Domain generalization across tumor types, laboratories, and species – Insights from the 2022 edition of the Mitosis Domain Generalization Challenge

Marc Aubreville[a], Nikolas Stathonikos[b], Taryn A. Donovan[c], Robert Klopfleisch[d], Jonathan Ganz[a], Jonas Ammeling[a], Frauke Wilm[e,f], Mitko Veta[g], Samir Jabari[h], Markus Eckstein[i], Jonas Annuscheit[j], Christian Krumnow[j], Engin Bozaba[k], Sercan Çayır[k], Hongyan Gu[l], Xiang 'Anthony' Chen[l], Mostafa Jahanifar[m], Adam Shephard[m], Satoshi Kondo[n], Satoshi Kasai[o], Sujatha Kotte[p], VG Saipradeep[p], Maxime W. Lafarge[q], Viktor H. Koelzer[q], Ziyue Wang[r], Yongbing Zhang[r], Sen Yang[s], Xiyue Wang[t], Katharina Breininger[f], Christof A. Bertram[u]

[a]*Technische Hochschule Ingolstadt, Ingolstadt, Germany*
[b]*Pathology Department, UMC Utrecht, The Netherlands*
[c]*Department of Anatomic Pathology, The Schwarzman Animal Medical Center, New York, USA*
[d]*Institute of Veterinary Pathology, Freie Universität Berlin, Berlin, Germany*
[e]*Pattern Recognition Lab, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany*
[f]*Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany*
[g]*Computational Pathology Group, Radboud UMC Nijmegen, The Netherlands*
[h]*Institute of Neuropathology, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany*
[i]*Institute of Pathology, University Hospital Erlangen, Friedrich-Alexander-Universität Erlangen-Nünberg, Erlangen, Germany*
[j]*University of Applied Sciences (HTW) Berlin, Berlin, Germany*
[k]*Artificial Intelligence Research Team, Virasoft Corporation, New York, USA*
[l]*University of California, Los Angeles, USA*
[m]*University of Warwick, United Kingdom*
[n]*Muroran Institute of Technology, Muroran, Japan*
[o]*Niigata University of Health and Welfare, Niigata, Japan*
[p]*TCS Research, Tata Consultancy Services Ltd, Hyderabad, India*
[q]*Department of Pathology and Molecular Pathology, University Hospital Zurich, University of Zurich, Zurich, Switzerland*
[r]*Harbin Institute of Technology, Shenzhen, China*
[s]*College of Biomedical Engineering, Sichuan University, Chengdu, China*
[t]*Department of Radiation Oncology, Stanford University School of Medicine, Palo Alto, USA*
[u]*Institute of Pathology, University of Veterinary Medicine, Vienna, Austria*

Abstract

Recognition of mitotic figures in histologic tumor specimens is highly relevant to patient outcome assessment. This task is challenging for algorithms and human experts alike, with deterioration of algorithmic performance under shifts in image representations. Considerable covariate shifts occur when assessment is performed on different tumor types, images are acquired using different digitization devices, or specimens are produced in different laboratories. This observation motivated the inception of the 2022 challenge on MItosis Domain Generalization (MIDOG 2022). The challenge provided annotated histologic tumor images from six different domains and evaluated the algorithmic approaches for mitotic figure detection provided by nine challenge participants on ten independent domains. Ground truth for mitotic figure detection was established in two ways: a three-expert consensus and an independent, immunohistochemistry-assisted set of labels. This work represents an overview of the challenge tasks, the algorithmic strategies employed by the participants, and potential factors contributing to their success. With an F_1 score of 0.764 for the top-performing team, we summarize that domain generalization across various tumor domains is possible with today's deep learning-based recognition pipelines. When assessed against the immunohistochemistry-assisted reference standard, all methods resulted in reduced recall scores, but with only minor changes in the order of participants in the ranking.

1. Introduction

Despite advances in molecular characterization of biological tumor behavior, morphological tumor classification using established histopathologic techniques remains an important factor in tumor prognostication (Makki, 2015; Soliman and Yussif, 2016). One criterion of particular interest within many tumor grading schemes is the density of cells undergoing division, which are visible as mitotic figures (MFs) in hematoxylin and eosin (H&E)-stained histopathological sections (Veta et al., 2015, 2019). The number of MFs within a specific tumor area is enumerated by experienced pathologists, resulting in the mitotic count (MC). Despite the prognostic relevance of the MC, low inter-rater consistency on an object level has been reported in many studies (Veta et al., 2016; Meyer et al., 2005, 2009; Malon et al., 2012; Bertram et al., 2021). The recommendation for pathologists is to select the region of the suspected highest mitotic activity, which is considered to be the best predictor of tumor behavior (Azzola et al., 2003; Meuten et al., 2008; Veta et al., 2015). Selection of this regions of interest (ROI) within the tumor has a great impact on the MC (Bertram et al., 2020), but is difficult for pathologists to reliably accomplish and is poorly reproducible (Aubreville et al., 2020; Bertram et al., 2021). While assessment of mitotic activity in the entire tumor section (or in the case of the digital image: the whole slide image (WSI)) would be preferable in order to identify those mitotic hotspot ROI, this is not feasible in current practice. Additionally, low inter-rater consistency on an object level within these selected ROI has been reported in many studies with the tendency of pathologists to overlook MFs (Veta et al., 2016; Meyer et al., 2005, 2009; Malon et al., 2012; Bertram et al., 2021).The combination of these circumstances and the recent availability of large-scale digital pathology solutions makes automatic detection of MFs desirable.

Unsurprisingly, MF detection was one of the earliest identified areas of research interest in computational pathology, with the first approaches in 2008 (Malon et al., 2008). The first challenge on MF detection in breast cancer (MITOS2012, (Roux et al., 2013)) was held at the International Conference on Pattern Recognition (ICPR) and resulted in the first publicly available MF dataset. While this gave rise to algorithm development in the field, it was also an example of questionable dataset quality, as the training and test sets were selected from the same histology slides (Roux et al., 2013). More recent challenges (MITOS2014 (Roux et al., 2014), AMIDA13 (Veta et al., 2015), TUPAC16 (Veta et al., 2019)) also comprised breast cancer and incorporated a higher number of cases, yet, were still limited by having the same digitization device for the training and test set.

As shown by prior research (Aubreville et al., 2021), the digitization device has a decisive influence on detection quality, as it coincides with a shift in image representation, leading to a domain shift in latent representation of the detection models (Stacke et al., 2020; Aubreville et al., 2023a). Investigation of these limitations was the main idea behind the MItosis DOmain Generalization (MIDOG) challenge, held as a one-time event at the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI) in 2021. This challenge, which was the first to directly target domain generalization in histopathology, evaluated the detection of MFs in ROIs of human breast cancer, digitized using various devices (WSI scanners).

Since MFs are not only of interest for human breast cancer, the 2022 MIDOG challenge extended the task of MF domain generalization to include further representation shifts of interest: In addition to the use of different WSI scanners, the training dataset was enhanced by including histological specimens from different tumor types as well as different species (human, canine, feline), processed by different laboratories. Each of these contributing factors defined a *tumor domain*. We define a tumor domain as a specific combination of tumor type, species, lab, and WSI scanner. We found that the domain gap between tumor types is substantial (Aubreville et al., 2023b) and seems to be more important than the domain gap between scanners, thus the cases used for the MIDOG 2022 challenge were primarily categorized by the tumor domain.

Challenge format and task

As in previous challenges on MF detection, we provided ROIs, selected by an experienced pathologist from a tumor region with the presumed highest mitotic activity and appropriate tissue and scan quality. MF candidates were identified and assessed by a blinded consensus vote of three experts. The training set, consisting of 405 tumor cases (corresponding to 405 patients) and featuring 9,501 MF annotations was released on April 20, 2022. These

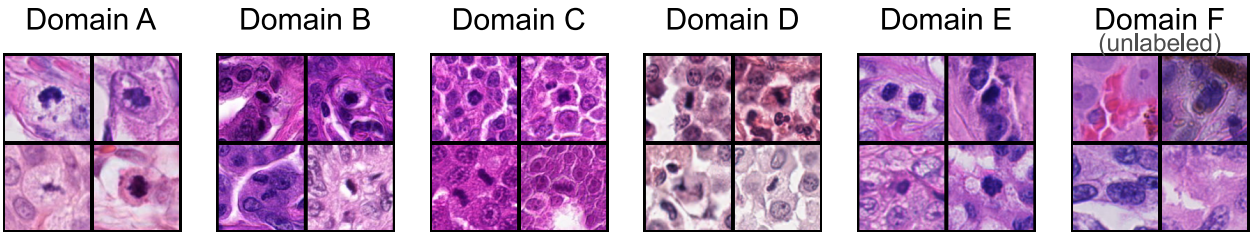

Figure 1: Random selection of crops of size $128 \times 128 px$, centered around annotated MFs from the six domains of the training set (A: human breast carcinoma, B: canine lung carcinoma, C: canine lymphosarcoma, D: canine cutaneous mast cell tumor, E: human pancreatic and gastrointestinal neuroendocrine tumor, F: human melanoma). Domain F was not labeled, hence the crops were selected at random.

cases were split across six tumor domains (see Fig. 1), out of which five were provided with labels and one was provided without labels as an additional data source for unsupervised domain adaptation techniques. An extended version of the training set, including two novel domains, was made available under a Creative Commons CC0 license post-challenge (Aubreville et al., 2023b).

The participants were required to package their algorithmic solution in the form of a docker container[1], which was subsequently evaluated on the test data on the grand-challenge.org platform[2] in a fully automatic manner, i.e., no participant had access to any of the test images during or after the challenge. To perform a technical validation of the docker containers, we provided an independent preliminary test set, consisting of four unseen tumor domains. During a preliminary test phase, which started on August 5, participants were allowed to perform one evaluation of an algorithmic approach per day. We explicitly made the participants aware that the four domains of the preliminary test set were disjointed from the tumor domains of the actual challenge test set, so overfitting to those domains by means of hyperparameter or model selection would not be meaningful. The final challenge submission phase started on August 26 and lasted until August 30. During this phase, participating teams were exclusively authorized to submit a single algorithmic approach.

The challenge provided two tracks: As multiple openly accessible datasets on MF detection already exist, we gave participants the choice to either use only data provided by the challenge (track 1) or additionally use publicly available data and labels (track 2). In the second track, participants also had the option to use in-house datasets under the condition that these datasets were made publicly available and announced on the challenge website up to one month prior to the challenge. We opted for this strategy to maximize the reproducibility of the challenge results. However, no participating team chose to use previously non-public datasets.

The structured challenge design includes details about the policies regarding participation, publication, awards, and results announcement, and was made available publicly (Aubreville et al., 2022). The challenge design was proposed and evaluated in a single-blinded peer review for admission to MICCAI 2022.

Main novelties over the predecessor

While the task (MFs detection on ROIs images) was identical to the preceding MIDOG 2021 challenge, we incorporated three major modifications in the 2022 challenge design that set it apart from its predecessor:

- We extended the sources of domain shift by not only including the imaging device and the inherent stain differences between cases but also by incorporating different laboratories (and hence tissue processing), different tumor types, and different species, minimizing the gap to real-world data variability.
- The evaluation was carried out on ten independent tumor domains, representing a wide variety of conditions and thus allowing for better generalization of

[1] A reference docker container for evaluation was made available to the participants at: `https://github.com/DeepMicroscopy/MIDOG_evaluation_docker`

[2] `https://midog2022.grand-challenge.org`

the assessment. The ten domains were additionally disjoint from the four independent domains of the preliminary test used for technical validation of the docker pipeline.

- We established the ground truth of the test set not only as the consensus of three experts on the H&E-stained sections (used for challenge evaluation and ranking) but also by additionally using an immunohistochemical (IHC) stain for Phospho-Histone H3 (PHH3) (specific for cells entering the mitotic cycle (Hendzel et al., 1997)), which was superimposed on the H&E image for assisted labeling aiming to object-level confusion, which is a main source of inter-rater disagreement (Veta et al., 2016).

2. Material and evaluation methods

For all tumor types included in our datasets, the MC has well-established prognostic relevance for discriminating patient outcome, either as a solitary prognostic test or as part of an established grading scheme. We retrieved human tissue samples from the diagnostic archives (DAs) of the Department of Pathology of the University Medical Center (UMC) Utrecht, The Netherlands, as well as the Institute of Neuropathology and the Institute of Pathology of the University Hospital Erlangen, Germany. All samples were prepared from paraffin-embedded tumour sections stained according to the standard procedures of the respective institutions. We received ethics approval from the UMC Utrecht (TCBio 20-776) and the ethics board of the medical faculty of FAU Erlangen-Nürnberg (AZ 92_14B, AZ 193_18B, 22_342_B). For samples taken from the DAs of veterinary pathology laboratories (Freie Universität Berlin (FUB), Germany and University of Veterinary Medicine Vienna (VMU), Austria), no ethics approval was required.

2.1. Challenge cohort and tumor domains

We included a total of 405 cases in our dataset (see Fig. 1):

- Domain A: Human breast carcinoma, retrieved from the DA of UMC Utrecht. 150 cases split across three scanners (Hamamatsu XR, Hamamatsu S360, Aperio Scanscope CS2, 50 each) at 40× magnification (0.23 to 0.25 μm/px), previously released as training set of the 2021 MIDOG challenge (Aubreville et al., 2023a). The MC is part of the College of American Pathologists guidelines for breast cancer (Fitzgibbons and Connolly, 2023).

- Domain B: Canine lung carcinoma, retrieved from the DA of VMU. 44 cases digitized with a 3DHistech Pannoramic Scan II at 40× magnification (0.25 μm/px). The MC is part of the grading scheme by McNiel et al. (1997).

- Domain C: Canine lymphosarcoma, retrieved from the DA of VMU. 55 cases digitized with a 3DHistech Pannoramic Scan II at 40× magnification (0.25 μm/px). MC is part of the grading scheme by Valli et al. (2013).

- Domain D: Canine cutaneous mast cell tumor, retrieved from the DA of FUB. 50 cases digitized with an Aperio ScanScope CS2 at 40× magnification (0.25 μm/px). MC is part of the grading scheme by Kiupel et al. (2011).

- Domain E: Human pancreatic and gastrointestinal neuroendocrine tumor, retrieved from the DA of UMC Utrecht. 55 cases digitized with a Hamamatsu XR (C12000-22) at 40× magnification (0.23 μm/px). MC is part of the 2022 WHO classification scheme of endocrine and neuroendocrine tumors (WHO Classification of Tumours Editorial Board, 2022).

- Domain F: Human melanoma, retrieved from the DA of UMC Utrecht. 51 cases digitized with a Hamamatsu XR (C12000-22) at 40× magnification (0.23 μm/px). MC is part of the staging and classification scheme of the AJCC for melanoma (Gershenwald et al., 2017). This domain was not labeled and only provided as an additional source of data diversity for unsupervised approaches.

While, ideally, a consecutive selection of cases would be desirable to provide representative samples, we intentionally deviated from this norm in this iteration of the challenge. Specifically, we ensured the inclusion of a

minimum number of mitotically active cases across all domains. This was done in order to ensure sufficient dataset support for MF objects in each respective domain.

We prepared a small preliminary test set to check the validity of the algorithmic approaches through the docker submission system. In this dataset, the following domains were included:

- Domain α: Human breast carcinoma, similar to the training set domain A, but scanned with a Hamamatsu RS2 scanner. Five cases, previously used as part of the preliminary test set of MIDOG 2021 (Aubreville et al., 2023a).

- Domain β: Canine osteosarcoma, retrieved from the DA of VMU. Five cases digitized with a 3DHistech Pannoramic Scan II at 40× magnification (0.25 μm/px).

- Domain γ: Human lymphoma, retrieved from the DA of UMC Utrecht. Five cases digitized with a Hamamatsu XR (C12000-22) at 40× magnification (0.23 μm/px).

- Domain δ: Canine pheochromocytoma, retrieved from the DA of VMU. Five cases digitized with a 3DHistech Pannoramic Scan II at 40× magnification (0.25 μm/px).

For the evaluation of the challenge, we constructed the so-called final test set, where only a single evaluation per team was permitted. The dataset included 10 cases per domain, encompassing the following domains, evenly divided between human and veterinary samples:

- Domain 1: Human melanoma, retrieved from the DA of UMC Utrecht, digitized using a Hamamatsu S360 (C13220) at 40× magnification (0.23 μm/px). MC is part of the staging and classification scheme of the AJCC for melanoma (Balch et al., 2009).

- Domain 2: Human astrocytoma, retrieved from the DA of the Institute of Neuropathology at University Hospital Erlangen, digitized with a Hamamatsu S60 at 40× magnification (0.22 μm/px). MC is part of the 2016 WHO grading scheme (Louis et al., 2016).

- Domain 3: Human bladder carcinoma, retrieved from the DA of the Institute of Pathology at University Hospital Erlangen, digitized with a 3DHistech Pannoramic Scan II at 40× magnification (0.25 μm/px). MC is used in the differentiation of tumor types according to (Epstein et al., 1998) and was recently confirmed to be prognostically significant by (Akkalp et al., 2016).

- Domain 4: Canine breast carcinoma, retrieved from the DA of VMU, digitized with a 3DHistech Pannoramic Scan II at 40× magnification (0.25 μm/px). MC is part of the grading scheme by Peña et al. (2013).

- Domain 5: Canine cutaneous mast cell tumor, retrieved from the DA of FUB, digitized with a Hamamatsu S360 (C13220) at 40× magnification (0.23 μm/px). MC is part of the grading scheme by Kiupel et al. (2011).

- Domain 6: Human meningioma, retrieved from the DA of the Institute of Neuropathology at University Hospital Erlangen, digitized with the Hamamatsu S60 at 40× magnification (0.22 μm/px). MC is part of the 2016 WHO grading scheme (Louis et al., 2016).

- Domain 7: Human colon carcinoma, retrieved from the DA of UMC Utrecht, digitized using a Hamamatsu S360 (C13220) at 40× magnification (0.23 μm/px). MC is not part of the grading scheme but was shown to predict survival for lymph-node negative colon carcinoma by Sinicrope et al. (1999).

- Domain 8: Canine splenic hemangiosarcoma, retrieved from the DA of VMU, digitized with a 3DHistech Pannoramic Scan II at 40× magnification (0.25 μm/px). MC is part of the grading scheme of Ogilvie et al. (1996).

- Domain 9: Feline (sub)cutaneous soft tissue sarcoma, retrieved from the DA of VMU, digitized with a 3DHistech Pannoramic Scan II at 40× magnification (0.25 μm/px). MC is part of the grading scheme of Dobromylskyj et al. (2021).

- Domain 10: Feline gastrointestinal lymphoma, retrieved from the DA of VMU, digitized with a

3DHistech Pannoramic Scan II at 40× magnification (0.25 μm/px). For cats, the MC is know to be correlated with the grade according to the National Cancer Institute working formulation (Valli et al., 2000).

While human melanoma (unlabeled) and canine cutaneous mast cell tumor were already part of the training set, the test set used different scanners for both tumor types.

2.2. Establishment of ground truth

The MC is typically assessed on an ROI of 10 high power fields, the size of which is dependent on the optical properties of the microscope (Fitzgibbons and Connolly, 2023). For digital microscopy, it is more sensible to directly define the area, calculated from the resolution of the digitization device, which we set in accordance with previous work (Veta et al., 2019, 2015) to 2 mm^2. The ROI was selected from each digitized WSI by a pathologist with expertise in tumor pathology (C.A.B.) as the area with appropriate tissue and scan quality and the perceived highest mitotic activity, which was considered to be more likely found in a region with high cellular density. This is in accordance with current guidelines (Donovan et al., 2021; Avallone et al., 2021; Ibrahim et al., 2022; Fitzgibbons and Connolly, 2023).

Given the well-known inter-rater disagreements in identification and annotation of MFs, strategic study design methods are essential to limit the effects of these factors on the ground truth for subsequent (ideally unbiased) evaluation. Two main annotation biases need to be considered: When presented with a MF, previously identified as such by another expert, an independent expert might be subject to a confirmation bias. Similarly, it has been reported that the chance of overlooking individual MFs, especially in densely populated cell areas or under suboptimal image quality, should not be neglected (Bertram et al., 2021). Our methods (described in more detail in Bertram et al. (2019)) take both factors into account by identifying MF candidates (i.e., MFs and non-mitotic figures/imposters), which are then classified by three experts in an blinded manner. This task was carried out by one expert (C.A.B.) with roughly equal number of MFs and non-mitotic figures labels, and, in a second step, was supported by a machine learning model aimed at identifying MFs with high recall. The model was trained on the initial manual annotation. This first expert also directly performed an initial classification of these additionally detected structures as MFs or imposters (non-mitotic figures). The second expert (R.K.) performed the same classification task (MF vs. non-mitotic figure) for all identified objects but was blinded to the decision of the first expert. In case of disagreement between both experts, a third expert (T.A.D.) rendered the final class label. All three experts have more than five years of experience in MF identification. This independent vote counteracts a confirmation bias, while use of the machine-learning support mitigates the omission of individual objects. Prior to the assessment, the experts agreed on common criteria for the identification of MFs (Donovan et al., 2021). The annotation of all parts of our dataset (training set, preliminary test set, and the final challenge test set) was carried out using the same methodology. This ground truth definition was used for performance evaluation and ranking of the participants during the MIDOG 2022 challenge.

PHH3-assisted ground truth

Due to the considerable degree of inter-rater disagreement, it is prudent to create a ground truth that relies less on the subjective judgments of multiple experts. Hence, as an alternative ground truth for the test set, we performed IHC staining for all slides of the test set for PHH3. This ground truth definition was not available during the MIDOG 2022 challenge and was developed for this summary paper to gain a better understanding of the algorithmic performance. Histone H3 is a protein that is phosphorylated in the early stages of the mitotic phase and represents a specific marker for mitosis (Hendzel et al., 1997; Bertram et al., 2020; Tellez et al., 2018). However, the specific stain is less pronounced in the last phase (telophase) of mitosis, a phase which is usually morphologically conspicuous with the H&E stain, and is already present in early prophase, which is usually not apparent based on H&E morphology. We hypothesized that the combination of these two staining techniques would increase label consistency. To evaluate H&E and PHH3 in the same cells, we de-stained the H&E-stained slides after digitization and re-stained them with an antibody for PHH3, combined with a secondary antibody equipped with a tailored enzyme that reacts with a substrate to yield a brown stain (see Fig. 2). After digitization of the IHC-stained slide and subsequent manual registration of both scans, a tool based on the EXACT annotation server was

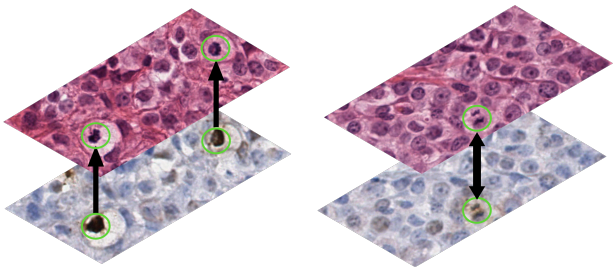

Figure 2: Correspondence between hematoxylin and eosin (H&E)-stained tissue (top) and immunohistochemistry stain against phosphohistone H3 (PHH3, bottom). The left panel shows two tumor cells (green circles) with clear immunopositivity against PHH3 conclusive for MFs, supporting H&E morphology. The right panel shows a mitotic figure in telophase where the PHH3-stain is less conclusive, but the morphology in the H&E is characteristic.

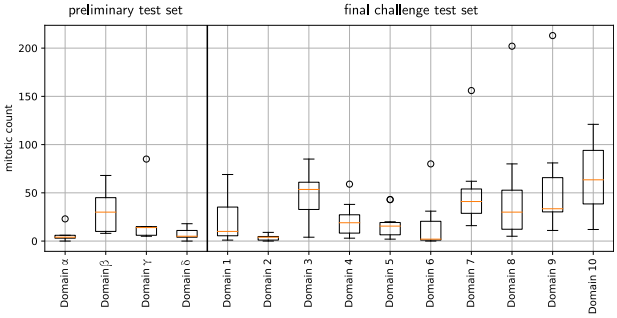

Figure 3: Box-whisker plot of the distribution of MC across the domains of the preliminary test set and the final challenge test set. Boxes indicate lower and upper quartile values, colored lines indicate median values.

employed by an expert (Marzahl et al., 2021), in which both scans could be superimposed with variable transparency. Hence, it was possible to simultaneously evaluate both the specific immunopositivity for PHH3 as well as the morphology in the H&E stain for each cell. In case of non-perfect registration between cells in the PHH3 and H&E stain, the expert annotated the exact coordinate of the MF the H&E stain. Out of 100 cases of the test set, we were able to register 98 to the respective ROIs in the H&E image. For two cases (068 and 100) restaining with PHH3 was not possible due to damage during tissue handling. Immunopositive cells lacking H&E morphology of MFs were not annotated (mostly early prophase MFs) as it is impossible to identify them in the H&E images to which algorithmic analysis and the primary ground truth were restricted.

2.3. Dataset statistics

The MC is expected to vary across tumor types and species. This expectation was confirmed in the distribution of MC shown in the histogram for the training set (Fig. 4) as well as in the box-whisker plots for the preliminary and the final challenge test set (Fig. 3). Tumor types with a comparatively high MC in our samples were canine lung cancer (domain B), canine lymphosarcoma (domain C), canine osteosarcoma (domain β), as well as human bladder carcinoma (domain 3), human colon carcinoma (domain 7), canine hemangiosarcoma (domain 8), and both feline tumors (domains 9 and 10). The mean MC of the training, preliminary test, and final challenge test set were 26.84, 18.00, and 34.74, respectively.

2.4. Reference approaches

For optimal familiarization, challenge participants were provided with three baseline approaches with algorithmic descriptions and preliminary test results. Out of these three approaches, two were based on the RetinaNet (Lin et al., 2017) single-stage object detection architecture and one was based on the Mask RCNN (He et al., 2017) architecture. The first RetinaNet-based approach used a domain-adversarial (Ganin et al., 2016) branch and was trained solely on the MIDOG 2021 training set (i.e., the identical setting as the reference approach for the MIDOG 2021 challenge) and the reference approach for the MIDOG 2021 challenge (Wilm et al., 2022). Considering that this approach was only trained on human breast cancer, we expected a considerable domain gap. The second RetinaNet-based approach was trained on the six domains of the training set (A-F) and used additional stain augmentation, based on Macenko's method for stain deconvolution (Macenko et al., 2009). As the top-performing approaches of MIDOG 2021 were all using (instance) segmentation, we also included the Mask RCNN for this purpose. This approach was, however, not trained with any specific domain-generalizing methods besides default image augmentation. We provided a detailed description of both approaches as part of the challenge proceedings (Ammeling et al., 2023).

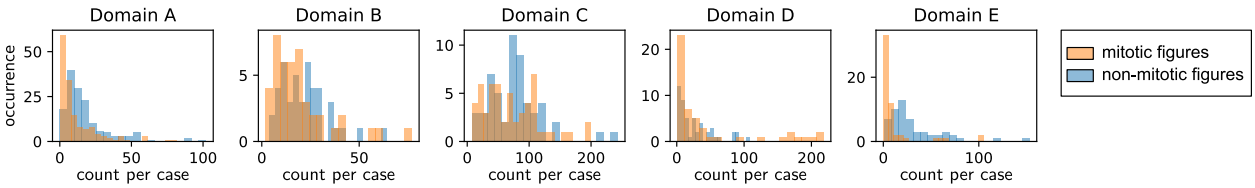

Figure 4: Histogram of MFs and non-mitotic figures in the training set of MIDOG 2022.

2.5. Evaluation methods and metrics

MF identification is a balanced pattern recognition problem in that both an over- and an underestimation of the MCs can lead to equally detrimental consequences: overestimation may lead to excessively aggressive treatment with significant side effects, whereas an underestimation may contribute to more conservative treatment, potentially diminishing the overall treatment outcome. As in prior challenges (Veta et al., 2019, 2016; Roux et al., 2013, 2014), we thus decided to use the F_1 score as our primary metric, as it represents the geometric mean between precision and recall and thus benefits from a good operating point set as a balance between both. To counter averaging effects from the strongly heterogeneous distribution of the MC, we opted to calculate the F_1 score across all cases/images from the summary of respective true positives, false positives, and false negatives over all slides. As the F_1 score is threshold-based, it is additionally insightful to see if competing approaches only chose an unsuitable decision threshold while having an otherwise proper pattern discrimination. Hence, we additionally evaluated the average precision (AP) metric, calculated as the mean precision for 101 linearly spaced recall values between 0 and 1. For the approaches by some teams, a meaningful calculation of AP was not possible due to missing below-threshold classification results. Further, we calculated the precision and recall for all algorithms.

To investigate the dependency on the selection of samples, we performed 1000-fold bootstrapping of the results of each test case, i.e., we randomly selected the same number of cases with replacement from the set of results per case before calculating the precision, recall, AP and F_1 values.

3. Overview of submitted methods

15 registered users from twelve teams submitted at least once to the preliminary test phase of the challenge. Out of those, nine also submitted to the final test phase. All models were based on methods of deep learning. The submitted methods were, as in previous challenges, discrepant in more than one key factor, which makes a direct identification of components for a successful MF detection method difficult. All teams submitted to track 1 (without additional data) of the challenge, while two teams opted to also submit approaches trained by utilizing additional data. In the following section, we will compare the algorithmic strategies of all teams and subsequently discuss the datasets that were additionally used in track 2.

3.1. Pattern recognition tasks

The majority of teams (5/9) chose to frame the task as an object detection task (see Table 1), partially with a second classification stage. Two teams used a semantic segmentation approach and two teams chose a classification-based detection. In particular, the approach by Jahanifar et al. (2022) used fixed-size disks around the centroid coordinate of the MFs to generate the segmentation masks for track 1 and a segmentation mask generated by the NuClick algorithm (Koohbanani et al., 2020) for track 2, while the approach by Yang et al. (2022) used the filled inner circle of the provided bounding box as segmentation target. In contrast, Lafarge and Koelzer (2023) used a classification of patches (78×78 px) with a sliding window like in the original works by Cireşan et al. (2013). Gu et al. (2023) framed object localization as a weakly-supervised learning task derived from class activation maps of medium-sized (240×240 px) patches that were classified as containing a MF or not.

team	tracks	1st stage	architecture	second stage	ensembling	TTA	augmentation				use of unlabeled domain
							geometric	stain	color	other	
Baseline 1 (Ammeling et al., 2023)	2	instance segmentation	Mask RCNN (He et al., 2017), ResNet50 backbone	–	✗	✗	✓	✗	✓	✗	–
Baseline 2 (Ammeling et al., 2023)	1	object detection	RetinaNet (Lin et al., 2017), ResNet18 backbone	–	✗	✗	✓	✓	✓	✗	domain-adversarial
Baseline MIDOG21 (Wilm et al., 2022)	1	object detection	RetinaNet (Lin et al., 2017), ResNet18 backbone	–	✗	✗	✓	✗	✓	✗	domain-adversarial
TIA Centre (Jahanifar et al., 2022)	1, 2	segmentation	Efficient-UNet (B0) (Jahanifar et al., 2021)	EfficientNet-B7	✓	✓	✓	✓	✓	sharpness	–
TCS Research (RnI) (Kotte et al., 2023)	1	object detection	DETR (Carion et al., 2020), ResNet50-DC5 backbone	EfficientNet-B7	✓	✗	✓	✗	✓	✗	–
USZ / UZH Zurich (ML) (Lafarge and Koelzer, 2023)	1	classification (sliding window)	P4-ResNet70 (Cohen and Welling, 2016; He et al., 2016)	–	✓	✗	✓	✗	✓	✗	hard-negative mining
UCLA-HCI (Gu et al., 2023)	1	classification (large tiles)	EfficientNet-B3	weakly supervised localization	✗	✗	✓	✓	✓	blur, noise, balanced-mixup	treated as negative
SKJP (Kondo et al., 2023)	1	object detection	EfficientDet (Tan et al., 2020), EfficientNet V2-L backbone	–	✗	✗	✓	✗	✗	stain normalization	–
HTW Berlin (Annuscheit and Krumnow, 2023)	1	object detection	YOLO v5 (Jocher et al., 2020)	–	✗	✓	✓	✓	✓	blur/sharpening	domain generalization
AI_medical (Yang et al., 2022; Wang et al., 2023)	1, 2	segmentation	SK-UNET (Wang et al., 2021), SE-ResNeXt50 encoder	–	✗	✗	✓	✗	✓	Fourier-domain augmentation	–
Virasoft (Bozaba et al., 2022)	1	object detection	YOLO v5 (Jocher et al., 2020), CSPDarknet53 stem	EfficientNet-B3	✗	✗	✓	✗	✗	mosaic	–
HITszCPATH (Wang et al., 2022)	1	object detection	RetinaNet (Lin et al., 2017), ResNet50 stem	–	✗	✗	✓	✗	✓	–	auxiliary classifier

Table 1: Overview of the submitted methods by all participating teams. TTA indicates test-time augmentation.

3.2. Architectures

The majority of submissions were derivatives of convolutional neural networks (CNNs), while one team designed their method based on the Detection Transformer (DETR) (Carion et al., 2020), which is an object detector derived from the vision transformer class of models and hence uses a CNN solely for feature extraction. Amongst the other approaches, EfficientNet (Tan and Le, 2019)-derived architectures were frequently used. Variants of EfficientNet were used as second stage in the approaches by Jahanifar et al. (2022), Kotte et al. (2023), and Bozaba et al. (2022), as classification approach by Gu et al. (2023), and as a stem of the mitosis detector by Jahanifar et al. (2022) and Kondo et al. (2023). Other researchers chose different well-established network stems, such as ResNet (He et al., 2016), SE-ResNeXt (Hu et al., 2020), or CSPDarknet53 (Bochkovskiy et al., 2020).

3.3. Ensembling and Test-Time Augmentation

While both ensembling and test-time augmentation (TTA) are strategies well-known to enhance model robustness, they were only employed by a minority of participants (see Table 1). Only the winning approach by Jahanifar et al. (2022) employed both ensembling and TTA. The runner-up approach by Kotte et al. (2023) employed a tailored ensembling of model scores of the first and second stages but only in cases where the score of an object in the first stage did not exceed a given threshold. The approach by Lafarge and Koelzer (2023) ensembled two models trained with different augmentation strategies, and integrated the effect of 90-degree rotation for TTA via the use of a rotation invariant model (Cohen and Welling, 2016). The approach by Annuscheit and Krumnow (2023) used four-fold TTA using mirroring of the images.

3.4. Augmentation

All participating teams used standard geometric image transformations like rotation, scaling, and elastic deformations. Additionally, the majority of teams opted to use one form of standard color augmentation that aims at manipulating the hue, brightness, and contrast. Additionally, multiple teams opted to use image perturbations such as blurring, sharpening, and noising. Bozaba et al. (2022) additionally employed mosaic augmentation. Bochkovskiy et al. (2020), and Gu et al. (2023) additionally used balanced mixup (Galdran et al., 2021). Besides those general computer vision augmentation strategies, specific stain augmentation strategies for H&E-stained

images were employed by three teams (Jahanifar et al., 2022; Gu et al., 2023; Annuscheit and Krumnow, 2023), while the approach by Yang and Soatto (2020) augmented images by performing a style-transfer in the frequency-domain.

3.5. Use of the unlabeled domain

The unlabeled domain F of the training set (human melanoma) was employed by four teams. Lafarge and Koelzer (2023) designed a hard-negative mining scheme that additionally employed the unlabeled domain by unmixing the stains into hematoxylin, eosin, and a residual component, and then extracted objects with high residual components (e.g., stain artifacts), which can be mistaken for MFs. Gu et al. (2023) used the surplus domain to treat all images as negatives and counteracted these noisy labels with a specifically crafted loss function. Annuscheit and Krumnow (2023) used the domain as an additional domain in a representation learning scheme for domain adaptation. Finally, Wang et al. (2022) used the additional data in an auxiliary domain classifier in a multi-task learning scheme.

3.6. Domain generalization methodologies

Besides augmentation, several teams employed specific strategies targeted at domain generalization. Annuscheit and Krumnow (2023) designed a domain adaptation scheme based on metric learning where the distance of each sample to prototypes of all domains was minimized to achieve domain generalization. (Wang et al., 2022) employed multi-task learning with two auxiliary tasks: an overall MF classification for the patch and a tumor domain classifier, likely regularizing the model (and hence counteracting domain overfitting). Similarly, (Yang et al., 2022) added a weight perturbation to the loss term, as this was shown to regularize the model and make it more robust to domain shifts (Wu et al., 2020).

3.7. Addional datasets used in track 2

In the second track of the challenge, it was permitted to use publicly available datasets. Yang et al. (2022) used a Hover-Net (Graham et al., 2019) which was trained on other histopathology datasets to generate more accurate segmentation masks. Similarly, the approach by Jahanifar et al. (2022) created enhanced MF segmentation masks by using NuClick (Koohbanani et al., 2020) and additionally by incorporating the TUPAC16 (Veta et al., 2019) dataset to the training dataset.

4. Results

The evaluation of track 1 on all ten tumor domains of the test set shows that the TIA Center approach (Jahanifar et al., 2022) yielded the best overall performance (F_1 = 0.764), closely followed by the approach from the TCS Research team (Kotte et al., 2023) (F_1 = 0.757). Breaking this down into the ten tumor domains, we find a similar overall picture, with both approaches scoring first or second in all domains (see Table 2). We also note that the two leading approaches chose different strategies when optimizing the operating point: While the TIA Center approach yielded a moderately lower recall value at a higher precision value, we found the opposite to be true for the TCS Research approach (see Fig. 8 and Fig. 7). The F1 score is roughly reflected in the precision recall curves of Fig. 9.

In the second track of the challenge, we find a clear superiority of the approach by (Jahanifar et al., 2022), further supported by having the leading edge in all tumor domains.

Comparing the performance in both tracks across tumor domains, we find that tumor domain 2 (human astrocytoma) and 6 (human meningioma), i.e, the neuropathological domains, seemed to have been particularly challenging, with overall maximum F_1 scores of 0.63 and 0.68, respectively (see Table 2). On the contrary, the domains 1 (human melanoma), 3 (human bladder carcinoma), 5 (canine cutaneous mast cell tumor), and 8 (canine splenic hemangiosarcoma) were the tumor domains to which the algorithms generalized best, achieving F_1 scores of up to 0.82, 0.81, 0.82 and 0.82, respectively.

Assessment on alternative (PHH3-assisted) ground truth

After the full annotation of 98 cases based on the joint information of the H&E and PHH3-stained images, we found an increase in the count of MF by 15.0%. Out of the mitotic figures identified aided by the PHH3-stained images, 28.78% (were previously not part of the consensus vote of the three experts based on the H&E stain. We performed a post-hoc analysis of all these cells, the results

Team	overall	Tumor 1	Tumor 2	Tumor 3	Tumor 4	Tumor 5	Tumor 6	Tumor 7	Tumor 8	Tumor 9	Tumor 10
Baseline 2 (Wilm)	0.714 [0.68,0.74]	0.74 [0.61,0.79]	0.48 [0.27,0.63]	0.75 [0.69,0.79]	0.68 [0.61,0.73]	0.81 [0.76,0.84]	0.66 [0.50,0.73]	0.72 [0.62,0.78]	0.77 [0.64,0.82]	0.69 [0.55,0.75]	0.66 [0.56,0.72]
Baseline 1 (Ammeling/Ganz)	0.654 [0.62,0.68]	0.72 [0.59,0.78]	0.32 [0.13,0.48]	0.72 [0.66,0.76]	0.56 [0.48,0.61]	0.76 [0.67,0.80]	0.60 [0.42,0.72]	0.67 [0.59,0.72]	0.70 [0.59,0.74]	0.57 [0.40,0.66]	0.64 [0.55,0.72]
Baseline MIDOG2021	0.513 [0.44,0.58]	0.73 [0.58,0.79]	0.38 [0.13,0.67]	0.74 [0.68,0.79]	0.23 [0.12,0.32]	0.69 [0.64,0.73]	0.62 [0.38,0.70]	0.69 [0.59,0.74]	0.50 [0.34,0.58]	0.32 [0.22,0.36]	0.08 [0.03,0.13]
TIA Centre	**0.764** [0.74,0.78]	**0.80** [0.74,0.84]	**0.65** [0.38,0.79]	**0.81** [0.78,0.83]	0.71 [0.62,0.78]	**0.83** [0.81,0.86]	**0.71** [0.52,0.78]	**0.75** [0.65,0.82]	0.79 [0.70,0.83]	0.70 [0.58,0.77]	**0.77** [0.70,0.81]
TCS Research (RnI)	0.757 [0.72,0.78]	0.76 [0.66,0.80]	0.48 [0.24,0.72]	0.79 [0.73,0.84]	**0.73** [0.66,0.76]	0.79 [0.74,0.83]	0.65 [0.41,0.75]	0.74 [0.64,0.80]	**0.84** [0.73,0.87]	**0.72** [0.62,0.77]	**0.77** [0.70,0.82]
USZ / UZH Zurich (ML)	0.696 [0.66,0.73]	0.76 [0.58,0.83]	0.28 [0.10,0.52]	0.75 [0.69,0.80]	0.66 [0.55,0.73]	0.73 [0.66,0.79]	0.64 [0.45,0.71]	0.71 [0.63,0.77]	0.78 [0.66,0.82]	0.64 [0.46,0.73]	0.66 [0.58,0.73]
UCLA-HCI	0.685 [0.65,0.71]	0.55 [0.39,0.67]	0.48 [0.20,0.74]	0.76 [0.70,0.80]	0.68 [0.62,0.72]	0.77 [0.74,0.79]	0.57 [0.46,0.65]	0.69 [0.61,0.75]	0.72 [0.62,0.76]	0.58 [0.44,0.66]	0.73 [0.67,0.79]
SKJP	0.671 [0.64,0.70]	0.76 [0.65,0.80]	0.51 [0.23,0.69]	0.75 [0.69,0.80]	0.60 [0.53,0.65]	0.70 [0.59,0.76]	0.65 [0.48,0.75]	0.67 [0.57,0.73]	0.71 [0.58,0.77]	0.60 [0.52,0.64]	0.63 [0.52,0.71]
HTW Berlin	0.666 [0.63,0.70]	0.75 [0.64,0.80]	0.57 [0.30,0.76]	0.72 [0.64,0.78]	0.57 [0.41,0.66]	0.77 [0.68,0.81]	0.66 [0.51,0.75]	0.64 [0.58,0.69]	0.69 [0.52,0.76]	0.56 [0.40,0.63]	0.68 [0.57,0.75]
AI_medical	0.659 [0.62,0.69]	**0.80** [0.70,0.84]	0.63 [0.33,0.83]	0.74 [0.67,0.80]	0.64 [0.52,0.73]	0.79 [0.72,0.84]	0.65 [0.49,0.72]	0.68 [0.62,0.72]	0.68 [0.56,0.75]	0.52 [0.43,0.60]	0.59 [0.51,0.65]
Virasoft	0.639 [0.61,0.67]	0.66 [0.50,0.74]	0.34 [0.15,0.65]	0.70 [0.66,0.72]	0.62 [0.54,0.67]	0.60 [0.54,0.65]	0.59 [0.35,0.65]	0.60 [0.49,0.67]	0.73 [0.64,0.77]	0.63 [0.50,0.68]	0.61 [0.54,0.66]
HITszCPath	0.630 [0.59,0.66]	0.75 [0.65,0.80]	0.38 [0.17,0.60]	0.73 [0.66,0.77]	0.55 [0.44,0.63]	0.72 [0.66,0.76]	0.57 [0.34,0.66]	0.62 [0.56,0.65]	0.70 [0.59,0.76]	0.49 [0.33,0.56]	0.61 [0.53,0.67]
TIA Centre (Task 2)	**0.749** [0.72,0.77]	**0.83** [0.77,0.86]	**0.70** [0.37,0.86]	**0.81** [0.78,0.84]	**0.71** [0.61,0.76]	**0.82** [0.80,0.84]	**0.69** [0.57,0.74]	**0.73** [0.64,0.79]	**0.78** [0.68,0.82]	**0.67** [0.52,0.73]	**0.73** [0.65,0.78]
AI_medical (Task 2)	0.708 [0.68,0.73]	0.82 [0.74,0.86]	0.61 [0.29,0.78]	0.78 [0.71,0.83]	0.65 [0.56,0.71]	0.78 [0.74,0.81]	0.66 [0.50,0.76]	0.71 [0.64,0.75]	0.73 [0.61,0.80]	0.58 [0.42,0.64]	0.71 [0.66,0.77]

Table 2: F_1 values across all tumor domains for all participants. Values in brackets indicate 95% confidence interval as a result of bootstrapping. The top group is the baselines, the middle group is the submissions in track 1 and the bottom group is the submissions in track 2 of the challenge.

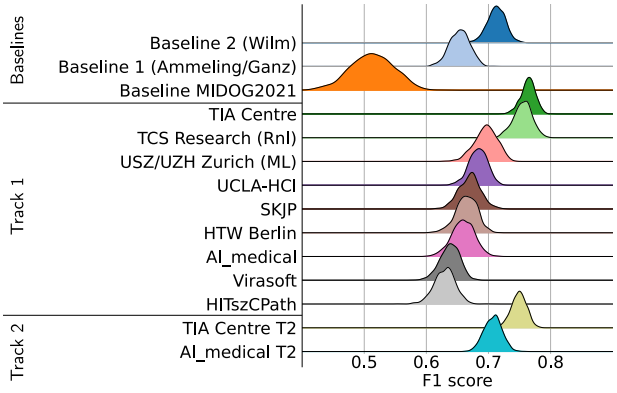

Figure 5: Distribution of the F_1 score as a result of bootstrapping.

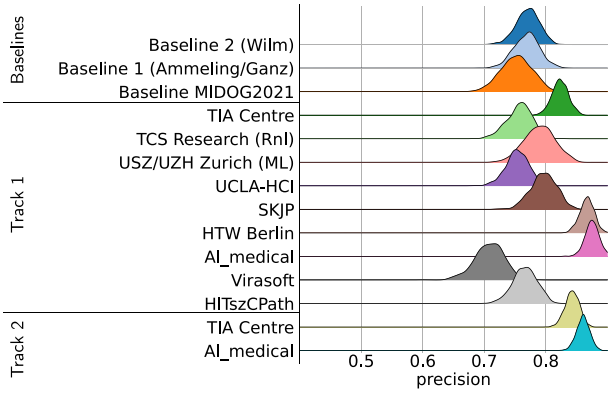

Figure 7: Distribution of precision as a result of bootstrapping.

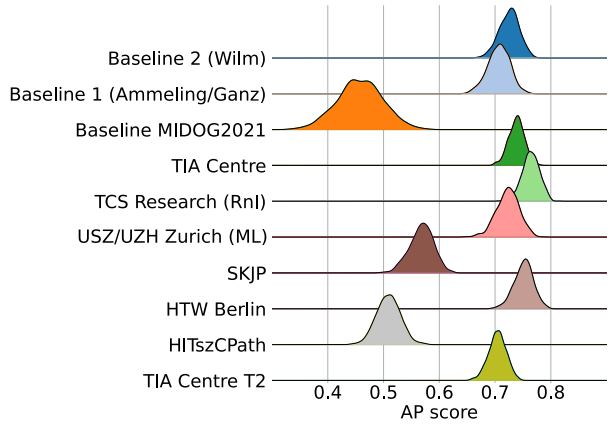

Figure 6: Distribution of the AP score as a result of bootstrapping. Only submissions that provided meaningful model scores are shown.

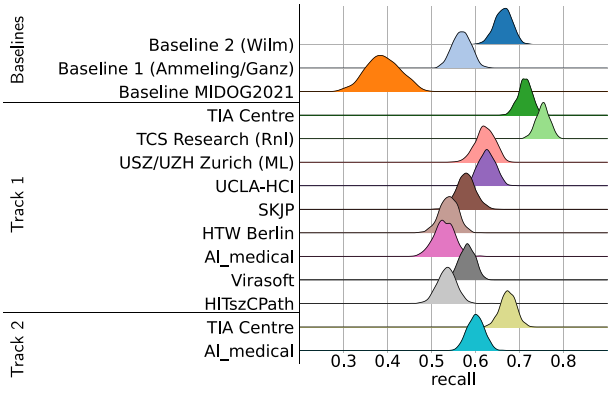

Figure 8: Distribution of recall as a result of bootstrapping.

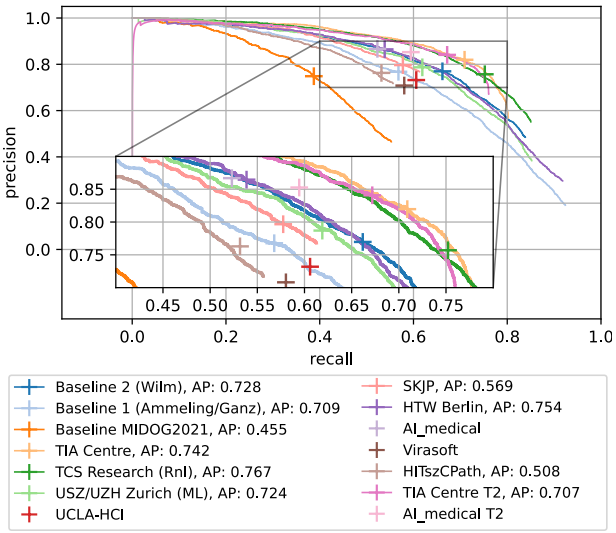

Figure 9: Precision-recall values and curves (for all participants where the model score per MFs was provided and consistent). The marker indicates operating point calculated by the thresholded detections of the participants. Minor mismatches may be explained by post-processing after thresholding.

of which are depicted in Table 3. Out of those MF only identified with help of the PHH3 stain, 20.97% were from the 9% of cases of feline lymphoma, which are generally difficult due to the small cell size resulting at low cellular details at the given image resolution. Over the complete test set, the primary reason for the discrepancy was a borderline mitotic figure morphology, which was hard to discriminate against imposters due to cells being out of focus or superimposed to other cells in thick tissue sections, poor tissue / image quality such as overstained chromatin structures, prophase morphology without obvious chromatin spikes that are difficult to differentiate from apoptotic cells or other, not further classified reasons. Less prominent were difficulties to delineate the MF from imposters due to a borderline cell cycle phase to the G2-phase with early membrane changes and G1-phase with formation of nuclear membranes of the two neighboring daughter cells. In 5.85% of cases the MFs were found to have an unusual morphology, while in 0.54% of cases we found an incomplete capture of the cell at the image borders. Only 1.08% of mitotic figures were considered labeling error in the HE-approach, as characteristic MF

category	subcategory	percentage
borderline morphology	prophase with resemblance to apoptosis	31.23 %
	out of focus (scan artefact or thick tissue section)	27.54 %
	poor image tissue quality (such as overstained)	5.67 %
	not further classified	23.13 %
	total	87.58 %
borderline to non mitotic phases	prophase with early membrane changes	4.05 %
	late telophase (with formation of nuclear membrane)	0.90 %
	total	4.95 %
untypical MF morphology		5.85 %
cut off at image border		0.54 %
overlooked (clear mitotic figure)		1.08 %

Table 3: Breakdown of mitotic figures that were additionally identified using the phosphohistone H3 (PHH3) stain.

morphology was apparent.

When evaluating with this alternative, IHC-assisted ground truth, we found overall lower recall values for all approaches, as shown in Fig. 10, also resulting in overall lower AP and F_1 values. However, the order of the approaches, when sorted by the F_1 value, was almost unaltered. Fig. 10 also shows the precision and recall values of both experts using the original H&E images when evaluated on the PHH3-assisted alternative ground truth, as well as the respective values for the three-expert consensus, indicating a good alignment between the consensus and the IHC-assisted GT. For expert 1 (C.A.B.), we found an overall precision, recall, and F_1 value of 0.926, 0.611, and 0.736, respectively, and for expert 2 (R.K.) we found an overall precision, recall, and F_1 value of 0.659, 0.747 and 0.700, respectively. The three expert consensus achieved a precision, recall, and F_1 value of 0.818, 0.711, and 0.761 respectively.

5. Discussion

The MIDOG 2022 challenge was the first to assess MF recognition across multiple tumor types. This extends the range of covariate shifts to the visual context of the MFs. In the previous challenge, the main domain shift could be attributed to changes in color, sharpness, and depth of field (caused by the differing scanners). In this challenge, the generalization to different tumor types and hence unknown tissue types that surround the MFs is harder to reflect in dedicated domain generalization strategies, e.g., domain augmentation. This may explain why the participants of this iteration of the challenge did not opt to formulate novel augmentation strategies. It is noteworthy that the top three team approaches to track 1 of the challenge used distinctively different strategies to address the

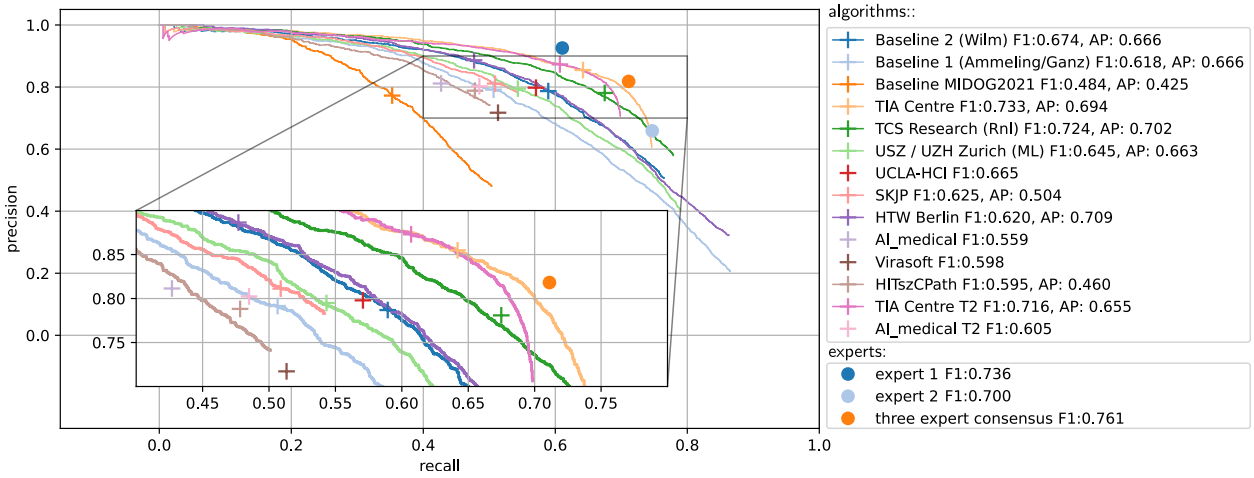

Figure 10: Precision and recall of all approaches and the experts, evaluated on the PHH3-assisted alternative ground truth. AP values and curves are only given for approaches where the model scores were provided and consistent. The expert scores represent the independent assessment of expert 1 and 2 on the hematoxylin and eosin-stained images, which was performed when establishing the original challenge ground truth, the 3 expert consensus represents the challenge ground truth.

pattern recognition problem (semantic segmentation followed by connected components analysis (Jahanifar et al., 2022), object detection (Kotte et al., 2023) and classification on a sliding window (Lafarge and Koelzer, 2023)), highlighting that the *how* (i.e., augmentation, sampling scheme, post-processing) of training was likely more important than the *what* (i.e., the neural network architecture). One commonality between the top three performing approaches to track 1 was, however, that they all used some form of ensembling technique, which has been reported as a strong determinant of success in biomedical challenges (Eisenmann et al., 2023), and likely contributed directly to domain robustness.

The use of containers for the algorithmic submission comes with an increased risk of unintended and unexpected technical failures for the participants. For this reason, we made an independent *preliminary* test set available to the participants. To avoid overfitting of hyperparameters to this set by the participants and, at the same time, to reduce the computational budget required to evaluate the containers, one daily execution was admitted during a two-week time frame prior to the submission. Since overfitting could still not be ruled out, in this version of the challenge, we opted to use four independent (disjointed from the challenge test set) domains in this phase.

The post-challenge evaluation on the alternative, PHH3-assisted ground truth yielded overall lower recall values for all approaches. We attribute this to the inclusion of multiple MFs having inconclusive morphological features in the H&E image, which could be identified with higher confidence in the IHC due to immunopositivity against PHH3-antibodies. Equivocal or inconclusive morphologies include the MF being out of focus due to the factual three-dimensionality of the sample as well as general difficulty in clearly differentiating some MF morphologies (particularly prometaphase MF) from imposters. In the PHH3 stain, however, these structures are clearly distinguishable due to immunoreactivity, which provides an unaltered high contrast, contributing to the overall higher number of MFs. Similar to the expert annotators of the H&E-approach, algorithms were trained (based on the ground truth used) to exclude these morphologically inconclusive structures, which explains the lower recall values of all approaches. The good agreement of the challenge ground truth (three-expert consensus) compared to the alternative and IHC-assisted ground truth highlights the benefits of multiple blinded expert ensembles for H&E-based MF annotations. The in-depth evaluation of mitotic figures that were only identifiable using the IHC stain as secondary source of information

(Table 3), however, also reveals the limitations of purely H&E-based ground truth definitions, as occurring borderline morphological patterns were found to represent the majority of IHC-positive MFs that were not found in the expert consensus of the H&E stain.

One insight from our challenge is the limitations of the AP metric, which averages the precision at defined recall values, as a challenge metric. Besides a high number of hyperparameters (such as the maximum number of detections, the interpolation method, and grid), the AP metric is used according to multiple different definitions (Hirling et al., 2023). Moreover, as can be seen in Fig. 9, none of the algorithms reached the zero value for precision, which penalized the approaches in the AP metric. We hypothesize that this is a result of all approaches using a detection threshold before the non-maximum suppression; a common procedure to reduce computational overhead for the matching of ground truth and candidates, which is an operation in $O(n^2)$. If no value can be meaningfully interpolated for high recall values (e.g., for the MIDOG 2021 baseline approach in Fig. 9 above a recall value of 0.6), the precision value is commonly extrapolated to 0, which penalizes the approach unjustly. Similarly, should the averaging be confined to the maximum achieved recall value, methods employing a high detection threshold would gain an unfair advantage. In particular, this is demonstrated when comparing the winning approach of Jahanifar et al. (2022) and the runner-up of (Kotte et al., 2023). While the precision-recall curve in Fig. 9 clearly indicates the superiority of the winning approach, the AP metric (see Fig. 6) benefits from the lower detection threshold of the approach by Kotte et al. (2023), giving a false impression that the latter approach has a higher decision-threshold independent performance. This provides additional evidence for the utility of the F_1 score as the primary challenge metric.

We found that the top algorithmic solutions of this challenge detected MFs at a level similar to that of the 2021 MIDOG challenge (top F_1 value of 0.748 in 2021 (Aubreville et al., 2023a) and 0.764 in 2022). Additionally, comparing these performances to published F_1 values for human experts (0.563 for human breast cancer (Aubreville et al., 2023a), 0.79 on canine cutaneous mast cell tumor (Bertram et al., 2021)) indicates that the automatic approaches are in the range of human experts. Nevertheless, it is worth pointing out that human experts typically perform this task not only on ROIs but on the entire slide, which was not the task of this challenge. We hence encourage the creation of further datasets and challenges incorporating annotations on the entire WSIs and thus also providing labels for a much more diverse set of tissue characteristics.

Acknowledgement

Computational resources and additional support for the challenge have been provided by grand-challenge.org. The challenge organizers would like to thank Siemens Healthineers and Tribun Health for donating the monetary prizes of the challenge, which have been awarded to the top three participants in each of the two tracks. The organizers would further like to thank Medical Data Donors e.V. for providing assistance in the organization of the awards. J.A. acknowledges support from the Bavarian Institute for Digital Transformation (project ReGInA). M.A. and R.K. acknowledge support by the German Research Foundation (project number 520330054).

6. Author contributions

The challenge was organized by Marc Aubreville, Katharina Breininger, Frauke Wilm, Christof A. Bertram, Samir Jabari, Nikolas Stathonikos, and Mitko Veta.

Frauke Wilm, Jonas Ammeling, and Jonathan Ganz provided the algorithmic reference approaches for the challenge.

The evaluations for this paper were carried out by Marc Aubreville. Christof A. Bertram and Marc Aubreville wrote the main text of this work. Taryn A. Donovan, as a native speaker, has made additional language corrections. All authors reviewed the manuscript.

Christof A. Bertram, Robert Klopfleisch, and Taryn A. Donovan served as expert pathologists in annotating the complete challenge dataset. Nikolas Stathonikos, Robert Klopfleisch, Samir Jabari, Christof Bertram, and Markus Eckstein provided samples for the challenge.

Jonas Annuscheit and Christian Krumnow (Team HTW Berlin), Engin Bozaba (Team Virasoft), Mostafa Jahanifar and Adam Shephard (Team TIA Centre), Satoshi Kondo and Satoshi Kasai (Team SKJP), Sujatha Kotte and VG Saipradeep (Team TCS Research), Maxime W. Lafarge

and Viktor H. Koelzer (Team USZ / UZH Zurich), Ziyue Wang and Yongbing Zhang (Team HITszCPATH), Sen Yang and Xiyue Wang (Team AI_medical) were participants of the challenge. All participants contributed to the overview of the submitted methods section.

References

Akkalp, A.K., OPnur, O., Tetikkurt, U.S., Tolga, D., Özsoy, S., Müslümanoğlu, A.Y., 2016. Prognostic significance of mitotic activity in noninvasive, low grade, papillary urothelial carcinoma. Anal Quant Cytopathol Histopathol 38, 23–30.

Ammeling, J., Wilm, F., Ganz, J., Breininger, K., Aubreville, M., 2023. Reference algorithms for the Mitosis Domain Generalization (MIDOG) 2022 Challenge, in: Mitosis Domain Generalization and Diabetic Retinopathy Analysis. Springer Nature Switzerland, pp. 201–205. doi:10.1007/978-3-031-33658-4_19.

Annuscheit, J., Krumnow, C., 2023. Radial prediction domain adaption classifier for the MIDOG 2022 challenge, in: Mitosis Domain Generalization and Diabetic Retinopathy Analysis. Springer Nature Switzerland, pp. 206–210. doi:10.1007/978-3-031-33658-4_20.

Aubreville, M., Bertram, C., Breininger, K., Jabari, S., Stathonikos, N., Veta, M., 2022. MItosis DOmain Generalization Challenge 2022. Structured description of the challenge design. doi:10.5281/zenodo.6362337.

Aubreville, M., Bertram, C., Veta, M., Klopfleisch, R., Stathonikos, N., Breininger, K., ter Hoeve, N., Ciompi, F., Maier, A., 2021. Quantifying the scanner-induced domain gap in mitosis detection, in: Medical Imaging with Deep Learning (MIDL), Lübeck, 2021, pp. 1–3.

Aubreville, M., Bertram, C.A., Marzahl, C., Gurtner, C., Dettwiler, M., Schmidt, A., Bartenschlager, F., Merz, S., Fragoso, M., Kershaw, O., et al., 2020. Deep learning algorithms out-perform veterinary pathologists in detecting the mitotically most active tumor region. Scientific Reports 10:16447, 1–11. doi:10.1038/s41598-020-73246-2.

Aubreville, M., Stathonikos, N., Bertram, C.A., Klopfleisch, R., Ter Hoeve, N., Ciompi, F., Wilm, F., Marzahl, C., Donovan, T.A., Maier, A., et al., 2023a. Mitosis domain generalization in histopathology images—the MIDOG challenge. Medical Image Analysis 84, 102699.

Aubreville, M., Wilm, F., Stathonikos, N., Breininger, K., Donovan, T.A., Jabari, S., Veta, M., Ganz, J., Ammeling, J., Van Diest, P.J., Klopfleisch, R., Bertram, C.A., 2023b. A comprehensive multi-domain dataset for mitotic figure detection. Scientific Data 10, 484. doi:10.1038/s41597-023-02327-4.

Avallone, G., Rasotto, R., Chambers, J.K., Miller, A.D., Behling-Kelly, E., Monti, P., Berlato, D., Valenti, P., Roccabianca, P., 2021. Review of histological grading systems in veterinary medicine. Vet. Pathol. 58, 809–828. doi:10.1177/0300985821999831.

Azzola, M.F., Shaw, H.M., Thompson, J.F., Soong, S.j., Scolyer, R.A., Watson, G.F., Colman, M.H., Zhang, Y., 2003. Tumor mitotic rate is a more powerful prognostic indicator than ulceration in patients with primary cutaneous melanoma. Cancer 97, 1488–1498. doi:10.1002/cncr.11196.

Balch, C.M., Gershenwald, J.E., Soong, S.j., Thompson, J.F., Atkins, M.B., Byrd, D.R., Buzaid, A.C., Cochran, A.J., Coit, D.G., Ding, S., et al., 2009. Final version of 2009 ajcc melanoma staging and classification. Journal of clinical oncology 27, 6199. doi:10.1200/JCO.2009.23.4799.

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., 2020. Computerized calculation of mitotic count distribution in canine cutaneous mast cell tumor sections: mitotic count is area dependent. Veterinary pathology 57, 214–226. doi:10.1177/0300985819890686.

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. doi:10.1038/s41597-019-0290-4.

Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M., 2020. Yolov4: Optimal speed and accuracy of object detection. arXiv:2004.10934.

Bozaba, E., Çayır, S., Tekin, E., Kukuk, S.B., Shah1, A.I., 2022. Mitosis detection using YOLOv5 and EfficientNet. doi:10.5281/zenodo.8315656.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-end object detection with transformers, in: European conference on computer vision, Springer. pp. 213–229. doi:10.1007/978-3-030-58452-8_13.

Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16, Springer. pp. 411–418. doi:10.1007/978-3-642-40763-5_51.

Cohen, T., Welling, M., 2016. Group equivariant convolutional networks, in: International conference on machine learning, PMLR. pp. 2990–2999. URL: https://proceedings.mlr.press/v48/cohenc16.html.

Dobromylskyj, M.J., Richards, V., Smith, K.C., 2021. Prognostic factors and proposed grading system for cutaneous and subcutaneous soft tissue sarcomas in cats, based on a retrospective study. Journal of Feline Medicine and Surgery 23, 168–174. doi:10.1177/1098612X20942393.

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. doi:10.1177/0300985820980049.

Eisenmann, M., Reinke, A., Weru, V., Tizabi, M.D., Isensee, F., Adler, T.J., Ali, S., Andrearczyk, V., Aubreville, M., Baid, U., et al., 2023. Why is the winner the best?, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19955–19966.

Epstein, J.I., Amin, M.B., Reuter, V.R., Mostofi, F.K., Committee, B.C.C., et al., 1998. The world health organization/international society of urological pathology consensus classification of urothelial (transitional cell) neoplasms of the urinary bladder. The American journal of surgical pathology 22, 1435–1448.

Fitzgibbons, P.L., Connolly, J.L., 2023. Protocol for the examination of resection specimens from patients with invasive carcinoma of the breast. CAP guidelines 4.8.1.0. URL: https://www.cap.org/cancerprotocols.

Galdran, A., Carneiro, G., González Ballester, M.A., 2021. Balanced-mixup for highly imbalanced medical image classification, in: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24, Springer. pp. 323–333. doi:10.1007/978-3-030-87240-3_31.

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.

Gershenwald, J.E., Scolyer, R.A., Hess, K.R., et al., 2017. Melanoma of the skin, in: Amin, M., Edge, S.B., Greene, F.L., et al. (Eds.), AJCC cancer staging manual, Springer.

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. doi:10.1016/j.media.2019.101563.

Gu, H., Haeri, M., Ni, S., Williams, C.K., Zarrin-Khameh, N., Magaki, S., Chen, X., 2023. Detecting mitoses with a convolutional neural network for midog 2022 challenge, in: Mitosis Domain Generalization and Diabetic Retinopathy Analysis. Springer Nature Switzerland, pp. 211–216. doi:10.1007/978-3-031-33658-4_21.

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. doi:10.1109/TPAMI.2018.2844175.

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. doi:10.1109/CVPR.2016.90.

Hendzel, M.J., Wei, Y., Mancini, M.A., Van Hooser, A., Ranalli, T., Brinkley, B., Bazett-Jones, D.P., Allis, C.D., 1997. Mitosis-specific phosphorylation of histone h3 initiates primarily within pericentromeric heterochromatin during g2 and spreads in an ordered fashion coincident with mitotic chromosome condensation. Chromosoma 106, 348–360. doi:10.1007/s004120050256.

Hirling, D., Tasnadi, E., Caicedo, J., Caroprese, M.V., Sjögren, R., Aubreville, M., Koos, K., Horvath, P., 2023. Segmentation metric misinterpretations in bioimage analysis. Nature Methods , 1–4doi:10.1038/s41592-023-01942-8.

Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E., 2020. Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 42, 2011–2023. doi:10.1109/TPAMI.2019.2913372.

Ibrahim, A., Lashen, A., Toss, M., Mihai, R., Rakha, E., 2022. Assessment of mitotic activity in breast cancer: Revisited in the digital pathology era. J. Clin. Pathol. 75, 365–372. doi:10.1136/jclinpath-2021-207742.

Jahanifar, M., Shephard, A., Zamanitajeddin, N., Raza, S.E.A., Rajpoot, N., 2022. Stain-robust mitotic figure detection for midog 2022 challenge. arXiv preprint arXiv:2208.12587 .

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. doi:10.1109/ICCVW54120.2021.00081.

Jocher, G., Stoken, A., Borovec, J., NanoCode012, ChristopherSTAN, Changyu, L., Laughing, tkianai, Hogan, A., lorenzomammana, yxNONG, AlexWang1900, Diaconu, L., Marc, wanghaoyang0106, ml5ah, Doug, Ingham, F., Frederik, Guilhen, Hatovix, Poznanski, J., Fang, J., Yu, L., changyu98, Wang, M., Gupta, N., Akhtar, O., PetrDvoracek, Rai, P., 2020. ultralytics/yolov5: v3.1 - Bug Fixes and Performance Improvements. doi:10.5281/zenodo.4154370.

Kiupel, M., Webster, J., Bailey, K., Best, S., DeLay, J., Detrisac, C., Fitzgerald, S., Gamble, D., Ginn, P., Goldschmidt, M., et al., 2011. Proposal of a 2-tier histologic grading system for canine cutaneous mast cell tumors to more accurately predict biological behavior. Vet. Pathol. 48, 147–155. doi:10.1177/0300985810386469.

Kondo, S., Kasai, S., Hirasawa, K., 2023. Tackling mitosis domain generalization in histopathology images with color normalization, in: Mitosis Domain Generalization and Diabetic Retinopathy Analysis. Springer Nature Switzerland, pp. 217–220. doi:10.1007/978-3-031-33658-4_22.

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. doi:10.1016/j.media.2020.101771.

Kotte, S., Saipradeep, V., Sivadasan, N., Joseph, T., Sharma, H., Walia, V., Varma, B., Mukherjee, G., 2023. A deep learning based ensemble model for generalized mitosis detection in h &e stained whole slide images, in: Mitosis Domain Generalization and Diabetic

Retinopathy Analysis. Springer Nature Switzerland, pp. 221–225. doi:10.1007/978-3-031-33658-4_23.

Lafarge, M.W., Koelzer, V.H., 2023. Fine-grained hard-negative mining: Generalizing mitosis detection with a fifth of the MIDOG 2022 dataset, in: Mitosis Domain Generalization and Diabetic Retinopathy Analysis. Springer Nature Switzerland, pp. 226–233. doi:10.1007/978-3-031-33658-4_24.

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. doi:10.1109/TPAMI.2018.2858826.

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. doi:10.1007/s00401-016-1545-1.

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. doi:10.1109/isbi.2009.5193250.

Makki, J., 2015. Diversity of breast carcinoma: histological subtypes and clinical relevance. Clinical medicine insights: Pathology 8, CPath–S31563.

Malon, C., Brachtel, E., Cosatto, E., Graf, H.P., Kurata, A., Kuroda, M., Meyer, J.S., Saito, A., Wu, S., Yagi, Y., 2012. Mitotic figure recognition: Agreement among pathologists and computerized detector. Analytical Cellular Pathology 35, 97–100. doi:10.3233/ACP-2011-0029.

Malon, C., Miller, M., Burger, H.C., Cosatto, E., Graf, H.P., 2008. Identifying histological elements with convolutional neural networks, in: Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology, pp. 450–456. doi:10.1145/1456223.1456316.

Marzahl, C., Aubreville, M., Bertram, C.A., Maier, J., Bergler, C., Kröger, C., Voigt, J., Breininger, K., Klopfleisch, R., Maier, A., 2021. EXACT: a collaboration toolset for algorithm-aided annotation of images with annotation version control. Scientific Reports 11:4343, 1–10. doi:10.1038/s41598-021-83827-4.

McNiel, E., Ogilvie, G., Powers, B., Hutchison, J., Salman, M., Withrow, S., 1997. Evaluation of prognostic factors for dogs with primary lung tumors: 67 cases (1985-1992). Journal of the American Veterinary Medical Association 211, 1422–1427.

Meuten, D., Moore, F., George, W., 2008. Appendix: Diagnostic Schemes and Algorithms, in: Meuten, D.J. (Ed.), Tumors in Domestic Animals. Iowa State Press, Ames, Iowa, USA, pp. 755–769.

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. doi:10.1038/modpathol.3800388.

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. doi:10.5858/133.11.1826.

Ogilvie, G.K., Powers, B.E., Mallinckrodt, C.H., Withrow, S.J., 1996. Surgery and doxorubicin in dogs with hemangiosarcoma. Journal of Veterinary Internal Medicine 10, 379–384. doi:10.1111/j.1939-1676.1996.tb02085.x.

Peña, L., Andrés, P.D., Clemente, M., Cuesta, P., Pérez-Alenza, M., 2013. Prognostic value of histological grading in noninflammatory canine mammary carcinomas in a prospective study with two-year follow-up: relationship with clinical and histological characteristics. Veterinary Pathology 50, 94–105. doi:10.1177/0300985812447830.

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.

Roux, L., Racoceanu, D., Loménie, N., Kulikova, M., Irshad, H., Klossa, J., Capron, F., Genestie, C., Le Naour, G., Gurcan, M., 2013. Mitosis detection in breast cancer histological images an icpr 2012 contest. Journal of Pathology Informatics 4, 8. doi:10.4103/2153-3539.112693.

Sinicrope, F.A., Hart, J., Hsu, H.A., Lemoine, M., Michelassi, F., Stephens, L.C., 1999. Apoptotic and mitotic indices predict survival rates in lymph node-negative colon carcinomas. Clinical cancer research 5, 1793–1804.

Soliman, N.A., Yussif, S.M., 2016. Ki-67 as a prognostic marker according to breast cancer molecular subtype. Cancer biology & medicine 13, 496.

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. doi:10.1109/JBHI.2020.3032060.

Tan, M., Le, Q., 2019. EfficientNet: Rethinking model scaling for convolutional neural networks, in: Chaudhuri, K., Salakhutdinov, R. (Eds.), Proceedings of the 36th International Conference on Machine Learning, PMLR. pp. 6105–6114. URL: https://proceedings.mlr.press/v97/tan19a.html.

Tan, M., Pang, R., Le, Q.V., 2020. Efficientdet: Scalable and efficient object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10781–10790. doi:10.1109/CVPR42600.2020.01079.

Tellez, D., Balkenhol, M., Otte-Höller, I., van de Loo, R., Vogels, R., Bult, P., Wauters, C., Vreuls, W., Mol, S., Karssemeijer, N., et al., 2018. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. IEEE transactions on medical imaging 37, 2126–2136. doi:10.1109/TMI.2018.2820199.

Valli, V., Jacobs, R., Norris, A., Couto, C.G., Morrison, W., McCaw, D., Cotter, S., Ogilvie, G., Moore, A., 2000. The histologic classification of 602 cases of feline lymphoproliferative disease using the national cancer institute working formulation. Journal of Veterinary Diagnostic Investigation 12, 295–306.

Valli, V., Kass, P.H., Myint, M.S., Scott, F., 2013. Canine lymphomas: Association of classification type, disease stage, tumor subtype, mitotic rate, and treatment with survival. Vet. Pathol. 50, 738–748. doi:10.1177/0300985813478210.

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. doi:10.1016/j.media.2019.02.012.

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. doi:10.1371/journal.pone.0161286.

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. doi:10.1016/j.media.2014.11.010.

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. doi:10.1109/JSEN.2021.3056131.

Wang, X., Zhang, J., Yang, S., Xiang, J., Luo, F., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X., 2023. A generalizable and robust deep learning algorithm for mitosis detection in multicenter breast histopathological images. Medical Image Analysis 84, 102703. doi:10.1016/j.media.2022.102703.

Wang, Z., Chen, Y., Fang, Z., Bian, H., Zhang, Y., 2022. Multi-task retinanet for mitosis detection, in: Mitosis Domain Generalization and Diabetic Retinopathy

Analysis. Springer Nature Switzerland, pp. 234–240. doi:10.1007/978-3-031-33658-4_25.

WHO Classification of Tumours Editorial Board, 2022. Who classification of endocrine and neuroendocrine tumours. .

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. doi:10.1007/978-3-030-97281-3_1.

Wu, D., Xia, S.T., Wang, Y., 2020. Adversarial weight perturbation helps robust generalization. Advances in Neural Information Processing Systems 33, 2958–2969. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1ef91c212e30e14bf125e9374262401f-Paper.pdf.

Yang, S., Luo, F., Zhang, J., Wang, X., 2022. SK-Unet Model with Fourier Domain and Weight Perturbation for Mitosis Detection. URL: https://doi.org/10.5281/zenodo.7035741, doi:10.5281/zenodo.7035741.

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. doi:10.1109/CVPR42600.2020.00414.