

Toward interpretable and generalized mitosis detection in digital pathology using deep learning

Hasan Farooq¹ , Saira Saleem² , Iffat Aleem², Ayesha Iftikhar²,
Umer Nisar Sheikh² and Hammad Naveed¹ 

Abstract

Objective: The mitotic activity index is an important prognostic factor in the diagnosis of cancer. The task of mitosis detection is difficult as the nuclei are microscopic in size and partially labeled, and there are many more non-mitotic nuclei compared to mitotic ones. In this paper, we highlight the challenges of current mitosis detection pipelines and propose a method to tackle these challenges.

Methods: Our proposed methodology is inspired from recent research on deep learning and an extensive analysis on the dataset and training pipeline. We first used the MiDoG'22 dataset for training, validation, and testing. We then tested the methodology without fine-tuning on the TUPAC'16 dataset and on a real-time case from Shaukat Khanum Memorial Cancer Hospital and Research Centre.

Results: Our methodology has shown promising results both quantitatively and qualitatively. Quantitatively, our methodology achieved an F1-score of 0.87 on the MiDoG'22 dataset and an F1-score of 0.83 on the TUPAC dataset. Qualitatively, our methodology is generalizable and interpretable across various datasets and clinical settings.

Conclusion: In this paper, we highlight the challenges of current mitosis detection pipelines and propose a method that can accurately predict mitotic nuclei. We illustrate the accuracy, generalizability, and interpretability of our approach across various datasets and clinical settings. Our methodology can speed up the adoption of computer-aided digital pathology in clinical settings.

Keywords

Mitosis detection, deep learning, interpretable AI, digital pathology, generalizability

Submission date: 26 September 2023; Acceptance date: 1 May 2024

Introduction

Cancer is a widespread disease that affects people all over the world. The mitotic activity index (MAI) is an important prognostic factor in the diagnosis of this disease.¹ In a standard diagnostic procedure, pathologists typically count the number of mitotic nuclei in 10 consecutive microscopic high-power fields (HPF) or in specific field areas like 1 mm² or 2 mm² under 40× magnification.² However, this process is tedious and time-consuming due to the large size of the hematoxylin and eosin slides and the infrequent occurrence of mitotic nuclei. Moreover, the heterogeneity

of patterns and the similarity between mitotic and non-mitotic cells can often be misinterpreted as mitotic activity. Thus, reproducibly counting mitotic cells can be

¹Computational Biology Research Lab, National University of Computer & Emerging Sciences, Islamabad, Pakistan

²Shaukat Khanum Memorial Cancer Hospital and Research Centre, Lahore, Pakistan

Corresponding author:

Hammad Naveed, Computational Biology Research Lab, National University of Computer & Emerging Sciences, Islamabad 44800, Pakistan.

Email: hammad.naveed@nu.edu.pk



challenging among clinical experts.³ This reproducibility of counting the number of mitotic cells is very important in certain neoplastic processes that are of the intermediate category as they can change into malignant categories based only on the number of mitosis in the particular area.

The advent of image acquisition techniques in the early 2000s enabled the digitization of glass slides, allowing pathology slide regions to be scanned and stored as digital images.⁴ This digitization of histopathology, combined with advancements in medical image processing and machine learning techniques, has conjured a new era of computer-aided pathology.⁵ The machine learning methods are capable of extracting patterns from the data and making predictions based on these patterns.⁶ In digital pathology, research has been carried out on various tasks: diagnosis of prostate cancer, microbiological diseases, and lymph node metastases, to name a few.^{7,8,9} The automation of mitosis detection is also one such task that can assist the pathologist in the diagnosis.

Related work

Mitosis detection methods have primarily utilized either handcrafted features, deep learning, or a combination of both. The initial methods used image processing techniques to extract features manually from slide images and trained classifiers to distinguish between mitotic and non-mitotic nuclei.^{10,11,12} These methods require manual feature design and selection. The features used in these studies included morphological features, such as area, perimeter, eccentricity, long and short axis length, diameter, as well as statistical features like the mean, median, variance of each color channel, color histogram features, and color scale.^{10,11,12} However, due to the diversity of mitotic nuclei shapes and textures, customizing these hand-crafted feature-based methods for mitosis detection tasks is difficult, and the outcomes are often unsatisfactory.¹³

In comparison, deep learning algorithms applied in the field of medical image processing have been shown to outperform the results of state-of-the-art handcrafted feature-based classification methods.¹⁴ Deep learning methods usually approach this task as a classification, object detection, or segmentation problem using convolutional neural networks (CNNs).^{13,15,16,17,18,19} However, deep learning-based methods have a significant limitation: their performance is known to deteriorate with a covariate shift in the images encountered during clinical diagnosis, making these models less robust and less trustworthy.²⁰ In digital pathology, this domain shift can be caused by the staining procedure (which can differ over time and/or across laboratories), the acquisition device (whole slide scanner), and the tumor type itself (different tumor cell morphology and tissue architecture).¹⁵ Despite these challenges, computer-assisted automated detection, particularly deep learning methods, has become an area of increasing interest

for researchers.⁴ These methods have the potential to reduce the pathologist's workload and improve diagnostic efficiency.

In recent biomedical imaging challenges (TUPAC, MiDoG), object detection was seen as a go to approach for mitosis detection.¹⁵ Dusenberry and Hu (2018) proposed a classification solution for mitosis detection.¹⁶ This work preprocessed whole slide images (WSIs) by creating patches of mitotic and non-mitotic nuclei. These patches go through the modified ResNet-50 model, and the predictions are smoothed with clustering and thresholding. This work achieved an F1 score of 0.60. Sohail et al. (2021) proposed DHE-Mit, a hybrid of classification and object detection.¹⁷ This work first used object detection architecture (Mask-RCNN) for candidate mitotic nuclei selection, further using an ensemble of varying classification architectures for classification. This enabled the extraction of features from different architectures on the same dataset. DHE-Mit achieved an F1 score of 0.77. However, in DHE-Mit, the loss factor from the first stage is omitted which, at times, may not be feasible.¹⁹ Khan et al. (2023) proposed SMDetector, an object detection pipeline for tackling the small sizes of mitotic nuclei.¹⁹ This study highlighted the issue of small object detection in mitosis detection. To tackle this problem, the authors modified the Faster R-CNN architecture by incorporating dilations in the ResNet-101 backbone model. The dilations help the convolutional neural network in covering larger spatial contexts. The authors further refined the pipeline by dividing the pipeline into two stages: the mitotic candidates are selected in the first stage with Faster R-CNN and fine-tuned in the second stage via fully connected layers. This staging makes the model more robust and reduces false positives. SMDetector achieved an F1 score of 0.64. Jahanifar et al. (2024) proposed MDfS, a robust two-stage hybrid pipeline.¹⁸ In the first stage, the mitosis candidates are segmented at a lower resolution. In the second stage, these are refined after classification at a higher resolution. This work utilizes EfficientNetB-0 and U-Net in the first stage and EfficientNetB-7 in the second stage. Although MDfS generalizes better across datasets and achieves an F1 score of 0.79, the authors have highlighted that the approach misses small mitotic nuclei.

What makes mitosis detection using deep learning challenging

Implementing deep learning -based methods is a challenging task due to several reasons. First, mitosis has four distinct phases, and each phase has a different shape and texture, making it difficult to distinguish mitotic nuclei from other cells, such as apoptotic cells and lymphocytes.^{10,11} Second, it is important to detect mitosis in cancerous cells only and not in normal cells or granulation

tissue. Third, the number of mitotic nuclei is significantly lower than non-mitotic nuclei, making it challenging to extract useful features for deep learning models due to the imbalance of positive and negative samples.²¹ Fourth, datasets for training deep learning models are limited, with most public datasets originating from research challenges like the MITosis DOmain Generalization Challenge (MIDOG'22) rather than directly from hospitals, resulting in differences in image quality, lab environments, tissue types, and availability of patches rather than WSIs.^{10,12,13} Lastly, the interpretability of deep learning models is one of the key challenges in the medical domain, especially in clinical settings.¹³ Hence, these datasets and models may not accurately represent the morphology and structure of mitotic nuclei across all pathological types.

Proposed method

Taking into account the abovementioned problems, instead of object detection, we have approached the task from a classification perspective, focusing on improving generalizability and interpretability. The reasons behind our classification approach are the inherent problems in the object detection architectures, including (a) small size of the object to be detected, (b) incomplete annotations, (c) imbalanced dataset (foreground–background, scaling, class), and (d) localization, especially in domain shifts.^{22,23,24,25,26} The task of mitosis detection is prone to the abovementioned issues as the nuclei are microscopic in size, have many more non-mitotic nuclei compared to mitotic ones, and have partially labeled data. For instance, in the recent MiDoG 2022 challenge, the reference models used object detection architectures, particularly RetinaNet, which can detect boxes of 32×32 px at the minimum. The mitotic nuclei, however, are usually of lesser sizes. In the classification space, the larger models tend to overparameterize, leading to overfitting.^{27,28}

Therefore, we have used a hybrid of the concepts inspired from the EfficientNet and ResNet families and dilated convolutions to build our small classification model, as shown in Figure 1.^{29,30,31} The EfficientNets have been designed considering the compound scaling of depth, width, and resolution of models, whereas the ResNets have been designed with the concern of smooth gradient flow and preservation of information.^{29,30} The dilated convolution increases the receptive field of the filters that capture a larger spatial context while preserving the resolution.³¹ To further improve the generalizability and interpretability of our method, we have done an extensive analysis on the data processing and training pipeline. Our method achieved F1 scores of 0.87 and 0.83 on the MIDOG'22 and TUPAC datasets, respectively. Lastly, we validated whether or not our methodology learned features that can be attributed to a particular class using visual interpretations generated by gradient-weighted class activation mapping (GradCam).³²

Methodology

In this section, we describe our methodology in detail. This includes the dataset, data extraction and preprocessing, training pipeline, inference, and implementation details.

Dataset

The dataset used in this study was taken from Mitosis Domain Generalization Challenge 2022 (MiDoG'22).³³ The dataset consists of 350 WSIs from different scanners (Hamamatsu NanoZoomer XR, Hamamatsu NanoZoomer S360, Aperio ScanScope CS2, 3DHISTECH Panoramic Scan II, and Leica Aperio GT 450), species (human, dog, and cat), and tumor types (canine lung cancer, human breast cancer, canine lymphoma, human neuroendocrine tumor, and canine cutaneous mast cell tumor).

The variations (Figure 2) in the images due to environments, tumor types, and species help build models that are robust and generalizable. The dataset has 9501 annotations with bounding boxes of 50×50 px. These include mitotic nuclei and hard negatives, which look similar to mitotic nuclei, but are not. Figure 3 shows the extracted bounding boxes.

2.2 Overview

Figure 4 presents the overall framework of the proposed method. The method is divided into five blocks: data extraction, data processing, classification, inference, and interpretability. In the first block, we extract patches from the annotated WSIs. These patches are normalized, enhanced, and augmented in the second block. In the third block, the data are trained, validated, and tested for classification of the mitotic and non-mitotic patches. Finally, these patches are predicted in the inference block. After the deep learning model is trained, the patches can be examined via a heat map in the interpretability block. Details of each block are given in the subsequent sections.

Data extraction. The size of the mitotic nuclei is typically 25×25 px in the MIDOG'22 dataset. The annotations come with bounding boxes of 50×50 px around the mitotic and hard negative areas, which look similar to mitotic nuclei, but are not. The mitotic annotations account for the mitotic class in our proposed method. For the non-mitotic class, we add the hard negatives given in the dataset and extract patches randomly from WSIs to add structures other than nuclei. These randomly extracted patches might contain mitotic areas as well, but this helps the model extract features that can differentiate between the two classes. Our customized dataset for classification consisted of 2293 mitotic, 2617 hard negatives, and 3000 randomly extracted negative patches. For data extraction,

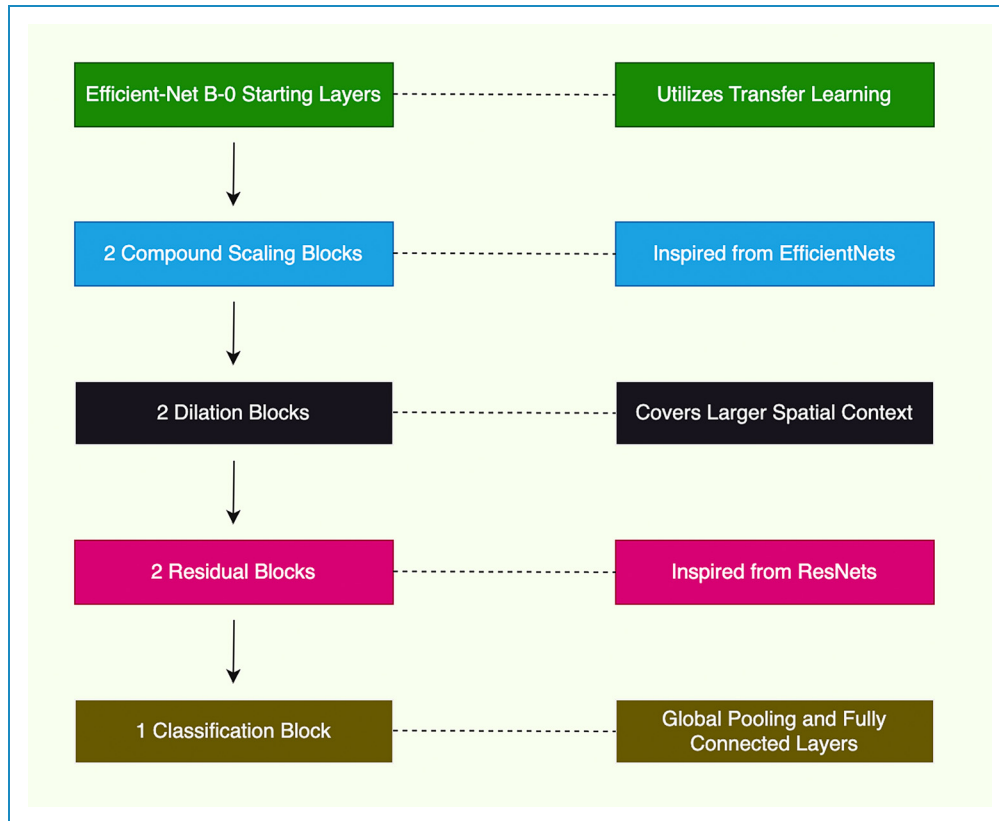


Figure 1 Architecture.

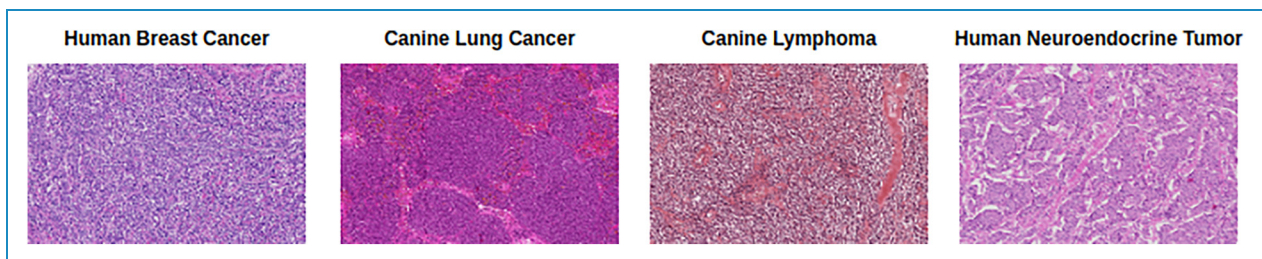


Figure 2 Tumor-wise variance in images from different lab environments and scanners.

we experimented with different patch sizes: 50×50 , 100×100 , 256×256 , and 512×512 .

Data processing. In the data processing block, we used stain normalization to tackle the issue of generalizability. Stain normalization techniques are used to transfer the color distributions of all WSIs to a reference image while preserving the structure and other information of the image. This ensures that WSI samples from different labs, tissues, and domains follow a base color distribution. These techniques can be broadly classified into two classes: conventional and deep learning-based methods. Our method utilized StainNet, a deep learning-based method that uses 1D-CNN with StainGAN to adjust color mappings in a pixel-to-pixel manner.³⁴

Classification. In the classification block, we divided the dataset into training, testing, and validation sets. We used image augmentations (flips, noise, sharpen, blur, equalization, contrast, and brightness) to add variations in the training set. The augmentations help the model recognize patterns not dominantly present in the training data. After augmentation, we trained our CNN model. During the model selection, we split our dataset in a rather non-conventional way. We trained the model on three (two human tissues and one animal tissue) out of five tumor types and validated/tested on samples from all tumor types/species. This ensured the generalizability of the model in capturing the mitotic areas during the training. However, the final model was trained on tissues from humans only (breast and neuroendocrine). Our final

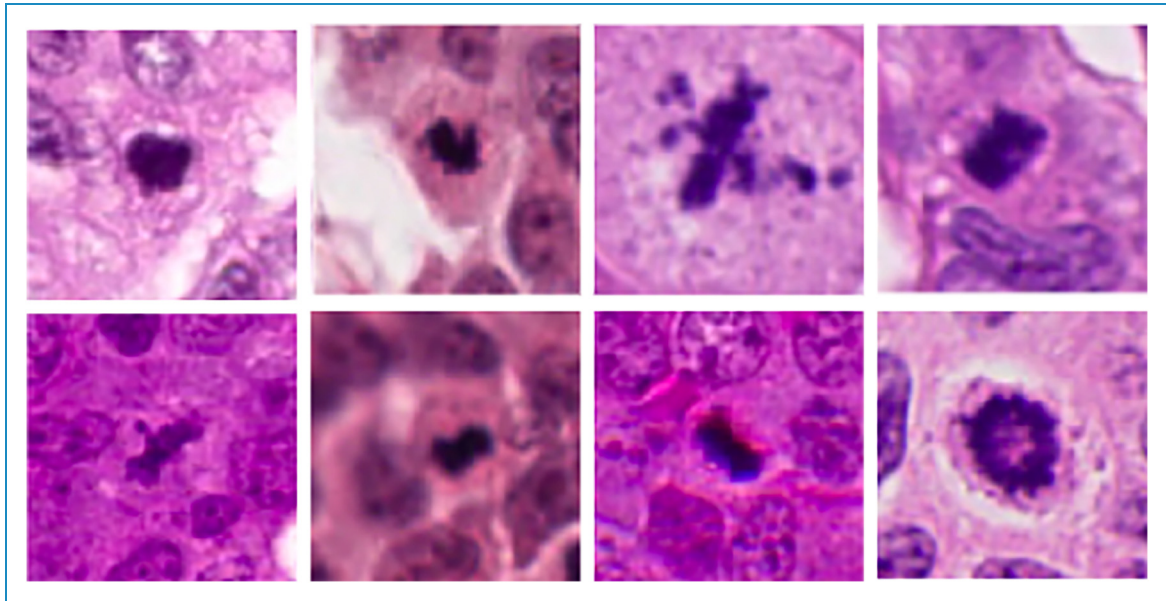


Figure 3 50 × 50 px bounding box annotations in the dataset.

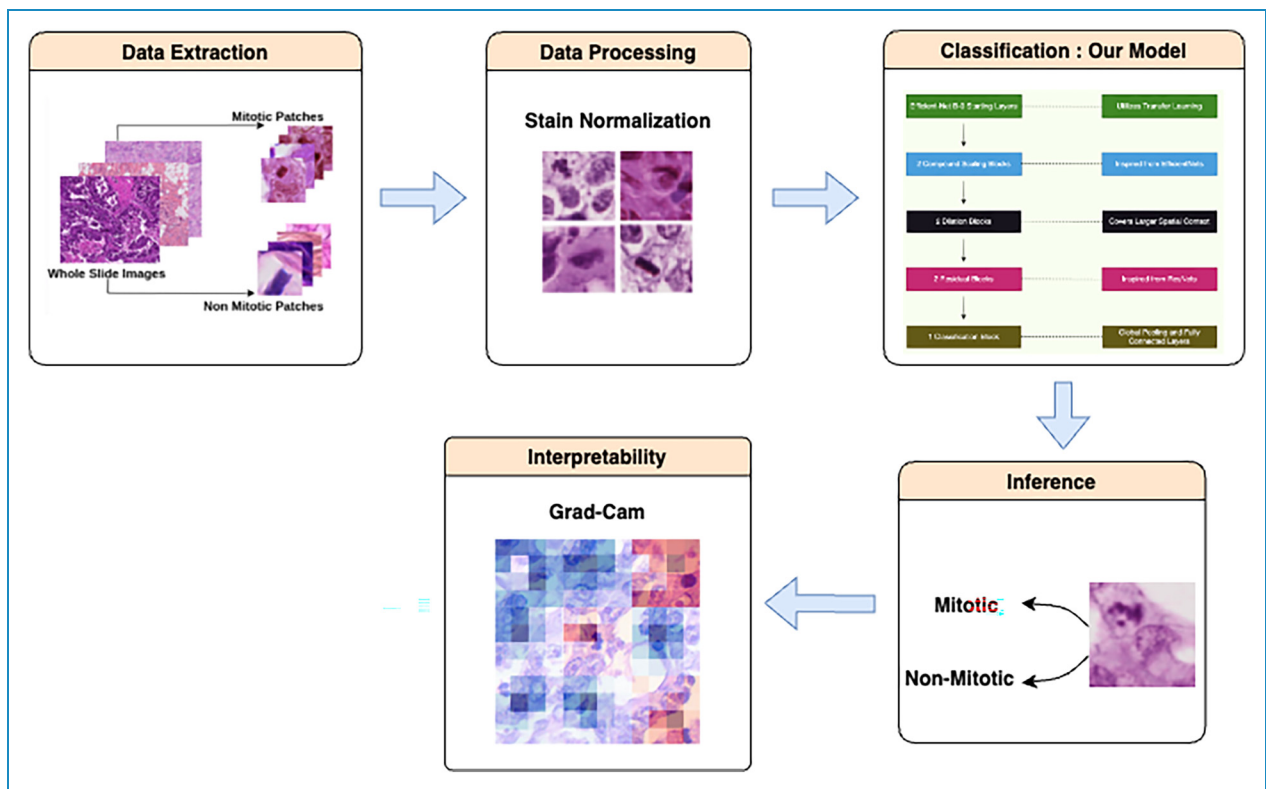


Figure 4 Method divided into five blocks: data extraction, data processing, classification, inference, and interpretability. The patches are extracted, normalized, enhanced, and augmented in the first two blocks. We then trained the model to distinguish between mitotic and non-mitotic patches. The predicted mitotic patches were validated for focus areas in the interpretability block.

model was inspired by EfficientNet, ResNet, and dilated convolutions, with the first few layers adapted from EfficientNet B-0 to take advantage of transfer learning.

Inference. We tested our model on (a) publicly available datasets and (b) on a local WSI of human colon cancer from Shaukat Khanum Memorial Cancer Hospital and

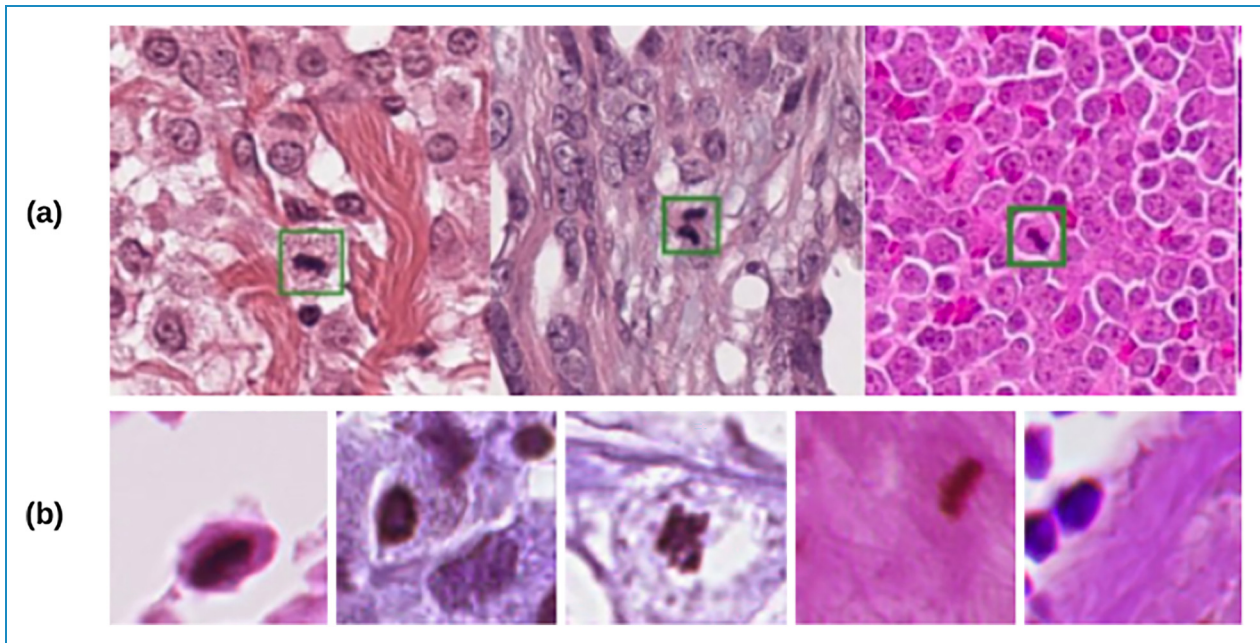


Figure 5 Comparison of mitotic nuclei with backgrounds. The top row (a) shows the size of the bounding box of the mitotic nuclei with respect to the background in large patches. The bottom row (b) shows the size of the bounding box of the mitotic nuclei with respect to the background in small patches.

Research Center, Lahore, Pakistan (SKMCH&RC). For the purpose of robustness and generalizability, we did not fine-tune our model on the TUPAC dataset and the local WSI. Instead, we used the pretrained weights of our model from the MiDoG'22 dataset.

- Publicly available human breast cancer dataset (TUPAC'16)³⁵: We followed the same procedure given in the training phase. This included extracting patches from given annotations and following all the steps until the classification step.
- Local human colon cancer whole slide image: The local WSI was scanned with the MoticEasyScan Pro 6 scanner at 40× resolution with a standard resolution setting. The WSI is available at (<https://drive.google.com/drive/folders/1PCHBVjaEy535OXXDD7KvbeWsFPxS0Q2m>). For the local WSI, we selected a region of interest and extracted 100×100 patches. The extracted patches went through all the steps mentioned above. The results were verified by a team of pathologists at SKMCH&RC. The purpose of this exercise was to validate the generalizability and interpretability of our model on tissues and environments not seen in the training data.

Interpretability. In the interpretability block, we validated the focus areas of the model inference with GradCAM.³² GradCAM is a technique used to understand the regions

of an image that are important for a prediction. It utilizes the gradients flowing into the final convolutional layer and maps it to the input. By visualizing these maps, we can observe the parts of the input image that were important in the prediction of the model. Section 4 discusses this in detail with illustrations.

Implementation details. For reproducibility concerns, we share our architecture details and hyperparameters. Our architecture follows the structure below in sequence:

- Starting layers with two blocks inspired from EfficientNet-B0
- Dilations blocks with a dilation factor of two
- Residual blocks with skip connections inspired from ResNet-18
- Two-dimensional average pooling, followed by fully connected layers

We used the cross-entropy loss function with weighted loss. Instead of a fixed constant, the weight is adjusted in proportion to the positive class. Our batch size was 32, and the Adam optimizer was used with an initial learning rate of 0.0002. Lastly, the starting layers of the model adapted from EfficientNet-B0 utilized the ImageNet pretrained weights. The implementation and customized dataset for classification can be found in this link (<https://drive.google.com/drive/folders/1xqBBuXrLUPpj7KXZfVn6jWLS0ctbG-Y8>).

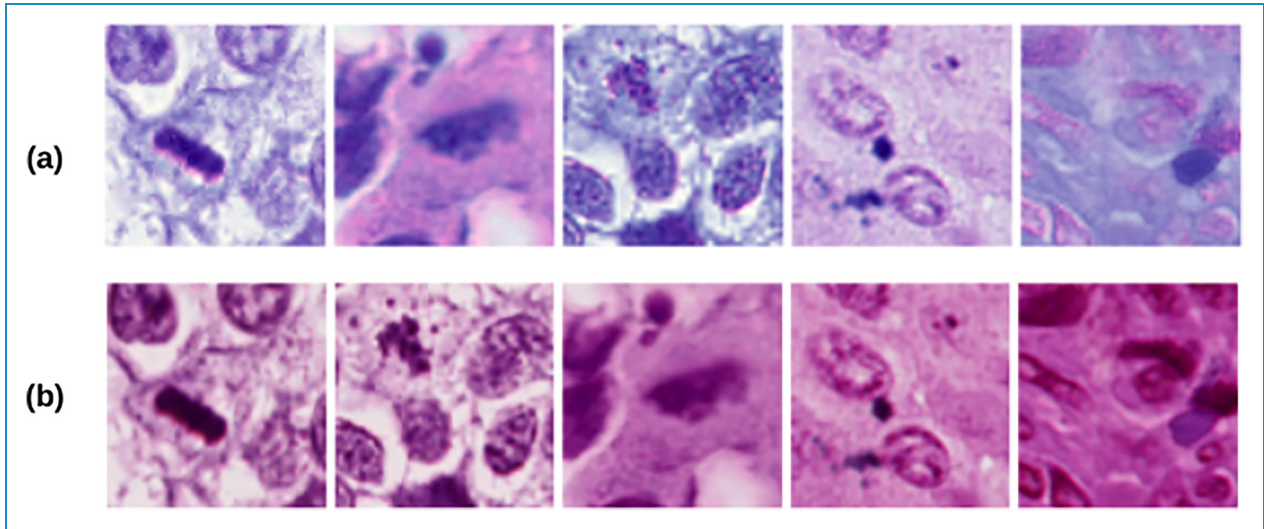


Figure 6 Top row (a) shows normalization using conventional methods, while the bottom row (b) shows normalization using StainNet.

Results and experimental analysis

In this section, we present the results and analysis of the experiments carried out for establishing our methodology. We divide these into two categories: qualitative and quantitative. The qualitative analysis is based on the feedback from pathologists, whereas the quantitative analysis is based on recent deep learning research studies.

Qualitative

The qualitative analysis is performed to find suitable resolution and stain normalization of the images. The feedback from pathologists helps in making a model that can accurately assist in the mitosis detection task.

Image resolution. Image resolution affects the training and robustness of the deep learning models.³⁶ Usually, image resolutions of 256, 512, or even 1024 are used for classification. However, the mitotic activity covers an average area of 25×25 px in our dataset. This microscopic nature of the mitotic activity can result in the model focusing on irrelevant details (i.e., background) for making predictions.³⁷ We found that smaller sizes produce better results; therefore, we used a patch size of 100×100 for our pipeline. Figure 5 shows the size comparison of the mitotic activities in large (512×512) and small (100×100) patches.

Stain normalization. MIDOG'21 and MIDOG'22 are domain generalization challenges that try to address the fact that computational models should work on images from different organs, scanners, and even species. One of the techniques of ensuring such generalization is stain normalization. For stain normalization, we experimented with both the conventional stain normalization techniques

(Macenko and Reinhard) and the generative adversarial network (GAN)-based stain normalization. Conventional methods normalize images through pixel-by-pixel color mapping on all the channels of the image and depend on a reference image to estimate the stain parameters, but it is hard for one reference image to cover all staining phenomena or represent all input images, resulting in an inaccurate estimation of stain parameters.³⁴ By contrast, deep learning-based methods mostly apply GANs to achieve stain normalization. GAN-based methods perform well in stain normalization; however, some issues in robustness have been reported.³⁴ StainNet is a deep learning-based method that utilizes 1D-CNN with StainGAN to adjust color mappings in a pixel-to-pixel manner. This makes StainNet more robust than the other deep learning architectures. The results of our experiments are illustrated in Figure 6.

Quantitative

The quantitative analysis addresses the challenges of deep learning in the medical domain. These include architecture design and hyperparameter selection.

Architecture design. We designed our model architecture keeping in view the recent findings. Recent studies have shown that larger models tend to overfit on datasets with low-resolution images, lesser classes, and smaller dataset size.^{27,38} Moreover, studies have also shown that certain architectures (based on their depth, width, and structure) are capable of extracting various patterns from the datasets.^{27,39} Based on these findings, we designed a hybrid architecture that is inspired from EfficientNet, ResNet,

Table 1 Results of the TUPAC'16 dataset. Our model outperformed other state-of-the-art methodologies by a significant margin without fine tuning the TUPAC dataset. *In these models, we estimated the number of parameters by the architecture details provided.

Models	Approach	F1 score	No. of parameters
EfficientNet B-0 ²⁹ , 2019	Classification	0.80	5.3 million
EfficientNet B-2 ²⁹ , 2019	Classification	0.78	9.2 million
EfficientNet B-4 ²⁹ , 2019	Classification	0.74	19 million
ResNet 18 ³⁰ 2016,	Classification	0.76	11 million
MAX Detector ¹⁶ , 2018	Classification	0.60	23 million*
RetinaNet ¹⁸ , 2017	Detection	0.72	36.4 million
EfficientDet-4 ¹⁸ , 2019	Detection	0.61	20.7 million
SMDetector ¹⁹ , 2023	Detection	0.64	61 million*
DHE-Mit ¹⁷ , 2021	Hybrid	0.77	56 million*
MDFS ¹⁸ , 2024	Hybrid	0.79	88 million*
Our model	Classification	0.83	3.1 million

and dilated convolutions. Table 1 shows a comparison of different architectures.

Hyperparameter selection. In this analysis, we focused on the effect of batch size, optimizer, and loss function on the model performance. Different measures can be used to evaluate the performance, including train/validation loss, precision, recall, sensitivity, specificity, accuracy, and ROC. However, to evaluate the performance, we considered the F1 scores that can be interpreted as the harmonic mean of precision and recall/sensitivity.

- **Batch size:** The batch size refers to the number of samples passed to model in an iteration. A study on the effect of batch sizes in histopathology suggests that smaller batch sizes can often result in overfitting, whereas larger batch sizes can make it difficult to achieve convergence.⁴¹ In our analysis, we experimented with the batch sizes of 16, 32, and 64.
- **Optimizer:** Optimizers are algorithms used to update the weights of the deep learning model. Based on a study on optimizers for convolutional neural networks, we experimented with Adam, RMSProp, and SGD.⁴²
- **Loss function:** Loss functions are functions that compute the error between the model output and the ground truth. These functions can be of various natures (i.e., probabilistic, distance based, and similarity based). In computer vision, cross entropy and its variations have been widely used for image classification.⁴³ We have also used the weighted version of cross entropy for training our model due to its simplicity.

Figure 7 shows our initial analysis on the hyperparameters. The adaptive optimizers (Adam, RMSProp) with a batch size of 32 yield the best results. To get better results, these selected hyperparameters (i.e., 32 batch size and Adam/

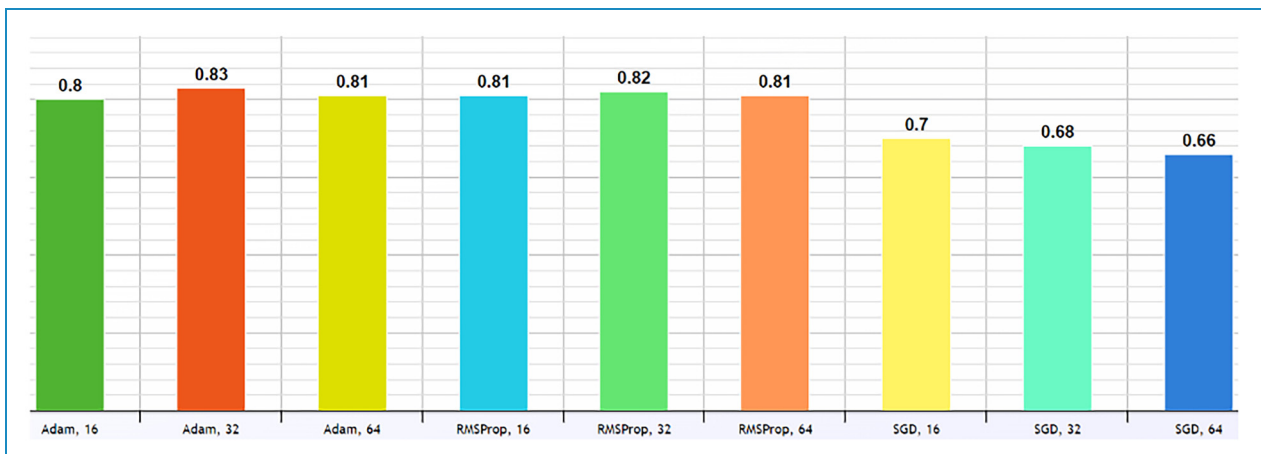


Figure 7 F1 scores of our hyperparameter analysis.

RMSProp optimizers) were further tuned after experimentation with learning rates and training for more epochs.

Table 2 shows the performance of our model on the MiDoG dataset. The results demonstrate the fact that the model established on the qualitative and quantitative analyses is able to predict the mitotic nuclei with a high F1 score (0.87).

In order to further explore the generalizability of our pipeline, we tested our model on the TUPAC dataset without fine-tuning. Table 1 shows our experimental results and comparison with other methodologies on the TUPAC dataset. We categorized the comparisons as classification, detection, and hybrid approaches. We also included the model size (number of parameters) in our comparisons to validate that the task of mitosis detection can be handled with smaller models without compromising on performance. In the classification space, we used three different architectures (EfficientNet, ResNet, and Max Detector) to validate that classification architectures are capable of handling the task of classifying mitotic nuclei accurately. We then provide results of recent detection and hybrid pipelines discussed in our literature review. Lastly, we provide the results of our methodology without fine-tuning the dataset. We were able to outperform other state-of-the-art methodologies with a significant margin.

Discussion

Our analysis shows that a hybrid of these concepts performs better than the individual models, even with a much smaller

Table 2 Results of our model on the MiDoG'22 dataset.

	Precision	Recall/ Sensitivity	F1 score	Specificity
Our model	0.92	0.79	0.87	0.93

model. The extensive analysis with the feedback from domain experts also makes the model capable of adapting to unseen data. The results are significantly better than those reported in previous studies. This is due to the fact that (a) we handled the problem as a classification task unlike most previous studies; (b) we applied a relatively small deep learning model that is appropriate for the amount of data available; and (c) we included feedback from the domain experts in our analysis. The deep learning architectures are made with real-life images in mind that are easy to gather and annotate, while in the medical domain, it is very hard to gather good quality data (due to privacy constraints) and even harder to annotate properly. We previously successfully used smaller models in similar medical problems.⁴⁰

Interpretability

In the medical domain, interpretability plays a key role in the trustworthiness of deep learning models in a clinical setting. Thus, we validated our approach using visual explanations with GradCam.³² Figures 8 and 9 show the results of our approach on mitotic nuclei in the MiDOG'22 and TUPAC'16 datasets, respectively.

We can see that (a) our approach is capable of focusing on the mitotic areas across the two datasets, and (b) normalizing the patches with StainNet helps the model generalize and focus better on the unseen data.

In the local WSI case, we first selected the region of interest, and then passed it through the model in 100×100 patches. The grid shown in Figure 10 highlighted the focus areas of our model both in the case of mitotic and non-mitotic patches (i.e., what areas are important when predicting mitotic or non-mitotic). Our results showed that the model is able to learn features similar to those that help humans differentiate between mitotic and non-mitotic nuclei.

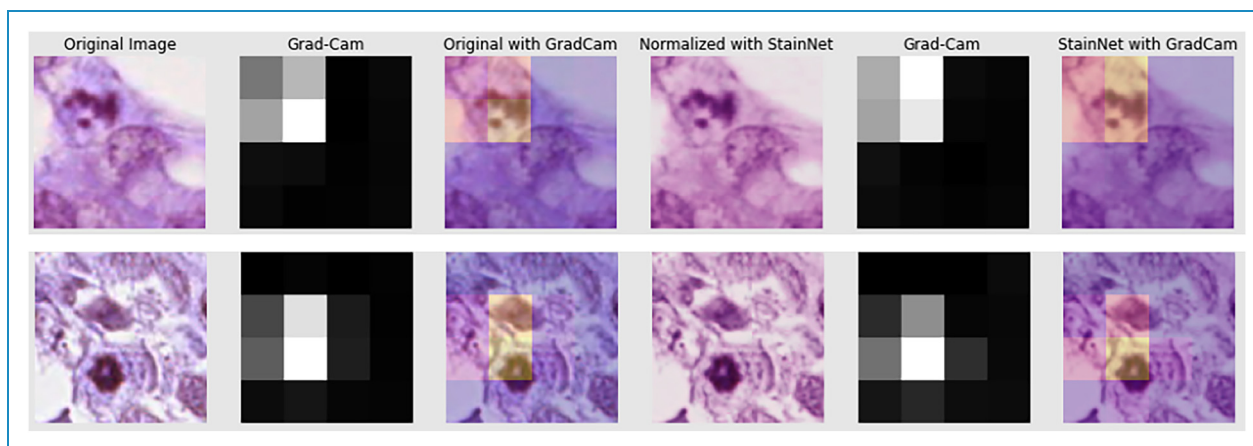


Figure 8 GradCam explanation of the model on the MiDoG'22 dataset. The whiter areas on the GradCam matrix are focused more, whereas the darker ones are less.

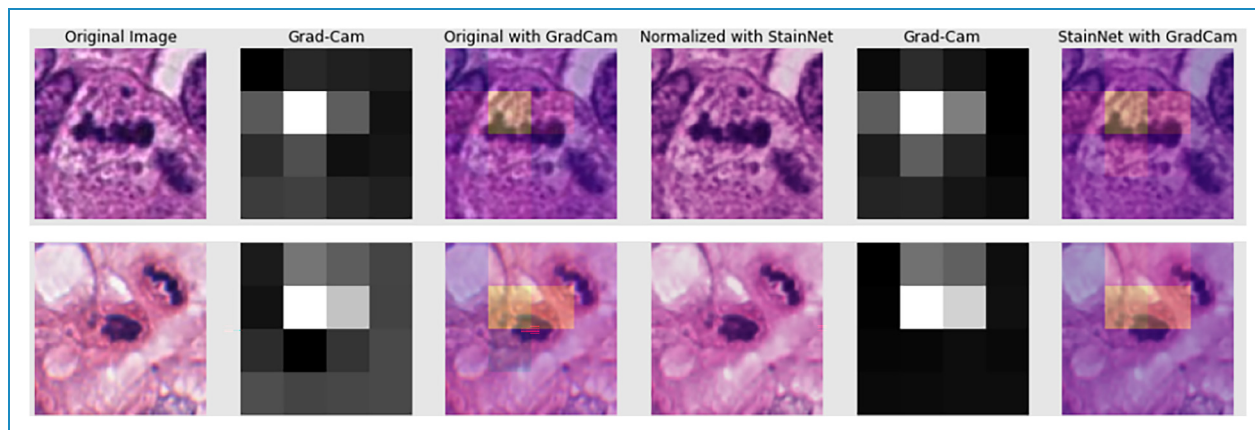


Figure 9 GradCam explanation of the model on the TUPAC'16 dataset. The whiter areas on the GradCam matrix are focused more, whereas the darker ones are less.

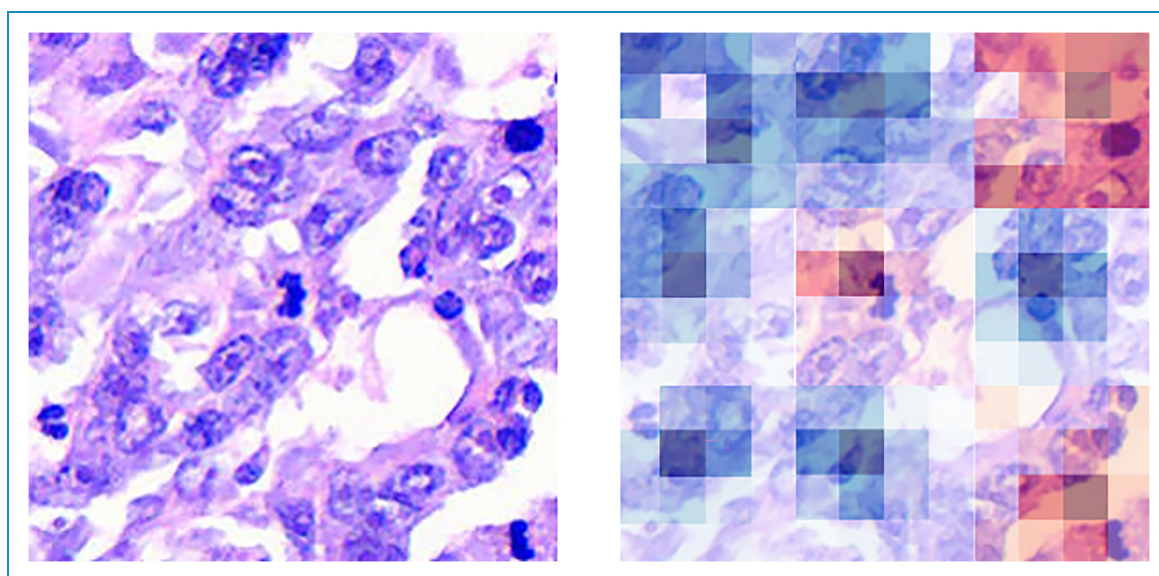


Figure 10 GradCam explanation of the model on a selected region of interest. The image on the left shows the original region. The image on the right shows the prediction and heat map of focus areas. The dark red color represents the focus area in the predicted mitotic patches, while the dark blue color represents the focus areas in the predicted non-mitotic patches.

Conclusion

In this study, we demonstrate that our small model-based classification method can tackle the challenges highlighted and accurately predict mitotic nuclei in two publicly available datasets. Moreover, we illustrate the generalizability and interpretability of our approach across various datasets and clinical settings. This study aims to speed up the adoption of computer-aided digital pathology in clinical settings.

Author contributions: Hasan Farooq, Saira Saleem, and Hammad Naveed contributed to the study conception, design, and analysis. Saira Saleem, Iffat Aleem, Ayesha Iftikhar, and

Umer Nisar Sheikh performed the local data collection. The first draft of the manuscript was written by Hasan Farooq, and all authors commented on the previous versions of the manuscript. All authors read and approved the final manuscript.

Data sharing statement: Data are available on the following link: <https://drive.google.com/drive/folders/1PCHBVjaEy535OXXDD7KvbeWsFPxS0Q2m>.


Declaration of conflicting interests: The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


Ethics approval: The use of local hospital data was approved by the Internal Review Board of Shaukat Khanum Memorial Hospital and Research Center. The study protocol of local data was approved by the institutional review board of Shaukat Khanum Memorial Hospital, and the requirement for informed consent was waived (Approval no. EX-05-01-21-08). The rest of the data is publicly available.

Funding: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the National Center in Big Data and Cloud Computing (NCBC), National Center of Artificial Intelligence (NCAI), and the National University of Computer and Emerging Sciences (NUCES-FAST), Islamabad, Pakistan.

Guarantor: HN.

ORCID iDs: Hasan Farooq  <https://orcid.org/0009-0005-3782-8105>

Saira Saleem  <https://orcid.org/0000-0001-8170-2984>

Hammad Naveed  <https://orcid.org/0000-0002-1867-974X>

References

1. Baak JPA, van Diest Paul J, Voorhorst Feja J, et al. Prospective multicenter validation of the independent prognostic value of the mitotic activity index in lymph nodes—negative breast cancer patients younger than 55 years. *J Clin Oncol* 2005; 23: 5993–6001. PMID: 16135467.
2. Meuten DJ, Moore FM and George JW. Mitotic count and the field of view area: time to standardize. *Vet Pathol* 2016; 53: 7–9. PMID: 26712813.
3. Elmore JG, Longton GM, Carney PA, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *JAMA* 2015; 313: 1122–1132.
4. Chan RCK, To CKC, Cheng KCT, et al. Artificial intelligence in breast cancer histopathology. *Histopathology* 2023; 82: 198–210. <https://onlinelibrary.wiley.com/doi/abs/10.1111/his.14820>
5. Zarella MD, Bowman D, Aeffner F, et al. A practical guide to whole slide imaging: a white paper from the digital pathology association. *Arch Pathol Lab Med* 2019; 143: 222–234.
6. Koteluk O, Wartecki A, Mazurek S, et al. How do machines learn? Artificial intelligence as a new era in medicine. *J Pers Med* 2021; 11: 32.
7. Marletta S, Eccher A, Martelli FM, et al. Artificial intelligence based algorithms for the diagnosis of prostate cancer: a systematic review. *Am J Clin Pathol* 2024; aqad182. doi:10.1093/ajcp/aqad182.
8. Marletta S, L'Imperio V, Eccher A, et al. Artificial intelligence based tools applied to pathological diagnosis of microbiological diseases. *Pathology* 2023; 243: 154362.
9. Caldonazzi N, Rizzo PC, Eccher A, et al. Value of artificial intelligence in evaluating lymph node metastases. *Cancers (Basel)* 2023; 15: 2491.
10. Tek FB. Mitosis detection using generic features and an ensemble of cascade adaboosts. *J Pathol Inform* 2013; 4: 12. <https://www.sciencedirect.com/science/article/pii/S2153353922006332>
11. Sommer C, Fiaschi L, Hamprecht FA, et al. Learning based mitotic cell detection in histopathological images. In: Proceedings of the 21st International Conference on Pattern Recognition, 2012, pp.2306–2309. <https://ieeexplore.ieee.org/document/6460626>.
12. Irshad H. Automated mitosis detection in histopathology using morphological and multi-channel statistics features. *J Pathol Inform* 2013; 4: 10. <https://www.sciencedirect.com/science/article/pii/S2153353922006319>
13. Wang X, Zhang J, Yang S, et al. A generalizable and robust deep learning algorithm for mitosis detection in multicenter breast histopathological images. *Med Image Anal* 2023; 84: 102703. PMID: 36481608.
14. Janowczyk A and Madabhushi A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J Pathol Inform* 2016; 7: 29. <https://www.sciencedirect.com/science/article/pii/S2153353922005478>
15. Aubreville M, Stathonikos N, Bertram CA, et al. Mitosis domain generalization in histopathology images—the MiDoG challenge. *Med Image Anal* 2023; 84: 102699. <https://www.sciencedirect.com/science/article/pii/S1361841522003279>
16. Dusenberry M and Hu F. *Deep learning for breast cancer mitosis detection*. 2018. <https://github.com/CODAIT/deep-histopath/blob/master/docs/tupac16-paper/paper.pdf>.
17. Sohail A, Khan A, Nisar H, et al. Mitotic nuclei analysis in breast cancer histopathology images using deep ensemble classifiers. *Med Image Anal* 2021; 72: 102121. <https://www.sciencedirect.com/science/article/pii/S1361841521001675>
18. Jahanifar M, Shephard A, Zamanitajeddin N, et al. Mitosis detection, fast and slow: robust and efficient detection of mitotic figures. *Med Image Anal* 2024; 94: 103132.
19. Khan HU, Raza B, Shah MH, et al. SMdetector: small mitotic detector in histopathology images using faster R-CNN with dilated convolutions in back-bone model. *Biomed Signal Process Control* 2023; 81: 104414. <https://www.sciencedirect.com/science/article/pii/S1746809422008680>
20. Dhar T, Dey N, Borra S, et al. Challenges of deep learning in medical image analysis—improving explainability and trust. *IEEE Trans Technol Soc* 2023; 4: 68–75.
21. Foucart A, Debeir O and Decaestecker C. Shortcomings and areas for improvement in digital pathology image segmentation challenges. *Comput Med Imaging Graph* 2023; 103: 102155. <https://www.sciencedirect.com/science/article/pii/S089561122001252>
22. Oksuz K, Cam BC, Kalkan S, et al. Imbalance problems in object detection: a review. *IEEE Trans Pattern Anal Mach Intell* 2021; 43: 3388–3415.
23. Zhao Z-Q, Zheng P, Xu S-T, et al. Object detection with deep learning: a review. *IEEE Trans Neural Netw Learn Syst* 2019; 30: 3212–3232.
24. Liu L, Ouyang W, Wang X, et al. Deep learning for generic object detection: a survey. *Int J Comput Vis* 2020; 128: 261–318.
25. Zhang H, et al. Solving missing-annotation object detection with back-ground recalibration loss. 2020. <https://arxiv.org/abs/2002.05274>.

26. Zhang D, Han J, Cheng G, et al. Weakly supervised object localization and detection: a survey. *IEEE Trans Pattern Anal Mach Intell* 2022; 44: 5866–5885
 27. Kondratyuk D, Tan M, Brown M, et al. When ensembling smaller models is more efficient than single large models. 2020. <https://arxiv.org/abs/2005.00570>.
 28. Alzubaidi L, Zhang J, Humaidi AJ, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 2021; 8: 53.
 29. Tan M and Le QV. Efficientnet: Rethinking model scaling for convolutional neural networks. 2019. <https://arxiv.org/abs/1905.11946>.
 30. He K, Zhang X, Ren S, et al. [Deep residual learning for image recognition](#). In: [2016 IEEE Conference on Computer Vision and Pattern Recognition \(CVPR\)](#), Las Vegas, NV, USA, 2016, pp.770–778. doi:10.1109/CVPR.2016.90.
 31. Lei X, Pan H and Huang X. A dilated CNN model for image classification. *IEEE Access* 2019; 7: 124087–124095.
 32. Selvaraju RR, Cogswell M, Das A, et al. [Grad-CAM: visual explanations from deep networks via gradient-based localization](#). In: [2017 IEEE International Conference on Computer Vision \(ICCV\)](#), pp.618–626, doi:10.1109/ICCV.2017.74.
 33. Aubreville M, et al. [Mitosis Domain Generalization Challenge 2022](#). In: [25th International Conference on Medical Image Computing and Computer Assisted Intervention \(MICCAI 2022\)](#). Zenodo, 2022. doi:10.5281/zenodo.6362337.
 34. Kang H, et al. StainNet: A fast and robust stain normalization network. *Front Med (Lausanne)* 2021; 8. <https://www.frontiersin.org/articles/10.3389/fmed.2021.746307>
 35. Veta M, Heng YJ, Stathonikos N, et al. Predicting breast tumor proliferation from whole-slide images: the TUPAC16 challenge. *Med Image Anal* 2019; 54: 111–121. <https://www.sciencedirect.com/science/article/pii/S1361841518305231>
 36. Tang S, Jing C, Jiang Y, et al. The effect of image resolution on convolutional neural networks in breast ultrasound. *Heliyon* 2023; 9: e19253.
 37. Ras G, Xie N, van Gerven M, et al. Explainable deep learning: a field guide for the uninitiated. *J Artif Int Res* 2022; 73: 329–397.
 38. Thanapol P, Lavangnananda K, Bouvry P, et al. [Reducing overfitting and improving generalization in training convolutional neural network \(CNN\) under limited sample sizes in image recognition](#). In: [2020 - 5th International Conference on Information Technology \(InCIT\)](#), 2020, pp.300–305. doi:10.1109/InCIT50588.2020.9310787.
 39. Chen F and Tsou JY. Assessing the effects of convolutional neural network architectural factors on model performance for remote sensing image classification: an in-depth investigation. *Int J Appl Earth Obs Geoinf* 2022; 112: 102865. <https://www.sciencedirect.com/science/article/pii/S156984322200067X>
 40. Ahmed S, Tariq M and Naveed H. PMNet: a probability map based scaled network for breast cancer diagnosis. *Comput Med Imaging Graph* 2021; 89: 101863. <https://www.sciencedirect.com/science/article/pii/S0895611121000112>
 41. Kandel I and Castelli M. The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT Express* 2020; 6: 312–315.
 42. Poojary R and Pai A. [Comparative study of model optimization techniques in fine-tuned CNN models](#). In: [2019 International Conference on Electrical and Computing Technologies and Applications \(ICECTA\)](#), 2019, pp.1–4. doi:10.1109/ICECTA48151.2019.8959681.
 43. Tian Y, Su D, Lauria S, et al. Recent advances on loss functions in deep learning for computer vision. *Neurocomputing* 2022; 497: 129–158.
-