research area (DNN) and practical application (quality monitoring).

## Chapter 3

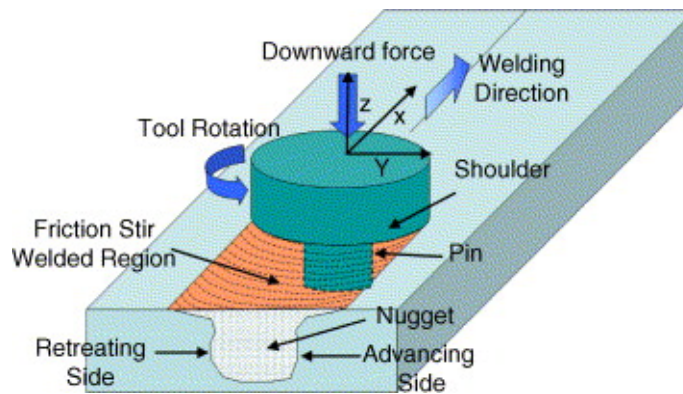
### Datasets

In this chapter, detail of the datasets employed in this study are discussed, which are central to the development and evaluation of the robustness of various calibration methods.

#### 3.1 Friction Stir Welding (FSW) Dataset

Since the exemplary use case for quality monitoring deals with an Friction Stir Welding (FSW) application, the basics of this process are explained in the following:

"Friction stir welding (FSW) is a solid-state welding technology. It is particularly suitable for joining metals with low melting temperatures such as aluminum and magnesium (an overview of the process can be seen in figure 3.1). The demand for high quality products and low-cost manufacturing has increased the need for the constant monitoring of the manufacturing process."HARTL et al. 2019



**Figure 3.1:** Illustration of friction stir welding MISHRA and MA 2005

Dataset from FSW dataset comprises a collection of 112 friction stir weld seam images. Each image has been subdivided into smaller cutouts, showcasing distinct sections of the weld seams. The weld seams are categorized into seven distinct groups based on the welding parameters employed during their fabrication, such as tool geometry. However, for weld seams numbered 17 and 18, group information is unavailable, which necessitates their exclusion from the dataset, reducing the number from 114 to 112.

The grouping facilitates the observation of concept drift when training on images from a subset of groups (1 to  $n$ , where  $n < 7$ ) and testing on the remaining groups. Labels for the dataset have been assigned based on various welding characteristics, including the temperature during welding, the presence of toe flash (*Grat* in German), and surface galling, which refers to small irregularities on the seam surface.

For data filtering and splitting for validation, the data from groups 1, 2, and 3 are filtered for cross-validation purposes, and the data from groups 5, 6, and 7 are used for Out-of-distribution (OD) data, which is limited to the first 500 entries. Cross-validation is a technique for assessing how the statistical analysis will generalize to an independent dataset. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.

For binary classification tasks, the default defect type is toe flash/grat, with the labels being "OK" for acceptable quality and "not OK" for defects, with an approximately equal distribution across these classes.

1. **toe\_flash**: with "OK" and "NOT OK" as possible classes.
2. **surface\_fault**: also with "OK" and "NOT OK" classes, but it is essential to note that this label is highly subjective and requires meticulous verification.

Below are 2 images 3.2, 3.3 displaying an image without defect and another with toe flash.



**Figure 3.2:** Cutout with toe flash



**Figure 3.3:** Cutout without toe flash

This dataset curated by HARTL et al. 2019 provides a comprehensive basis for training and assessing deep learning models aimed at the automated visual inspection of friction stir welds, contributing significantly to the field of quality monitoring in manufacturing processes.

### 3.2 CIFAR-10 and CIFAR-10-C Dataset

The CIFAR-10 dataset consists of 60,000 32x32 color images in 10 different classes, with 6,000 images per class. The dataset is divided into 50,000 training images and 10,000 test images. The training images contain 5,000 images from each class, whereas the test images contain 1,000 from each class. The ten classes in CIFAR-10 represent airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. These classes are mutually exclusive and

cover a range of living creatures and vehicles. CIFAR-10 has been a benchmark for algorithms in image recognition, machine learning, and computer vision.

CIFAR-10-C includes the original CIFAR-10 test set images that have been algorithmically altered with 15 different types of corruption, such as noise, blur, weather, and digital effects. Each type of corruption is applied at five levels of severity, resulting in a comprehensive test suite for assessing model performance across a spectrum of challenging conditions. This dataset provides a stringent benchmark for machine learning practitioners to test the resilience of their models beyond the sanitized environments typically used in training. It serves as a tool for understanding and improving how machine learning systems can withstand and adapt to the unpredictability of the real world.

In this project the CIFAR-10 and CIFAR-10-C datasets are obtained from HENDRYCKS and DIETTERICH 2019.

## Chapter 4

# Investigation of Robustness of Calibration Techniques

This chapter investigates the robustness of various calibration methods on machine learning models using the FSW dataset benchmarking on the CIFAR dataset. The ResNet50 model and Swin transformer are introduced, alongside data filtering and validation techniques. The Expected Calibration Error (ECE) PAKDAMAN NAEINI et al. 2015 is employed as the principal metric. Results for various calibration methods are presented, including an analysis of concept drift.

### 4.1 Experiment with FSW Dataset

#### 4.1.1 Methodologies

**Model Choice:** The project utilizes ResNet50, a state-of-the-art deep learning model that is pre-trained. ResNet50 belongs to the residual network family, and is known for its effectiveness in a wide range of image recognition tasks HE et al. 2015. ResNet50 has been previously trained on a large dataset (usually ImageNet) and has learned a variety of features that can be transferred to the specific task at hand HE et al. 2015, which in this case, is quality monitoring.

**Dataset:** The dataset used is FSW dataset.

**Calibration Measuring Metrics** To quantitatively evaluate the calibration of the predictive models utilized in this study, we employ ECE as a principal metric. ECE is frequently used for quantifying miss-calibration. ECE is a scalar summary measure estimating miss-calibration by approximating equation (4.1) as follows. In the first step, confidence scores  $\hat{p}$  of all samples are partitioned into  $M$  equally sized bins of size  $1/M$ , and secondly, for each bin  $B_m$ , the respective mean confidence and the accuracy is computed based on the ground truth class  $y$ . Finally, the ECE is estimated by calculating the mean difference between confidence and accuracy over all bins:

$$ECE^d = \sum_{m=1}^M \frac{|B_m|}{N} \|\text{acc}(B_m) - \text{conf}(B_m)\|_d \quad (4.1)$$

with  $d$  usually set to 1 (L1-norm). TOMANI et al. 2023



### 4.1.2 Results and Discussion

#### Results for Multiple Calibration Methods:

In this part, the result on the efficacy of various calibration methods when applied to the FSW datasets is displayed. The calibration methods are evaluated based on their ECE for both In-distribution (ID) and OD data.

#### Baseline:

As a point of comparison, we established a baseline using the FSW datasets without the application of any calibration method. The performance metrics obtained are as follows:

- ECE for ID: 0.0069
- ECE for OD: 0.1044

This initial analysis serves as a benchmark for subsequent calibration endeavors. It is worth noting that modern calibration methods often exhibit sensitivity to hyperparameter settings, necessitating careful hyperparameter tuning to ensure optimal performance. Concept drift is a phenomenon where the statistical properties of the target variable, which the model is trying to predict, change over time in unforeseen ways. This can cause model performance to degrade. In our experiments, concept drift was observed and is exemplified by the changes in test accuracy and discrepancy in ECE for ID and OD data.

For ID data, the model achieved a test accuracy of 98.19% with an ECE of 0.0069. However, for OD data, there was a notable decrease in performance, with the test accuracy dropping to 81.00% and an ECE of 0.1044. This discrepancy in performance metrics is indicative of concept drift, suggesting that the model's ability to generalize to new, unseen data is compromised.

The accuracy and ECE of the various calibration methods after training, testing and calibration can be seen in table 4.1. Bear in mind that only the calibration methods with high accuracy and low ECE are of interest.

**Table 4.1:** Calibration results for FSW dataset

Method	ECE for ID	ECE for OD	ID Accuracy	OD Accuracy
<b>Baseline</b>	0.0069	0.1044	98.19	81.00
<b>DAC + Temperature Scaling</b>	0.0383	0.0235	97.47	85.60
<b>MC Dropout</b>	0.0918	0.1068	92.42	77.40
<b>Mixup</b>	0.0318	0.0318	99.10	86.20
<b>Label Smoothing</b>	0.0596	0.0596	99.82	82.60
<b>Focal Loss</b>	0.0115	0.1159	99.28	82.20
<b>Correctness Ranking Loss (CRL)</b>	0.0053	0.1240	99.46	83.20

## 4.2 Benchmarking with Public Dataset CIFAR-10

In this section, we detail the approach and methodologies employed in the benchmarking process using the public dataset CIFAR-10 and CIFAR-10-C obtained from HENDRYCKS and DIETTERICH 2019.

**Importance of Benchmarking with CIFAR-10:** Benchmarking with the CIFAR-10 dataset serves as a pivotal step in validating the robustness and generalization capabilities of our machine learning models. Another purpose of benchmarking with CIFAR dataset is to discover calibration methods that are consistently robust across domains.

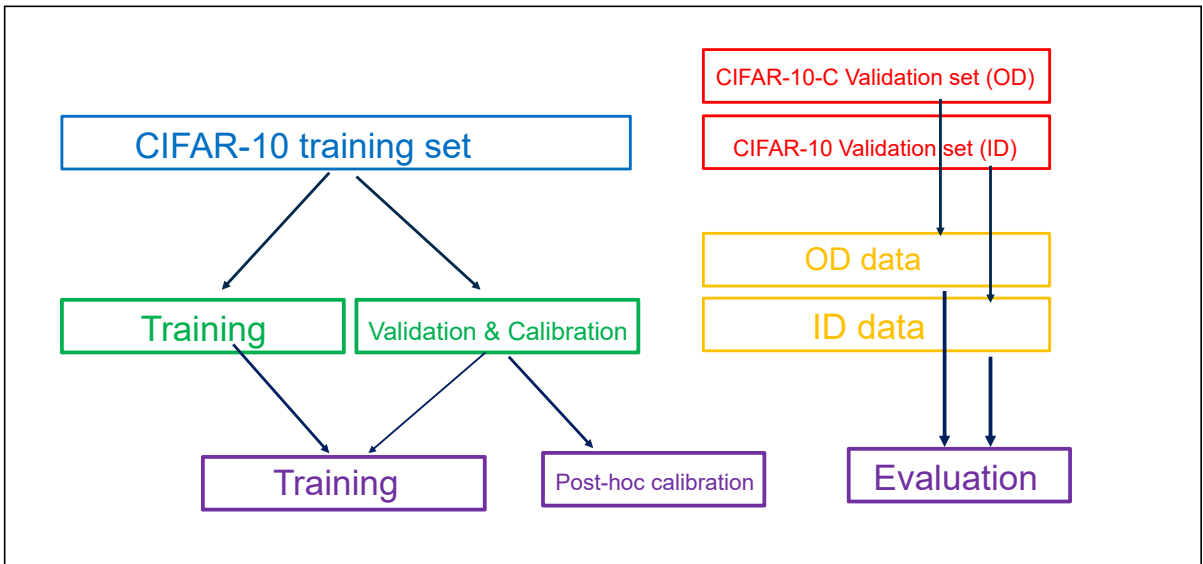
#### 4.2.1 Methodologies

**Workflow:** The workflow for our calibration project is centered around the strategic splitting of the CIFAR dataset for targeted training, calibration, and validation. We initiate with a training phase on the CIFAR-10 dataset to establish a foundational model performance. Following this, the model is calibrated using a subset of the CIFAR-10 data, fine-tuning the model's probabilistic outputs for accuracy.

The calibrated model then undergoes a validation process. It is first validated on a separate subset of the CIFAR-10 dataset to assess performance on ID data. Subsequently, the model's robustness is challenged against the CIFAR-10-C dataset, which provides OD data with a variety of image corruptions. This validation aims to test the model's resilience to data that deviates from the distribution seen during training.

The final model evaluation is conducted on both ID and OD data to ensure robust and accurate performance. This systematic process ensures that the CIFAR dataset is utilized effectively to develop a model that is well-calibrated and generalizes well to new, unseen data conditions.

An illustration of the data split and workflow is shown in figure 4.1.



**Figure 4.1:** Cifar dataset split

#### 4.2.2 Results and Discussion

Results and the corresponding methods can be found in table 4.2

**Table 4.2:** Calibration results for CIFAR dataset

Method	ECE for ID	ECE for OD	ID Accuracy	OD Accuracy
Baseline	0.0325	0.0944	94.16	86.60
DAC + Temperature Scaling	0.0207	0.0207	95.42	87.30
MC Dropout	0.0120	0.0720	91.06	77.90
Mixup	0.0699	0.0960	94.00	94.00
Focal Loss	0.0081	0.0209	95.44	86.10
Label Smoothing	0.0281	0.0409	93.30	93.30
Correctness Ranking Loss (CRL)	0.0338	0.0790	92.34	83.20

An overview of the performance comparison between FSW dataset and CIFAR dataset is also displayed in table 4.3.

**Table 4.3:** Calibration results for CIFAR and FSW datasets

	CIFAR Dataset		FSW Dataset	
Method	ECE for ID	ECE for OD	ECE for ID	ECE for OD
Baseline	0.0325	0.0944	0.0069	0.1044
DAC + Temperature Scaling	0.0207	0.0207	0.0383	0.0235
MC Dropout	0.0120	0.0720	0.0918	0.1068
Mixup	0.0699	0.0960	0.0318	0.0318
Focal Loss	0.0081	0.0209	0.0115	0.1159
Label Smoothing	0.0281	0.0409	0.0596	0.0596
Correctness Ranking Loss (CRL)	0.0338	0.0790	0.0053	0.1240

Based on the calibration results for the CIFAR and FSW datasets, we can observe the following about the model's performance:

- **Baseline Performance:**

- The baseline model has higher ECE for OD as compared to ID in both datasets, indicating less calibration when handling out-of-distribution data.
- There is a lower ECE for ID in the FSW dataset compared to CIFAR, suggesting better calibration for in-distribution data in the FSW dataset.

- **DAC + Temperature Scaling:**

- This method significantly reduces ECE for both ID and OD in the CIFAR dataset, achieving equal ECE for ID and OD, which is a substantial improvement.
- For the FSW dataset, there is an improvement, but OD error remains higher than ID error.

- **MC Dropout:**

- MC Dropout is more effective in reducing ECE for ID in the CIFAR dataset but less so for OD.
- In the FSW dataset, MC Dropout does reduce ECE for ID but to a lesser degree, and the ECE for OD is considerably higher than for ID.

- **Mixup:**

- Mixup increases the ECE for both ID and OD in the CIFAR dataset, suggesting a possibility of overfitting or poor calibration.
- Conversely, in the FSW dataset, Mixup achieves equal and low ECE for both ID and OD, indicating better suitability for this dataset.
- **Focal Loss:**
  - Focal Loss exhibits superior calibration on the CIFAR dataset with the lowest ECE for ID (0.0081) compared to other methods, indicating highly accurate probability estimates.
  - While not the lowest for ECE for OD (0.0209), it significantly outperforms the baseline and most other methods, suggesting good generalization to out-of-distribution data. Its focus on difficult-to-classify examples likely contributes to its effectiveness in both in-distribution and out-of-distribution settings.
- **Label Smoothing:**
  - Label Smoothing consistently improves calibration across both datasets, with a more noticeable effect on the OD ECE in the CIFAR dataset.
  - It also improves ECE for both ID and OD in the FSW dataset, albeit to a lesser extent.
- **Correctness Ranking Loss (CRL):**
  - CRL improves ECE for ID in the CIFAR dataset but increases ECE for OD.
  - In the FSW dataset, CRL achieves the lowest ECE for ID but with the highest ECE for OD among all the methods.

*Note:* While calibration plays a vital role in the performance of models, particularly in terms of the reliability of their probability estimates, it should not be the sole metric for assessing a model's overall quality. Other performance metrics, such as accuracy, must also be considered. Despite thorough hyperparameter tuning during the experiment, we cannot discount the possibility of hyperparameter tuning impacting the results.

### 4.3 Combination of Calibration Methods with Swin Transformer on CIFAR-10

**Motivation for Combination of Calibration Methods:** Calibration methods, each with their unique approach to aligning predicted probabilities with empirical frequencies, collectively contribute to a model's robustness and reliability. Combining these methods aims to leverage their complementary strengths, thereby addressing different aspects of calibration, such as temperature scaling's adjustment of confidence scores and label smoothing's mitigation of overfitting. In the following experiment, the calibration methods that yield good results across two datasets are collected. The technique of combining calibration methods is still a research gap that has not been widely explored so far.

**Advantage of Swin Transformer:** The choice to utilize the Swin Transformer model (inspired by the work of LIU et al. 2021) is informed by its cutting-edge architecture. The shifted window scheme brings greater efficiency by limiting self-attention computation to non-overlapping local windows, while also allowing for cross-window connection. This hierarchical architecture has the flexibility to model at various scales and has linear computational complexity with respect to image size. These qualities of Swin Transformer make it compatible with a broad range of vision tasks LIU et al. 2021. Swin Transformer achieves strong performance on the recognition tasks of image classification, object detection and semantic segmentation LIU et al. 2021. This capability is especially beneficial in datasets like CIFAR-10, where the diversity of subjects and contexts demands a model that can discern subtle, yet critical, features within the data. These characteristics position the Swin Transformer as an ideal candidate for pushing the boundaries of what is achievable with advanced calibration methods, setting a new benchmark for accuracy and trustworthiness in model predictions.

This section of the report encapsulates our forward-thinking approach in machine learning model development, where the combination of sophisticated calibration methods and transformative neural network architectures like the Swin Transformer could potentially spearhead advancements in the field of deep learning.

### Evaluation of Baseline Performance

Our initial evaluation establishes a baseline for the model’s calibration using the CIFAR dataset. This baseline serves as a reference point against which the efficiency of subsequent calibration methods is measured. The results can be found in table 4.4. These baseline figures are noteworthy as they already demonstrate an improvement over the initial results obtained with the ResNet50 architecture, suggesting that the model’s calibration is reasonably well-aligned with the empirical distribution of the data even prior to the application of targeted calibration techniques. So far, still, the same as all the previous experiments, density-aware calibration combined with temperature scaling, have the best performance, the ECE’s are even lower than the experiment with ResNet50, this indicates that the transformer based model, Swin, have better performance.

**Table 4.4:** Calibration results for combined methods with Swin

Method	ECE ID	ECE OD	ID Accuracy	OD Accuracy
Baseline	0.0113	0.0268	96.98	91.60
DAC with Temperature Scaling	0.0125	0.0125	97.02	93.80
DAC, Temperature Scaling, Label Smoothing	0.0125	0.0125	97.02	93.80
DAC, Temperature Scaling, Mixup, Label Smoothing	0.1442	0.1442	94.70	91.10

Contrary to expectations, this complex calibration approach did not yield the anticipated enhancement in model performance. Both in-distribution and out-of-distribution ECE values experienced a significant increase, suggesting that the combination of these specific methods might introduce conflicting calibration signals to the model. This result highlights the intricate balance required in calibration techniques and the possibility of overfitting the calibration methods. It also suggests that further hyperparameter tuning may be necessary to optimize the calibration strategy and reconcile the competing influences of the individual methods, because modern calibration methods are often very sensitive to hyperparameter tuning, and each calibration method reaches optimal performance under different hyperparameters. Fine-tuning these parameters could be the key to unlocking the potential benefits of combining these powerful calibration techniques.

The insights gained from these experiments are instrumental in guiding future calibration efforts. They emphasize the need for a nuanced understanding of how different calibration methods interact with each other, and the importance of methodical experimentation in identifying optimal calibration strategies for deep neural networks.

# Chapter 5

## Summary and Outlook

This project delved into the calibration of deep neural networks, with a specific focus on image classification tasks within the manufacturing domain. Various calibration methods were systematically explored using both the industrial FSW dataset and benchmarking against the public CIFAR datasets.

The study underscored the critical role of model calibration, particularly in contexts prioritizing prediction accuracy. While certain calibration methods, such as density-aware calibration combined with temperature scaling, exhibited potential for improving model reliability, the integration of multiple techniques did not consistently yield performance enhancements, highlighting the nuanced nature of calibration.

The application of calibration methods to the FSW dataset demonstrated their relevance in real-world manufacturing scenarios, where erroneous predictions can carry substantial consequences. The project's outcomes lay the groundwork for future research directions:

- Exploring advanced calibration techniques, including Bayesian approaches and ensemble methods, for finer adjustments of model confidence.
- Adapting calibration methods to even more complex and noisy manufacturing scenarios to test their practical utility.
- Investigating the role of transfer learning and domain adaptation in enhancing calibration for diverse tasks.
- Optimizing calibration methods for efficiency, particularly under hardware constraints.
- Developing automated calibration pipelines that intelligently select and combine methods based on model and dataset characteristics.
- Benchmarking on more public datasets for more robustness validation, across different applicational domain.

In conclusion, this project underscores the significance of model calibration in the realm of deep learning for manufacturing and its potential to enhance the dependability of DNN in industrial contexts. As the field progresses, calibration is poised to become an integral aspect of the training process, guiding the evolution of more robust and reliable DNN models for industrial applications.

# Appendix A

## Digital Appendix

### Overview of the Code Repository

The code repository associated with this thesis includes all the scripts, data, and additional resources that were used in the computational analysis presented in this work. The repository is structured as follows:

- **Data Annotation Files:** These files contain annotations for the datasets used in the analysis.
- **Jupyter Notebooks:** Jupyter notebooks with detailed code, comments, and results visualization.

### Accessing the Repository

The code repository is hosted on Gitlab, providing version control and easy access for review and collaboration. You can access the repository at the following URL:

<https://gitlab.lrz.de/00000000014B345E/PytorchCalibration>

### Using the Repository

To use the code repository, follow the instructions below:

1. Clone the repository to your local machine using the command:  
`gitlab.lrz.de:00000000014B345E/PytorchCalibration.git`
2. Navigate to the specific Notebook for the analysis you are interested in.
3. Run the Jupyter notebooks or scripts



## Software Requirements

The analysis code is written in Python. To run the Jupyter notebooks, you will need to have the following software installed:

- Python 3.8 or higher
- Jupyter Lab or Jupyter Notebook
- Necessary Python libraries as specified in every notebook

For further details, please refer to the README file in the repository.

# Bibliography

- ARENDT, P. D., APLEY, D. W., and CHEN, W., (Sept. 2012). “Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability”. In: *Journal of Mechanical Design* 134.10, p. 100908. ISSN: 1050-0472. DOI: 10.1115/1.4007390. eprint: [https://asmedigitalcollection.asme.org/mechanicaldesign/article-pdf/134/10/100908/5761063/100908\\_1.pdf](https://asmedigitalcollection.asme.org/mechanicaldesign/article-pdf/134/10/100908/5761063/100908_1.pdf). URL: <https://doi.org/10.1115/1.4007390>.
- ASHFAHANI, A., PRATAMA, M., LUGHOFFER, E., and YEE, E. Y. K., (2021). “Autonomous Deep Quality Monitoring in Streaming Environments”. In: *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. DOI: 10.1109/IJCNN52387.2021.9534461.
- EL BILALI, A., LAMANE, H., TALEB, A., and NAFII, A., (2022). “A framework based on multi-variate distribution-based virtual sample generation and DNN for predicting water quality with small data”. In: *Journal of Cleaner Production* 368, p. 133227. ISSN: 0959-6526. DOI: <https://doi.org/10.1016/j.jclepro.2022.133227>. URL: <https://www.sciencedirect.com/science/article/pii/S0959652622028141>.
- FEURER, M. and HUTTER, F., (2019). “Hyperparameter Optimization”. In: *Automated Machine Learning: Methods, Systems, Challenges*. Ed. by HUTTER, F., KOTTHOFF, L., and VANSCHOREN, J. Cham: Springer International Publishing, pp. 3–33. ISBN: 978-3-030-05318-5. DOI: 10.1007/978-3-030-05318-5\_1. URL: [https://doi.org/10.1007/978-3-030-05318-5\\_1](https://doi.org/10.1007/978-3-030-05318-5_1).
- GAL, Y. and GHAHRAMANI, Z., (20–22 Jun 2016). “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning”. In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by BALCAN, M. F. and WEINBERGER, K. Q. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, pp. 1050–1059. URL: <https://proceedings.mlr.press/v48/gal16.html>.
- GUO, C., PLEISS, G., SUN, Y., and WEINBERGER, K. Q., (June 2017). “On Calibration of Modern Neural Networks”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by PRECUP, D. and TEH, Y. W. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 1321–1330. URL: <https://proceedings.mlr.press/v70/guo17a.html>.
- HARTL, R., LANDGRAF, J., SPAHL, J., BACHMANN, A., and ZAEH, M. F., (2019). “Automated visual inspection of friction stir welds: a deep learning approach”. In: *Multimodal Sensing: Technologies and Applications*. Ed. by STELLA, E. Vol. 11059. International Society for Optics and Photonics. SPIE, p. 1105909. DOI: 10.1117/12.2525947. URL: <https://doi.org/10.1117/12.2525947>.
- HE, K., ZHANG, X., REN, S., and SUN, J., (2015). “Deep Residual Learning for Image Recognition”. In: *CoRR abs/1512.03385*. arXiv: 1512.03385. URL: <http://arxiv.org/abs/1512.03385>.
- HEBBALAGUPPE, R., PRAKASH, J., MADAN, N., and ARORA, C., (June 2022). “A Stitch in Time Saves Nine: A Train-Time Regularizing Loss for Improved Neural Network Calibration”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16081–16090.

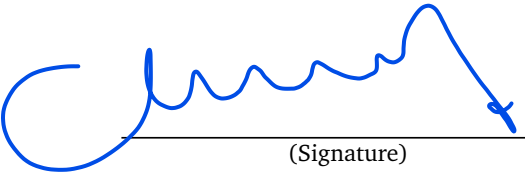
- HENDRYCKS, D. and DIETTERICH, T., (2019). “Benchmarking Neural Network Robustness to Common Corruptions and Perturbations”. In: *Proceedings of the International Conference on Learning Representations*.
- HOENS, T. R., POLIKAR, R., and CHAWLA, N. V., (Apr. 1, 2012). “Learning from streaming data with concept drift and imbalance: an overview”. In: *Progress in Artificial Intelligence* 1.1, pp. 89–101. ISSN: 2192-6360. DOI: 10.1007/s13748-011-0008-0. URL: <https://doi.org/10.1007/s13748-011-0008-0>.
- LIN, T.-Y., GOYAL, P., GIRSHICK, R., HE, K., and DOLLAR, P., (Oct. 2017). “Focal Loss for Dense Object Detection”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- LIU, Z., LIN, Y., CAO, Y., HU, H., WEI, Y., ZHANG, Z., LIN, S., and GUO, B., (Oct. 2021). “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022.
- LOQUERCIO, A., SEGU, M., and SCARAMUZZA, D., (2020). “A General Framework for Uncertainty Estimation in Deep Learning”. In: *IEEE Robotics and Automation Letters* 5.2, pp. 3153–3160. DOI: 10.1109/LRA.2020.2974682.
- MISHRA, R. and MA, Z., (2005). “Friction stir welding and processing”. In: *Materials Science and Engineering: R: Reports* 50.1, pp. 1–78. ISSN: 0927-796X. DOI: <https://doi.org/10.1016/j.mser.2005.07.001>. URL: <https://www.sciencedirect.com/science/article/pii/S0927796X05000768>.
- MOHAMMADZADEH, S., PRACHASEREE, P., and LEJEUNE, E., (2023). “Investigating deep learning model calibration for classification problems in mechanics”. In: *Mechanics of Materials* 184, p. 104749. ISSN: 0167-6636. DOI: <https://doi.org/10.1016/j.mechmat.2023.104749>. URL: <https://www.sciencedirect.com/science/article/pii/S0167663623001953>.
- MOON, J., KIM, J., SHIN, Y., and HWANG, S., (13–18 Jul 2020). “Confidence-Aware Learning for Deep Neural Networks”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by III, H. D. and SINGH, A. Vol. 119. Proceedings of Machine Learning Research. PMLR. URL: <https://proceedings.mlr.press/v119/moon20a.html>.
- NANNAPANENI, S., MAHADEVAN, S., and RACHURI, S., (2016). “Performance evaluation of a manufacturing process under uncertainty using Bayesian networks”. In: *Journal of Cleaner Production* 113, pp. 947–959. ISSN: 0959-6526. DOI: <https://doi.org/10.1016/j.jclepro.2015.12.003>. URL: <https://www.sciencedirect.com/science/article/pii/S0959652615018144>.
- PAKDAMAN NAEINI, M., COOPER, G., and HAUSKRECHT, M., (Feb. 2015). “Obtaining Well Calibrated Probabilities Using Bayesian Binning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 29.1. DOI: 10.1609/aaai.v29i1.9602. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/9602>.
- PSAROS, A. F., MENG, X., ZOU, Z., GUO, L., and KARNIADAKIS, G. E., (2023). “Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons”. In: *Journal of Computational Physics* 477, p. 111902. ISSN: 0021-9991. DOI: <https://doi.org/10.1016/j.jcp.2022.111902>. URL: <https://www.sciencedirect.com/science/article/pii/S0021999122009652>.
- ROŽANEC, J. M., BIZJAK, L., TRAJKOVA, E., ZAJEC, P., KEIZER, J., FORTUNA, B., and MLADENIĆ, D., (Mar. 16, 2023). “Active learning and novel model calibration measurements for automated visual inspection in manufacturing”. In: *Journal of Intelligent Manufacturing*.
- SEIFFER, C., ZIEKOW, H., SCHREIER, U., and GERLING, A., (2021). “Detection of Concept Drift in Manufacturing Data with SHAP Values to Improve Error Prediction”. In: URL: <https://api.semanticscholar.org/CorpusID:249423528>.

- SZEGEDY, C., VANHOUCKE, V., IOFFE, S., SHLENS, J., and WOJNA, Z., (June 2016). “Rethinking the Inception Architecture for Computer Vision”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 2823.
- TANAKA, T., INUI, T., and KAWAI, S., (Nov. 2022). “DNN-based optical performance monitoring and its application for soft failure localization by multipoint estimation”. In: *J. Opt. Commun. Netw.* 14.11, pp. 894–902. DOI: 10.1364/JOCN.461422. URL: <https://opg.optica.org/jocn/abstract.cfm?URI=jocn-14-11-894>.
- TOMANI, C., GRUBER, S., ERDEM, M. E., CREMERS, D., and BUETTNER, F., (June 2021). “Post-Hoc Uncertainty Calibration for Domain Drift Scenarios”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10124–10132.
- TOMANI, C., WASEDA, F. K., SHEN, Y., and CREMERS, D., (23–29 Jul 2023). “Beyond In-Domain Scenarios: Robust Density-Aware Calibration”. In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by KRAUSE, A., BRUNSKILL, E., CHO, K., ENGELHARDT, B., SABATO, S., and SCARLETT, J. Vol. 202. *Proceedings of Machine Learning Research*. PMLR, pp. 34344–34368. URL: <https://proceedings.mlr.press/v202/tomani23a.html>.
- WANG, J., MA, Y., ZHANG, L., GAO, R. X., and WU, D., (2018). “Deep learning for smart manufacturing: Methods and applications”. In: *Journal of Manufacturing Systems* 48. Special Issue on Smart Manufacturing, pp. 144–156. ISSN: 0278-6125. DOI: <https://doi.org/10.1016/j.jmsy.2018.01.003>. URL: <https://www.sciencedirect.com/science/article/pii/S0278612518300037>.
- WANG, Z., WU, B., GARIKIPATI, K., and HUAN, X., (2020). “A perspective on regression and Bayesian approaches for system identification of pattern formation dynamics”. In: *Theoretical and Applied Mechanics Letters* 10.3, pp. 188–194. ISSN: 2095-0349. DOI: <https://doi.org/10.1016/j.taml.2020.01.028>. URL: <https://www.sciencedirect.com/science/article/pii/S2095034920300325>.
- WUEST, T., IRGENS, C., and THOBEN, K.-D., (Mar. 16, 2023). “An approach to monitoring quality in manufacturing using supervised machine learning on product state data”. In: *Journal of Intelligent Manufacturing* 25.5, pp. 1167–1180. DOI: 10.1007/s10845-013-0761-y. URL: <https://doi.org/10.1007/s10845-013-0761-y>.
- ZHANG, C.-B., JIANG, P.-T., HOU, Q., WEI, Y., HAN, Q., LI, Z., and CHENG, M.-M., (2021). “Delving Deep Into Label Smoothing”. In: *IEEE Transactions on Image Processing* 30, pp. 5984–5996. DOI: 10.1109/TIP.2021.3089942.
- ZHANG, H., CISSE, M., DAUPHIN, Y. N., and LOPEZ-PAZ, D., (2018). *mixup: Beyond Empirical Risk Minimization*. arXiv: 1710.09412 [cs.LG].

## Disclaimer

I hereby declare that this thesis is entirely the result of my own work except where otherwise indicated. I have only used the resources given in the list of references.

Garching, April 6, 2024



---

(Signature)