

# Combatting out-of-distribution errors using model-agnostic meta-learning for digital pathology

Freja Fagerblom, Karin Stacke and Jesper Molin

The self-archived postprint version of this conference paper is available at Linköping University Institutional Repository (DiVA):

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-174913>

N.B.: When citing this work, cite the original publication.

Fagerblom, F., Stacke, K., Molin, J., (2021), Combatting out-of-distribution errors using model-agnostic meta-learning for digital pathology, *Proceedings of SPIE Medical Imaging*, 116030S.  
<https://doi.org/10.1117/12.2579796>

Original publication available at:

<https://doi.org/10.1117/12.2579796>

Copyright: SPIE - International Society for Optical Engineering  
<https://www.spiedigitallibrary.org/>



# Combating Out-of-distribution Errors using Model-Agnostic Meta-Learning for Digital Pathology

Freja Fagerblom<sup>a</sup>, Karin Stacke<sup>a,b</sup>, and Jesper Molin<sup>a</sup>

<sup>a</sup>Sectra AB, Linköping, Sweden

<sup>b</sup>Linköping University, Linköping, Sweden

## ABSTRACT

Clinical deployment of systems based on deep neural networks is hampered by sensitivity to domain shift, caused by, e.g., new scanners or rare events, factors usually overcome by human supervision. We suggest a correct-then-predict approach, where the user labels a few samples of the new data for each slide, which is used to update the network. This few-shot meta-learning method is based on Model-Agnostic Meta-Learning (MAML), with the goal of training to adapt quickly to new tasks. Here we adapt and apply the method to the histopathological setting by identifying a task as a whole-slide image with its corresponding classification problem. We evaluated the method on three datasets, while purposefully leaving out-of-distribution data out from the training data, such as whole-slide images from other centers, scanners or with different tumor classes. Our results show that MAML outperforms conventionally trained baseline networks on all our datasets in average accuracy per slide. Furthermore, MAML is useful as a robustness mechanism to out-of-distribution data. The model becomes less sensitive to difference between whole-slide images and is viable for clinical implementation when used with the correct-then-predict workflow. This offers a reduced need for data annotation when training networks, and a reduced risk of performance loss when domain shift data occurs after deployment.

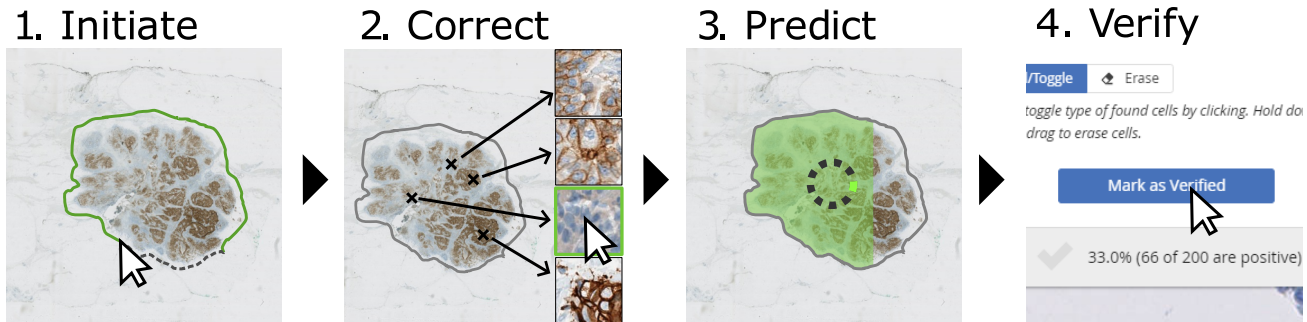
**Keywords:** Meta learning, digital pathology, active learning, deep learning, few-shot learning, domain shift, clinical implementation

## 1. INTRODUCTION

Clinical adoption of deep learning-based image-analysis systems within digital pathology has been held back by a number of different factors. Existing systems are typically sensitive to rare events and so-called domain shifts, which may occur when the deep learning network is given an image from a new scanner or with new staining characteristics. In addition, it is hard to predict both how and when domain shift occurs.

It is therefore common to use one or more safety nets when deep-learning predictions are used in a clinical setting. Recently published recommendations for the use of HER2-scoring algorithms from the College of American Pathologists<sup>1</sup> recommend for example that quality assurance procedures are used continuously to monitor system performance and that a pathologist visually verify the result. This *predict-then-correct* way to create human-in-the-loop systems uses a few minutes of the pathologist’s time to achieve better robustness compared to a fully automatic operation. In this study, we instead evaluate a *correct-then-predict* approach, where we first ask the pathologist to provide a small number of ground truth samples for each new case, and then use them to facilitate the prediction task, as shown in Figure 1. This results in increased accuracy of the initial predictions and reduction of needed corrections.

The workflow allowing users to interact and affect a model’s performance can be built using a number of machine learning techniques, such as active learning, few-shot learning and meta learning. Active learning aims to optimize training and minimize training samples needed by giving suggestive data samples for labeling during training. Yang et al.<sup>2</sup> evaluated this approach in the domain of histopathology. They were able to reduce the needed training data with 50%, but this still meant thousands of training samples. In few-shot learning,<sup>3</sup> typically only 1–20 annotated samples are needed to adapt a pre-trained model, which is a more feasible task for a pathologist. Medela et al.<sup>4</sup> investigated few-shot learning for histopathology based on Siamese networks, but still required 60 training images for adaption to new tissue types.



**Figure 1:** Envisioned *correct-then-predict* workflow where the user initiates the analysis and immediately is asked to correct a small random sample of image patches. These corrected image patches are fed into the prediction module that is capable of using this information to deliver better whole-slide predictions.

In meta learning, or learning to learn, a model is trained to quickly adapt to the new task, using only a few labeled samples. Different methods of meta learning have been applied to histopathology. Wen et al.<sup>5</sup> used a LSTM-based model for image classification, and Gamper et al.<sup>6</sup> used meta-learning to do one-class classification for abnormality detection. However, to our knowledge, no one has utilized meta learning for online adoption *per slide*. In this paper, we propose using Model-Agnostic Meta-Learning (MAML)<sup>7</sup> by handling each slide as a separate task, and quickly fine-tune a model for high-accuracy classification per slide. Using this approach, the algorithm becomes less sensitive to differences between slides commonly seen in clinical practice.

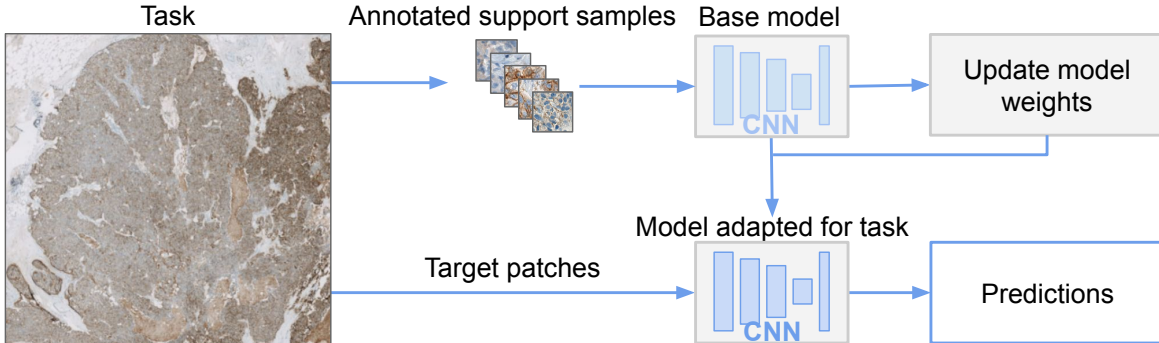
This work is based on the master thesis by Fagerblom,<sup>8</sup> where this submission expands the MAML implementation and extends the number of datasets evaluated. The contributions of this work consists of adapting a state-of-the-art meta-learning method to be used in a correct-then-predict scenario with histopathology images, as shown in Figure 1. We evaluate the performance of this setup on multiple datasets where we simulate out-of-distribution events, and show that MAML outperforms both vanilla CNN models and kNN based transfer learning using only 20 labeled samples, making this approach feasible for clinical implementation.

## 2. METHOD

The meta-learning framework Model-Agnostic Meta-Learning<sup>7</sup> (MAML) is a few-shot learning technique, with the purpose of finding good initialization parameters for a *base model*, which are able to update quickly for new tasks with only the information from a few *support samples*. The updated model is then used to classify the rest of the samples of the task, i.e., the *target samples*. Fine-tuning in transfer learning requires many data points to avoid overfitting to the training samples of the new domain. In contrast, MAML base-model parameters are being meta trained across several tasks while at the same time being optimized task-wise with fine-tuning using only a few data points from each task. During inference, the base-model parameters are initially the same for each task before being fine-tuned with labeled support samples.

In order to adapt MAML to the histopathology setting, we considered each whole-slide image (WSI) as a separate task, with a corresponding classification problem and a number of randomly sampled support and target patches. In many histopathology datasets, not all classes are present in all slides. Handling each slide as a separate task meant that there was an unknown number of classes present for each task, forcing the base model to learn parameters which quickly could adapt to tasks with different number of classes. Slide-specific adaption was done by using a small support set of randomly selected labeled patches from one WSI, simulating the realistic scenario where a pathologist is asked to classify a small number of random samples before model inference. The MAML prediction process in a histopathological setting is shown in Figure 2.

The MAML implementation as introduced by Finn et al.<sup>7</sup> has been known to be very sensitive to the choice of model architecture and hyper-parameters. We therefore use the suggested improvements made by Antoniou et al.,<sup>9</sup> with the difference of using warm restarts for the cosine annealing meta-optimizer learning rate scheduler to make the number of trained epochs more flexible.



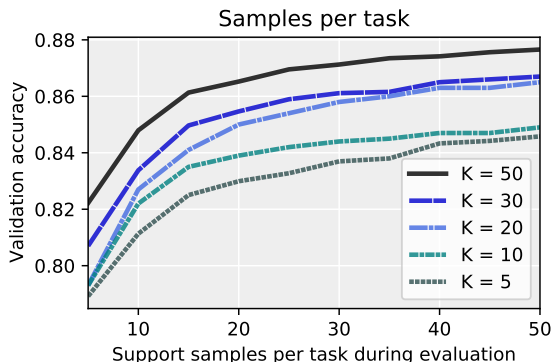
**Figure 2:** Prediction framework. For a task (WSI) a set of support patches are sampled randomly and passed through the base model. The information from the task specific loss is used to update the initial weights of the base model, which is then used to make predictions for the rest of the patches in the WSI.

## 2.1 Experiments

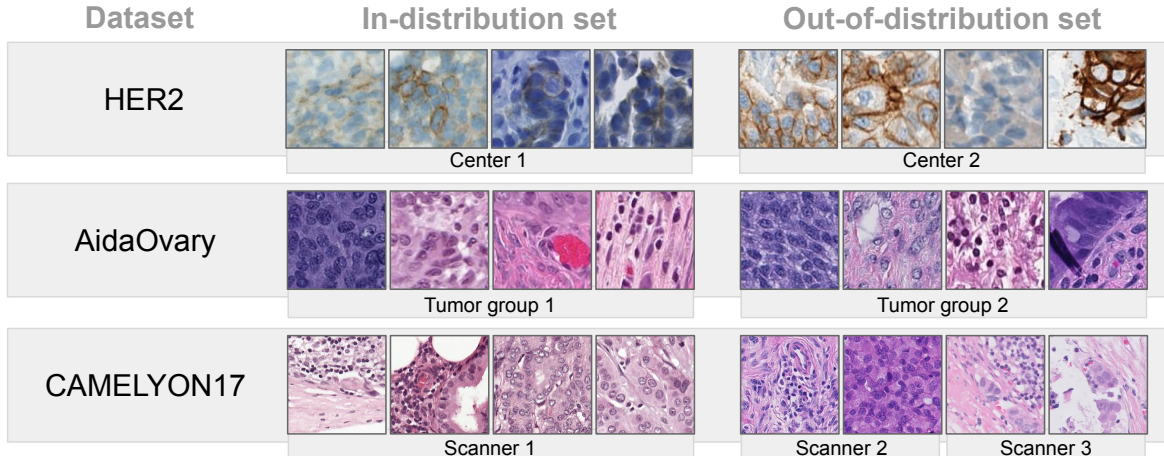
The aim of this work was to evaluate how slide-specific model adoption using MAML can handle out-of-distribution data, in comparison to conventional training. We therefore applied our method on three datasets, simulating different events where new characteristics were introduced during evaluation. The events were *out-of-center*, *out-of-malignancy* and *out-of-scanner*, i.e., data from different medical centers, of different malignancy types or from different scanners. Each dataset was separated into in-distribution and out-of-distribution sets according to these characteristics, where the in-distribution data was used for training (including validation) as well as evaluation and the out-of-distribution data only was used for evaluation.

Our MAML base model followed the same architecture as Antoniou et al.,<sup>9</sup> a 4-layer convolutional network ending with a fully connected layer. In all training sessions, learning rate was varied up to  $5 \cdot 10^{-5}$  using a cosine annealing scheduling, with warm restarts every 20<sup>th</sup> epoch. Adam optimizer and cross-entropy loss were used for all models. During training, random augmentation with horizontal and vertical flips, rotation by 180 degrees, scale (0.9–1.1) and color jitter (0.1) was applied.

As baseline, a non-meta-learning approach was evaluated. A model with the same architecture as the MAML base model was trained using supervised learning on the in-distribution datasets. Other larger model architectures were also evaluated, such as ResNet50<sup>10</sup> and Inception v3,<sup>11</sup> but did not show any significant improvements and were therefore not included in the results. Furthermore, a naive task adoption scheme was evaluated using the k-Nearest Neighbor (kNN) classifier, based on features extracted from the baseline model’s final convolutional layer. The implementation from Scikit-learn<sup>12</sup> was used, with three-neighbors’ vote and Euclidean distance



**Figure 3:** Validation accuracy depending on the number of patches sampled for task adaption.  $K$  support samples per task during MAML training on the HER2 dataset. More support patches during training yield higher validation accuracy even if the number of support samples is lower when evaluating the model.



**Figure 4:** Example patches from the three datasets and their respective in- and out-of-distribution sets. Characteristics differ depending on scanning machine, center for scanning or type of tumor.

metric. As fine-tuning the baseline model directly has been shown to easily overfit when trained on only a few samples,<sup>13</sup> the kNN approach was evaluated as a more robust method.

A MAML experiment consisted of training for 40 epochs with 200 iterations each, where an iteration was made up of a task batch of 16 WSIs (i.e., 16 tasks), with 50 support patches and 50 target patches sampled from each slide. The number of support and target patches as well as task batch size was limited due to GPU memory restrictions. To perform these experiments, we used 4 GPUs of a NVIDIA DGX-2 system. The performance of the model was evaluated on 100 validation tasks after each epoch. A specific WSI can occur several times during evaluation, but the support and target samples will be randomly selected each time.

Baseline and MAML models from the best three epochs respectively were evaluated again on a larger validation set for 1000 tasks, and the top-performing model was tested with 5000 tasks from the in- and out-of-distribution sets respectively. During the final evaluation, 20 support patches were sampled as we found that this is a reasonable trade-off between prediction performance and the correction effort that we expect a pathologist to find acceptable. The number of target patches was set to the smallest number of available patches over all slide images in the dataset, with an upper limit of 500 patches per slide due to computational restrictions. For the kNN classifier, the features of the 20 support samples extracted from the baseline model were used as training samples.

## 2.2 Datasets

The data used for the *out-of-center* test scenario was an in-house dataset with 800 HER2 stained slides from two medical centers, one of which was chosen as in-distribution and used for training, henceforth denoted HER2. The patches were labeled into five classes by varying staining strength, with annotation quality verified by an experienced pathologist. For the *out-of-malignancy* test scenario, the dataset consisting of 160 ovary slides were used, denoted AidaOvary, provided by Analytic Imaging Diagnostics Arena,<sup>14</sup> of different tumor types. Areas were annotated as tumor, normal tissue or artifact. The slides were separated based on tumor type, with the six largest types divided into the in-distribution set together with the vascular invasion type, a total of 70% of the slides (Tumor group 1), and the other 30% to the out-of-distribution set (Tumor group 2) with no patient overlap. We sampled patches from the annotated areas, excluding the artifact class, and then relabeled as either normal or malignant, creating a binary classification problem. This formulation of the problem may be less clinically relevant, but it allowed us to test MAML adaptive capacity on previously unseen morphology. We used the 50 lymph node slides containing tumor from the CAMELYON17<sup>15</sup> dataset. These slides come from five different medical centers, scanned on three different scanners. For the *out-of-scanner* dataset, the in-distribution set was WSIs scanned on one scanner origin from three centers, and two out-of-distribution sets were selected from the other two centers, each scanned with a different scanner. Patch size of  $128 \times 128$  pixels were used in all experiments, sampled at 0.5 microns per pixel resolution. Examples images are shown in Figure 4.

### 3. RESULTS

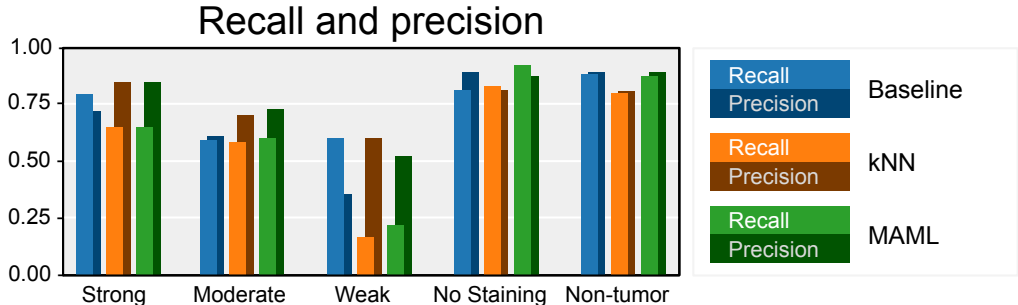
**Table 1:** Average patch accuracy per task and standard error. Training was done with 50 support samples per task and evaluation on 5000 tasks with 20 support samples per task. Datasets marked with † are out-of-distribution datasets corresponding to a realistic scenario where the training distribution does not cover all data that may be introduced to machine learning systems in digital pathology.

Dataset		Baseline	kNN	MAML
<b>HER2</b>	Center 1	84.39 ± 0.21%	81.96 ± 0.20%	<b>88.66 ± 0.15%</b>
	Center 2†	76.09 ± 0.27%	81.39 ± 0.22%	<b>84.46 ± 0.19%</b>
<b>AidaOvary</b>	Tumor group 1	87.00 ± 0.19%	<b>98.30 ± 0.10%</b>	97.65 ± 0.14%
	Tumor group 2†	84.77 ± 0.20%	93.17 ± 0.19%	<b>97.43 ± 0.13%</b>
<b>CAMELYON17</b>	Scanner 1	91.93 ± 0.13%	93.47 ± 0.09%	<b>96.48 ± 0.05%</b>
	Scanner 2†	50.73 ± 0.34%	94.27 ± 0.08%	<b>94.85 ± 0.09%</b>
	Scanner 3†	21.23 ± 0.41%	92.12 ± 0.14%	<b>95.84 ± 0.06%</b>

Our results show that MAML can be used as a robustness mechanism for several classification problems in histopathology. In Table 1, MAML outperforms the baseline model consistently while also dropping less in performance when domain-shift data is introduced. The simple baseline task adaption using kNN classification displays stable performance even on the out-of-distribution set but is unable to acquire the best accuracy on most datasets. This shows that the baseline model learns relevant features from the image patches but is unable to perform the final classification well in an out-of-distribution setting, where MAML shows the most notable improvement over the baseline. It should be noted that stain normalization possibly could improve the performance of baseline on the CAMELYON17 out-of-distribution sets, but MAML can handle the color differences even without.

The weakness of both MAML and kNN is that they depend heavily on the fine-tuning samples, which were picked randomly. The weakness of both MAML and kNN is that they heavily depend on the samples included in the support set, since these are used to fine-tune the model. The support set is selected randomly, and if no samples of a class is included in the support set, the performance of the model will be poor. One example of this is shown in Figure ?? . Here, recall and precision are shown for the out-of-distribution dataset of the HER2 dataset (center 2). Recall is significantly lower for the "weak" class, a class with generally very few samples per slide.

This is also indicated in Figure 3, where the effect of the number of support samples used during training and evaluation was investigated – more support samples always gave better results. Therefore, 50 support samples were used during training, but the number of samples were restricted to 20 during evaluation, as we wanted to mimic a more realistic scenario where these samples would be annotated by the pathologist. Still, with this restriction, MAML is able to consistently achieve high performance on all evaluated datasets, as shown in Table 1.



**Figure 5:** Class-wise recall and precision for the out-of-distribution HER2 dataset (Center 2). Recall is considerably lower for the weak staining class, which typically had few patches represented in WSIs.

## 4. DISCUSSION

From the results, it is clear that a non-meta-learning approach has limitations in clinical applications as out-of-distribution data most certainly will occur at some point. It is hard to predict when such domain shifts will take place, and they are therefore important to take into consideration when designing a workflow for clinical use.

Both the proposed correct-then-predict workflow with a MAML-based model, and the naive task adaption approach using kNN showed promising results to combat problems with out-of-distribution data, with MAML achieving higher accuracy overall. Even so, it is necessary to consider what classification problems in histopathology this framework is suitable for. First, since the technique require human labeling before prediction, it is not suitable for unsupervised analysis of scanned slides to e.g., prioritize cases in a clinical workflow before review. Instead, the technique should work better for tasks involving quantifying tissue or cells. From our results, MAML appear to work well for classifying anomalies, such as identifying tumorous tissue when there is enough of it to ensure that a few patches will be included in the support set. Rare events not occurring for many patients could be classified this way. It is however more difficult to classify events that are rare within a slide. It is difficult to achieve class balanced accuracy when certain classes are not well represented within a slide (even more so for the kNN method). This was often the case for example with the weak staining class in the HER2 dataset. It might be possible to investigate solutions to this by incorporating an active sampling approach, such as sample suggestions given by the base model or by manual sample selection. However, this needs to be studied further.

The proposed suggestions by Antoniou et al.<sup>9</sup> resulted in easier hyper-parameter tuning compared to the original MAML.<sup>7</sup> Although some changes, e.g., not using second order derivative for all epochs, resulted in shortened training time, others increased the computational complexity, such as across step and across parameter learning rates. Overall, the training process is heavily computationally demanding, which forced us to use a cluster of GPUs to achieve these results. However, the MAML adaptation is light weight and only adds milliseconds to inference cost, which is an important aspect for the intended clinical use. Furthermore, the effort to manually label support patches is low compared to a predict-then-correct approach. No formal experiments have been performed, but small internal experiments and data from usage logs indicate that it takes a pathologist roughly 1-2 second per patch to label it. Labeling 20 patches would therefore be expected to take around 30 seconds in total to perform in a clinical setting.

The MAML approach is shown to be a robustness mechanism, and is applied continuously when the model is deployed. Any domain shift that may arise over time can be handled directly, without special interventions such as stain color normalization or re-training with more data. This simple correct-then-predict workflow can make integration of AI tools in real-world applications more viable.

## 5. CONCLUSIONS

In this paper, we show that by asking a pathologist to only sample a handful of samples, it is possible to adapt a model to be robust against out-of-distribution changes. The method was evaluated on three datasets with different staining techniques and various classification tasks. It was evaluated on several out-of-distribution scenarios, with consistently high performance in all cases. Using this approach, the algorithm becomes less sensitive of differences between slides, regardless of whether the differences are introduced by morphology, scanner artifacts, new scanner hardware, or changes in slide preparation. Together with the *correct-then-predict* workflow, more robust model predictions can be made, resulting in a feasible solution for clinical implementation.

## ACKNOWLEDGMENTS

Freja Fagerblom would like to thank supervisors Abdelrahman Eldesokey and Michael Felsberg for their support during her master thesis work. The authors thank Region Gävleborg and University Medical Center Utrecht for contributing to the HER2 dataset used in this work.

## REFERENCES

- [1] Bui, M. M. et al., “Quantitative Image Analysis of Human Epidermal Growth Factor Receptor 2 Immunohistochemistry for Breast Cancer: Guideline From the College of American Pathologists,” *Archives of Pathology & Laboratory Medicine* **143**, 1180–1195 (Oct. 2019).
- [2] Yang, L. et al., “Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation,” *arXiv:1706.04737 [cs]* (June 2017). arXiv: 1706.04737.
- [3] Wang, Y. et al., “Generalizing from a Few Examples: A Survey on Few-Shot Learning,” *arXiv:1904.05046 [cs]* (Mar. 2020). arXiv: 1904.05046.
- [4] Medela, A. et al., “Few Shot Learning in Histopathological Images: Reducing the Need of Labeled Data on Biological Datasets,” in [*2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*], 1860–1864 (Apr. 2019). ISSN: 1945-8452.
- [5] Wen, Q. et al., “A meta-learning method for histopathology image classification based on LSTM-model,” in [*Tenth International Conference on Graphics and Image Processing (ICGIP 2018)*], **11069**, 110691H, International Society for Optics and Photonics (May 2019).
- [6] Gamper, J. et al., “Meta-SVDD: Probabilistic Meta-Learning for One-Class Classification in Cancer Histology Images,” *arXiv:2003.03109 [cs, eess]* (Mar. 2020). arXiv: 2003.03109.
- [7] Finn, C., Abbeel, P., and Levine, S., “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks,” *arXiv:1703.03400 [cs]* (July 2017). arXiv: 1703.03400.
- [8] Fagerblom, F., [*Model-Agnostic Meta-Learning for Digital Pathology*] (2020).
- [9] Antoniou, A., Edwards, H., and Storkey, A., “How to train your MAML,” *arXiv:1810.09502 [cs, stat]* (Mar. 2019). arXiv: 1810.09502.
- [10] He, K. et al., “Deep Residual Learning for Image Recognition,” *arXiv:1512.03385 [cs]* (Dec. 2015). arXiv: 1512.03385.
- [11] Szegedy, C. et al., “Rethinking the Inception Architecture for Computer Vision,” *arXiv:1512.00567 [cs]* (Dec. 2015). arXiv: 1512.00567.
- [12] Pedregosa, F. et al., “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
- [13] Ravi, S. and Larochelle, H., “Optimization as a Model for Few-Shot Learning,” *Proceedings of 5th International Conference on Learning Representations (ICLR)*, 11 (2017).
- [14] Lindman, K., F. Rose, J., Lindvall, M., and Bivik Stadler, C., “Ovary data from the Visual Sweden project DROID,” (2019).
- [15] Litjens, G. et al., “1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset,” *GigaScience* **7** (June 2018).