Mitosis detection, fast and slow: Robust and efficient detection of mitotic figures

Mostafa Jahanifar [a], Adam Shephard [a], Neda Zamanitajeddin [a], Simon Graham [a,b],
Shan E. Ahmed Raza [a], Fayyaz Minhas [a], Nasir Rajpoot [a,b,*]

[a] *Tissue Image Analytic (TIA) Center, Department of Computer Science, University of Warwick, UK*
[b] *Histofy Ltd, Birmingham, UK*

A B S T R A C T

Counting of mitotic figures is a fundamental step in grading and prognostication of several cancers. However, manual mitosis counting is tedious and time-consuming. In addition, variation in the appearance of mitotic figures causes a high degree of discordance among pathologists. With advances in deep learning models, several automatic mitosis detection algorithms have been proposed but they are sensitive to *domain shift* often seen in histology images. We propose a robust and efficient two-stage mitosis detection framework, which comprises mitosis candidate segmentation (*Detecting Fast*) and candidate refinement (*Detecting Slow*) stages. The proposed candidate segmentation model, termed *EUNet*, is fast and accurate due to its architectural design. EUNet can precisely segment candidates at a lower resolution to considerably speed up candidate detection. Candidates are then refined using a deeper classifier network, EfficientNet-B7, in the second stage. We make sure both stages are robust against domain shift by incorporating domain generalization methods. We demonstrate state-of-the-art performance and generalizability of the proposed model on the three largest publicly available mitosis datasets, winning the two mitosis domain generalization challenge contests (MIDOG21 and MIDOG22). Finally, we showcase the utility of the proposed algorithm by processing the TCGA breast cancer cohort (1,124 whole-slide images) to generate and release a repository of more than 620K potential mitotic figures (not exhaustively validated).

1. Introduction

Mitosis, a key cell-life cycle process, involves chromosome replication and separation into two nuclei, resulting in two identical cells (Cheeseman and Desai, 2008). Detection and counting of mitotic figures, particularly relevant for tumor analysis in various cancers (Cree et al., 2021), have demonstrated a strong correlation with cell proliferation, serving as a key parameter in tumor grading systems (Paul and Mukherjee, 2015; Rakha et al., 2008). However, the diversity in mitotic figure appearances and resemblance of imposters/mimicker cells often lead to significant inter-rater variability (see Fig. 1 for examples of mitoses and mimickers).

The rise of digital pathology (DP), driven by whole-slide scanners, has fostered the growth of Computational Pathology (CPath), which facilitates analysis of multi-gigapixel Whole-Slide Images (WSIs) (Graham et al., 2019; Shephard et al., 2021; Alemi Koohbanani et al., 2018, 2019). Generally, CPath enhances objectivity and reproducibility in histopathology tasks (Cruz-Roa et al., 2017; Bizzego et al., 2019; Djuric

et al., 2017), with Deep Learning (DL) methods providing promising avenues for automated mitotic figure detection/counting (Mathew et al., 2021; Aubreville et al., 2022b). Nonetheless, applying machine learning to clinical practice poses challenges. Models must be robust to WSI appearance variations, stemming from differences in sample preparation, tissue types, and scanner hardware (Asif et al., 2021; Aubreville et al., 2022b). This variability introduces *domain shift* in WSIs from different scanners and sites (see Fig. 1.a for an example of variation caused by two different scanners on the same sample).

The MItosis DOmain Generalization challenge (MIDOG21) (Aubreville et al., 2022b) offered a testing ground for mitotic figure detection algorithms amidst domain shift, specifically in human breast cancer. Yet, domain shift also emerges from tissue type and species differences, affecting mitotic figure appearance (Bertram et al., 2019). Thus, robust tools for different cancer types, species, scanners, or preparation protocols are desirable. MIDOG22 (Aubreville et al., 2022a) expanded on this by considering domain shifts from different tumor types and species.

Fig. 1. Two common challenges with automatic mitosis detection task.

On an average WSI, among around 100,000 nuclei, 100–1000 mitotic figures are 'rare events' requiring high-resolution manual counting. Such rare event detection at high resolution (40× magnification) is taxing for humans and algorithms alike (He et al., 2016; Lin et al., 2017; He et al., 2017). To ease this, pathologists usually scout at a lower magnification to select a 'mitotic hotspot' based on cell density and morphology, and then count mitoses at a higher magnification (Ellis et al., 2005). An effective mitosis detection algorithm should maintain accuracy while speeding up the counting process, especially when processing large WSIs.

In this work, we aim to resolve the accuracy and speed challenges by introducing a novel mitosis detection method, inspired by the multi-magnification workflow used by pathologists. In particular, our approach consists of two steps: (1) segmentation of mitotic candidates (Detecting Fast) using a novel model architecture, and (2) candidate refinement by passing the candidates to a deeper classifier network to differentiate between mitotic figures and mimickers (Detecting Slow). To enable mitosis segmentation, we propose to generate mitosis masks from point annotations using NuClick (Alemi Koohbanani et al., 2020). For the first time, we investigate the effect of different combinations of three well-known domain generalization techniques for mitosis segmentation. We use this knowledge to design a robust model to counter the domain shift caused by using different scanners. We show that our proposed method outperforms all other state-of-the-art (SOTA) algorithms in literature and achieved the first rank in both the MIDOG21 and MIDOG22 challenges. Furthermore, we showcase the practicality of our algorithm by detecting mitotic figures in the breast cohort of the TCGA dataset (TCGA-BRCA). In summary, the main contributions of this paper include:

1. An efficient two-stage mitosis detection method based on a novel segmentation model architecture and a deep classification model.
2. A self-supervised training method to pretrain both encoder and decoder parts of a segmentation model.
3. Investigation of the effect of different combinations of domain generalization techniques on the mitosis segmentation task.
4. Release of the AI-generated results for TCGA-BRCA cohort consisting of over 620 K mitotic figures (available at https://zenodo.org/records/10245707).
5. Outperforming other SOTA algorithms in cross-validation experiments while being considerably faster as well as ranking 1st in MIDOG21 and MIDOG22 challenges.

2. Related works

Since 2012 (Krizhevsky et al., 2012) convolutional neural networks (CNNs) paved the way for transformative advancements in computer vision, with impressive results in image segmentation, detection, and classification tasks (Li et al., 2021). Their subsequent ubiquity in CPath made CNNs a cornerstone of various segmentation and classification tasks, including the detection of mitotic figures (Mathew et al., 2021; Dif and Elberrichi, 2020). The research community has responded to this phenomenon with multiple AI challenges centered around mitotic figure detection (Ludovic et al., 2013; Veta et al., 2019; Aubreville

et al., 2021, 2022a). The first publicly available mitosis detection challenges and datasets were ICPR2012 (Ludovic et al., 2013) and its subsequent challenge MITOS-ATYPIA 2014, both of which consisted of a limited number of cases and training images. At this stage, DL-based methods were less prevalent, and the small image set was manageable. This was followed by the TUPAC16 challenge (Veta et al., 2019), where participants were tasked with counting mitotic figures and predicting a WSI tumor proliferation score. More recently, the MIDOG21 and MIDOG22 challenges focused on detecting mitotic figures in histology images from various scanners to address domain shift (Aubreville et al., 2021, 2022a). In these contests, all participants utilized CNNs for mitotic figure detection.

2.1. Mitosis detection

Mitotic figure detection through DL usually involves three main approaches. The first employs patch-based classification, dividing regions of interest (ROIs) or WSIs into small patches for CNN classification. The second involves detection models that predict bounding boxes or centroid points for the mitotic figures. The third approach uses segmentation models to semantically delineate targets before determining the mitotic centroid via post-processing.

Initial DL-based methods handled this task as a classification problem. Notably, Akram et al. (2018) used ResNet-12 for the task on the TUPAC16 dataset (Veta et al., 2019) and improved their model by fine-tuning it with additional mined mitotic figures. An ensemble of CNNs was proposed by Tellez et al. (2018), with knowledge distillation reducing computational needs and 'HED stain augmentation' increasing the range of realistic H&E stain variations for better model training. Despite their success, the severe limitation of these models lies in their inefficiency, as they need to iterate through every high-resolution WSI patch, with patches being small to contain a single mitosis.

Bounding box detection models, such as RetinaNet (Lin et al., 2017), Cascade R-CNN (Razavi et al., 2021), and EfficientDet (Tan et al., 2020) are more efficient for mitosis detection than patch-based models due to their ability to process larger images, capturing more context, and enabling faster predictions. For example, both Wilm et al. (2021b) and Chung et al. (2021) used two different versions of RetinaNet for mitosis detection while improving the domain generalization capability of their models by incorporating domain adversarial training (Ganin et al., 2016) and style transfer augmentation techniques, respectively.

Several approaches treated mitosis detection as a segmentation task using methods like Mask R-CNN (He et al., 2017) or fully convolutional networks (FCN) such as U-Net (Ronneberger et al., 2015), which use large image patches to reduce processing times. Notably, Kausar et al. (2020), Fick et al. (2021), and Sebai et al. (2020) optimized Mask R-CNN for this purpose. Li et al. (2019)'s FCN model, SegMitosis, used point mitosis annotations (weak labels) to form concentric circles, achieving SOTA results on the TUPAC dataset using a concentric loss in training. Additionally, Yang et al. (2021) proposed SK-Unet, an improved U-Net model with selective kernels, that achieved the joint first rank on the MIDOG21 challenge leaderboard. Interestingly, the top three MIDOG21 entries turned the detection task into instance-level segmentation. Yang et al. (2021) used HoVer-Net (Graham et al., 2019) to generate nuclear segmentation masks and filter non-mitotic figures, creating ground truth masks. Fick et al. (2021) manually segmented 100 mitotic figures to train an initial Mask-RCNN model, generating pixel-level segmentation. Despite their superiority, segmentation approaches require exhaustive annotations and can be computationally costly. Our pipeline mitigates these issues by using an interactive model to generate reliable mitosis masks (Alemi Koohbanani et al., 2020) and performing segmentation in a lower resolution.

Lastly, multi-stage methods, such as the ones used by Nateghi and Pourakpour (2021), Liang et al. (2021) and Mahmood et al. (2020), have gained traction. These typically involve finding mitotic candidates using a bounding box detection model, followed by classification. Fick

et al. (2021) implemented a two-stage process involving a Mask-RCNN for segmenting mitotic figures, followed by classification through an ensemble of DenseNet201 and ResNet50. Similarly, Kondo (2021) utilized thresholding on blue ratio images for candidate mitotic region extraction before classification with a ResNet model. While these methods improved performance, they also considerably increased computational cost. We introduce a two-stage method with an adept segmentor for better generalization, reducing false candidates while preserving high sensitivity. By operating on downscaled images, our segmentation module considerably lowers computational costs and enhances pipeline efficiency.

2.2. Domain generalization

To address domain shift resulting from the varied scanner/source use, diverse approaches have been proposed, with most employing some form of color augmentation during training for enhanced algorithm generalizability amidst domain shift (Yang et al., 2021; Razavi et al., 2021; Kondo, 2021). Techniques such as the histology-specific 'HED stain augmentation'–which deconvolves an image into Hematoxylin and Eosin stain channels and perturbs them–are shown to be effective (Tellez et al., 2019; Nateghi and Pourakpour, 2021). Although the effectiveness of this approach on permuting image color information during model training to cover potential stain variations has been investigated before (Tellez et al., 2019), its impact on mitosis segmentation remains unexplored. Furthermore, stain normalization methods like (Vahadane et al., 2016) are widely utilized (Razavi et al., 2021; Liang et al., 2021) to reduce the domain shift caused by sample preparation and scanner variance.

Some methods harnessed unlabeled images from varied scanners through image synthesis techniques, generating new image variations for training (Fick et al., 2021; Chung et al., 2021). Techniques like Fourier domain mixing were also employed, swapping low-frequency domain information between images for unsupervised stain normalization and augmentation, potentially increasing model generalizability (Yang et al., 2021).

Other strategies combating domain shift in CPath include model pretraining (Alemi Koohbanani et al., 2021; Vuong et al., 2022) and domain-adversarial training (Ganin et al., 2016; Wilm et al., 2021b; Lafarge et al., 2019), though their effectiveness may be task-specific (Stacke et al., 2020; Jahanifar et al., 2023). The efficacy of these domain generalization techniques and their combinations on mitosis segmentation has yet to be explored. This study comprehensively investigates the impact of key domain generalization techniques on mitosis segmentation, aiming to identify optimal strategies.

3. Methodology

3.1. Overview

The 'Thinking, Fast and Slow' theory by Kahneman (2011) considers a dichotomy between two systems of thought where system 1, or 'Thinking Fast', makes decisions faster and instinctive while system 2, or 'Thinking Slow', usually takes a more deliberate process to arrive at a logical conclusion. Kahneman (2011) discusses the benefits and properties of each system and describes their importance. Many research studies in artificial intelligence have been motivated by this theory to come up with effective solutions for hard problems (Miech et al., 2021).

Metaphorically inspired by this theory, we propose the 'Mitosis Detection, Fast and Slow' (MDFS) framework (shown in Fig. 2(b) consisting of two main parts: (1) 'Detecting Fast', which is responsible for finding mitosis candidates as fast and much as possible and (2) 'Detecting Slow', where a deeper model refines mitosis candidates to eliminate mimickers. We approach candidate detection as a segmentation problem where mitosis masks are acquired by leveraging an interactive segmentation model, called NuClick (Jahanifar et al., 2019;
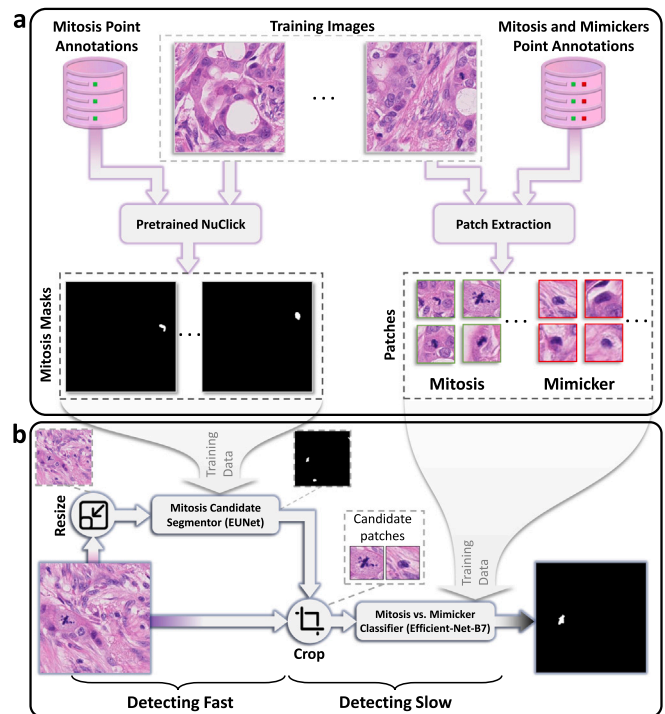

Fig. 2. The overview of the proposed mitosis detection method: (a) data pre-processing steps, where mitosis masks and mitosis/mimicker patches are generated, and (b) the proposed 'Detecting Fast' and 'Detecting Slow' systems for candidate segmentation and refinement.

Alemi Koohbanani et al., 2020). Because the goal of the 'Detecting Fast' system is to detect plausible mitotic candidates as fast as possible with high sensitivity, we also propose a down-sampling step in the fast system to further improve efficiency. Then, in the 'Detecting Slow' system, we extract small patches around mitotic candidates at full resolution and assess them using a deeper CNN. To make the entire framework more robust, we also include special considerations to counter the domain shift problem. We outline the proposed techniques in detail in the following sections. It is important to clarify that our use of the 'Fast and Slow' framework is metaphorical and does not directly align with Kahneman's cognitive systems although many attributes of these systems are similar. In our view, the 'Fast and Slow' metaphor provides an intuitive way to conceptualize such two-stage mitosis detection methods, without strict adherence to its delineation of cognitive processes.

3.2. Detecting fast: mitosis candidate segmentation

3.2.1. Network architecture

We propose an efficient segmentation model architecture, called *EUNet*, for the mitosis detection, which follows the encoder–decoder design of U-Net (Ronneberger et al., 2015). Here, we use a pre-trained EfficientNet-B0 model (Tan and Le, 2019) as the encoder and an inverse design of that for the decoder part (using upsampling blocks instead of down-sampling). In other words, we replaced the standard convolution layers in the standard UNet architecture with the Mobile Inverted Residual blocks coupled with a Squeeze-and-Excitation mechanism (MIRSE block).

The exact design of the MIRSE block is depicted in Fig. 3a, where a sequence of a 1×1 2D convolution layer, a $K \times K$ depth-wise convolution layer (to make the network lighter), a squeeze-and-excitation (S&E) layer (Hu et al., 2018), another 1×1 2D convolution layer, and a residual connection (to improve back-propagation and avoid vanishing gradients) are incorporated. In all layers, the parameters K and F
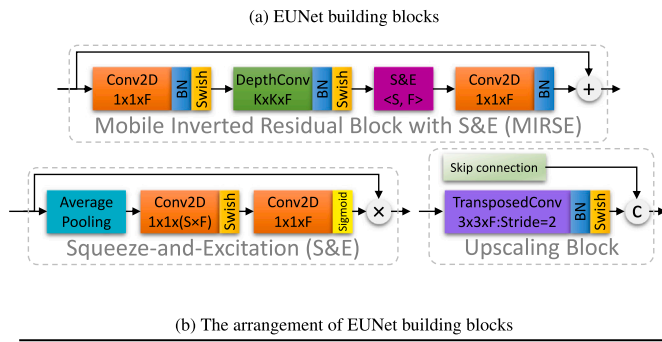
(a) EUNet building blocks

Fig. 4. Mitosis segmentation post-processing steps.

(b) The arrangement of EUNet building blocks

Encoder	Decoder
Conv2D (K=3, F=32, Stride=2)	Upscale (F=320) MIRSE (K=5, F=192, R=3)
MIRSE (K=3, F=16, R=1)	MIRSE (K=5, F=112, R=3)
MIRSE (K=3, F=24, R=2)*	Upscale (F=112) MIRSE (K=3, F=80, R=3)
MIRSE (K=5, F=40, R=2)*	Upscale (F=80) MIRSE (K=5, F=40, R=2)
MIRSE (K=3, F=80, R=3)*	Upscale (F=40) MIRSE (K=3, F=24, R=2)
MIRSE (K=5, F=112, R=3)	Upscale (F=24)
MIRSE (K=5, F=192, R=5)*	MIRSE (K=3, F=16, R=2)
MIRSE (K=3, F=320, R=1)	Conv2D (K=1, F=1)

Fig. 3. Architecture of the proposed **EUNet** for mitosis candidate segmentation. For each operation, its parameters are outlined in the parenthesis where K, F, and R denote kernel size, the number of feature maps, and the number of repetitions, respectively.

denote the kernel size and number of feature maps, respectively. It is worth noting that batch normalization (BN) (He et al., 2016) and Swish activation function (Ramachandran et al., 2017) are applied on the output of all convolution layers in the MIRSE block (except for the last one that only contains BN). The S&E layers provide a self-attention mechanism inside each layer of the network to calculate the importance of different feature maps and weight them accordingly. The squeeze parameter, S, of the S&E layer in this work is set to be 0.25. The 'Upscaling Block' in the proposed network architecture is a 3×3 transposed convolution layer with a stride of 2 to increase the spatial size of input feature maps by a scale of 2. Also, this block concatenates the resulting feature maps from the same level of the encoder (retrieved via 'Skip connection') with the upsampled feature maps to benefit from the high-resolution information available in the encoder part. The motivation behind incorporating elements such as Swish activations and other architectural modifications is to leverage advancements and proven benefits from related studies. These choices have been demonstrated to enhance the performance and efficiency of deep learning models in various tasks, including image classification and detection (Ramachandran et al., 2017; Tan and Le, 2019; Tan et al., 2020; Sandler et al., 2018).

The overall architecture design of the proposed EUNet model is described in Fig. 3b where the order of different building blocks, their design parameters (K and F), and the number of repetitions in each level (R) of the encoder and decoder parts are provided. In the star-marked MIRSE blocks of the encoder path, the first convolution layer is applied with a stride of 2 to decrease the spatial resolution of feature maps by a factor of 2.
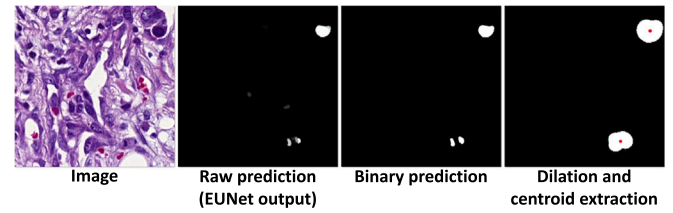
3.2.2. Mitosis masks generation

In order to provide stronger supervision and train our proposed segmentation model, we use NuClick (Alemi Koohbanani et al., 2020; Jahanifar et al., 2019) to obtain mitosis masks from mitosis point annotations. NuClick is an open-source[1] interactive segmentation model that can generate nuclei masks from input images using point annotations as guiding signals. The outputs of NuClick have proved to be reliable in various applications (Alemi Koohbanani et al., 2020; Shephard et al., 2021; Graham et al., 2021; Gamper et al., 2020). Using NuClick, we convert all the point annotations into mitosis masks to be used during the training of our segmentation models (as shown in Fig. 2a).

3.2.3. Training

Our segmentation model is trained on image tiles of size 512×512 pixels in all experiments. However, there is a severe class imbalance when comparing patches with mitotic figures (positive patches) to those with no mitotic figures (negative patches). This may lead to a decline in model sensitivity, owing to the model seeing many more negative patches during training. To mitigate this effect, we incorporated on-the-fly under-sampling of negative patches where each batch is forced to have an equal number of positive and negative patches.

In all experiments, model training was done in two phases. In the first phase, we froze all the encoder layers, only training the decoder, for 10 epochs. For the second phase, we trained the whole network for 50 epochs. We used the Jaccard loss function (Jahanifar et al., 2018) and the Adam optimizer (Kingma and Ba, 2014) with learning rates of 0.003 and 0.0004 to optimize the model during the first and second phases, respectively.

3.2.4. Post-processing

For post-processing, we performed simple thresholding of the prediction map y_p to obtain a binary mitosis mask. This threshold is set based on the results from cross-validation experiments (see Appendix A). Then, to merge the prediction masks for mitotic figures in the anaphase or telophase, we apply a morphological dilatation operation with a disk structuring element of 18 pixels radius because daughter cells of a mitotic figure are usually closer than that radius in these phases (see Fig. 4 for an example). Following this, we extracted the centroids of connected components in the processed mask as mitosis candidates.

3.3. Detecting slow: mitosis candidate refinement

Within the 'Detecting Fast' system of our framework, we use a mitotic candidate segmentor, operating on a down-scaled input image to improve the processing speed. Since we do not use the full-resolution input image, it is expected that the detected candidates will not be of sufficient quality. To compensate for this, we consider a deeper CNN classifier in the 'Detecting Slow' system of the MDFS framework to accurately classify those candidates into mitoses or mimickers. We use EfficientNet-B7 (Tan and Le, 2019) as the mitotic candidate classifier

---

[1] https://github.com/mostafajahanifar/nuclick_torch.

which used 128×128 candidate patches extracted from full-resolution images as input and classify them as either mitotic figures or mimickers.

To deal with the problem of class imbalance between mitosis and mimicker categories, we incorporate under-sampling of the mimicker class as well as a weighted cross-entropy loss (where the mitosis class is given twice the weight) during the model training. These techniques allow us not to lose many real mitotic figures in the refinement phase and keep the overall sensitivity of the proposed method as high as possible.

3.4. Addressing domain shift

3.4.1. Stain normalization

Stain normalization is one of the mainstream techniques to address stain variation in digital pathology and various methods have been investigated to achieve it (Roy et al., 2018). In this study, we investigate the effect of using the Vahadane stain normalization method (Vahadane et al., 2016) on mitosis segmentation. The Vahadane method has been selected as it preserves the structural properties of stained tissue samples and is robust to stain sparsity that may be found in pathology images.

In particular, Vahadane et al. (2016) uses Sparse Non-negative Matrix Factorization (SNMF) to estimate the stain matrix $\mathbf{S}$ and concentration matrix $\mathbf{C}$ from the source and target images. Then, it scales the concentration map of the source image and combines it with the stain matrix of the target image to normalize the source image (Vahadane et al., 2016).

3.4.2. Stain augmentation

We incorporate HED stain augmentation in the training of our models by randomly changing the concentration of the H&E stains in the source image. We first use the SNMF algorithm to extract the source stain matrix $\mathbf{S}$ and concentrations matrix $\mathbf{C}$ and then we scale and shift the stain concentrations and finally convert the altered stain information back to RGB space, thus attaining an augmented image $\hat{\mathbf{I}}$:

$$\hat{\mathbf{I}} = \mathbf{I_0} \exp\left(-\mathbf{S}\left(\alpha\mathbf{C} + \beta\right)\right), \tag{1}$$

where $\mathbf{I_0}$ is the incident intensity of the light source driven from the source image ($\mathbf{I}$), $\alpha \sim U(0.75, 1.25)$ and $\beta \sim U(-0.2, 0.2)$ are stain concentration scale and shift factors randomly selected from uniform distributions. It is important to note that we could simultaneously perform stain normalization to a target stain matrix and stain augmentation by setting the $\mathbf{S}$ matrix in Eq. (1) to a pre-extracted target stain matrix. In this work, we use the TIAToolbox (Pocock et al., 2022) implementation of both stain normalization and stain augmentation algorithms. Hereafter, by stain augmentation, we mean HED stain augmentation.

3.4.3. Self-supervised learning

While labeled mitosis datasets are scarce, unlabeled WSIs and histology images presenting vast stain and scan quality variations are abundant. Ideally, these vast datasets can be utilized to enhance performance on limited mitosis datasets. In this pursuit, self-supervised learning (SSL) algorithms have shown success in extracting relevant visual features from unlabeled data (Jing and Tian, 2020). This study explores the effects of three pretraining algorithms on the mitosis segmentation task using an unlabeled histology image dataset. Until now, SSL has mainly been used in simpler CPath tasks, like patch classification (Stacke et al., 2020; Vuong et al., 2022; Alemi Koohbanani et al., 2021).

To steer network learning towards histology-relevant features, we propose a self-supervised histology learning (SSHL) method to pretrain the entire segmentation network (both encoder and decoder). This method, inspired by Alemi Koohbanani et al. (2021), utilizes self-supervision for two histology-related tasks: (1) image magnification
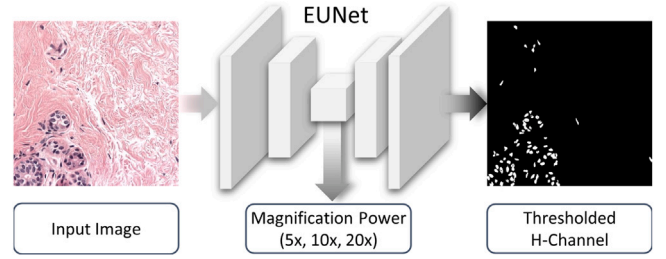

Fig. 5. Self-supervised Histology Learning (SSHL) for pretraining.

power prediction and (2) Hematoxylin channel (H-Channel) segmentation as depicted in Fig. 5. Here, two output branches are engaged, with the first branch leveraging magnification power labels for a cross-entropy loss function, and the second branch using on-the-fly generated Hematoxylin segmentation maps. These maps are obtained as follows: first, the H-channel of the input image ($\mathbf{H}$) is extracted using the Vahadane method, then a threshold $\tau = 0.7 \cdot p_{98} + p_2$ is calculated for the binary conversion to $\mathbf{B} = \mathbf{H} \leq \tau$ where p_{98} and p_2 are the 98th and 2nd percentiles of $\mathbf{H}$, respectively. The binarized H-Channel, $\mathbf{B}$, represents the Hematoxylin-rich areas in the image. Post binarization, morphological opening operations are performed on the resultant H-Channel to eliminate spurious small objects. The training procedure is similar to the one explained in Section 3.2.3, combining the loss functions of the classification and segmentation tasks:

$$\mathcal{L}_{SSHL} = \mathcal{L}_{Jaccard}\left(\mathbf{B}, \mathbf{B_p}\right) - \sum_{i=1}^{3} m_i \log\left(m'_i\right), \tag{2}$$

where $\mathbf{B_p}$, m'_i, and m_i are the predicted map of the binarized H-Channel, magnification power prediction and ground truth at one of 3 categories $\{5\times, 10\times, 20\times\}$, respectively.

For comparison, we additionally pretrain the segmentation model encoder on the same dataset using a self-supervised contrastive learning (SSCL) algorithm, SimCLR (Chen et al., 2020), and a supervised contrastive learning method, SCL (Khosla et al., 2020), predicting the magnification of the input image while augmenting it extensively. For all pertaining tasks, we attained more than 250,000 tiles of size 512×512 pixels extracted at three different levels of magnification ($5\times$, $10\times$, $20\times$) from the training set of the Camelyon16 dataset (Bejnordi et al., 2017).

3.5. Mitosis detection in WSIs

The standard mitotic count or score within a $2\,\text{mm}^2$ hotspot Region of Interest (ROI) serves as a proxy for overall mitotic activity throughout a WSI, due to the impracticality of manual counting across an entire tissue sample (Rakha et al., 2008). This approach, however, carries inherent subjectivity in ROI selection, impacting the final mitotic score. This subjectivity can be minimized by leveraging DP for mitosis detection across the entire WSI, necessitating an efficient method for accurately detecting mitotic figures within a reasonable time frame.

We thus propose a WSI processing pipeline where TIAToolbox's tissue segmentation model (Pocock et al., 2022) is used to identify the tissue region, which operates on image tiles of 1024×1024 pixels extracted at 8 microns-per-pixel resolution (on average, the model takes around 23 s to process a WSI). From the tissue region, tiles of 512×512 pixels in size with 50-pixel overlap are extracted at 0.25 microns-per-pixel resolution (approximately $40\times$ objective magnification). The proposed method then detects mitoses within these extracted patches. In this stage, the 'Detecting Fast' system is trained on down-scaled images (with a scaling factor of 0.75) to identify candidates (tiles are resized accordingly). The 'Detecting Slow' system refines candidates at full resolution (0.25 μm/pixel) following the MDFS method to ensure high-quality detection while minimizing processing time (refer

Section 4.3.1 and Table 1). Once mitosis detection is completed, the mitotic hotspot region can be deterministically identified through an overlapping window search across the WSI, selecting the window with the maximum mitosis count. Subsequently, the mitotic score for the hotspot can be computed as detailed below.

Clinically, particularly in breast cancer cases, a mitotic score (MS) is estimated from the Mitotic Count (MC) in a hotspot region exhibiting high tumor activity. Specifically, the Nottingham breast cancer grading system proposes counting mitoses in a 2 mm^2 region to derive the MS across three categories ($MS1, MS2, MS3$) (Ellis et al., 2005):

$$MS = \begin{cases} 1, & MC \leq 8 \\ 2, & 9 \leq MC \leq 16 \\ 3, & MC \geq 17 \end{cases} \tag{3}$$

4. Materials and experiments

4.1. Datasets

MIDOG. The MIDOG21 challenge dataset (Aubreville et al., 2022b), employed for tuning parameters and model evaluation, provides a training set with 150 labeled and 50 unlabeled cases (images covering 2 mm^2 regions), hosting 1721 mitotic figures and 2714 mimickers. Evaluation of the associated 80-case MIDOG21-test set is achieved by submitting to the challenge leaderboard, as the data is not publicly accessible. Notably, the MIDOG datasets, acquired via various scanners, exhibit domain shifts. This property informed our 3-fold *leave-one-domain-out* cross-validation strategy, ensuring each fold contained images from a distinct scanner. Furthermore, we train and validate the proposed MDFS method on MIDOG22 training set (Aubreville et al., 2022a) which contains 354 labeled images from canine lung cancer, human breast cancer, canine lymphoma, human neuroendocrine tumor, and canine cutaneous cast cell tumor (3-fold cross-validation on training domains with pooled subsets for model selection). The external test set for MIDOG22 contains 100 images from different tumor types that are not provided in the training set, making mitosis detection more challenging and requiring the MDFS to capture more generalizable features. The mitosis masks for both MIDOG21 and MIDOG22 datasets are obtained using the approach explained in Section 3.2.2.

TUPAC. The TUPAC dataset (Veta et al., 2019) offers a publicly accessible 73-case training set from three centers, hosting 1599 mitotic point annotations. A test set of 34 cases with no available labels exists for comparison with other methodologies (Akram et al., 2018; Kausar et al., 2020; Sebai et al., 2020; Li et al., 2019; Mahmood et al., 2020). Evaluation metrics for the final test set are derived by submitting results to the TUPAC challenge organizers.

ICPR2012. ICPR2012, a widely-cited mitosis dataset, is the sole dataset providing mitosis mask annotations for its five cases (Ludovic et al., 2013). Comprising 226 and 103 mitotic figures in training and test sets respectively, only the test set is used here for generalizability experiments.

4.2. Evaluation metrics

4.2.1. Detection

To avoid over-representation of cases with no mitoses in the evaluation metric, overall F1 scores (F1) were used to rank the different mitosis detection methods (Aubreville et al., 2022b; Veta et al., 2015). We also reported overall recall/sensitivity (Rec) and precision (Prc) metrics to compare different methods more thoroughly in all cross-validation and external tests on all datasets.

4.2.2. Mitotic score estimation

Mitotic score (MS) estimation in WSIs is crucial for accurate cancer grading, particularly in breast cancer, where inaccurate MS estimation can misinform treatment planning (Elston and Ellis, 1991). Although other cancer types possess unique scoring systems, exploring these is beyond the study's scope. However, understanding the error in breast cancer MS estimation is essential for this work.

While neither MIDOG nor TUPAC datasets provide ground truth MS information, all MIDOG images and 50 TUPAC images (cases 24–73) cover approximately 2 mm^2 sample area. Using available mitotic count (MC) data (from GT annotations), we estimate each image's expected MS via Eq. (3).

Since MS estimation is an ordinal regression task, it requires an appropriate evaluation metric. Quadratic Weighted Kappa (QWK) is commonly employed for similar tasks (Veta et al., 2019), but data population imbalance in available datasets and real-world data render it unsuitable for this task. We, therefore, propose a mitotic score error – an average category-based mean squared error – to evaluate algorithm performance on the MS estimation task:

$$\overline{ME} = \frac{1}{3} \sum_{s=1}^{3} \sum_{n \in \mathbf{T}_s} \frac{(\widehat{MS_n} - MS_n)^2}{N_s}, \tag{4}$$

where $\widehat{MS_n}$ and MS_n are GT and predicted mitotic scores for case n, respectively. $\mathbf{T}_s = \{n | \widehat{MS_n} = s\}$ and N_s are the set of all cases belonging to each mitotic score category $s \in \{1, 2, 3\}$ and their respective population. The squared error term emphasizes catastrophic prediction errors ($MS3$ predicted as $MS1$ or vice versa) and calculating the mean squared error for each category separately mitigates bias towards higher population categories.

4.3. Validation experiments

4.3.1. Internal cross-validation

The proposed MDFS method's performance was evaluated through cross-validation on the MIDOG21 and TUPAC training datasets. Results for the proposed EUNet segmentation model ('Detecting Fast' system) and the full pipeline (MDFS) are shown in Table 1 for the MIDOG21 dataset, with F1 scores of 0.754 and 0.785, respectively. Our method surpasses UNet (Ronneberger et al., 2015), RetinaNet (with ResNet-50 backbone) (Lin et al., 2017), and EfficientDet (with Efficient-Net-B4 backbone) (Tan et al., 2020) by 18%, 6%, and 5% in F1, respectively. Similar improvements are evident in both recall and precision metrics.

MDFS outperformed all SOTA mitosis detection methods when cross-validated on the TUPAC dataset, achieving an overall F1 of 0.767 (Table 2). This score outstrips the strongest reported TUPAC results by 5% (Li et al., 2019). Additionally, the proposed segmentation model, EUNet, demonstrated a high standalone detection accuracy.

It is important to mention that based on the results in Table 1, we select the scaling factor (Scl) of 0.75 for the rest of the validation experiments reported in this study (except for the results reported in ablation studies – Section 4.6). Please refer to Section 4.4 to find more details about the added value of the proposed MDFS method.

For mitotic score estimation, we assessed five algorithms on the MIDOG21 and TUPAC datasets. Confusion matrices and correlation plots for the mitotic count and mitotic score estimations are shown in Fig. 6, along with the proposed mitotic score error $\overline{ME}$ and Pearson's correlation coefficient (r). The MDFS method produced the lowest $\overline{ME}$ values of 0.183 and 0.066 for the MIDOG21 and TUPAC datasets, respectively, significantly outperforming the baseline (RetinaNet). Our method's predicted mitosis counts also exhibited a strong correlation with GT counts, with Pearson's r values of 0.97 and 0.98 for the MIDOG21 and TUPAC datasets, respectively.

The data in Fig. 6 shed light on ROI scoring performance. While baseline RetinaNet and our method achieve F1 scores of 0.720 and

Table 1
Results of internal cross-validation experiments on the MIDOG21 training set as well as the effect of down-scaling the image on the performance of the candidate segmentation and the full detection pipeline (Time of EUNet at segmentation scale (Scl) of 1 is the reference for Speed gain calculation). The reported time is the average ROI processing duration measured in seconds.

	Scl	F1	Rec	Prc	Time	Speed
UNet	1	0.601 ± 0.132	0.531 ± 0.154	0.693 ± 0.075	9.26	0.29×
RetinaNet	1	0.720 ± 0.122	0.726 ± 0.095	0.714 ± 0.110	9.88	0.28×
EfficientDet	0.5	0.581 ± 0.138	0.642 ± 0.122	0.531 ± 0.131	0.76	3.61×
	0.75	0.608 ± 0.134	0.691 ± 0.114	0.543 ± 0.126	1.71	1.60×
	1	0.726 ± 0.115	0.731 ± 0.107	0.722 ± 0.088	3.05	0.90×
EfficientDet+EfficientNet	0.5	0.612 ± 0.126	0.635 ± 0.118	0.592 ± 0.127	0.83	3.31×
	0.75	0.620 ± 0.108	0.581 ± 0.148	0.666 ± 0.095	1.78	1.54×
	1	0.739 ± 0.015	0.691 ± 0.110	0.795 ± 0.041	3.11	0.88×
EUNet	0.5	0.714 ± 0.006	0.714 ± 0.053	0.715 ± 0.053	**0.70**	**3.94×**
	0.75	0.740 ± 0.013	0.782 ± 0.039	0.704 ± 0.057	1.46	1.88×
	1	0.754 ± 0.033	**0.824** ± 0.035	0.695 ± 0.069	2.75	1.00×
MDFS	0.5	0.768 ± 0.011	0.744 ± 0.049	0.794 ± 0.035	0.76	3.59×
	0.75	0.781 ± 0.006	0.764 ± 0.025	0.799 ± 0.024	1.53	1.80×
	1	**0.785** ± 0.004	0.771 ± 0.020	**0.801** ± 0.021	2.80	0.98×

Table 2
Results of internal cross-validation experiments on the TUPAC training set.

Method	F1	Rec	Prc
SmallMitosis (Kausar et al., 2020)	0.599	**0.873**	0.456
EfficientDet (Tan et al., 2020)	0.608	0.682	0.549
Mahmood et al. (2020)	0.641	0.642	0.641
MaskMitosis (Sebai et al., 2020)	0.660	0.689	0.633
Akram et al. (2018)	0.690	0.661	**0.722**
SegMitos (Li et al., 2019)	0.717	–	–
EUNet	0.738 ± 0.051	0.852 ± 0.064	0.651 ± 0.033
MDFS (proposed)	**0.767** ± 0.023	0.833 ± 0.016	0.712 ± 0.030

Table 3
Results of external validation experiments on the MIDOG21 test set for the baseline and the top seven algorithms from the MIDOG21 leaderboard.

Method	Details	F1	Rec	Prc
RetinaNet	Baseline detection with stain normalization and data aug.	0.698	0.699	0.696
Razavi et al. (2021)	Cascade R-CNN	0.706	0.706	0.708
Liang et al. (2021)	Fused Detector and Deep Ensemble Classification	0.706	0.686	0.727
Wilm et al. (2021b)	MIDOG reference algorithm: RetinaNet with Domain Adversarial branch	0.710	0.731	0.690
Chung et al. (2021)	RetinaNet-101 with Style transfer augmentation	0.724	0.678	**0.776**
Fick et al. (2021)	Mask R-CNN + ensemble of ResNet and DenseNet	0.736	0.709	0.764
Yang et al. (2021)	Sk-UNet Model with Fourier Domain for Mitosis Detection	**0.747**	0.741	0.753
MDFS (proposed)	EUNet (Detecting Fast) + EfficientNetB7 (Detecting Slow)	**0.747**	**0.762**	0.733

0.785 respectively on the MIDOG21 dataset, the difference in correlation scores is only 2%—the standard metric for ROI-based mitosis assessment (Veta et al., 2019, 2015, 2016). In contrast, our proposed mitotic score error produced $\overline{ME}$ values of 0.351 and 0.183, respectively, emphasizing its efficacy for mitotic score estimation by showing more differentiation. This difference is further highlighted by comparing three methods (Point-EUNet, Mask-EUNet, and MDFS) on the TUPAC dataset in Fig. 6b. Despite achieving identical correlation coefficients (0.98), the $\overline{ME}$ metric differentiates the methods' performance with values of 0.15, 0.13, and 0.07, respectively. These results underscore the proposed metric's ability to accurately evaluate mitotic score estimation performance, suggesting its suitability for future comparative studies.

4.3.2. External validation

Our proposed method was tested on 34 TUPAC dataset images and results were submitted for evaluation. The external validation results are detailed in Table 4, comparing our proposed EUNet model and MDFS pipeline performance against the 11 leading methods in TUPAC's 2016 Task 3 (mitosis detection) challenge. Only three out of these methods have detailed algorithm descriptions (Radboud (Tellez et al., 2018), Warwick (Akram et al., 2018), and SegMitos (Li et al., 2019)), with the rest summarized in the TUPAC challenge paper (Veta et al., 2019). TUPAC challenge participants reported solely overall F1 values.

When evaluated on the TUPAC test set (without external data), our method outperforms all other methods (achieving an F1 of 0.675). This underscores the efficacy of our 'Detecting Slow' system where the F1 improves by 5% primarily by boosting precision upon the addition of a deep classifier atop our EUNet candidate segmentation model. This enhancement is made at a slight computational speed sacrifice (Table 1, Section 4.3.1). Hence, our method outperforms the challenge winner, Lunit, which managed an F1 of 0.652. Subsequent post-challenge submissions, specifically the SegMitos method (Li et al., 2019), yield comparable results to ours (F1 = 0.669). Nevertheless, our method significantly surpasses all other techniques, which typically merge patch classifiers with hard-negative mining.

The MIDOG test set comprises 80 images from diverse breast tumor instances scanned using four different scanners, two of which were utilized for the MIDOG training set acquisition. Methods are evaluated by submission to the challenge platform, similar to TUPAC. Our method's results, alongside the top eight performing methods in the MIDOG21 challenge's final testing, are displayed in Table 3. Our algorithm ranks top (tied with Yang et al. (2021)), winning the challenge with an F1 of 0.747 and the highest recall of 0.762 (precision of 0.733). It proves superior to all region proposal-based algorithms such as bounding box detection algorithms (Lin et al., 2017; Razavi et al., 2021; Liang et al., 2021; Wilm et al., 2021b; Chung et al., 2021) and Mask-RCNNs (Fick et al., 2021), raising the detection F1 score by roughly 5% compared
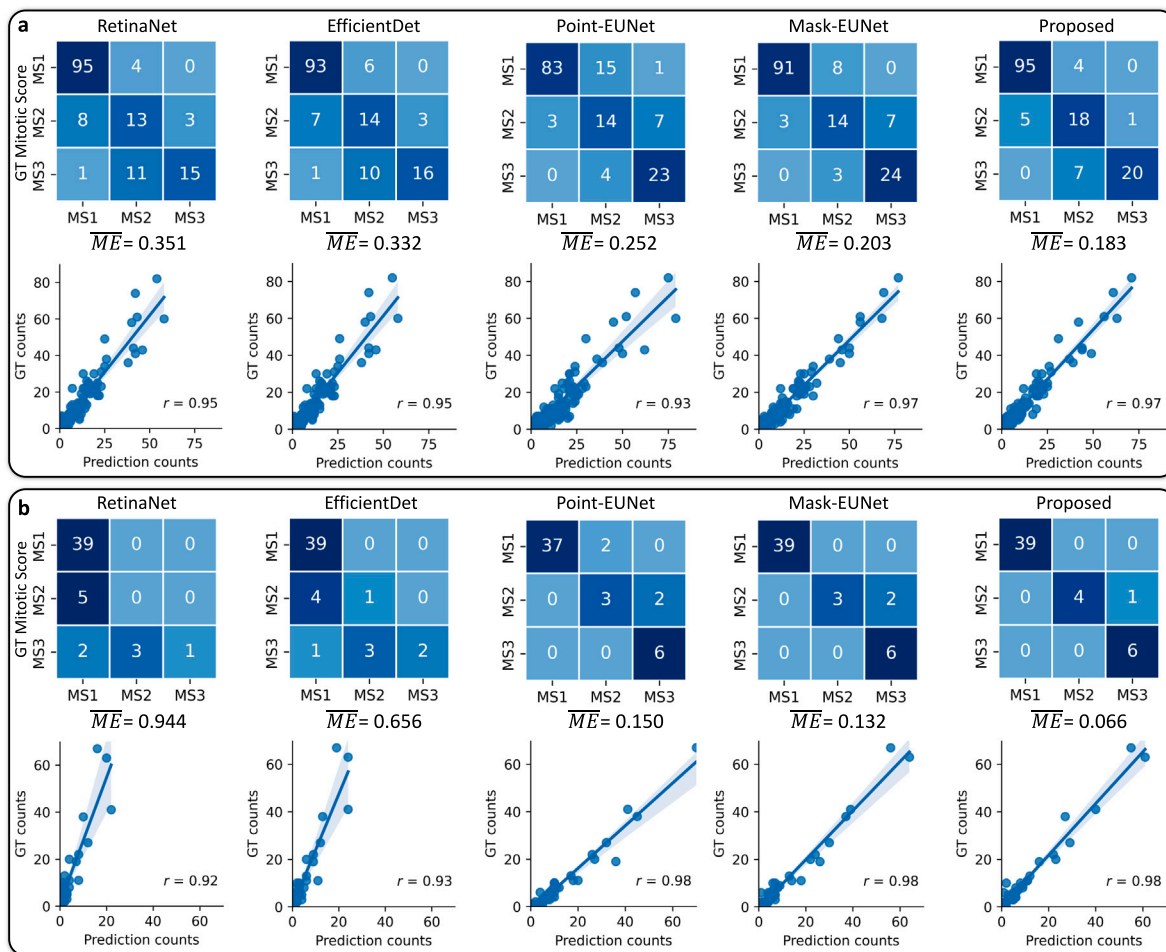
Fig. 6. Confusion matrices for mitosis-score estimations (MS1, MS2, MS3) and mitosis score error ($\overline{ME}$), with correlation plots and Pearson's correlation coefficient, are reported for mitosis-count predictions from five different methods when evaluated on the MIDOG21 (a) and TUPAC (b) datasets.
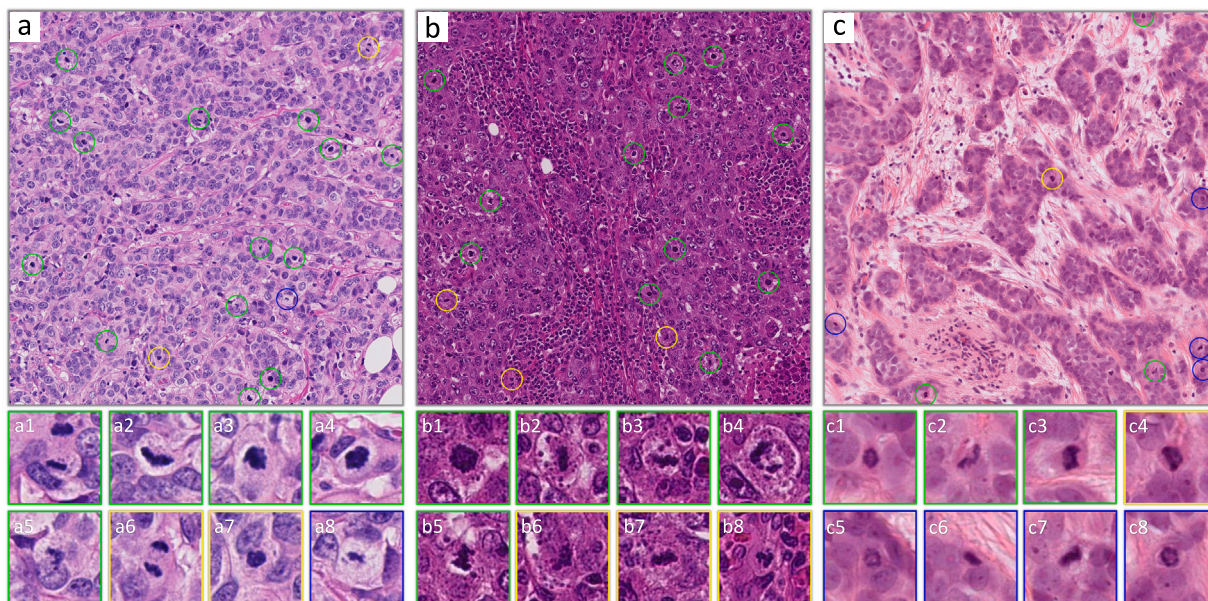

Fig. 7. Mitosis detection results of the proposed method on three different images (images in panels 'a,c' from MIDOG21, and panel 'b' from TUPAC dataset in the top) with the zoomed-in patches of some of the detected mitoses/mimickers in them (panels a#,b#, and c# in the bottom). Circles or patch borders of the color green, blue, or yellow indicate true positive, false negative, and false positive predictions concerning ground truth annotations, respectively.

Table 4
Results of external validation experiments on the TUPAC test set.

Methods	Details	F1	Rec	Prc
Heidelberg[a]	Combine residual networks with Hough voting and hard-negative mining	0.481	–	–
Tellez et al. (2018)	Multi-CNN patch classifiers – Use private auxiliary data	0.541	–	–
PIEAS[a]	CNN patch classifier using auxiliary ICPR2012/2014 datasets	0.571	–	–
Microsoft research[a]	Shallow CNN patch classifier using auxiliary ICPR2012/2014 datasets	0.596	–	–
Contextvision[a]	CNN pixel classifier similar to Cireşan et al. (2013) and hard-negative mining	0.616	–	–
CUHK[a]	Custom CNN patch classifier using auxiliary ICPR2012/2014 datasets	0.620	–	–
HUST[a]	Custom CNN patch classifier with hard-negative mining	0.626	–	–
Akram et al. (2018)	ResNet patch classifier trained on TUPAC+external dataset	0.640	0.671	0.613
IBM research[a]	Patch classification using Wide residual network and hard negative mining	0.648	–	–
Lunit[a]	ResNet patch classifier with hard-negative mining	0.652	–	–
Li et al. (2019)	Weakly (point) supervised mitosis segmentation concentric loss	0.669	0.700	0.640
EUNet	Only segmentation part (Detecting Fast)	0.629	0.794	0.521
MDFS (proposed)	EUNet (Detecting Fast) + EfficientNetB7 (Detecting Slow)	0.675	0.770	0.600

[a] No further information/citation is found for these entries except for the provided details.

to the RetinaNet baseline. Notably, other two-stage methods that used a bounding box detector for candidate detection (like Liang et al. (2021) in Table 3 and Nateghi and Pourakpour (2021) where authors used a combination Faster R-CNN and EfficientNet-B0 and achieved F1 of 0.676) performed considerably worse than the proposed method that uses EUNet (a segmentation model) in the first stage. This is in line with our finding in internal cross-validation experiments in Table 1 where the combination of EUNet and EfficientNet (MDFS) worked considerably better than the combination of EfficientDet and EfficientNet.

The bar chart of F1-scores of top methods over the cases from the four scanners (domains) in the test set is presented in Fig. 8 to examine the method's performance across different domains. Only images from 'Scanner A' are used in training, with the other three classified as 'out-of-domain' scanners. Our method consistently performs well across all scanners, especially 'Scanner A' and 'Scanner E' (achieving F1 of 0.837 and 0.808 respectively). With an F1 of 0.677, 'Scanner D' results fall outside the top 3 performers. Conversely, Yang et al. (2021) achieved F1 of 0.726 potentially due to their unique Fourier domain-mixing algorithm, incorporating unlabeled 'Scanner D' images into training. Finally, we computed the averaged domain-level F1 score by independently averaging the F1 for each scanner. This yielded a superior domain-level F1 of 0.747 ± 0.08 across all scanners, compared to Yang et al. (2021) (0.734 ± 0.07) and other methods. The averaged domain-level F1 treats each scanner equally, whereas the overall F1 introduced in Section 4.2.1 could be influenced by scanners with more images or mitotic instances.

4.3.3. Independent cross-validation

Finally, we aim to demonstrate the generalizability of the proposed model on external data from different dataset sources. To this end, we train our MDFS method on either TUPAC or MIDOG21 data, before testing it on the training set of the other dataset. In addition, we report the results of TUPAC and MIDOD21 models on the ICPR dataset. The images of the test sets in this experiment are not only from different sources but also have been annotated with different annotation protocols. The results for these experiments are reported in Table 5. We observe that the model trained on the TUPAC data shows good generalizability to other datasets, achieving F1 scores in the range of 0.758 and 0.745 on the MIDOG training set and ICPR2012 test set, respectively. Similarly, the model trained on the MIDOG dataset also demonstrates competitive performance with F1 scores, especially on the ICPR2012 test set. It is important to note that due to variations in annotation protocols (posterior shift) and potential prior shifts (variations in the distribution of labels in different datasets), we refrain from making direct comparisons regarding the superiority of one model over the other.
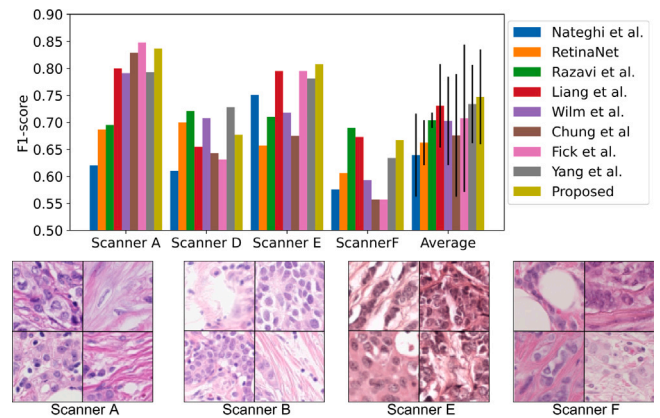

Fig. 8. Results of the top performing methods of the MIDOG21 challenge on images from different scanners.

Table 5
Independent cross-validation experiments where the model is trained on the 'Source Dataset' and tested on the 'Target Dataset'.

Source dataset	Target dataset	F1	Rec	Prc
TUPAC	MIDOG21	0.758	0.714	0.807
MIDOG21	TUPAC	0.697	0.708	0.686
MIDOG21	ICPR2012	0.736	0.685	0.796
TUPAC	ICPR2012	0.745	0.695	0.803

4.4. Added value of 'Detecting Fast' and 'Detecting Slow'

The 'Detecting Fast' system's efficacy in our MDFS framework was assessed by testing various down-sampling scales for the resizing module preceding the mitosis candidate segmentor (Fig. 2b). Stain-normalized images were used without stain augmentation for these experiments to prevent potential random effects due to data variability, ensuring fair comparisons with RetinaNet and EfficientDet. Consider the 0.75 down-scaling ratio from Table 1 for an example, although performance metrics seem reduced (F1 = 0.740) compared to using full image resolution (F1 = 0.754), the F1 is only 0.004 lower after applying 'Detecting Slow' (i.e., MDFS pipeline) (0.785 for full resolution and 0.781 when down-scaled by 0.75), while being about 1.8 times faster. We use the EUNet algorithm at full resolution (scale = 1) as a time baseline and indicate other algorithms' speed gain relative to it in Table 1. The proposed EUNet and full pipeline outpace EfficientDet and RetinaNet bounding box detection models, although deeper encoders may provide better F1 scores for bounding box detection models at increased computation time.

The enhancement offered by a second-stage candidate classifier extends beyond just EUNet. As demonstrated in Table 1, integrating
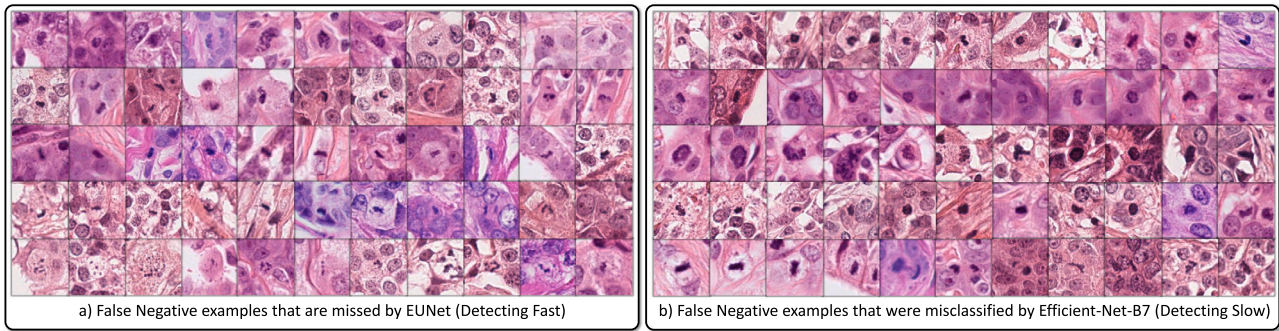
a) False Negative examples that are missed by EUNet (Detecting Fast)

b) False Negative examples that were misclassified by Efficient-Net-B7 (Detecting Slow)

Fig. 9. Challenging mitotic figures that were missed by MDFS framework either in candidate segmentation phase (a) or candidate refinement phase (b). Images are collected from the results on the MIDOG21 dataset.

the EfficientNet classifier as the secondary mitosis refinement to the EfficientDet candidate detector ('EfficientDet+EfficientNet') results in improved mitosis detection performance across all scales compared to using only EfficientDet. Nevertheless, it appears that EfficientDet might not serve as an optimal choice for the initial detection stage when compared to EUNet. EfficientDet performance notably declines with lower-resolution images (from 0.726 to 0.581 in F1, over 14% decrease in performance), preventing the realization of both speed enhancement and sustained high accuracy in a two-stage pipeline, a capability that is very well realized by EUNet.

Moreover, comparing the running time of the EUNet (with approximately 11M parameters) and speed gain with the full MDFS pipeline at each scale reveals that adding 'Detecting Slow' atop 'Detecting Fast' incurs minimal computational overhead (for instance, running time only increases about 50 ms for scale 1) while significantly enhancing detection F1 (about 5% for scale 0.5). This is primarily because the proposed pipeline's second system only processes a lower number of small patches of mitosis candidates although this system uses a deeper and slower network (with approximately 65M parameters). Benchmarking experiments were run on a Nvidia DGX-2 device with one Tesla V100 GPU.

4.5. Qualitative assessment

In our study, we also conduct a qualitative evaluation of our mitosis detection approach. We randomly select three images, images a and c from the MIDOG dataset and image b from the TUPAC dataset, for in-depth analysis (Fig. 7). In these larger images, TP, FP, and FN detections are denoted by green, yellow, and blue circles, respectively. We pick eight unique detections from each image for detailed visualization. We randomly choose TPs and display all FPs and FNs beneath the associated larger image, color-coding the detection boundaries to match the circles above.

Our examination of Fig. 7 revealed inconsistencies in mitotic figure annotations in both TUPAC and MIDOG21 datasets. We confirmed these inconsistencies by consulting pathologists who evaluated the selected detections (highlighted below the larger images in Fig. 7). Though their expert opinion was sought, we acknowledge mitosis interpretation is not definitive and can vary among pathologists (Veta et al., 2016; Ibrahim et al., 2022; Alkhasawneh et al., 2015). We mitigated bias by blinding the pathologists to the patches' association with our algorithm's detections and providing full images for context. The pathologists affirmed that all patches, excluding c6, contained mitotic figures. This indicates that the original annotators may have overlooked six mitoses in these ROIs. Interestingly, our pathologists identified an additional mitotic figure missed by both the original annotators and our algorithm. However, their validation was limited to typical examples that are usually recognizable by breast pathologists.

Our observations corroborate existing literature on the difficulties of mitotic figure detection and inter-observer variability implications (Veta et al., 2016; Ibrahim et al., 2022; Alkhasawneh et al.,

2015; Saldanha et al., 2020; Bertram et al., 2020; Molenaar et al., 2000; Lashen et al., 2021; Ibrahim et al., 2023). They also underscored the need for continuous refinement of annotation process (by using PHH3 IHC staining (Alkhasawneh et al., 2015; Ibrahim et al., 2023; Tellez et al., 2018), using well-established annotations protocols/guidelines (Lashen et al., 2021; Ibrahim et al., 2022), or consensus of pathologists observations (Wilm et al., 2021a; Bertram et al., 2019)) and detection algorithms to account for the inherent limitations and complexities of datasets. This nuanced understanding informs the development and evaluation of mitotic detection methodologies, emphasizing both the strengths and potential areas of improvement.

Despite the decrease in the false positive rate for the EUNet segmentation model after the hard-negative mining phase (see Tables 1 and 2 where MDFS precision increases over EUNet), there remains a risk of discarding true positives during classification. In Fig. 9, we present 110 false negative figures overlooked by EUNet in the 'Detecting Fast' system or misclassified by the EfficientNet-B7 model in the 'Detecting Slow' system. Most missed samples in Fig. 9 are small, particularly those overlooked in the first phase (fourth row of Fig. 9a). Also, anaphase mitotic figures with distantly spaced daughter cells pose challenges for segmentation and classification tasks (second row of Fig. 9a). Some mitotic figures resemble inflammatory cells and are thus misclassified in the second phase (first row of Fig. 9b), and some atypical mitotic figures are even difficult for the 'Detecting Slow' system (third row of Fig. 9b).

4.6. Ablation studies

To optimize our framework and assess its different aspects, we conducted ablation studies using the MIDOG21 training set and 3-fold cross-validation as detailed in Section 4.1. In these experiments, the scaling factor in the 'Detection Fast' system was set to 1 to fully utilize the data.

4.6.1. Supervision for EUNet

We trained our EUNet alongside RetinaNet (Lin et al., 2017) and EfficientDet (Tan et al., 2020) on candidate detection, using dilated point annotations as ground truth masks (Point-EUNet). It achieved an F1 score, recall, and precision of 0.731, 0.775, and 0.693, respectively. This lets us compare supervisory signal performance. The EUNet trained on mitosis masks (Mask-EUNet) attained an F1 of 0.754, significantly outperforming both bounding box detection methods (see Table 1). Intriguingly, Point-EUNet outperformed both bounding box models but was still inferior to Mask-EUNet. We propose that using a segmentation model for mitosis candidate detection is a suitable approach.

Table 6
Effect of using different pretraining methods on the performance of the mitosis candidate segmentation model.

Method	Pretrain	F1	Rec	Prc
ImageNet	Encoder	0.722	0.760	0.687
SimCLR	Encoder	0.716	0.735	0.698
SCL	Encoder	0.716	0.703	**0.729**
Proposed-SSHL	Encoder–Decoder	**0.741**	**0.807**	0.685

Table 7
Effect of using different combinations of domain generalization techniques on the mitosis candidate segmentation model's performance.

Ex.	Pretraining	SN	SA	F1	Rec	Prc
1	ImageNet	✗	✗	0.722	0.760	0.687
2	ImageNet	✗	✓	0.765	0.807	0.728
3	ImageNet	✓	✗	0.754	**0.824**	0.695
4	ImageNet	✓	✓	**0.773**	0.805	**0.744**
5	SSHL	✗	✗	0.741	0.807	0.685
6	SSHL	✗	✓	0.758	0.788	0.730
7	SSHL	✓	✗	0.748	0.775	0.724
8	SSHL	✓	✓	0.762	0.805	0.724

4.6.2. Effect of domain generalization techniques

We examine three different techniques to tackle the domain-shift issue in histology images: stain normalization (SN), stain augmentation (SA), and encoder pre-training (see Section 3.4). Firstly, we test our candidate segmentation model pretrained with various self-supervised learning methods on the MIDOG21 dataset. The methods include *ImageNet*, SimCLR (Chen et al., 2020), and SCL (Khosla et al., 2020) for the model encoder, and the proposed SSHL method for both the encoder and decoder. No other domain generalization techniques are used here to ensure optimal pretraining method selection. As shown in Table 6, SSHL notably outperforms both SimCLR and SCL pretraining algorithms by 2.5%, achieving an F1 of 0.741. Thus, we only consider the SSHL pretraining method in subsequent domain generalization investigation experiments (ImageNet weights are also included as the standard approach).

In Table 7, we present the effect of SN, SA, and pretraining techniques on the mitosis candidate segmentation task. As expected, without using any of the proposed techniques, the model performs the worst (Ex.1 with F1 of 0.722). With ImageNet pretrained weights, the model performance consistently improved with the addition of SN and SA. A similar pattern was observed with SSHL pretraining, though SN and SA had less impact on the final F1. Interestingly, when combined with SN and SA, ImageNet pretrained weights outperformed the SSHL pretrained model (Ex.4 with F1 of 0.773). These results suggest that SSL techniques are beneficial for introducing domain-invariance to mitotic segmentation models. However, when combined with SN and SA techniques, they might be unnecessary. Thus, we used ImageNet weights for encoder pretraining and a combination of SN and SA for training and inference on benchmark datasets. Note that SN was excluded during WSI inference to reduce computational load.

4.7. Adaptation to other tissues and species

The diversity of tissue types and species in the MIDOG22 dataset provides an ideal test bed for the generalization capability of our MDFS algorithm. However, the MIDOG22 challenge's main track prohibits the use of external resources, precluding the integration of NuClick for mitosis mask generation as discussed in Section 3.2.2. Instead, we employ Point-EUNet (utilizing dilated points as mitosis GT, see Section 4.6.1) within the 'Detecting Fast' system, which can be trained using the original point annotation. Furthermore, unlike the MIDOG21 challenge, we opted not to use stain normalization in MIDOG22. We conducted cross-validation experiments on the MIDOG22 training set, and to evaluate MDFS on the external MIDOG22 test set, we submitted

Table 8
Cross-validation and external test results for mitosis detection on MIDOG22 dataset.

Method	Set	F1	Rec	Prc
Yang et al. (2021)	Test	0.658	0.528	**0.875**
Wilm et al. (2021b)[a]	Test	0.713	0.663	0.771
Kotte et al. (2022)	Test	0.751	**0.764**	0.738
Saipradeep et al. (2022)	Test	0.756	0.754	0.759
MDFS (proposed)	Test	**0.764**	0.712	0.823
MDFS (proposed)	Train	0.816	0.831	0.801

[a] This is MIDOG reference algorithm which did not compete in the challenge.

our algorithm to the challenge. Table 8 reports our cross-validation results and the top-performing methods on the MIDOG22 test set. Remarkably, MDFS secured the 1st rank on the final test leaderboard.[2]

Considering that the MIDOG22 test set comprises 100 cases from 10 different unseen tumor types (human melanoma, human astrocytoma, human bladder carcinoma, canine breast cancer, canine cutaneous mast cell tumor, human meningioma, human colon carcinoma, canine hemangiosarcoma, feline soft tissue sarcoma, and feline lymphoma), the high F1 score of 0.764 achieved by the MDFS method demonstrates its robustness and adaptability to domain shifts caused by varying scanners, labs, species, and tumor types. Notably, MDFS outperformed SOTA attention-based transformer models (Kotte et al., 2022; Saipradeep et al., 2022), reaffirming the superiority of our proposed method based on mitosis segmentation over standard bounding box detection models for mitosis detection.

4.8. Large-scale mitosis detection on TCGA WSIs

To showcase the capability of our proposed method for the efficient processing of WSIs, we processed the entire breast cohort of the TCGA dataset (TCGA-BRCA) with an improved version of the MIDOG22 model (see Appendix B). Over 620K mitotic figures were detected in 1124 WSIs, with the candidate segmentation and refinement parts of the algorithm requiring around 2.5 (±6) min and 6.4 (±4.5) s, respectively. Thus, each slide was processed in under 3 min. Given its high efficiency in processing large WSIs and its robustness to scanner-induced variations, our algorithm is a viable tool for research uses that demand WSI-level mitosis detection.

We have made the output of our mitosis detection algorithm publicly accessible for research purposes at https://zenodo.org/records/10245707 to facilitate mitosis-related down-stream tasks, such as biomarker discovery and survival prediction for breast cancer. The 'TCGA-BRCA Mitosis Dataset' also includes mitotic hotspot regions, hotspot mitotic counts, and hotspot mitotic scores. For more details, refer to Appendix B. While strong correlations between pathologist and algorithmic mitotic counts were observed in hotspot regions of 757 cases ($r = 0.8$, $p<0.001$), we acknowledge the necessity to explicitly mention that all detections were not reviewed nor revised, cautioning users regarding downstream analysis due to potential errors inherent in the dataset.

5. Discussion

Various algorithms for automatic mitosis detection have been proposed (Mathew et al., 2021), with many aimed at enhancing detection in mitotic hotspots or regions of interest (ROIs). This paper outlines an efficient, generalizable algorithm for mitosis detection, designed to be resilient to domain shifts caused by scanner variability, cancer types, or species. This two-stage algorithm initially segments lower-resolution

---

[2] MIDOG 2022 challenge leaderboard is accessible at https://midog2022.grand-challenge.org/.

mitosis candidates and then refines them at a higher resolution for improved speed and accuracy. We introduce a new metric, mitotic score error $\overline{ME}$, to better assess the performance of mitotic score estimation methods.

5.1. Generalizability of MDFS

Our proposed method's robustness is evaluated through external validation experiments (Sections 4.3.1, 4.3.2 and 4.7). These trials highlight the MDFS's adaptability to unseen domains, including those arising from variations in staining, scanner use, tissue types, case species, or annotation protocols. Our method achieved first place in both the MIDOG2021 and MIDOG22 grand challenges on mitosis domain generalization and outperformed all other techniques on the TUPAC-mitosis challenge leaderboard.

Notably, the MIDOG21 dataset's test set comprises images from four different scanners. Our algorithm achieves an F1 of 0.837 when tested on the 'Scanner A' subset (Fig. 8)—a 5% improvement over internal cross-validation experiments. We attribute this high generalizability to the use of domain generalization techniques (Section 3.4) and model design. Future research should continue utilizing encoder–decoder models and stain-augmentation techniques to combat domain shifts in mitosis detection. While we did not observe any added value in pretraining the segmentation model for our current work, it could potentially prove beneficial when handling small-scale datasets. Furthermore, despite potential advantages, we caution against the use of stain normalization due to potential inconsistencies, particularly when original images contain staining components other than Hematoxylin and Eosin Vu et al. (2022).

5.2. Efficiency of MDFS

The proposed algorithm outperforms other region-proposal-based techniques in processing 2 mm^2 sample images (Section 4.3.1). This efficiency owes to the method's dual detection systems. The 'Detecting Fast' system quickly identifies potential mitotic figures in downscaled images, while the 'Detecting Slow' system accurately classifies candidates using full-resolution patches. Despite the initial loss of high-resolution information due to down-sampling, our two-step approach regains much of the lost performance (Table 1). The 'Detecting Slow' system can improve the F1 score by 3%–5%, while minimally impacting computational cost (Section 4.4).

As seen in Table 1, the proposed method maintains mitosis detection accuracy using down-scaled images in the 'Detecting Fast' system, while significantly enhancing speed. With a down-sampling scale of 0.75, the proposed method sustains an F1 value above 0.78, reducing 2 mm^2 ROI processing time from 2.8 s to 1.53 s. On average, WSIs are processed in about 3 min. This demonstrates the MDFS system's practical efficiency and the generalizability of EUNet with lower-magnification images. Contrarily, methods such as EfficientDet markedly deteriorate in performance at lower resolutions and show sensitivity to threshold selection (as illustrated in Fig. A.1 of Appendix A), rendering it ill-suited for employment in the 'Detecting Fast' system.

5.3. Limitations and future work

As mentioned in Section 4.5, one of the main drawbacks of the MDFS method is missing very small/faint or inflammatory-like mitoses. It appears, however, many of the false negatives (FNs) in Fig. 9 are mislabeled mimickers, not actual mitoses. For example, dissolved nuclear materials in the samples of bottom row in Fig. 9a exhibit traits of dead or dying cells, which are called Karyorrhectic cells (Ibrahim et al., 2022). In general, Ibrahim et al. (2022) proposed guidelines on recognizing common mimickers in breast cancer listed as apoptotic bodies, tissue artifact (pigmentation), hyperchromatic malignant cells, foamy macrophages, karyorrhectic cells, and out-of-focus lymphocytes/fibroblast cells which are challenging for both pathologists and AI to be distinguished.

An additional issue is the considerable number of true positive candidates identified by the 'Detecting Fast' system pruned during refinement. Thus, a highly sensitive and specific classifier is necessary. Ensembling multiple classifiers has shown efficacy in enhancing classifier performance (Liang et al., 2021; Kotte et al., 2022), although at the cost of computational power and decreased algorithmic speed. Essentially, for optimal functionality of 'Detecting Fast and Slow', both segmentation and classification components must perform well.

A crucial facet of automated mitosis detection is determining its potential to enhance survival prediction, as patient prognosis is ultimately the aim of mitotic counting in slides. Given that our algorithm can perform on par with experienced pathologists on mitosis detection in ROIs (Aubreville et al., 2022b), AI-assisted mitotic scoring is anticipated to improve survival prediction accuracy. Nonetheless, this topic and evaluation of mitosis detection accuracy on the WSI level have not been probed yet and exceed the current study's scope. Further, it would be insightful to scrutinize the performance of our proposed mitosis detection algorithm across other cancer types. Automation of mitosis detection in such instances could contribute to developing objective prognostic measures for patients.

6. Conclusions

This paper presented a two-stage algorithm for effective mitosis detection in breast histology images and WSIs, involving an initial 'Detecting Fast' phase to segment mitotic candidates, followed by a 'Detecting Slow' phase to refine these candidates using a deeper CNN. We introduced the EUNet model for mitosis segmentation and utilized EfficientNet (Tan and Le, 2019) for candidate classification. We demonstrated that the 'Detecting Fast' phase could employ lower-resolution images to substantially boost algorithm speed without compromising accuracy. Our investigation into the effect of three domain generalization techniques on the mitosis detection task indicated that a combination of stain normalization and augmentation techniques yielded optimal results. Self-supervised pretraining of the encoder model, even with a novel preprocessing method capable of joint encoder and decoder pretraining, was found to be unnecessary with our method and mid-to large-scale datasets. Our approach outperformed all other SOTA methods for mitosis detection on MIDOG and TUPAC datasets, known for significant domain shifts. This performance advantage is evident in terms of both traditional detection metrics and the recently proposed mitosis score error, $\overline{ME}$, which assesses the mitosis detection model's performance on mitosis score estimation in hotspot regions.

Our algorithm secured first places in the MIDOG 2021 and 2022 mitosis detection challenges and outperformed all other methods on the TUPAC dataset. Furthermore, we processed 1124 WSIs of the TCGA-BRCA cohort using our efficient method and generated over 620 K mitotic figures. This dataset, along with the mitosis masks produced for the TUPAC and MIDOG datasets, is made publicly available to aid in the development of mitosis detection models and mitosis-based survival analysis for breast cancer.

CRediT authorship contribution statement

Mostafa Jahanifar: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Adam Shephard:** Writing – review & editing, Software, Methodology, Data curation. **Neda Zamanitajeddin:** Writing – review & editing, Software, Methodology, Data curation. **Simon Graham:** Writing – review & editing, Methodology. **Shan E. Ahmed Raza:** Writing – review & editing, Methodology. **Fayyaz Minhas:** Writing – review & editing, Methodology. **Nasir Rajpoot:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.
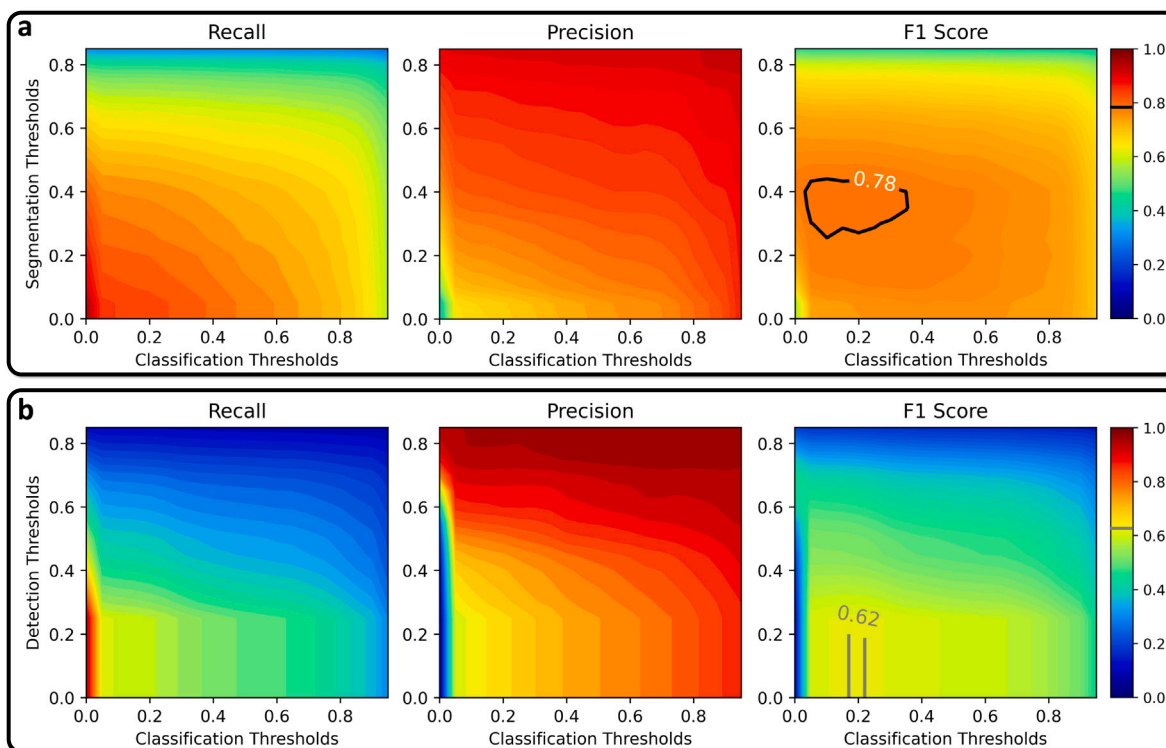
Fig. A.1. Threshold analysis experiment on MIDOG training set using EUNet (a) or EfficientDet (b) as the candidate detector in the proposed two-stage detection pipeline. Recall, Precision, and F1 values are highlighted against various selections of segmentation (detection) and classification thresholds during the post-processing step. Black and gray contours show peak performance region.

Declaration of competing interest

Data availability

The data used in this research is publicly available and links/publications related to them are cited in the manuscript.

Acknowledgments

Appendix A. Threshold selection

The post-processing design is impactful on the final output and should be carefully designed. The simple segmentation and classification design of our model allows the incorporation of a fast post-processing method (Section 3.2.4). There are two main parameters in the proposed post-processing and the entire prediction pipeline, which are segmentation and classification thresholds. We set these parameters by conducting a grid search on a range of thresholds for both tasks. In particular, we sweep the segmentation threshold between 0 and 0.85, and for each segmentation threshold, we test a range of classification thresholds from 0 to 0.95 (both inclusive). For each threshold configuration, we calculate the F1, Recall, and Precision of the results against the GT and plot the results as surface maps in Fig. A.1. These experiments are done using an image scale of 0.75 in the first stage of detection. The best-performing point is chosen based on the F1 map, where a good combination of recall and precision leads to high values of the F1.

Through careful examination of the heatmaps Fig. A.1 one can observe the complex trade-offs between sensitivity and specificity in the initial stage of our two-stage detection process. High detection thresholds in the first stage yield higher specificity but lower sensitivity, resulting in fewer but more accurate candidate detections. However, this stringent selection might miss some true positive cases. By contrast, lower detection thresholds increase sensitivity, at the risk of passing more false positives to the subsequent classification stage. To compensate for the lower specificity in the first stage, the second stage employs a range of classification thresholds. With the help of the classification heatmap, it is possible to select an optimal classification threshold that best balances precision and recall, effectively refining the set of candidates passed on from the first stage. This allows us to maintain an excellent F1 score, despite the trade-offs made in the initial stage. Therefore, by appropriately tuning the thresholds in both stages, we can achieve efficient and robust performance in mitotic figure detection.

A benefit of our proposed method is that the resulting F1 map (shown in Fig. A.1a) is almost flat for all spans of segmentation and classification thresholds and this shows the robustness of the proposed method against threshold selection. In particular, the peak performance (where the F1 is higher than 0.78 and highlighted by a black contour in Fig. A.1a) can be achieved by selecting the segmentation threshold in the range of [0.3, 0.45] and the classification threshold anywhere
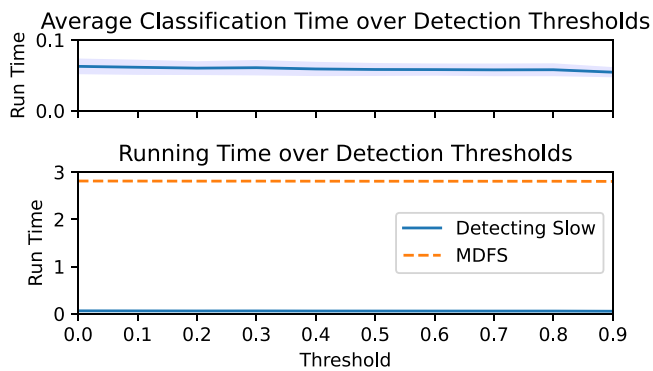
Fig. A.2. Analysis of 'Detecting Slow' and 'MDFS' run times (in seconds) over varying detection thresholds. The upper plot shows the average classification time for 'Detecting Slow', with shading indicating standard deviation. The lower plot juxtaposes running times for both systems. The *x*-axis is similar in both plots and represents thresholds used in the 'Detecting Fast' system.

between $[0.1, 0.4]$ which is indicative of the robustness of our method against threshold selection. That said, our approach does allow for flexibility. If a user seeks to optimize for sensitivity (at the cost of precision), they can select lower thresholds. Conversely, if precision is the priority (at the cost of sensitivity), higher thresholds can be selected. Our method's distinctive dual advantage of robustness and flexibility in threshold selection, demonstrated by consistent performance across varying thresholds yet allowing adjustments for a desired balance of sensitivity and precision, underscores its potential to be a practical tool for mitotic figure detection. Please note that this figure is related to the experiment where we used stain-normalized images and trained our models with the stain-augmentation technique. For other experiments (using different image scales in the first stage of MDFS), these maps might be slightly different but it has been seen that the landscape of F1 values in relation to segmentation and classification thresholds is almost always flat with different variations of our model. Nonetheless, for each cross-validation experiment, we repeated a similar threshold selection procedure to select the segmentation and classification thresholds that best suit the configuration of the method to make fair comparisons in the paper.

On the other hand, we repeated similar experiments for the pipeline when 'EfficientDet' was used as candidate detection in the first stage of MDFS. The new results depicted in Fig. A.1b unequivocally reveal that utilizing EfficientDet as a candidate detector not only diminishes the F1 score (from 0.87 for EUNet+EfficientNet to 0.62 for EfficientDet+EfficientNet) but also restricts the optimal performance range concerning threshold selection for each stage. Specifically, only a detection threshold of 0.2 and a classification threshold between 0 and 0.25 yield the best results.

In addition, to address concerns about computational efficiency, we have plotted the runtime impact of varying detection thresholds on the classification stage and the whole MDFS pipeline in Fig. A.2. This has been done using full-resolution images in the first stage of the pipeline. Our findings suggest that while the detection threshold significantly influences the number of candidates presented to the classification stage, the added computational load does not substantially affect overall runtime. This observation underscores the efficiency of our two-stage approach, even when adjusting for higher sensitivity in the initial detection phase.

Appendix B. Details of TCGA-BRCA mitosis dataset

The WSI processing pipeline is explained in Section 3.5. In our processing of the TCGA-BRCA dataset, we relied on the TIAToolbox software, which effectively manages image scaling. We extracted tiles

from WSIs at a resolution of 0.25 mpp, equivalent to roughly a 40x magnification, then fed them into the MDFS pipeline which downscales the image tile by a factor of 0.75 in the 'Detecting Fast' system. When encountering WSIs scanned at higher magnifications, such as 80x, TIAToolbox performed the necessary downscaling of tiles to match our desired resolution. As for the 20x slides, we ensured high-quality detection by utilizing segmentation and classification models trained on half-scaled (20× images).

In order to deal with the large variability of mitoses and usual artifacts in WSIs (such as pen markings and stain residues), we fine-tuned the MIDOG22 classifier model on a manually curated dataset of common artifacts in histopathology. Small (128×128 pixels) artifact patches were extracted from a selection of TCGA and an in-house dataset to form a collection of 22,600 artifact images. Doing this has made our 'Detecting Slow' system more robust against obvious artifacts in WSIs.

In total, our released 'TCGA-BRCA Mitosis Dataset' comprises 1124 mitosis detection files in GeoJSON format containing more than 0.67 million candidates (initially detected by the segmentation model), of which 622,528 are confirmed to be mitoses by the 'Detecting Slow' classifier. We have released both mitosis figures and proxy figures (instances that were pruned out by the 'Detecting Slow' system but had a mid-range probability of being mitosis) to further aid in developing better mitosis detection models and downstream analysis in the future. For each WSI in the dataset, we release the candidates' centroid, bounding box, hotspot location, hotspot mitotic count, and hotspot mitotic score. This dataset can be found at https://zenodo.org/records/10245707. It should be noted that we did not conduct a comprehensive review of all mitotic figures within each WSI, and we do not purport these to be free of errors. Nonetheless, a pathologist examined the resultant hotspot regions of interest from 757 WSIs within the TCGA-BRCA Mitosis Dataset where we found strong correlations between pathologist and MDFS mitotic counts (r = 0.8, p<0.001). Furthermore, MDFS-derived mitosis scores are shown to be as prognostic as pathologist-assigned mitosis scores (Ibrahim et al., 2024). This examination was also aimed at verifying the quality of the selections, ensuring excessive false detections or artifacts did not primarily drive them and were in a plausible location in the tumor landscape.

References

Akram, S.U., Qaiser, T., Graham, S., Kannala, J., Heikkilä, J., Rajpoot, N., 2018. Leveraging unlabeled whole-slide-images for mitosis detection. In: Computational Pathology and Ophthalmic Medical Image Analysis. Springer, pp. 69–77.

Alemi Koohbanani, N., Jahanifar, M., Gooya, A., Rajpoot, N., 2018. Nuclei detection using mixture density networks. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 241–248.

Alemi Koohbanani, N., Jahanifar, M., Gooya, A., Rajpoot, N., 2019. Nuclear instance segmentation using a proposal-free spatially aware deep learning framework. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 622–630.

Alemi Koohbanani, N., Jahanifar, M., Tajadin, N.Z., Rajpoot, N., 2020. NuClick: a deep learning framework for interactive segmentation of microscopic images. Med. Image Anal. 65, 101771.

Alemi Koohbanani, N., Unnikrishnan, B., Khurram, S.A., Krishnaswamy, P., Rajpoot, N., 2021. Self-path: Self-supervision for classification of pathology images with limited annotations. IEEE Trans. Med. Imaging 40 (10), 2845–2856.

Alkhasawneh, A., Reith, J.D., Toro, T.Z., Ayed, A.O., Lu, X., George, T.J., Duckworth, L.V., 2015. Interobserver variability of mitotic index and utility of PHH3 for risk stratification in gastrointestinal stromal tumors. Am. J. Clin. Pathol. 143 (3), 385–392.

Asif, A., Rajpoot, K., Snead, D., Minhas, F., Rajpoot, N., 2021. Towards launching AI algorithms for cellular pathology into clinical & pharmaceutical orbits. arXiv preprint arXiv:2112.09496.

Aubreville, M., Bertram, C., Breininger, K., Jabari, S., Stathonikos, N., Veta, M., 2022a. Mitosis domain generalization challenge 2022. http://dx.doi.org/10.5281/zenodo.6362337.

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

Aubreville, M., Stathonikos, N., Bertram, C.A., Klopleisch, R., ter Hoeve, N., Ciompi, F., Wilm, F., Marzahl, C., Donovan, T.A., Maier, A., et al., 2022b. Mitosis domain generalization in histopathology images–the MIDOG challenge. arXiv preprint arXiv:2204.03742.

Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318 (22), 2199–2210.

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. Sci. Data 6 (1), 1–9.

Bertram, C.A., Veta, M., Marzahl, C., Stathonikos, N., Maier, A., Klopfleisch, R., Aubreville, M., 2020. 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.

Bizzego, A., Bussola, N., Chierici, M., Maggio, V., Francescatto, M., Cima, L., Cristoforetti, M., Jurman, G., Furlanello, C., 2019. Evaluating reproducibility of AI algorithms in digital pathology with DAPPER. PLoS Comput. Biol. 15 (3), e1006269.

Cheeseman, I.M., Desai, A., 2008. Molecular architecture of the kinetochore–microtubule interface. Nat. Rev. Mol. Cell Biol. 9 (1), 33–46.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. PMLR, pp. 1597–1607.

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

Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 411–418.

Cree, I.A., Tan, P.H., Travis, W.D., Wesseling, P., Yagi, Y., White, V.A., Lokuhetty, D., Scolyer, R.A., 2021. Counting mitoses: SI (ze) matters! Mod. Pathol. 34 (9), 1651–1657.

Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N.N., Tomaszewski, J., González, F.A., Madabhushi, A., 2017. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. Sci. Rep. 7 (1), 46450.

Dif, N., Elberrichi, Z., 2020. Deep learning methods for mitosis detection in breast cancer histopathological images: A comprehensive review. In: Artificial Intelligence and Machine Learning for Digital Pathology. Springer, pp. 279–306.

Djuric, U., Zadeh, G., Aldape, K., Diamandis, P., 2017. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. NPJ Precis. Oncol. 1 (1), 22.

Ellis, I., Pinder, S., Bobrow, L., Buley, I., Coyne, J., Going, J., Humphreys, S., Jasani, B., Lakhani, S., Lowe, J., et al., 2005. Pathology reporting of breast disease. NHS Publ. (58).

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 (5), 403–410.

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

Gamper, J., Koohbanani, N.A., Benes, K., Graham, S., Jahanifar, M., Khurram, S.A., Azam, A., Hewitt, K., Rajpoot, N., 2020. Pannuke dataset extension, insights and baselines. arXiv preprint arXiv:2003.10778.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., Lempitsky, V., 2016. Domain-adversarial training of neural networks. J. Mach. Learn. Res. 17 (59), 1–35.

Graham, S., Jahanifar, M., Azam, A., Nimir, M., Tsang, Y.-W., Dodd, K., Hero, E., Sahota, H., Tank, A., Benes, K., et al., 2021. Lizard: a large-scale dataset for colonic nuclear instance segmentation and classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 684–693.

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. Med. Image Anal. 58, 101563. http://dx.doi.org/10.1016/j.media.2019.101563, URL: https://www.sciencedirect.com/science/article/pii/S1361841519301045.

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. CVPR.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7132–7141.

Ibrahim, A., Jahanifar, M., Wahab, N., Toss, M.S., Makhlouf, S., Atallah, N., Lashen, A.G., Katayama, A., Graham, S., Bilal, M., et al., 2024. Artificial intelligence-based mitosis scoring in breast cancer: Clinical application. Mod. Pathol. 37 (3), 100416.

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 (6), 365–372.

Ibrahim, A., Toss, M.S., Makhlouf, S., Miligy, I.M., Minhas, F., Rakha, E.A., 2023. Improving mitotic cell counting accuracy and efficiency using phosphohistone-H3 (PHH3) antibody counterstained with haematoxylin and eosin as part of breast cancer grading. Histopathology 82 (3), 393–406.

Jahanifar, M., Alemi Koohbanani, N., Rajpoot, N., 2019. Nuclick: From clicks in the nuclei to nuclear boundaries. In: MICCAI 2019 Workshop COMPAY.

Jahanifar, M., Raza, M., Xu, K., Vuong, T., Jewsbury, R., Shephard, A., Zamani-tajeddin, N., Kwak, J.T., Raza, S.E.A., Minhas, F., Rajpoot, N., 2023. Domain generalization in computational pathology: Survey and guidelines. arXiv:2310.19656.

Jahanifar, M., Tajeddin, N.Z., Alemi Koohbanani, N., Gooya, A., Rajpoot, N., 2018. Segmentation of skin lesions and their attributes using multi-scale convolutional neural networks and domain specific augmentations. arXiv preprint arXiv:1809.10243.

Jing, L., Tian, Y., 2020. Self-supervised visual feature learning with deep neural networks: A survey. IEEE Trans. Pattern Anal. Mach. Intell. 43 (11), 4037–4058.

Kahneman, D., 2011. Thinking, Fast and Slow. Macmillan.

Kausar, T., Wang, M., Ashraf, M.A., Kausar, A., 2020. SmallMitosis: small size mitotic cells detection in breast histopathology images. IEEE Access 9, 905–922.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D., 2020. Supervised contrastive learning. Adv. Neural Inf. Process. Syst. 33, 18661–18673.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kondo, S., 2021. Multi-source domain adaptation using gradient reversal layer for mitotic cell detection. http://dx.doi.org/10.48550/ARXIV.2109.01503, URL: https://arxiv.org/abs/2109.01503.

Kotte, S., Saipradeep, V., Sivadasan, N., Joseph, T., Sharma, H., Walia, V., Varma, B., Mukherjee, G., 2022. A deep learning based ensemble model for generalized mitosis detection in H&E stained whole slide images. http://dx.doi.org/10.5281/zenodo.7030876.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25.

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

Lashen, A., Ibrahim, A., Katayama, A., Ball, G., Mihai, R., Toss, M., Rakha, E., 2021. Visual assessment of mitotic figures in breast cancer: a comparative study between light microscopy and whole slide images. Histopathology 79 (6), 913–925.

Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J., 2021. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Trans. Neural Netw. Learn. Syst..

Li, C., Wang, X., Liu, W., Latecki, L.J., Wang, B., Huang, J., 2019. Weakly supervised mitosis detection in breast histopathology images using concentric loss. Med. Image Anal. 53, 165–178.

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

Ludovic, R., Daniel, R., Nicolas, L., Maria, K., Humayun, I., Jacques, K., Frédérique, C., Catherine, G., et al., 2013. Mitosis detection in breast cancer histological images an ICPR 2012 contest. J. Pathol. Inform. 4 (1), 8.

Mahmood, T., Arsalan, M., Owais, M., Lee, M.B., Park, K.R., 2020. Artificial intelligence-based mitosis detection in breast cancer histopathology images using faster R-CNN and deep CNNs. J. Clin. Med. 9 (3), 749.

Mathew, T., Kini, J.R., Rajan, J., 2021. Computational methods for automated mitosis detection in histopathology images: A review. Biocybern. Biomed. Eng. 41 (1), 64–82.

Miech, A., Alayrac, J.-B., Laptev, I., Sivic, J., Zisserman, A., 2021. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9826–9836.

Molenaar, W.M., Plaat, B.E., Berends, E.R., te Meerman, G.J., 2000. Observer reliability in assessment of mitotic activity and MIB-1—Determined proliferation rate in pediatric sarcomas. Ann. Diagn. Pathol. 4 (4), 228–235.

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

Paul, A., Mukherjee, D.P., 2015. Mitosis detection for invasive breast cancer grading in histopathological images. IEEE Trans. Image Process. 24 (11), 4041–4054.

Pocock, J., Graham, S., Vu, Q.D., Jahanifar, M., Deshpande, S., Hadjigeorghiou, G., Shephard, A., Bashir, R.M.S., Bilal, M., Lu, W., Epstein, D., Minhas, F., Rajpoot, N.M., Raza, S.E.A., 2022. TIAToolbox as an end-to-end library for advanced tissue image analytics. Commun. Med. 2 (1), 120. http://dx.doi.org/10.1038/s43856-022-00186-5, URL: https://www.nature.com/articles/s43856-022-00186-5.

Rakha, E.A., El-Sayed, M.E., Lee, A.H., Elston, C.W., Grainge, M.J., Hodi, Z., Blamey, R.W., Ellis, I.O., 2008. Prognostic significance of nottingham histologic grade in invasive breast carcinoma. J. Clin. Oncol. 26 (19), 3153–3158.

Ramachandran, P., Zoph, B., Le, Q.V., 2017. Swish: a self-gated activation function. p. 5, arXiv preprint arXiv:1710.05941 7.

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

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 234–241.

Roy, S., kumar Jain, A., Lal, S., Kini, J., 2018. A study about color normalization methods for histopathology images. Micron 114, 42–61.

Saipradeep, V., Kotte, S., Sivadasan, N., Joseph, T., Sharma, H., Walia, V., Varma, B., Mukherjee, G., 2022. Attention based transformers for mitosis domain generalization in H&E stained WSIs. http://dx.doi.org/10.5281/zenodo.7029973.

Saldanha, G., Ali, R., Bakshi, A., Basiouni, A., Bishop, R., Colloby, P., Craig, P., Da Forno, P., Edward, S., Espinosa de Los Monteros, O., et al., 2020. Global and mitosis-specific interobserver variation in mitotic count scoring and implications for malignant melanoma staging. Histopathology 76 (6), 803–813.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C., 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4510–4520.

Sebai, M., Wang, X., Wang, T., 2020. MaskMitosis: a deep learning framework for fully supervised, weakly supervised, and unsupervised mitosis detection in histopathology images. Med. Biol. Eng. Comput. 58 (7), 1603–1623.

Shephard, A.J., Graham, S., Bashir, S., Jahanifar, M., Mahmood, H., Khurram, A., Rajpoot, N.M., 2021. Simultaneous nuclear instance and layer segmentation in oral epithelial dysplasia. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 552–561.

Stacke, K., Eilertsen, G., Unger, J., Lundström, C., 2020. Measuring domain shift for deep learning in histopathology. IEEE J. Biomed. Health Inform. 25 (2), 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.

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.

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 Trans. Med. Imaging 37 (9), 2126–2136.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., Van Der Laak, J., 2019. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med. Image Anal. 58, 101544.

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 Trans. Med. Imaging 35 (8), 1962–1971.

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., Cireşan, D.C., Schmidhuber, J., Giusti, A., Gambardella, L.M., Tek, F.B., Walter, T., Wang, C.W., Kondo, S., Matuszewski, B.J., Precioso, F., Snell, V., Kittler, J., de Campos, T.E., Khan, A.M., Rajpoot, N.M., Arkoumani, E., Lacle, M.M., Viergever, M.A., Pluim, J.P., 2015. Assessment of algorithms for mitosis detection in breast cancer histopathology images. Med. Image Anal. 20 (1), 237–248.

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. Med. Image Anal. 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 (8), e0161286.

Vu, Q.D., Jewsbury, R., Graham, S., Jahanifar, M., Raza, S.E.A., Minhas, F., Bhalerao, A., Rajpoot, N., 2022. Nuclear segmentation and classification: On color & compression generalization. In: International Workshop on Machine Learning in Medical Imaging. Springer, pp. 10–19.

Vuong, T.T.L., Vu, Q.D., Jahanifar, M., Graham, S., Kwak, J.T., Rajpoot, N., 2022. Impash: A novel domain-shift resistant representation for colorectal cancer tissue classification. In: Medical Computer Vision Workshop. MCV.

Wilm, F., Bertram, C.A., Marzahl, C., Bartel, A., Donovan, T.A., Assenmacher, C.-A., Becker, K., Bennett, M., Corner, S., Cossic, B., et al., 2021a. Influence of inter-annotator variability on automatic mitotic figure assessment. In: Bildverarbeitung für Die Medizin 2021: Proceedings, German Workshop on Medical Image Computing, Regensburg, March 7-9, 2021. Springer, pp. 241–246.

Wilm, F., Breininger, K., Aubreville, M., 2021b. Domain adversarial retinanet as a reference algorithm for the mitosis domain generalization (midog) challenge. In: Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis, MICCAI 2021 Challenges L2R, MIDOG and MOOD. Springer, pp. 5–13.

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