Non-Parametric Calibration for Classification

Jonathan Wenger
TU München, KTH Royal Institute of Technology

July 5th, 2019

Outline

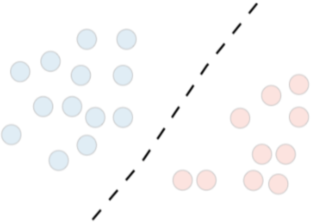
Knowing When We Don't

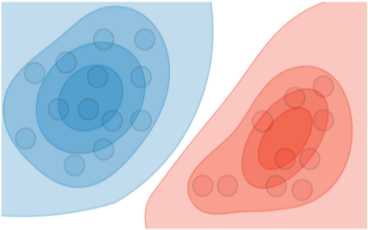
Figure 1: Segmented scenery of Tübingen from the cityscapes data set [1], illustrating a typical classification task in computer vision.

Research Question

How can **prediction uncertainty** of a multi-class classifier, applied to computer vision problems, be **accurately represented** independent of model specification?

Uncertainty Representation





Definitions and Notation

- $f_{X,Y}$ joint probability density of inputs and labels
- f classification model
- $\boldsymbol{z} = f(\boldsymbol{x})$ confidence score
- $\hat{y} = \arg\max_i(\boldsymbol{z_i})$ class prediction
- $\hat{z} = \max_i(\boldsymbol{z_i})$ confidence in prediction

Misrepresentation of Uncertainty

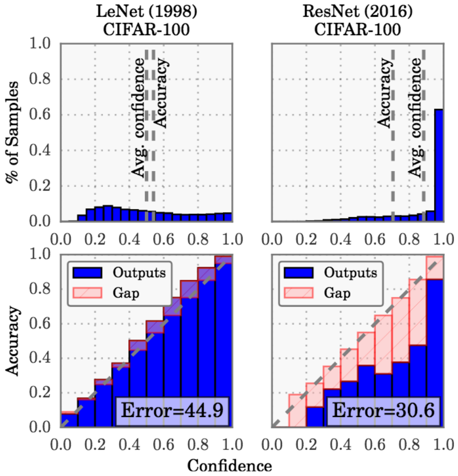


Figure 2: Confidence histograms and reliability diagrams for a simple and a modern NN architecture [2].

Calibration

Definition

A classifier is called **calibrated** [3, 4] if its confidence in its class prediction matches the probability of its prediction being correct, i.e.

$$\mathbb{E}\left[1_{\hat{y}=y} \mid \hat{z}\right] = \hat{z}.$$

Let $1 \leq p < \infty$, then

$$\mathrm{ECE}_p = \mathbb{E}\left[\left|\hat{z} - \mathbb{E}\left[1_{\hat{y}=y} \mid \hat{z}\right]\right|^p\right]^{\frac{1}{p}}$$

is called the **expected calibration error** [5].

Active Learning

Idea

- Labelled samples are expensive to obtain
- Query most **informative samples** (e.g. uncertainty sampling [6])

Figure 3: Illustration of active learning [7].

Over- and **underconfidence** [8] relate to query quality:

$$o(f) = \mathbb{E}\left[\hat{z} \mid \hat{y} \neq y\right] \qquad u(f) = \mathbb{E}\left[1 - \hat{z} \mid \hat{y} = y\right]$$

Relationship to calibration

Theorem

Let $1 \leq p < q \leq \infty$, then the following relationship between over-, underconfidence and the expected calibration error holds:

$$|o(f)\mathbb{P}(\hat{y} \neq y) - u(f)\mathbb{P}(\hat{y} = y)| \leq \mathrm{ECE}_p \leq \mathrm{ECE}_q.$$

Corollary

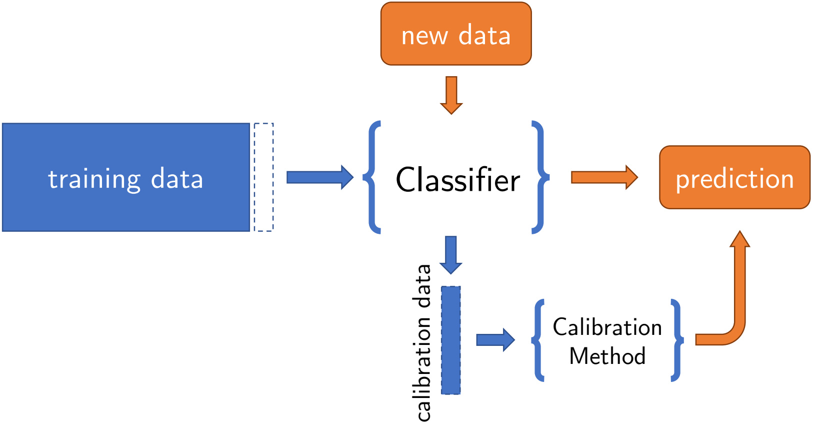
Assume f is calibrated and $\mathbb{P}(\hat{y} \neq y) \notin \{0, 1\}$, then

$$\frac{o(f)}{u(f)} = \frac{\mathbb{P}(\hat{y} = y)}{\mathbb{P}(\hat{y} \neq y)},$$

i.e. the **odds** of making a correct prediction determine the **ratio** between over- and underconfidence.

Probability Calibration

Improve uncertainty representation **post-hoc** by using a subset of the training data for calibration.



Existing Methods of Calibration

Binary Methods

- Platt Scaling [9, 10]
- Beta Calibration [11, 12]
- Isotonic Regression [13]
- Bayesian Binning into Quantiles (BBQ) [5]

Multi-class Methods

- One-vs-all [13]
- Temperature Scaling [2]

Limitations

- Binary methods not applicable for multi-class problems
- Temperature Scaling designed for NNs

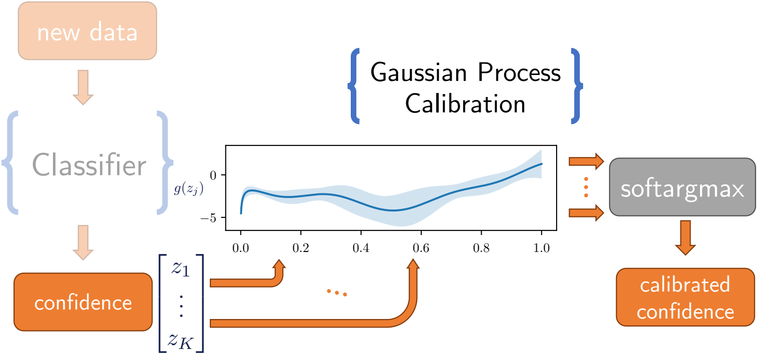
Gaussian Process Calibration

Requirements

- **Multi-class** classifiers
- **Arbitrary classifiers** $\implies$ non-parametric
- Incorporation of **prior knowledge** $\implies$ "don't fix what isn't broken"

Definition

- **Latent function**: $g \sim \mathcal{GP}\left(\mu(\cdot), k(\cdot\,,\cdot \mid \theta)\right)$
- **Inverse link function**: $\sigma(g(\boldsymbol{z}))_j = \frac{\exp(g(\boldsymbol{z}_j))}{\sum_{k=1}^{K} \exp(g(\boldsymbol{z}_k))}$
- **Likelihood**: $\mathrm{Cat}(y|\sigma(g(\boldsymbol{z})))$



Inference and Prediction

Inference of Parameters

- adjusted **scalable variational** Gaussian Processes (SVGP) [14]
  - sparse representation $p(\boldsymbol{u} \mid \boldsymbol{y})$ instead of $p(\boldsymbol{g} \mid \boldsymbol{y})$ due to $\mathcal{O}((NK)^3)$
  - approximate $p(\boldsymbol{u} \mid \boldsymbol{y})$ by $q(\boldsymbol{u}) \sim \mathcal{N}(m, S)$
- optimize all parameters jointly
  - variational parameters m, S
  - locations of inducing inputs
  - kernel parameters θ

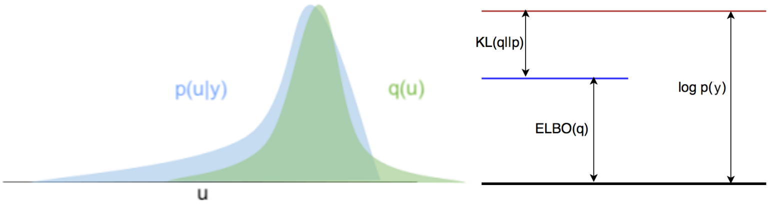


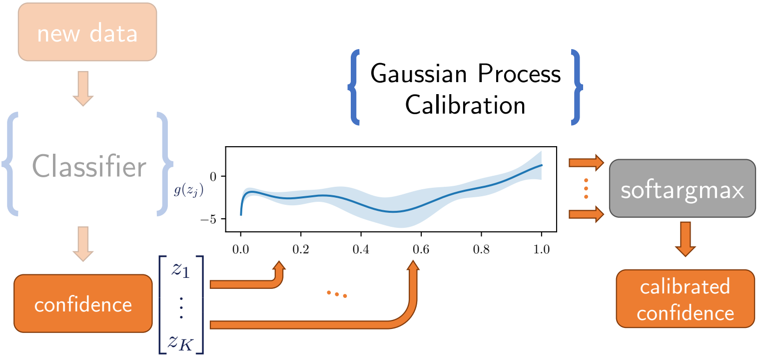
Figure 4: Illustration of variational inference [15, 16].

Inference and Prediction

Prediction of Confidence

Calibrated confidence for new input $\boldsymbol{z}_*$ via **Monte-Carlo** integration:

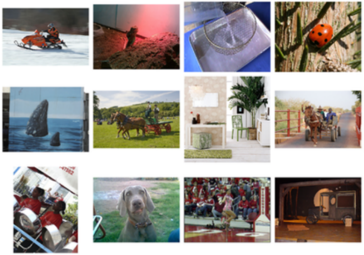
$$p(\boldsymbol{y}_* \mid \boldsymbol{y}) = \int p(\boldsymbol{y}_* \mid \boldsymbol{g}_*) \underbrace{p(\boldsymbol{g}_* \mid \boldsymbol{y})}_{\approx \int p(\boldsymbol{g}_* | \boldsymbol{u}) q(\boldsymbol{u}) \, d\boldsymbol{u}} \, d\boldsymbol{g}_*$$



Experiments

Benchmark Data Sets

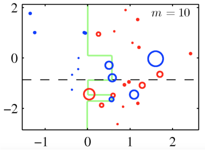
- **MNIST** [17]: Handwritten digit recognition
  - 10 classes
  - dimension 28×28
  - train: 60000, calibration: 1000, test: 9000
- **ImageNet 2012** [18]: Image database of natural objects and scenes
  - 1000 classes
  - varying dimension
  - train: 1.2 million, calibration: 1000, test: 9000

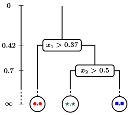


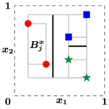
Experiments

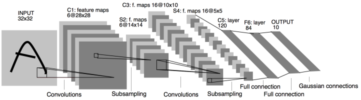
Classifiers

- **Boosting:** AdaBoost [19, 20], XGBoost [21]
- **Forests:** Mondrian Forest [22], Random Forest [23]
- **Convolutional Neural Networks:**
  - AlexNet [24]
  - VGG19 [25]
  - ResNet50, ResNet152 [26]
  - DenseNet121, DenseNet201 [27]
  - Inception v4 [28]
  - SE ResNeXt50, SE ResNeXt101[29, 30]









Experiments: Results

Table 1: **Average ECE$_1$** of ten Monte-Carlo cross validation folds on multi-class benchmark data sets.

			one-vs-all					
Data Set	Model	Uncal.	Platt	Isotonic	Beta	BBQ	Temp.	GPcalib
MNIST	AdaBoost	.6121	.2267	.1319	.2222	.1384	.1567	**.0414**
MNIST	XGBoost	.0740	.0449	.0176	.0184	.0207	.0222	**.0180**
MNIST	Mondr. Forest	.2163	.0357	.0282	.0383	.0762	**.0208**	.0213
MNIST	Rand. Forest	.1178	.0273	.0207	.0259	.1233	**.0121**	.0148
MNIST	1 layer NN	.0262	**.0126**	.0140	.0168	.0186	.0195	.0239
ImageNet	AlexNet	.0354	.1143	.2771	.2321	.1344	**.0336**	.0354
ImageNet	VGG19	.0375	.1018	.2656	.2484	.1642	**.0347**	.0351
ImageNet	ResNet50	.0444	.0911	.2632	.2239	.1627	**.0333**	**.0333**
ImageNet	ResNet152	.0525	.0862	.2374	.2177	.1665	**.0328**	.0336
ImageNet	DenseNet121	.0369	.0941	.2374	.2277	.1536	.0333	**.0331**
ImageNet	DenseNet201	.0421	.0923	.2306	.2195	.1602	**.0319**	.0336
ImageNet	Inception v4	.0311	.0852	.2795	.1628	.1569	.0460	**.0307**
ImageNet	SE ResNeXt50	.0432	.0837	.2570	.1723	.1717	.0462	**.0311**
ImageNet	SE ResNeXt101	.0571	.0837	.2718	.1660	.1513	.0435	**.0317**

Experiments: Results

Table 2: **Average ECE_1 and standard deviation** of ten Monte-Carlo cross validation folds on multi-class benchmark data sets.

Data Set	Model	Uncal.	Temp.	GPcalib
MNIST	AdaBoost	.6121	.1567 ± .0122	**.0414** ± .0085
MNIST	XGBoost	.0740	.0222 ± .0015	**.0180** ± .0014
MNIST	Mondr. Forest	.2163	**.0208** ± .0012	.0213 ± .0020
MNIST	Rand. Forest	.1178	**.0121** ± .0012	.0148 ± .0021
MNIST	1 layer NN	.0262	.0195 ± .0060	.0239 ± .0023
ImageNet	AlexNet	.0354	**.0336** ± .0038	.0354 ± .0024
ImageNet	VGG19	.0375	**.0347** ± .0036	.0351 ± .0042
ImageNet	ResNet50	.0444	**.0333** ± .0032	**.0333** ± .0024
ImageNet	ResNet152	.0525	**.0328** ± .0030	.0336 ± .0032
ImageNet	DenseNet121	.0369	.0333 ± .0034	**.0331** ± .0038
ImageNet	DenseNet201	.0421	**.0319** ± .0029	.0336 ± .0040
ImageNet	Inception v4	.0311	.0460 ± .0061	**.0307** ± .0017
ImageNet	SE ResNeXt50	.0432	.0462 ± .0028	**.0311** ± .0033
ImageNet	SE ResNeXt101	.0571	.0435 ± .0061	**.0317** ± .0031

Experiments: Active Learning

- **KITTI** [31, 32]: Stream-based urban traffic scenes
  - 8 classes
  - features [33] from segmented 3D point clouds
  - dimension 60

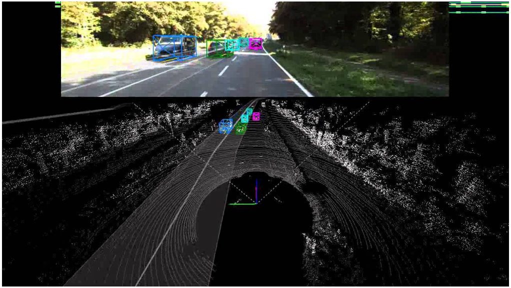
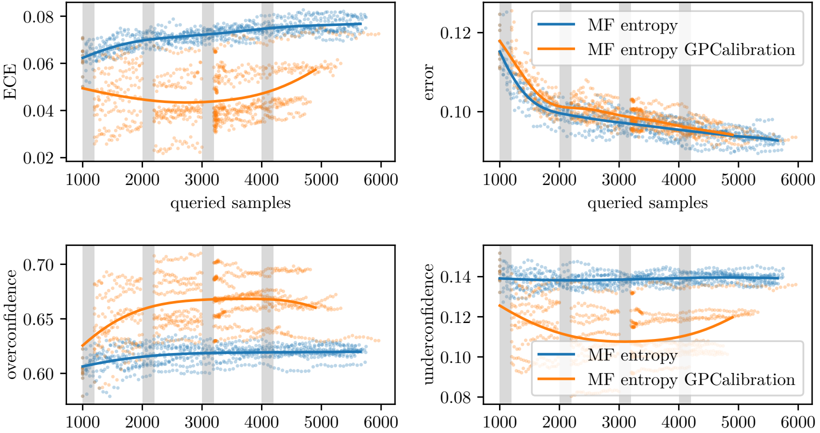


Figure 5: Example traffic scene showing the original image, ground truth bounding boxes, captured point clouds and a road overlay.

Experiments: Active Learning

- **KITTI** [31, 32]: Stream-based urban traffic scenes



Conclusion

Summary

- Accurate **uncertainty representation** is important
- Calibration, over- and underconfidence are linked
- GPcalib: **multi-class** calibration method for **arbitrary classifiers**

Future Work

- Theoretical framework for calibration [34]
  - Accuracy and uncertainty estimation
  - Calibration set size
- Extension of GP calibration
  - monotone latent process [35] $\implies$ accuracy guarantee
  - online calibration [36]
- Calibration and active learning
  - Switching strategy training and calibration
  - “Active calibration”

Preprint and Implementation

Non-Parametric Calibration for Classification

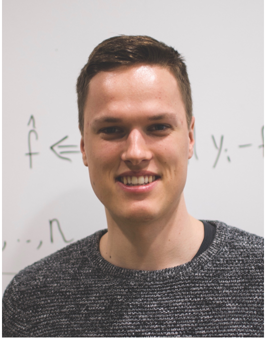
Jonathan Wenger
Technical University of Munich (TUM)
KTH Royal Institute of Technology
j.wenger@tum.de

Hedvig Kjellström
KTH Royal Institute of Technology
hedvig@kth.se

Rudolph Triebel
Technical University of Munich (TUM)
German Aerospace Center (DLR)
triebel@in.tum.de

- **Preprint** [37]: https://arxiv.org/abs/1906.04933
- **Code**: https://github.com/JonathanWenger/pycalib

Questions?



Jonathan Wenger
j.wenger@tum.de

Dr. habil. Rudolph Triebel (TUM, DLR)

Prof. Dr. Hedvig Kjellström (KTH)

- **Preprint** [37]: `https://arxiv.org/abs/1906.04933`
- **Code**: `https://github.com/JonathanWenger/pycalib`

References I

[1] Marius Cordts et al. "The Cityscapes Dataset for Semantic Urban Scene Understanding". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[2] Chuan Guo et al. "On calibration of modern neural networks". In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 2017.

[3] Allan H. Murphy. "A New Vector Partition of the Probability Score". In: *Journal of Applied Meteorology (1962-1982)* 12.4 (1973), pp. 595–600.

[4] Morris H. DeGroot and Stephen E. Fienberg. "The Comparison and Evaluation of Forecasters". In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 32.1/2 (1983), pp. 12–22.

[5] Mahdi Pakdaman Naeini et al. "Obtaining Well Calibrated Probabilities Using Bayesian Binning". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.* Ed. by Blai Bonet and Sven Koenig. AAAI Press, 2015, pp. 2901–2907.

[6] Burr Settles. *Active learning literature survey*. Tech. rep. 55-66. University of Wisconsin, Madison, 2010, p. 11.

[7] Stefan Hosein. *Active Learning: Curious AI Algorithms*. 2018. URL: `https://www.datacamp.com/community/tutorials/active-learning` (visited on 06/27/2019).

[8] D. Mund et al. "Active online confidence boosting for efficient object classification". In: *2015 IEEE International Conference on Robotics and Automation (ICRA)*. 2015, pp. 1367–1373.

References II

[9] John C. Platt. "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods". In: *Advances in Large-Margin Classifiers*. MIT Press, 1999, pp. 61–74.

[10] Hsuan-Tien Lin et al. "A note on Platt's probabilistic outputs for support vector machines". In: *Machine learning* 68.3 (2007), pp. 267–276.

[11] Meelis Kull et al. "Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration". In: *Electronic Journal of Statistics* 11.2 (2017), pp. 5052–5080.

[12] Meelis Kull et al. "Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*. Vol. 54. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, 2017, pp. 623–631.

[13] Bianca Zadrozny and Charles Elkan. "Transforming Classifier Scores into Accurate Multiclass Probability Estimates". In: *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '02. Edmonton, Alberta, Canada: ACM, 2002, pp. 694–699.

[14] James Hensman et al. "Scalable Variational Gaussian Process Classification". In: *Proceedings of AISTATS*. 2015.

[15] Evan Jang. *A Beginner's Guide to Variational Methods: Mean-Field Approximation*. 2016. URL: `https://blog.evjang.com/2016/08/variational-bayes.html` (visited on 06/27/2019).

References III

[16] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[17] Yann LeCun et al. "Gradient-Based Learning Applied to Document Recognition". In: *Proceedings of the IEEE*. Vol. 86/11. 1998, pp. 2278–2324.

[18] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252.

[19] Yoav Freund and Robert E Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Journal of computer and system sciences* 55.1 (1997), pp. 119–139.

[20] Trevor Hastie et al. "Multi-class adaboost". In: *Statistics and its Interface* 2.3 (2009), pp. 349–360.

[21] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 785–794.

[22] Balaji Lakshminarayanan et al. "Mondrian Forests: Efficient Online Random Forests". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'14. Montreal, Canada: MIT Press, 2014, pp. 3140–3148.

[23] Leo Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32.

References IV

[24] Alex Krizhevsky et al. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'12. Lake Tahoe, Nevada: Curran Associates Inc., 2012, pp. 1097–1105.

[25] S. Liu and W. Deng. "Very deep convolutional neural network based image classification using small training sample size". In: *3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. 2015, pp. 730–734.

[26] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778.

[27] Gao Huang et al. "Densely connected convolutional networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

[28] Christian Szegedy et al. "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning". In: *AAAI*. 2016.

[29] Saining Xie et al. "Aggregated Residual Transformations for Deep Neural Networks". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 5987–5995.

[30] Jie Hu et al. "Squeeze-and-Excitation Networks". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[31] Andreas Geiger et al. "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

References V

[32] Alexander Narr et al. "Stream-based active learning for efficient and adaptive classification of 3d objects". In: *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE. 2016, pp. 227–233.

[33] Michael Himmelsbach et al. "Real-time object classification in 3D point clouds using point feature histograms". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2009, pp. 994–1000.

[34] Juozas Vaicenavicius et al. "Evaluating model calibration in classification". In: *Proceedings of Machine Learning Research*. Vol. 89. Proceedings of Machine Learning Research. PMLR, 2019, pp. 3459–3467.

[35] Jaakko Riihimäki and Aki Vehtari. "Gaussian processes with monotonicity information". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 2010, pp. 645–652.

[36] Thang D Bui et al. "Streaming sparse Gaussian process approximations". In: *Advances in Neural Information Processing Systems*. 2017, pp. 3299–3307.

[37] Jonathan Wenger et al. "Non-Parametric Calibration for Classification". In: *arXiv preprint arXiv:1906.04933* (2019). arXiv: `1906.04933`. URL: `https://github.com/JonathanWenger/pycalib`.