UNIVERSITY OF LEEDS

This is a repository copy of *Mitosis domain generalization in histopathology images — The MIDOG challenge*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/194640/

Version: Accepted Version

Article:
Aubreville, M, Stathonikos, N, Bertram, CA et al. (38 more authors) (2023) Mitosis domain generalization in histopathology images — The MIDOG challenge. Medical Image Analysis, 84. 102699. ISSN 1361-8415

https://doi.org/10.1016/j.media.2022.102699

© 2022 Elsevier B.V. All rights reserved. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/.

Reuse
This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can’t change the article in any way or use it commercially. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose
university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

arXiv:2204.03742v1 [eess.IV] 6 Apr 2022

Mitosis domain generalization in histopathology images - The MIDOG challenge

Marc Aubreville[a,*], Nikolas Stathonikos[b], Christof A. Bertram[c], Robert Klopfleisch[d], Natalie ter Hoeve[b], Francesco Ciompi[e], Frauke Wilm[f], Christian Marzahl[f], Taryn A. Donovan[g], Andreas Maier[f], Jack Breen[h], Nishant Ravikumar[h], Youjin Chung[i], Jinah Park[i], Ramin Nateghi[j], Fattaneh Pourakpour[k], Rutger H.J. Fick[l], Saima Ben Hadj[l], Mostafa Jahanifar[m], Nasir Rajpoot[m], Jakob Dexl[n], Thomas Wittenberg[n], Satoshi Kondo[o], Maxime W. Lafarge[p], Viktor H. Koelzer[p], Jingtang Liang[q], Yubo Wang[q], Xi Long[r], Jingxin Liu[s], Salar Razavi[t], April Khademi[t], Sen Yang[u], Xiyue Wang[v], Mitko Veta[w,**], Katharina Breininger[x,**]

[a] *Technische Hochschule Ingolstadt, Ingolstadt, Germany*
[b] *Pathology Department, UMC Utrecht, The Netherlands*
[c] *Institute of Pathology, University of Veterinary Medicine, Vienna, Austria*
[d] *Institute of Veterinary Pathology, Freie Universität Berlin, Berlin, Germany*
[e] *Computational Pathology Group, Radboud UMC Nijmegen, The Netherlands*
[f] *Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany*
[g] *Department of Anatomic Pathology, Schwarzman Animal Medical Center, New York, USA*
[h] *CISTIB Center for Computational Imaging and Simulation Technologies in Biomedicine, School of Computing, University of Leeds, UK*
[i] *Korea Advanced Institute of Science and Technology, Daejeon, South Korea*
[j] *Electrical and Electronics Engineering Department, Shiraz University of Technology, Shiraz, Iran*
[k] *Iranian Brain Mapping Biobank (IBMB), National Brain Mapping Laboratory (NBML), Tehran, Iran*
[l] *Tribun Health, Paris, France*
[m] *Tissue Image Analytics Centre, Department of Computer Science, University of Warwick, UK*
[n] *Fraunhofer-Institute for Integrated Circuits IIS: Erlangen, Germany*
[o] *Muroran Institute of Technology, Hokkaido, Japan*
[p] *Department of Pathology and Molecular Pathology, University Hospital and University of Zurich, Zurich, Switzerland*
[q] *School of Life Science and Technology, Xidian University, Shannxi, China*
[r] *Histo Pathology Diagnostic Center, Shanghai, China*
[s] *Xi'an Jiaotong-Liverpool University, Suzhou, China*
[t] *Image Analysis in Medicine Lab (IAMLAB), Electrical, Computer and Biomedical Engineering, Ryerson University, Toronto, ON, Canada*
[u] *Tencent AI Lab, Shenzhen 518057, China*
[v] *College of Computer Science, Sichuan University, Chengdu 610065, China*
[w] *Medical Image Analysis Group, TU Eindhoven, The Netherlands*
[x] *Department of Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany*

Abstract

The density of mitotic figures within tumor tissue is known to be highly correlated with tumor proliferation and thus is an important marker in tumor grading. Recognition of mitotic figures by pathologists is known to be subject to a strong inter-rater bias, which limits the prognostic value. State-of-the-art deep learning methods can support the expert in this assessment but are known to strongly deteriorate when applied in a different clinical environment than was used for training. One decisive component in the underlying domain shift has been identified as the variability caused by using different whole slide scanners. The goal of the MICCAI MIDOG 2021 challenge has been to propose and evaluate methods that counter this domain shift and derive scanner-agnostic mitosis detection algorithms. The challenge used a training set of 200 cases, split across four scanning systems. As a test set, an additional 100 cases split across four scanning systems, including two previously unseen scanners, were given. The best approaches performed on an expert level, with the winning algorithm yielding an F_1 score of 0.748 (CI95: 0.704-0.781). In this paper, we evaluate and compare the approaches that were submitted to the challenge and identify methodological factors contributing to better performance.

1. Introduction

Deep learning has revolutionized the field of digital histopathology in recent years, as methods continue to emerge that perform on par or even surpass the performance of human experts in specific tasks (Levine et al., 2019; Karimi et al., 2019; Aubreville et al., 2020b). The application of Artificial Intelligence (AI) methods in a computer-aided diagnostic workflow is especially beneficial for quantitative routine tasks, allowing for a faster diagnostic process, or for tasks with a known high inter-rater variability to increase diagnostic reproducibility by reducing diagnostic bias.

One task for which both of these conditions are met is tumor grading, i.e. the assessment of the malignant potential of a tumor from histological specimens (Veta et al., 2015; Balkenhol et al., 2019). Many tumor grading schemes rely on the identification and the counting of cells in the process of cell division (mitotic figures). The density of mitotic figures in an area (mitotic activity) is known to be highly correlated with proliferation of the tumor (Baak et al., 2008). Yet, identification of mitotic figures is known to suffer from a significant inter-rater variability (Meyer et al., 2005, 2009), which might be the dominant limiting factor for the prognostic value. Studies have shown that by using AI-based methods both reproducibility and accuracy can be increased (Bertram et al., 2021; Balkenhol et al., 2019).

One major limitation of the state-of-the-art deep learning-based methods is that their performance is known to significantly deteriorate with a covariate shift of the images, i.e. a change in visual representation between images that the model was trained upon and those that it encounters during inference in a clinical diagnostic workflow. Contrary to machine learning models, humans can often adapt seamlessly to this shift (Stacke et al., 2020; Aubreville et al., 2021b). The main causes for such a domain shift in histopathology are the staining procedure (which can differ over time and/or across laboratories), the acquisition device (whole slide scanner), and the tumor type itself (different tumor cell morphology and tissue architecture). While a limitation of an algorithmic aid to a specific subclass of tissues may be an acceptable restriction, a limitation towards factors that characterize a laboratory environment (tissue preparation, staining procedure, scanner) prevents the use of such methods in diagnostic practice across labs.

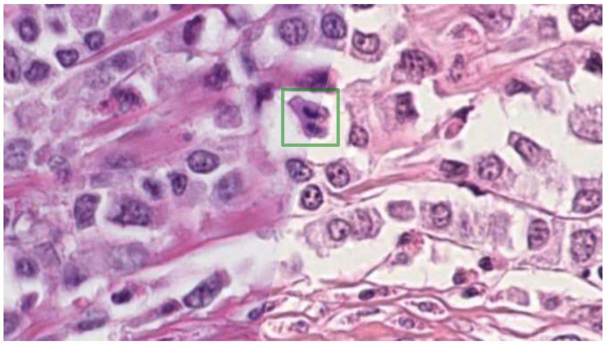

Figure 1: Breast cancer tissue, acquired using a Hamamatsu NanoZoomer XR (left, Scanner A) and a Hamamatsu NanoZoomer S360 (right, Scanner B). Besides a color shift, the depth of field is also affected by using a different scanner, caused by different optical properties of the objective. A mitotic figure in anaphase is indicated by a green box.

This motivated the design of the MItosis DOmain Generalization (MIDOG) 2021 challenge. As the Whole Slide Image (WSI) scanner used for digitization was identified to cause a strong domain shift (Aubreville et al., 2021b), presumably even stronger than the domain shift caused by the tissue preparation and staining procedure (Aubreville et al., 2020a), the challenge focused on the task of generalizing against scanner-induced domain shift for the identification of mitotic figures in histopathology images. Detection of mitotic figures is prone to shifts of color representation of the digital image, since both color and structural patterns are required for a proper identification. Besides color representation, which is influenced by the light source and sensors of the WSI scanner and proprietary color calibration schemes, the optical parameters of the microscope, such as the numerical aperture, also influence the representation of mitotic figures in the digital image (see Fig. 1).

Challenge format and task

The challenge was held in conjunction with the 2021 International Conference on Medical Image Computing

[*]Corresponding author: Email: marc.aubreville@thi.de;
[**]Authors contributed equally.

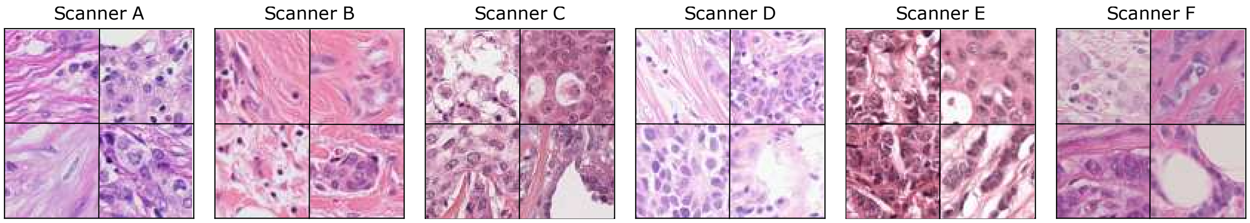

Figure 2: Random crops of breast cancer tissue from all six scanners of the training and test set, showing the visual variability between the scanners.

and Computer Assisted Intervention (MICCAI) conference as a one-time event. The structured challenge description (Aubreville et al., 2021a) was accepted after a single-blind peer review process. Participants were able to register and obtain the training data five months prior to the challenge submission deadline, allowing for sufficient time to develop and evaluate their algorithms. After registration, the participants were provided with the training set data and a description of the images and annotations, together with a Jupyter notebook showcasing how to work with the data and how to train an example object detector (RetinaNet) with it.

The task of the challenge was the detection of mitotic figures on regions of interests (ROIs) of a predefined size. For the training set, 200 cases of breast cancer (representing 200 patients) were retrieved, the ROI was selected by a trained pathologist and the WSIs were digitized using four different WSI scanners (50 cases each). The test set consisted of 80 cases digitized by four scanners (20 each) out of which two scanners were part of the training set and two were unseen. The challenge dataset represents a good trade-off between capturing the naturally occurring variability of WSIs and time invested for annotation (Aubreville et al., 2021a). The number of cases per scanner allows for a realistic estimation of the performance. Due to the task being about generalization, prior knowledge about the images of the test set needed to be excluded. Thus, the participants had no knowledge of the nature of the test scanners, no access to the test images and had to submit a Docker container to be evaluated automatically on the test data. For this, a reference algorithm (Wilm et al., 2022) embedded into a Docker container[1] was made available to the participants alongside a textual description and video tutorials on how to use it.

All participating teams submitted their fully automatic algorithm containers on the grand-challenge.org website. To check container functionality and algorithmic validity, a preliminary test set (20 cases, 5 per scanner, same scanners as in the test set) was made available for automatic evaluation on the platform, two weeks prior to the submission deadline. The evaluation container for the challenge was also made available on github[2]. Participants were not permitted to use other sources of images to train their models, besides general purpose datasets such as ImageNet. In order to prevent overfitting to the characteristics of the preliminary test seft, the number of submissions was limited to one per day. After 15 days, the preliminary test phase was closed and the final test phase was started, to which the participants were able to only submit once. Alongside with the submission for evaluation on the final test set, participating teams had to provide a short paper, describing their approach and their preliminary results, on a publicly available pre-print server. Further details about the challenge, including details about the submission instructions, the publication policy and the timetable can be found in the structured challenge description (Aubreville et al., 2021a).

Initially, 237 individuals registered on the challenge website[3] and 161 users joined the challenge on the grand-challenge.org platform. Members of the organizers' institutes were not allowed to parcitipate in the challenge. 46 users submitted at least one docker container to the challenge. At the end, 17 teams made a submission to

[1] https://github.com/DeepPathology/MIDOG_reference_docker

[2] https://github.com/DeepPathology/MIDOG_evaluation_docker/

[3] https://imig.science/midog/

the final test set. Single-blind peer review was carried out on all submitted short papers. After the peer review, we invited all teams that passed peer review and where the approach exceeded a minimum score of $F_1 \geq 0.6$ to the workshop (12/17 teams, acceptance rate 70.6%). The approaches presented in the workshop also are compared in this paper.

2. Material and methods

The main design principles of the challenge were:

1. To have a generally representative dataset of a relevant disease and an important diagnostic task.
2. To reflect a real clinical use case by ensuring a truly independent hold out set with unknown characteristics.
3. To achieve high label quality to ensure an accurate evaluation.

With around 2.3 million cases in 2020, breast cancer is one of the cancer types with the highest prevalence (Wild et al., 2020). Patients can benefit from adjuvant therapies significantly, However, aggressive therapies also carry the risk of serious side effects and thus should be restricted to patients with unfavorable prognostic markers, such as high tumors proliferation (Van Diest et al., 2004). One marker strongly correlated with proliferation is the mitotic count, i.e., the assessment of cells undergoing cell division (mitosis) in a defined area (commonly 10 high-power fields, here defined as $2.0mm^2$) (Elston and Ellis, 1991). This ROI is selected by an expert pathologist within the tumor area with the presumed highest mitotic activity in a hematoxylin and eosin-stained digitized microscopy slide. Since this task is part of many tumor grading schemes (e.g., meningioma (Louis et al., 2016) or lung adenocarcinoma (Moreira et al., 2020)), it is highly relevant for general tumor prognostication and was chosen as target task for our challenge.

In order to ensure representativeness of the dataset, we chose a rigorous inclusion scheme for the challenge cohort. For the challenge to yield trustworthy results and especially to avoid a methodological overfitting, the independent test set needed to stay a true hold out, i.e. completely hidden to the participants.

To ensure the quality of our evaluation, the labeling quality was improved by using multiple experts and, additionally, a machine-learning-augmented annotation pipeline (see below).

2.1. Challenge cohort

The challenge dataset consists of 300 breast cancer cases and was curated from a retrospective, consecutive selection taken from the diagnostic archive of the University Medical Center (UMC) Utrecht, The Netherlands. All samples were resected solely for diagnostic purposes. Inclusion criteria were the availability of the microscopy (glass) slide, a confirmed breast cancer excision (as documented in the patient record), the availability of a pathology report with a documented mitotic count and that the patients did not opt out for the use of their data in research projects. Prior to handing over to project partners within the organization committee, all slides and clinical meta data were anonymized. Additionally, for the use in the challenge we obtained approval by the institutional review board of UMC Utrecht (reference: TCBio 20-776). The specimen was preprocessed according to clinical standard routine and stained with hematoxylin and eosin (H&E) dye. Subsequently, the cases were randomly split into the training set (200 cases) and the preliminary (20 cases) and final (80 cases) test set. Within each of those sets, we performed another random split, to assign the cases to the scanners, i.e. we split up the 200 cases of the training set into 50 cases for each of the four training scanners (A,B,C,D, see below), and the 80+20 cases of the test sets into 20+5 for each of the the four test scanners (A,D,E,F, see below). By this procedure, we can expect no significant biases in any of the subsets of the dataset and assume a high degree of representativeness.

2.2. Image acquisition

We used four different scanners for the digitization of the training set:

- Scanner A: Hamamatsu NanoZoomer XR (C12000-22, Hamamatsu, Hamamatsu City, Japan), optical resolution: 0.23 microns/px at 40x magnification

- Scanner B: Hamamatsu NanoZoomer S360 (Hamamatsu, Hamamatsu City, Japan), optical resolution: 0.23 microns/px at 40x magnification

- Scanner C: Aperio Scanscope CS2 (Leica Biosystems, Nussloch, Germany), optical resolution: 0.25 microns/px at 40x magnification

- Scanner D: Leica Aperio GT 450 (Leica Biosystems, Nussloch, Germany), optical resolution: 0.26 microns/px at 40x magnification, custom optics by Leica Microsystems for native 40x scanning with 1 mm field of view (FOV)

Scanner A is the scanner that is used in clinical practice to digitize all slides at UMC Utrecht. Therefore, all slides of our dataset are also available scanned by this scanner and we use this as our reference scanner to counter a possible scanner-caused bias in the region of interest selection.

For the test set, we used a mix of known and unknown scanners to test simultaneously for performance on in-domain scanners and for generalization to out-of-domain scanners. In addition to the scanners A and D, the test set was scanned with:

- Scanner E: 3DHISTECH Panoramic 1000 (3DHISTECH, Budapest, Hungary), optical resolution: 0.24 microns/px at 20x magnification, Plan-Apochromat objective, numerical aperture of lens: 0.8

- Scanner F: Hamamatsu NanoZoomer 2.0RS (C10730-12, Hamamatsu, Hamamatsu City, Japan), optical resolution: 0.23 microns/px at 40x magnification, numerical aperture of lens: 0.75

For scanner D, which was part of the training and the test set, no labels were provided as part of the training set. Hence, this scanner was included for the sole purpose of providing data for approaches performing unsupervised domain adaptation.

Each of the slides was scanned by the assigned WSI scanner as well as by the clinical reference scanner (Scanner A). To reduce bias which might be caused by different tissue representation on other scanners, a trained pathologist (C.B.) selected a region of interest spanning 2.0 mm^2 on the reference scans. For streamlined dataset creation, all WSIs were uploaded to a central server running a collaborative annotation software (Marzahl et al., 2021a). There, the reference scans were registered to the respective image acquired by different scanners using a quadtree-based WSI registration method by Marzahl *et al.*(Marzahl et al., 2021b). The registration was manually fine-tuned and quality checked subsequently. Finally, the defined ROIs were extracted from the images acquired by all scanners.

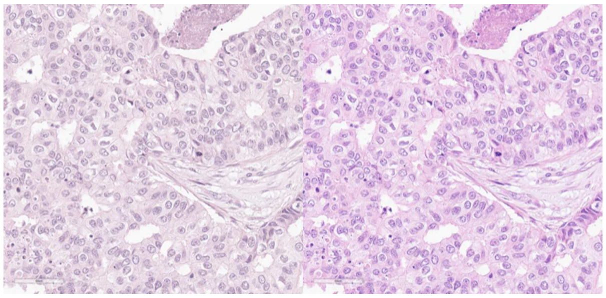

Figure 3: Image from Scanner D in the original (left) scanner color space and in the standard RGB color space (right).

For the Leica and Aperio scanner, the scanners color profile was available from the WSIs, which enabled the organizers to convert those images into the standard RGB color space. Regardless of this step, significant differences in color can be observed (see Fig. 3).

2.3. Labeling

In the process of labeling, three expert pathologists (C.B., R.K., T.D.) were involved. All experts work at different institutions (in different countries) and have 6+ years of professional experience and a demonstrably high level of expertise in mitotic figure recognition. Additionally, all experts agreed on common criteria for mitotic figure identification (Donovan et al., 2021).

The inter-rater variability of mitosis identification can be attributed to two main factors: Most notably, experts disagree on individual mitotic figures (object-level disagreement) due to morphological overlap with imposters (Veta et al., 2016). For this reason, most mitotic figures datasets were annotated as consensus voting between multiple experts (Veta et al., 2015, 2019; Roux et al., 2014; Bertram et al., 2019, 2020b; Aubreville et al., 2020a). Additionally, as mitotic figures can be considered sparse events in most WSIs and sometimes faint structures, experts tend to miss especially less recognizable objects when screening the image (Bertram et al., 2021, 2019).

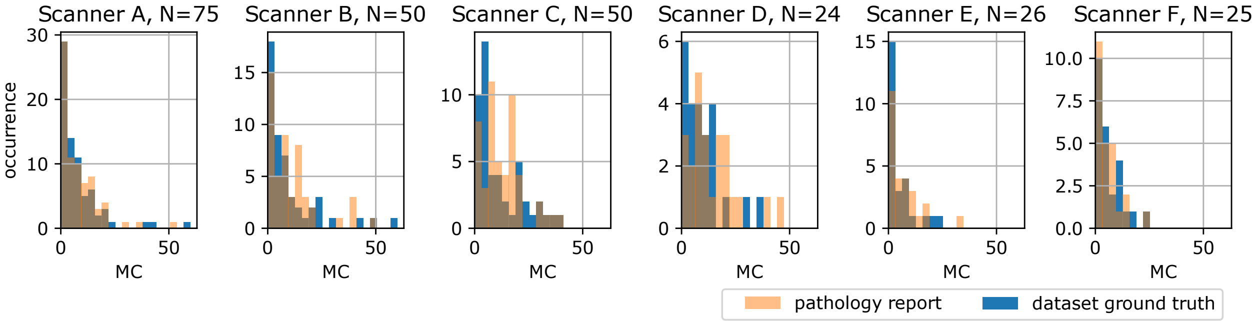

Figure 4: Distribution (histogram) of mitotic count (MC) across cases and scanners, according to the the pathology report (orange) and the MIDOG dataset (blue).

To account for both effects and create a high-quality dataset, we employed a machine learning-aided strategy (Bertram et al., 2020b): Initially, a single experienced expert (C.B.) screened all ROIs for mitotic figures and for a roughly equal number of imposters / hard negatives (non-mitotic cells with morphological similarity to mitotic figures). We trained a customized (Marzahl et al., 2020) RetinaNet (Lin et al., 2017) to identify cells that were missed in this initial labeling (mitotic figure candidates). Using a very low detection threshold on the model output, we ensured that this approach yielded high sensitivity (and low specificity). All mitotic figure candidates were then handed to the primary expert, again, to sort into missed mitotic figures and non-mitotic structures (which comprised the vast majority).

To reduce bias in the labeling process introduced by the first expert, all manual cell labels and detections from the algorithmic augmentation step were class-blinded and handed to a secondary expert (R.K.). The secondary expert then sorted all cells into mitotic figures and non-mitotic cells. In the case of agreement, the label was accepted as ground truth. In case of disagreement, the cells were again class-blinded and given to a third expert (T.D.), who then made the final decision.

In the dataset, the challenge organizers provided squared approximated bounding boxes of equal size (50px) to the participants. The non-mitotic cells (hard negatives) were provided alongside the true mitotic figures to enable the use within sampling schemes.

2.4. Dataset statistics

As Fig. 4 shows, the mitotic count (MC) (i.e., the number of mitoses per 2.0 mm^2) follows a similar distribution over all scanners in the original pathology report and the MIDOG dataset. Differences between the MC of the pathology report and the MIDOG dataset were not significant (two-sided paired t-test: p-value: 1.000, scipy stats package version 1.7.3). The intraclass correlation (ICC, two-way, single measurements, random raters) of the MC ratings indicate a substantial agreement (ICC2=0.684, pingouin package version 0.5.1) as of the definition by Hallgren (Hallgren, 2012), which can be attributed to different reasons. Firstly, it cannot be guaranteed that the slides that were included in this dataset were the same that were used for clinical MC assessment, as oftentimes multiple slides per case exist. Secondly, the annotation methodology used here differs substantially from how the MC is determined in a clinical setting. Lastly, even if the same slides were used, the MC is known to be highly area-dependent, which might further contribute to the discordance (Bertram et al., 2020a).

As depicted in Fig. 4, there is a high number of potentially low-grade tumors within the dataset, which is, however, reflective of the general population of cases at a tertiary hospital.

2.5. Reference approach

As a starting point, and to familiarize the participants with the submission process, the organizers provided a repository including an exemplary Docker container to all participants. The container provided a trained model including model weights and all scripts to run inference with it in the test environment. Alongside, the organizers made a description of the reference approach available to all participants during the challenge. The approach (Wilm et al., 2022) was based upon a customized (Marzahl et al., 2020) RetinaNet (Lin et al., 2017) implementation, where object classification and bounding box regression are solely performed at the highest resolution of the feature pyramid network. At the end of the encoder, a domain classification head was attached to the network. The task of the domain classification head is the discrimination of the four scanners of the training set. In between both, a gradient reversal layer (Ganin et al., 2016) was attached. This layer acts as a unity transform in the forward pass of the network, but inverts the networks' gradients, weighted by a constant, in the backward pass. This has the effect of adversarial training, i.e. of making the features less discriminative for the domain, and effectively reducing the domain covariate shift in feature space (Lafarge et al., 2019). The model uses a combined loss with terms for domain discrimination, object classification, and bounding box regression, using focal loss (Lin et al., 2017) for both classification tasks and a smooth L1 loss for the regression task. For scanner D (i.e., the scanner without annotations), only the domain-adversarial part of the loss is active. Besides this domain adaptation technique, the model only uses standard image augmentation techniques (affine transforms, flipping, contrast and brightness adjustments). For model selection, the area under the precision recall curve (AUPRC) was calculated for a validation set, consisting of a selection of ten cases of the scanners from the training set where annotation data was available (scanners A,B, and C). The model with the highest average AUPRC value was selected to be run on the test set. There was no access to any of the test sets during the development of the reference approach and no knowledge of the scanners selected for testing.

Additionally, Wilm et al. (2022) provide a baseline just using standard augmentation, which we will refer to as CNN baseline in the following. This model is using the same network topology and training scheme as the Reference model, but missing domain-augmentation techniques besides standard image augmentation (brightness, contrast, random zoom, flipping and rotation).

2.6. Evaluation methods

All participants were required to submit their approach as a docker container to the grand-challenge.org platform. There, the containers were automatically evaluated for each image of the test set independently. The participants had no access to any of the test images during the challenge and detailed detection results were also not available to them. This was done in order to ensure a fair comparison of the approaches and to discourage overfitting or manual fine-tuning towards the test set.

To ensure proper functionality of the automatically evaluated container images, we provided evaluation results on the preliminary test set. The main metrics (F_1 score, precision and recall) of submitted approaches were made available on a public leaderboard.

For the overall rating, the F_1 score was the primary metric. We calculated the F_1 score over all N processed slides as

$$F_1 = \frac{2\sum_k^N \mathrm{TP}_k}{2\sum_k^N \mathrm{TP}_k + \sum_k^N \mathrm{FN}_k + \sum_k^N \mathrm{FP}_k}$$

where TP_k, FN_k, and FP_k are the number of true positive, false negative and false positive detections on slide k and N is the total number of slides ($N = 80$ for the final test set and $N = 20$ for the preliminary test set).

The F_1 score was chosen as the main evaluation metric because it is defined as harmonic mean of precision and recall, and both underestimation as well as overestimation of the MC are equally severe for the diagnostic process. We did opt to calculate the overall F_1 instead of the mean F_1 of all slides, since slides with a low prevalence of mitotic figures would be overrepresented in this average.

We defined a detection to be a true positive whenever the Euclidean distance between a mitotic figure annotation and the detection was less than $7.5\,\mu m$. This value corresponds to the average size of mitotic figures in our dataset and provides a reasonable tolerance for misalignment of detection and ground truth labels. All detections not within $7.5\,\mu m$ of a ground truth mitotic figure annotation were considered false positives. Multiple detections of an already detected object were also counted as false positives, since they would introduce a positive bias to the MC. All ground truth mitotic figures without a detection within a proximity of $7.5\,\mu m$ are considered false negatives.

Table 1: Overview of methods submitted to the MIDOG challenge.

Team name	Core method/architecture	Multi-stage	Ensemble or TTA	Data augmentation					Domain adaptation		Used unlabelled domain	Used additional labels
				Color	Staining	Brightness	Contrast	Synthesis	Staining normalization	Other		
Reference method (Wilm et al., 2022)	RetinaNet Lin et al. (2017) with ResNet-18 He et al. (2016) encoder	✗	✗	✗	✗	✓	✓	✗	✗	Domain-adversarial training Pasqualino et al. (2021)	✓	✗
CNN baseline (Wilm et al., 2022)	RetinaNet Lin et al. (2017) with ResNet-18 He et al. (2016) encoder	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗
AI medical (Yang et al., 2022)	SK-UNet Wang et al. (2021)	✗	✗	✓	✗	✓	✗	Fourier domain mixing Yang and Soatto (2020)	✗	✗	✓	Mitosis segmentations
TIA Centre (Jahanifar et al., 2022)	Efficient-UNet Jahanifar et al. (2021) (candidate segmentation), Efficient-Net-B7 Tan and Le (2019) (candidate classification)	✗	✓	✓	✗	✓	✓	✗	Vahadane et al. Vahadane et al. (2016)	✗	✗	Mitosis segmentations
Tribun Health-care (Fick et al., 2022)	Mask-RCNN He et al. (2017) (candidate detection), ResNet-50 He et al. (2016) and DenseNet-201 Huang et al. (2017) (candidate classification)	✓	✓	✗	✗	✗	✗	CycleGAN de Bel et al. (2021)	✗	✗	✓	Mitosis segmentations
CGV (Chung et al., 2022)	RetinaNet Lin et al. (2017) with ResNet-101 He et al. (2016) encoder	✗	✗	✗	✗	✗	✗	StarGAN Choi et al. (2018)	✗	✗	✓	✗
XidianU-OUC (Liang et al., 2022)	DetectorRS Qiao et al. (2021) (candidate detection), ensemble of 5 models for candidate classification	✓	✓	✗	✗	✓	✓	✗	Macenko et al. Macenko et al. (2009)	✗	✗	✗
IAMLAB (Razavi et al., 2022)	Cascade R-CNN Cai and Vasconcelos (2018) with ResNet-101 He et al. (2016) encoder	✓	✗	✓	✗	✗	✓	✗	Macenko et al. Macenko et al. (2009)	✗	✗	✗
No.0 (Long et al., 2022)	Cascade R-CNN Cai and Vasconcelos (2018) with ResNet-50 He et al. (2016) encoder	✓	✗	✗	✓	✗	✗	✗	✗	Domain-adversarial training with PatchGAN Isola et al. (2017)	✓	✗
jdex (Dexl et al., 2022)	RetinaNet Lin et al. (2017) with Efficient-Net-B0 Tan and Le (2019) encoder	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗
Leeds (Breen et al., 2022)	UNet with ResNet-152 He et al. (2016) encoder	✗	✗	✓	✗	✓	✓	✗	✗	✗	✗	✗
PixelPath-AI (Nateghi and Pourakpour, 2022)	Faster-RCNN Ren et al. (2015) (candidate detection), Efficient-Net-B0 Tan and Le (2019) (candidate classification)	✓	✓	✗	✓	✓	✓	✗	✗	✗	✗	✗
SK (Kondo, 2022)	Thresholding of the blue ratio image (candidate detection) Chang et al. (2012), ResNet He et al. (2016) (candidate classification)	✓	✗	✓	✗	✓	✓	✗	✗	Domain-adversarial training Ganin et al. (2016)	✓	✗
ML (Lafarge and Koelzer, 2022)	Rotation-invariant CNN Lafarge et al. (2021) with 7 trainable layers	✗	✓	✓	✗	✓	✓	✗	✗	✗	✗	✗

For the evaluation in this paper, we also calculated the F_1 metric for each scanner. Further, we performed bootstrapping process (Hall, 1994) where we randomly selected M cases with replacement (where M is the number of images available per scanner or overall). This process was repeated 10,000 times to be able to derive a statistical distribution for the F_1 metric. This was also used to estimate the 5% and 95% confidence intervals for the F_1 score.

2.7. Post-challenge ensembling of approaches

We were also interested in discriminating parts of the dataset that were particularly easy or difficult to detect. In the same way, we wanted to see if there were hard negative candidates that confused a significant portion of the models.

For this, in a first step the detection results of approaches were matched against the ground truth, to yield the list of positives (false negatives and true positives) and false positives. For the matching we used a KDTree-based approach (Marzahl et al., 2020), where the detection and ground truth centroids were not allowed to exceed an euclidean distance of 7.5 μm.

The list of false positives was then, again, grouped using the same proximity criterion using a KD-tree. This was done in order to avoid counting false positives with slightly differing centroid coordinates multiple times, since the false positives can't be assigned to ground truth reference coordinates. We then assigned to each unique false positive (i.e., set of false detections within the distance of 7.5μm) the number of models that opted in favor of this detection. Using this methodology, we were able to calculate the number of votes for each detected cell in the set of detections from all models, and thus judge hard examples from easy ones for the ensemble of all models.

Additionally, we wanted to evaluate this for the best methods in the field. If we can assume the error to be independent between models, we can hypothesize that the ensemble of the top methods (all performing similarly) could outperform the individual methods. This is also interesting to investigate for ceiling effects in the evaluation caused by label noise in the test set: If the model significantly outperforms the individual models, we can assume that the performance evaluation is currently not limited by such labeling inconsistencies. For this reason, we also constructed an ensemble consisting of all approaches exceeding the baseline, and the baseline. This ensemble of models represents the top five of approaches in the challenge and is denoted as **top5 ensemble** in the remainder of this work. Since model scores of the individual approaches were unavailable, the ensemble was a simple majority vote.

3. Overview of the submitted methods

All submitted methods used convolutional neural networks (CNNs) for the task. Table 1 gives an overview of the network architectures, augmentation, and normalization strategies that were employed. The remainder of this chapter discusses several aspects of the strategies that were employed.

The detailed descriptions of all methods are published in the proceedings (Aubreville et al., 2022). Here, we want to report on common trends, interesting differences and strategies across the methods to provide readers with an insight on how the task of mitosis detection under domain shift can be tackled. Except for the aspect of translating the task into an (instance) segmentation approach with pixel-level masks, we did not see clear “winning” strategies. Instead, we believe that each approach found a strategy that put together matching operators, with some more interchangeable than others.

3.1. Single-stage and multi-stage detection approaches

Five out of the twelve teams submitted a multi-stage approach for detecting mitotic figures with the first stage generating a list of candidates (targeting a high recall with all mitotic figures included). The second stage then classified whether the extracted patches contained a mitotic figure or not. The first stage was either based on image features (Kondo, 2022), or used an object detection network (Faster R-CNN (Nateghi and Pourakpour, 2022), Mask-R-CNN (Fick et al., 2022), U-Net (Jahanifar et al., 2021), DetectorRS (Liang et al., 2022))). To refine these candidates, ResNet (ResNet50 or larger) or EfficientNet architectures (B0 / B7) were used. Two approaches used Cascade-R-CNN architectures (Long et al., 2022; Razavi et al., 2022), which are inherently multi-stage with sequentially trained detectors and which may therefore be seen as a way to automate the multi-staging. The remaining five teams used a RetinaNet architecture like the reference approach (Chung et al., 2022; Dexl et al., 2022), the predictions of a U-Net directly (Yang et al., 2022; Breen et al., 2022) or a rotation-invariant CNN (Lafarge and Koelzer, 2022).

3.2. Sampling strategies, loss functions and training mechanisms

Mitotic figure detection is a highly imbalanced problem and the ROIs provided in the MIDOG dataset are no exceptions. Most ROIs contained only 20 or fewer mitotic figures (see Fig. 4) and large areas of highly variable background. Additionally, nuclei, debris or necrotic cells may look extremely similar to mitotic cells (imposters) and therefore pose a considerable challenge whereas other regions can be easily disregarded. This typically prohibits random sampling of the input data and instead requires strategies to deal with this imbalance. Within the submitted approaches, different methods were employed for this, including a focal loss (Lin et al., 2017) which adjusts the loss for easy samples and was used by Yang et al. (2022); Fick et al. (2022); Breen et al. (2022), and Dexl et al. (2022). For segmentation-based approaches, typically a Dice loss or a Jaccard loss was used. Alternatively or additionally, most teams opted for targeted data sampling to ensure sufficient coverage of mitotic figures in each batch, e.g., by randomly undersampling non-mitotic regions (Jahanifar et al., 2022), excluding regions without mitotic figures (Fick et al., 2022), or filtering out easy patches directly (Lafarge and Koelzer, 2022).

In addition to approaching the task differently (i.e., detection vs. segmentation vs. classification), some teams opted to enrich the task with a domain-adversarial training mechanism (Long et al., 2022; Kondo, 2022) similar to the reference approach to encourage domain-independent feature extraction. This was also one of two strategies of using the unlabeled scanner provided in the training set,

Team	overall	Scanner A	Scanner D	Scanner E	Scanner F
Reference	0.718 [0.665,0.762]	0.791 [0.673,0.818]	0.708 [0.620,0.766]	0.718 [0.631,0.811]	0.593 [0.551,0.719]
CNN baseline	0.698 [0.639,0.745]	0.687 [0.650,0.812]	0.700 [0.621,0.770]	0.657 [0.519,0.747]	0.606 [0.521,0.717]
Top5 ensemble	**0.773** [0.722,0.813]	0.796 [0.748,0.874]	**0.745** [0.667,0.780]	0.787 [0.744,0.861]	0.642 [0.581,0.761]
AI medical	**0.748** [0.704,0.781]	0.793 [0.729,0.830]	**0.728** [0.643,0.780]	0.781 [0.708,0.843]	0.634 [0.583,0.732]
TIA Centre	0.747 [0.693,0.790]	0.837 [0.692,0.857]	0.677 [0.625,0.759]	**0.808** [0.683,0.837]	0.667 [0.578,0.768]
Tribvn Healthcare	0.736 [0.670,0.792]	**0.848** [0.731,0.875]	0.631 [0.608,0.768]	0.795 [0.633,0.848]	0.557 [0.498,0.712]
CGV	0.724 [0.657,0.779]	0.829 [0.750,0.867]	0.643 [0.547,0.728]	0.675 [0.639,0.836]	0.557 [0.521,0.697]
XidianU-OUC	0.707 [0.633,0.768]	0.800 [0.703,0.863]	0.655 [0.528,0.696]	0.795 [0.586,0.814]	0.673 [0.487,0.696]
IAMLAB	0.706 [0.650,0.748]	0.695 [0.646,0.809]	0.721 [0.642,0.757]	0.710 [0.496,0.824]	**0.690** [0.493,0.681]
No.0	0.701 [0.637,0.752]	0.826 [0.652,0.837]	0.698 [0.562,0.718]	0.757 [0.571,0.777]	0.632 [0.553,0.696]
jdex	0.696 [0.639,0.739]	0.782 [0.737,0.820]	0.682 [0.575,0.751]	0.667 [0.542,0.741]	0.430 [0.459,0.688]
Leeds	0.686 [0.620,0.737]	0.774 [0.624,0.795]	0.549 [0.503,0.699]	0.696 [0.594,0.786]	0.632 [0.547,0.742]
PixelPath-AI	0.676 [0.615,0.723]	0.620 [0.636,0.788]	0.610 [0.542,0.721]	0.751 [0.607,0.816]	0.576 [0.453,0.667]
SK	0.671 [0.607,0.716]	0.630 [0.636,0.783]	0.644 [0.563,0.730]	0.582 [0.519,0.714]	0.637 [0.464,0.693]
ML	0.632 [0.536,0.713]	0.738 [0.586,0.821]	0.375 [0.267,0.592]	0.755 [0.482,0.820]	0.514 [0.459,0.674]

Table 2: F_1 score for all participating approaches. Numbers in square brackets indicate 95% confidence interval as determined by bootstrapping.

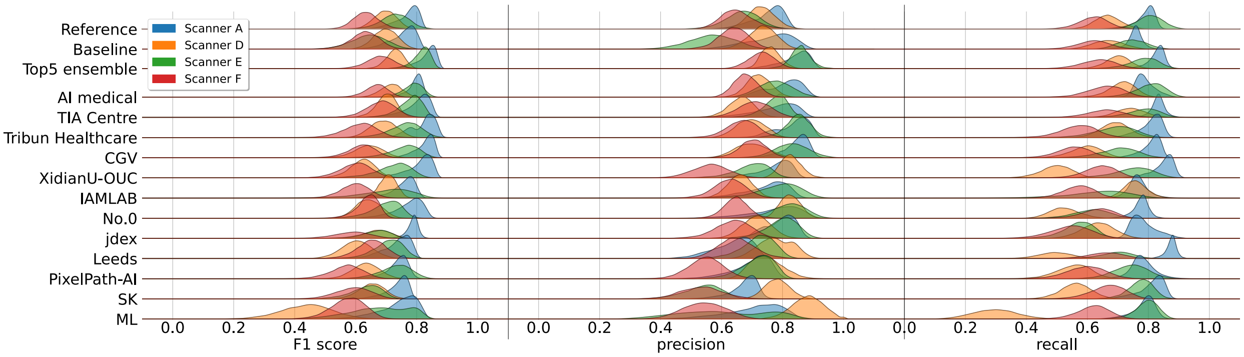

Figure 5: F_1, precision and recall scores of all participants across scanners (bootstrapping result, 20 random case draws with replacement, 10,000 repetitions).

as alternative to using it for data augmentation via image synthesis.

3.3. Instance label generation

The task of the challenge was to find the centroid coordinates of all mitotic figures, which was solved using object detection networks as well as semantic segmentation approaches by most participants. However, three teams (Jahanifar et al., 2022; Fick et al., 2022; Yang et al., 2022) chose to enhance the given set of labels (consisting only of approximate bounding boxes of the mitotic figures) by providing segmentation masks on pixel level for each mitotic figure instance to the training process.

The approach by Jahanifar et al. (2022) used an CNN-based interactive segmentation model (Koohbanani et al., 2020) that is targeted at generating segmentation masks from cell centroid coordinates. The tool is available as an open source tool on GitHub, was trained on publicly available data, and was only used to define segmentation masks for the labels given by the ground truth.

In a similar way, Yang et al. (2022) used another publicly available approach (Hover-Net, Graham et al. (2019)) that is aimed at nuclei instance segmentation on the dataset. From the output of this tool, they filtered out the segmentation masks of mitoses by thresholding the intersection over union between nuclei detected and ground truth mitotic figure bounding boxes. These segmentation masks were subsequently used as ground truth in a modified U-Net (SK-Unet, Wang et al. (2021)).

A third strategy was employed in the approach by Fick et al. (2022): They generated the segmentation masks for the instances by manually annotating approximately 100 mitoses across the dataset, and then fine-tuned a pre-trained Mask R-CNN on this small dataset to run inference on the remaining annotations and hence derive a segmentation mask for each mitotic figure. The generated instance masks were subsequently used to train another Mask-RCNN model with subsequent secondary classification stage.

3.4. Data augmentation

All participants employed some sort of data augmentation or domain adaptation technique during training and/or at test time in order to increase the robustness of their model against the unseen scanner within the final test set. These techniques where divided in roughly three type of groups:

- Standard data augmentation such as: Color, contrast, brightness, geometric
- Stain normalization techniques
- Generative adversarial networks

The standard approach was to apply color, geometric, contrast augmentation during training which should guarantee some level of robustness against unseen data. Groups such as Dexl et al. (2022); Breen et al. (2022); Lafarge and Koelzer (2022) relied solely on these methods. Some approaches used stain normalization techniques either as data augmentation during training or as a way to normalize all of the dataset according to some common stain and then apply data augmentation. Groups such as Jahanifar et al. (2022); Liang et al. (2022); Razavi et al. (2022) applied stain normalization to the entire dataset and then proceeded to apply standard data augmentation during training.

Another approach encountered was using generative adversarial networks (GANs) to generate images that simulated different scanners and different styles. Groups like Chung et al. (2022) and Fick et al. (2022) used a StarGAN and a Residual CycleGAN, respectively, as a data augmentation technique. The benefit of generative adversarial techniques is that they are configurable to simulate a multitude of potential scanner styles, however, the complexity and the hyperparameter search space are increased.

Even though all approaches shared some common characteristics, there was one implementation that stood out in its approach and was, at the end, quite successful in terms of performance: The group Yang et al. (2022) used, apart from standard color and geometric transformations of data augmentation, a Fourier-domain transform adaptation approach that separated high-frequent (i.e., structural) from low-frequent (i.e., color) components to transfer the stain information between images, acting as a stain normalization technique without relying on a specific stain matrix transformation.

3.5. Ensembling and test-time augmentation

Ensembling combines the outputs of multiple models, either with the same structure or even different model architectures, and is known to improve model robustness as well as overall performance. At the same time, using multiple large parallel models increases the compute time (and also the carbon footprint) at times significantly. While ensembling is common amongst the participants of competitions, we noted much less use of the technique in the MIDOG 2021 challenge.

Still, five teams employed (moderate) ensembling, mostly using a multi-fold or cross-validation setup on the provided training data and ensembling the resulting models, with different fusion strategies, e.g., simply setting a threshold of necessary detections (Lafarge and Koelzer, 2022), averaging of classifier predictions (Jahanifar et al., 2022), or weighted boxes fusion for the detection stage (Nateghi and Pourakpour, 2022). One team (Fick et al., 2022) assembled two models for the final classification stage (ResNet50 and DenseNet201) whereas another (Liang et al., 2022) put an ensemble of five different models in the center of their classification model. Test time augmentation was less frequently used, but for example by Jahanifar et al. (2022) for improving classifier performance and by Fick et al. (2022) for improving the results of the mask generation during model development.

4. Results

The majority of approaches were able to provide better results than the CNN baseline. The domain-adversarial reference method (Wilm et al., 2022) yielded a competitive F_1 score of 0.718 on the test set and was outperformed by only four approaches (see Table 2). With an F_1 score of 0.748, the overall best performance was reached by Yang *et al.*, utilizing segmentation and Fourier-domain mixing as augmentation (Yang et al., 2022). It is worth noting that this approach was amongst the only approaches which was able to achieve good performance on Scanner D (together with the approach by Razavi et al. (2022)), while other approaches performed better on other scanners. Even though the approach did not have the best performance on each individual scanner, it was the most consistently well-performing approach across all scanners, and thus the most generalizing solution. The runner-up approach (Jahanifar et al., 2022) had an almost identical F_1 score compared to the best approach, supported also by a good overall performance across all scanners,

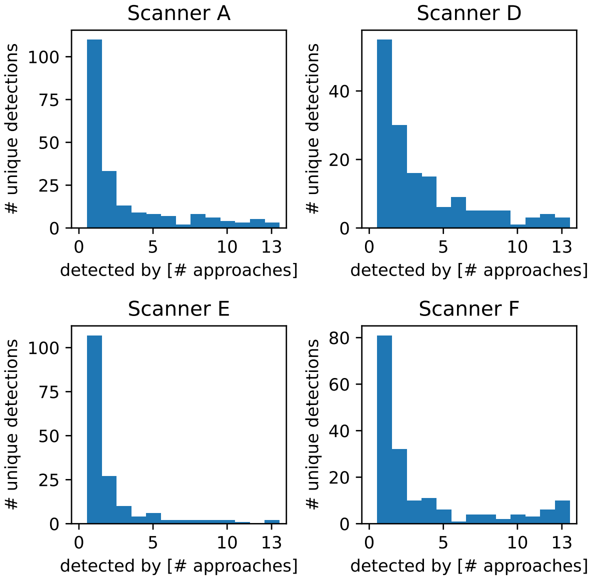

Figure 6: Histogram of false positives across scanners. Note that at least one false detection is necessary for an object to count as false positive.

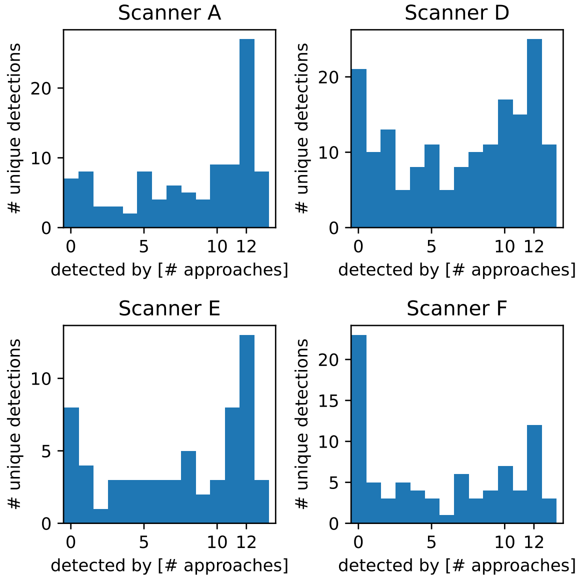

Figure 7: Histogram of detection of ground truth mitotic figures (true positives + false negatives) across scanners, showing how many mitoses have been detected by how many approaches.

with some minor weaknesses on scanner D, but it was the best performing algorithm on Scanner E. On Scanner F, the best performing approach was the multi-stage cascaded RCNN approach by Razavi et al. (2022), and on the scanner A, which was part of the training set, the best solution came from Fick et al. (2022).

4.1. *Post-Challenge Ensembling*

The top5 ensemble outperformed the leading approaches by Yang et al. (2022) and Jahanifar et al. (2022) in terms of overall F_1 score considerably. As Table 2 shows, this can be mainly attributed to a boost in performance for Scanner D, where the margin to the runner-up approach is the largest. For the other scanners, the top5 ensemble is on par with the respectively best approach for the scanner (especially when the 95% confidence intervals are considered). This ensembling method also yielded the overall highest precision (see Fig. 5) on all scanners.

4.2. *Object-level agreement*

To investigate the diversity in detections on object level, we assessed the agreement between the approaches on false positives (non-mitotic figure objects found by one or multiple approaches) and on the ground truth mitotic figures. The histogram of false positives is given in Fig. 6. It shows that the vast majority of unique false detections was only detected by a small number of approaches. There is little difference in this behavior across scanners. Looking at the histogram of false negatives in Fig. 7, we see a different behavior for missed detections: While most out of the total set of mitoses were only missed by a small number of methods for the seen scanner A (as visible in the high counts for mitoses that were found by >10 models), the number of mitotic figures that were missed by the majority or even the totality of detectors increased for the unlabeled/unseen scanners D and F, and also for scanner E (note that this scanner had a lower overall MC as of the ground truth labels). This is also underlined by the generally lower recall for those scanners compared to the seen scanner A (see Fig. 5 and Table 2). The total number of mitoses detected by all of the approaches was 25, i.e. the vast majority of mitoses was missed by at least one approach. In Fig. 8 and Fig. 9, we give examples for false positives and ground truth mitosis (true positives and false negatives), stratified by the number of models that detected those. It becomes obvious that some of those might be borderline mitotic figures and might even have a different label when re-evaluated by the same or a different set of experts (e.g., examples A1, D1, E3, or even F5 in Fig. 8). On the other hand, the low contrast of scanner D seemed to be a major obstacle for many approaches, even though sample images (without labels) of the scanner were available within the training set. Also atypical visual representations of clear mitoses, such as the examples D3 (atypical) or F3 (late telophase) were apparently hard to detect, as well as cells with unclear cell boundaries (A1, A2, D2).

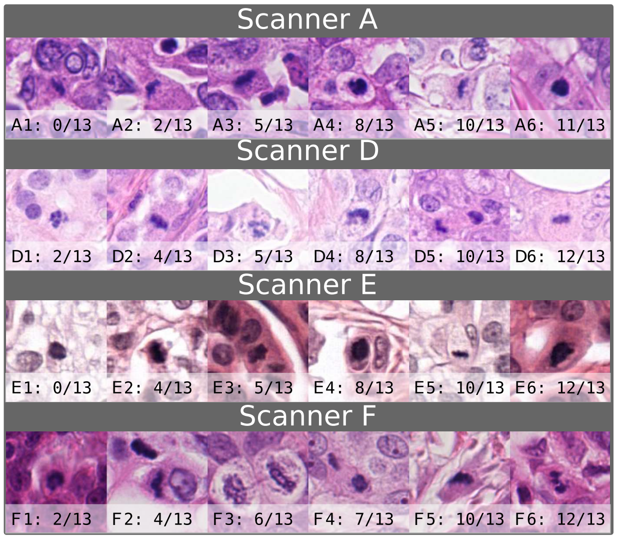

Figure 8: Examples of ground truth mitotic figures (true positives and false negatives), ordered by the count of models voting for it. The numbers (x/13) indicate, how many models voted for this cell to be a mitotic figure. The rows are stratified by the number of models to give examples for the complete distribution in Fig. 7.

Looking at some of the examples of false positives in Fig. 9, we see that especially the cells detected by most models (e.g., A3, A4, A6, D6, E6, F6) can be considered missed annotations of the dataset. Cellular objects incorporating bar-like structures, such as in the examples E5, F4) can be considered hard negatives for most of the approaches.

As depicted in Fig. 5, the top performing models excelled at finding a good compromise between precision and recall across all scanners. The unknown color distri-

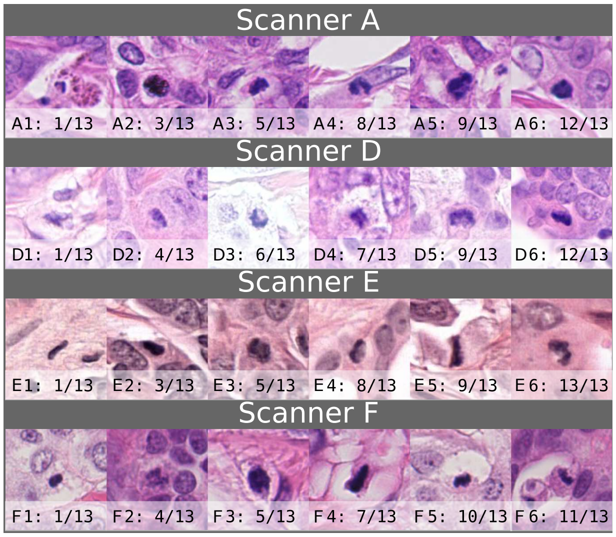

Figure 9: Examples of false positives, ordered by the count of models voting for it. The numbers (x/13) indicate, how many models voted for this cell to be a true mitotic figure.

bution of the unseen scanners can be particularly expected to have an influence on the calibration of the model output and thus result in underestimated or overestimated model scores, an effect which can be observed especially for the results on scanner D for the approaches by Liang et al. (2022), Razavi et al. (2022), Long et al. (2022), Kondo (2022), and Lafarge and Koelzer (2022).

It is worth noting that all three best performing approaches included an auxiliary task for mitosis segmentation, which apparently increased general performance. From the results, it can, however, not be determined if the auxiliary task also helped in domain generalization. Five out of the seven top performing approaches were utilizing methods of ensembling or test-time augmentation. Further, we see a slight trend that the better performing methods had larger classification networks.

5. Discussion

The results of the MIDOG 2021 challenge indicate that, using proper augmentation strategies and deep learning architectures, domain shift between whole slide imaging scanners can be compensated for to a high degree. The results by the best approaches in the field were in the range of well-performing human experts on the same task (see Bertram et al. (2021)). It must be noted, however, that the mitosis detection task was only performed on selected ROIs. In contrast, WSIs will have a much higher variability in tissue quality, including out-of-focus areas, and areas with necrosis, and thus have a variety of hard negative examples for the algorithms that have not been evaluated in this challenge.

All three expert pathologists were highly familiar with mitotic figure identification, however, a bias in annotation can not be excluded completely. Ultimately, a prediction of outcome, such as survival or recurrence, based on mitotic figure detection on WSIs, complemented with other morphological factors, would be the clinical target for an automated tumor grading. Yet, since this was not the scope of the MIDOG challenge, this evaluation is considered future work.

The best achieved F_1 score on Scanners A (fully supervised) and E (unknown) was in the order of state-of-the-art approaches trained fully supervised in-domain Aubreville et al. (2020a); Bertram et al. (2019). In contrast, the best results on Scanner D (image only, no labels provided) and Scanner F (unknown) were considerably weaker. While the good performance on Scanner A underlines the high consistency of labels due to the computer-aided approach and high level of expertise of our pathologists, the weak performance on scanners D and F might be related to the domain shift not being covered completely by the algorithms. On the other hand, it might be influenced by the image quality of the scanners (which might make mitosis detection in general more challenging for humans and algorithms alike) or even non-familiarity of the experts with the color and structural patterns representing mitoses within tissue imaged by those scanners. Thus, while this challenge evaluated mitosis detection on the largest set of scanners with controlled staining conditions, the evaluation on this subset of scanners might still have its limitations and not generalize to other, yet unseen scanners.

Even though most of the higher ranked approaches used ensembling or test-time augmentation, it is unclear if those methods were a success factor in our challenge setup or if there is a mere interrelation between participants utilizing ensembling / TTA and a higher algorithmic development effort, reflected in a higher score. However, the superior performance of the top5 ensemble (as an ensemble of approaches) points into the direction that, in general, ensembling techniques are a success factor for these kind of tasks.

The MIDOG challenge was the first in the field of generalization of mitosis detection to unseen domains, and thus an important step towards clinical applications. And yet, we can observe, that there are many challenges ahead on route to a clinical application: While, as mentioned, application on WSIs is a very different challenge, required for clinical application, so is the generalization to further tissue and cancer types, where mitosis detection plays an equally important role in the respective grading system. In fact, we can expect a substantial domain shift between tumor types, which is why this task will be the focus of the successor event of this challenge.

References

Aubreville, M., Bertram, C., Veta, M., Klopfleisch, R., Stathonikos, N., Breininger, K., ter Hoeve, N., Ciompi, F., Maier, A., 2021a. Mitosis domain generalization challenge, in: 24th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2021). doi:`10.5281/zenodo.4573978`.

Aubreville, M., Bertram, C., Veta, M., Klopfleisch, R., Stathonikos, N., Breininger, K., ter Hoeve, N., Ciompi, F., Maier, A., 2021b. Quantifying the scanner-induced domain gap in mitosis detection, in: Medical Imaging and Deep Learning (MIDL).

Aubreville, M., Bertram, C.A., Donovan, T.A., Marzahl, C., Maier, A., Klopfleisch, R., 2020a. A completely annotated whole slide image dataset of canine breast cancer to aid human breast cancer research. Scientific data 7:417, 1–10.

Aubreville, M., Bertram, C.A., Marzahl, C., Gurtner, C., Dettwiler, M., Schmidt, A., Bartenschlager, F., Merz, S., Fragoso, M., Kershaw, O., et al., 2020b. Deep learning algorithms out-perform veterinary pathologists in detecting the mitotically most active tumor region. Scientific Reports 10:16447, 1–11.

Aubreville, M., Zimmerer, D., Heinrich, M. (Eds.), 2022. Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis. Springer, Cham. doi:`10.1007/978-3-030-97281-3`.

Baak, J.P.A., Gudlaugsson, E., Skaland, I., Guo, L.H.R., Klos, J., Lende, T.H., Søiland, H., Janssen, E.A.M., zur Hausen, A., 2008. Proliferation is the strongest prognosticator in node-negative breast cancer: significance, error sources, alternatives and comparison with molecular prognostic markers. Breast Cancer Res Tr 115, 241–254.

Balkenhol, M.C., Tellez, D., Vreuls, W., Clahsen, P.C., Pinckaers, H., Ciompi, F., Bult, P., van der Laak, J.A., 2019. Deep learning assisted mitotic counting for breast cancer. Laboratory investigation 99, 1596–1606.

de Bel, T., Bokhorst, J.M., van der Laak, J., Litjens, G., 2021. Residual cyclegan for robust domain transformation of histopathological tissue slides. Medical Image Analysis 70, 102004.

Bertram, C.A., Aubreville, M., Donovan, T.A., Bartel, A., Wilm, F., Marzahl, C., Assenmacher, C.A., Becker, K., Bennett, M., Corner, S., Cossic, B., Denk, D., Dettwiler, M., Gonzalez, B.G., Gurtner, C., Haverkamp, A.K., Heier, A., Lehmbecker, A., Merz, S., Noland, E.L., Plog, S., Schmidt, A., Sebastian, F., Sledge, D.G., Smedley, R.C., Tecilla, M., Thaiwong, T., Fuchs-Baumgartinger, A., Meuten, D.J., Breininger, K., Kiupel, M., Maier, A., Klopfleisch, R., 2021. Computer-assisted mitotic count using a deep learning–based algorithm improves interobserver reproducibility and accuracy. Veterinary Pathology doi:`10.1177/03009858211067478`.

Bertram, C.A., Aubreville, M., Gurtner, C., Bartel, A., Corner, S.M., Dettwiler, M., Kershaw, O., Noland, E.L., Schmidt, A., Sledge, D.G., et al., 2020a. Computerized calculation of mitotic count distribution in canine cutaneous mast cell tumor sections: mitotic count is area dependent. Veterinary pathology 57, 214–226.

Bertram, C.A., Aubreville, M., Marzahl, C., Maier, A., Klopfleisch, R., 2019. A large-scale dataset for mitotic figure assessment on whole slide images of canine cutaneous mast cell tumor. Scientific data 6, 1–9.

Bertram, C.A., Veta, M., Marzahl, C., Stathonikos, N., Maier, A., Klopfleisch, R., Aubreville, M., 2020b. Are pathologist-defined labels reproducible? comparison of the tupac16 mitotic figure dataset with an alternative set of labels, in: Interpretable and Annotation-Efficient Learning for Medical Image Computing. Springer, pp. 204–213.

Breen, J., Zucker, K., Orsi, N.M., Ravikumar, N., 2022. Assessing domain adaptation techniques for mitosis detection in multi-scanner breast cancer histopathology images, in: Biomedical Image Registration, Domain Generalization and Out-of-Distribution Analysis, MICCAI 2021 Challenges L2R, MIDOG and MOOD, Springer, Cham. pp. 14–22.

Cai, Z., Vasconcelos, N., 2018. Cascade r-cnn: Delving into high quality object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6154–6162.

Chang, H., Loss, L.A., Parvin, B., 2012. Nuclear segmentation in h&e sections via multi-reference graph cut (mrgc), in: International symposium biomedical imaging.

Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J., 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8789–8797.

Chung, Y., Cho, J., Park, J., 2022. Domain-robust mitotic figure detection with style transfer, in: Biomedical Image Registration, Domain Generalization and Out-of-Distribution Analysis, MICCAI 2021 Challenges L2R, MIDOG and MOOD, Springer, Cham. pp. 23–31.

Dexl, J., Benz, M., Bruns, V., Kuritcy, P., Wittenberg, T., 2022. MitoDet: Simple and robust mitosis detection, in: Biomedical Image Registration, Domain Generalization and Out-of-Distribution Analysis, MICCAI 2021 Challenges L2R, MIDOG and MOOD, Springer, Cham. pp. 53–57.

Donovan, T.A., Moore, F.M., Bertram, C.A., Luong, R., Bolfa, P., Klopfleisch, R., Tvedten, H., Salas, E.N., Whitley, D.B., Aubreville, M., et al., 2021. Mitotic figures—normal, atypical, and imposters: A guide to identification. Veterinary pathology 58, 243–257.

Elston, C.W., Ellis, I.O., 1991. pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology 19, 403–410.

Fick, R.H., Moshayedi, A., Roy, G., Dedieu, J., Petit, S., Hadj, S.B., 2022. Domain-specific cycle-gan augmentation improves domain generalizability for mitosis detection, in: Biomedical Image Registration, Domain Generalization and Out-of-Distribution Analysis, MICCAI 2021 Challenges L2R, MIDOG and MOOD, Springer, Cham. pp. 40–47.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. The journal of machine learning research 17, 2096–2030.

Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T., Rajpoot, N., 2019. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical Image Analysis 58, 101563.

Hall, P., 1994. Methodology and theory for the bootstrap. Handbook of econometrics 4, 2341–2381.

Hallgren, K.A., 2012. Computing inter-rater reliability for observational data: an overview and tutorial. Tutorials in quantitative methods for psychology 8, 23.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, pp. 2961–2969.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134.

Jahanifar, M., Shepard, A., Zamanitajeddin, N., Bashir, R.S., Bilal, M., Khurram, S.A., Minhas, F., Rajpoot, N., 2022. Stain-robust mitotic figure detection for the Mitosis Domain Generalization Challenge, in: Biomedical Image Registration, Domain Generalization and Out-of-Distribution Analysis, MICCAI 2021 Challenges L2R, MIDOG and MOOD, Springer, Cham. pp. 48–52.

Jahanifar, M., Tajeddin, N.Z., Koohbanani, N.A., Rajpoot, N.M., 2021. Robust interactive semantic segmentation of pathology images with minimal user input, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 674–683.

Karimi, D., Nir, G., Fazli, L., Black, P.C., Goldenberg, L., Salcudean, S.E., 2019. Deep learning-based gleason grading of prostate cancer from histopathology images—role of multiscale decision aggregation and data augmentation. IEEE journal of biomedical and health informatics 24, 1413–1426.

Kondo, S., 2022. Multi-source domain adaptation using gradient reversal layer for mitotic cell detection, in: Biomedical Image Registration, Domain Generalization and Out-of-Distribution Analysis, MICCAI 2021 Challenges L2R, MIDOG and MOOD, Springer, Cham. pp. 58–61.

Koohbanani, N.A., Jahanifar, M., Tajadin, N.Z., Rajpoot, N., 2020. Nuclick: a deep learning framework for interactive segmentation of microscopic images. Medical Image Analysis 65, 101771.

Lafarge, M., Koelzer, V., 2022. Rotation invariance and extensive data augmentation: a strategy for the MItosis DOmain Generalization (MIDOG) Challenge, in: Biomedical Image Registration, Domain Generalization and Out-of-Distribution Analysis, MICCAI 2021 Challenges L2R, MIDOG and MOOD, Springer, Cham. pp. 62–67.

Lafarge, M.W., Bekkers, E.J., Pluim, J.P., Duits, R., Veta, M., 2021. Roto-translation equivariant convolutional networks: Application to histopathology image analysis. Medical Image Analysis 68, 101849.

Lafarge, M.W., Pluim, J.P., Eppenhof, K.A., Veta, M., 2019. Learning domain-invariant representations of histological images. Frontiers in medicine 6, 162.

Levine, A.B., Schlosser, C., Grewal, J., Coope, R., Jones, S.J., Yip, S., 2019. Rise of the machines: advances in deep learning for cancer diagnosis. Trends in cancer 5, 157–169.

Liang, J., Wang, C., Cheng, Y., Wang, Z., Wang, F., Huang, L., Yu, Z., Wang, Y., 2022. Detecting mitosis against domain shift using a fused detector and deep ensemble classification model for MIDOG challenge, in: Biomedical Image Registration, Domain Generalization and Out-of-Distribution Analysis, MICCAI 2021 Challenges L2R, MIDOG and MOOD, Springer, Cham. pp. 68–72.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, pp. 2980–2988.

Long, X., Cheng, Y., Mu, X., Liu, L., Liu, J., 2022. Domain adaptive cascade R-CNN for MItosis DOmain Generalization (MIDOG) Challenge, in: Biomedical Image Registration, Domain Generalization and Out-of-Distribution Analysis, MICCAI 2021 Challenges L2R, MIDOG and MOOD, Springer, Cham. pp. 73–76.

Louis, D.N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W.K., Ohgaki, H., Wiestler, O.D., Kleihues, P., Ellison, D.W., 2016. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathologica 131, 803–820.

Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E., 2009. A method for normalizing histology slides for quantitative analysis, in: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, IEEE. pp. 1107–1110.

Marzahl, C., Aubreville, M., Bertram, C.A., Maier, J., Bergler, C., Kröger, C., Voigt, J., Breininger, K., Klopfleisch, R., Maier, A., 2021a. Exact: a collaboration toolset for algorithm-aided annotation of images with annotation version control. Scientific reports 11, 1–11.

Marzahl, C., Aubreville, M., Bertram, C.A., Stayt, J., Jasensky, A.K., Bartenschlager, F., Fragoso-Garcia, M., Barton, A.K., Elsemann, S., Jabari, S., et al., 2020. Deep learning-based quantification of pulmonary hemosiderophages in cytology slides. Scientific Reports 10, 1–10.

Marzahl, C., Wilm, F., Tharun, L., Perner, S., Bertram, C.A., Kröger, C., Voigt, J., Klopfleisch, R., Maier, A., Aubreville, M., et al., 2021b. Robust quad-tree based registration on whole slide images, in: MICCAI Workshop on Computational Pathology, PMLR. pp. 181–190.

Meyer, J.S., Alvarez, C., Milikowski, C., Olson, N., Russo, I., Russo, J., Glass, A., Zehnbauer, B.A., Lister, K., Parwaresch, R., 2005. Breast carcinoma malignancy grading by Bloom-Richardson system vs proliferation index: Reproducibility of grade and advantages of proliferation index. Modern Pathol 18, 1067–1078.

Meyer, J.S., Cosatto, E., Graf, H.P., 2009. Mitotic index of invasive breast carcinoma. Achieving clinically meaningful precision and evaluating tertial cutoffs. Arch Pathol Lab Med 133, 1826–1833.

Moreira, A.L., Ocampo, P.S., Xia, Y., Zhong, H., Russell, P.A., Minami, Y., Cooper, W.A., Yoshida, A., Bubendorf, L., Papotti, M., et al., 2020. A grading system for invasive pulmonary adenocarcinoma: a proposal from the international association for the study of lung cancer pathology committee. Journal of Thoracic Oncology 15, 1599–1610.

Nateghi, R., Pourakpour, F., 2022. Two-step domain adaptation for mitosis cell detection in histopathology images, in: Biomedical Image Registration, Domain Generalization and Out-of-Distribution Analysis, MICCAI 2021 Challenges L2R, MIDOG and MOOD, Springer, Cham. pp. 32–39.

Pasqualino, G., Furnari, A., Signorello, G., Farinella, G.M., 2021. An unsupervised domain adaptation scheme for single-stage artwork recognition in cultural sites. Image and Vision Computing 107, 104098.

Qiao, S., Chen, L.C., Yuille, A., 2021. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10213–10224.

Razavi, S., Dambandkhameneh, F., Androutsos, D., Done, S., Khademi, A., 2022. Cascade R-CNN for MIDOG challenge, in: Biomedical Image Registration, Domain Generalization and Out-of-Distribution Analysis, MICCAI 2021 Challenges L2R, MIDOG and MOOD, Springer, Cham. pp. 81–85.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28, 91–99.

Roux, L., Racoceanu, D., Capron, F., Calvo, J., Attieh, E., Le Naour, G., Gloaguen, A., 2014. Mitos & atypia. Image Pervasive Access Lab (IPAL), Agency Sci., Technol. & Res. Inst. Infocom Res., Singapore, Tech. Rep 1, 1–8.

Stacke, K., Eilertsen, G., Unger, J., Lundström, C., 2020. Measuring domain shift for deep learning in histopathology. IEEE journal of biomedical and health informatics 25, 325–336.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks, in: International Conference on Machine Learning, PMLR. pp. 6105–6114.

Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I., Navab, N., 2016. Structure-preserving color normalization and sparse stain separation for histological images. IEEE transactions on medical imaging 35, 1962–1971.

Van Diest, P.J., van der Wall, E., Baak, J.P., 2004. Prognostic value of proliferation in invasive breast cancer: a review. Journal of clinical pathology 57, 675–681.

Veta, M., Heng, Y.J., Stathonikos, N., Bejnordi, B.E., Beca, F., Wollmann, T., Rohr, K., Shah, M.A., Wang, D., Rousson, M., et al., 2019. Predicting breast tumor proliferation from whole-slide images: the tupac16 challenge. Medical image analysis 54, 111–121.

Veta, M., Van Diest, P.J., Jiwa, M., Al-Janabi, S., Pluim, J.P., 2016. Mitosis counting in breast cancer: Object-level interobserver agreement and comparison to an automatic method. PloS one 11, e0161286.

Veta, M., Van Diest, P.J., Willems, S.M., Wang, H., Madabhushi, A., Cruz-Roa, A., Gonzalez, F., Larsen, A.B., Vestergaard, J.S., Dahl, A.B., et al., 2015. Assessment of algorithms for mitosis detection in breast cancer histopathology images. Medical image analysis 20, 237–248.

Wang, X., Yang, S., Fang, Y., Wei, Y., Wang, M., Zhang, J., Han, X., 2021. Sk-unet: An improved u-net model with selective kernel for the segmentation of lge cardiac mr images. IEEE Sensors Journal 21, 11643–11653.

Wild, C., Weiderpass, E., Stewart, B.W., 2020. World cancer report: cancer research for cancer prevention. IARC Press.

Wilm, F., Marzahl, C., Breininger, K., Aubreville, M., 2022. Domain adversarial retinanet as a reference algorithm for the midog challenge, in: Biomedical Image Registration, Domain Generalization and Out-of-Distribution Analysis: The MICCAI Challenges L2R, MIDOG and MOOD, Springer, Cham. pp. 5–13.

Yang, S., Luo, F., Zhang, J., Wang, X., 2022. Sk-unet model with fourier domain for mitosis detection, in: Biomedical Image Registration, Domain Generalization and Out-of-Distribution Analysis, MICCAI 2021 Challenges L2R, MIDOG and MOOD, Springer, Cham. pp. 86–90.

Yang, Y., Soatto, S., 2020. Fda: Fourier domain adaptation for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4085–4095.

Acknowledgement

Computational resources for the challenge have been donated by NVIDIA and AWS. The authors would like to thank the team from grand-challenge.org for their expertise and support. The challenge organizers would like to thank Siemens Healthineers for donating the monetary prizes of the challenge, who have been awarded to the top three participants.

Conflicts of interest

The authors declare no conflicts of interest.

Data usage statement

All data of the training set was released under the Creative Commons 4.0 BY (attribution) NC-ND (non-commercial, non-derivative) license.

Appendix A. Author contributions

The challenge was organized by Katharina Breininger, Natalie ter Hoeve, Christof A. Bertram, Francesco Ciompi, Robert Klopfleisch, Andreas Maier, Nikolas Stathonikos, Mitko Veta, and Marc Aubreville.

Frauke Wilm, Christian Marzahl, Katharina Breining and Marc Aubreville provided the algorithmic reference approach for the challenge.

The core writing group of this paper consisted of Marc Aubreville, Nikolas Stathonikos, Christof A. Bertram, Katharina Breininger, and Mitko Veta.

Taryn A. Donovan, Robert Klopfleisch and Christof A. Bertram served as expert pathologists in annotating the complete challenge data set.

Jack Breen and Nishant Ravikumar (Team Leeds), Youjin Chung and Jinah Park (Team CGV), Ramin Nateghi and Fattaneh Pourakpour (Team PixelPath-AI), Rutger H.J. Fick and Saima Ben Hadj (Team Tribun Healthcare), Mostafa Jahanifar and Nasir Rajpoot (Team PixelPath-AI), Jakob Dexl and Thomas Wittenberg (Team jdex), Satoshi Kondo (Team SK), Maxime W. Lafarge and Viktor H. Koelzer (Team ML), Jingtang Liang and Yubo Wang (Team XidianUOUC), Xi Long and Jingxin Liu (Team No. 0), Salar Razavi and April Khademi (Team IAMLAB), and Sen Yang and Xiyue Wang were the respective first and last authors of the participant's challenge papers, and thus contributed the algorithmic approaches to the challenge.

All authors reviewed the manuscript.