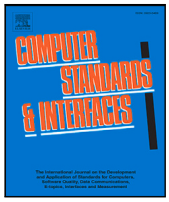
Computer Standards & Interfaces 91 (2025) 103889

Contents lists available at ScienceDirect

Computer Standards & Interfaces

journal homepage: www.elsevier.com/locate/csi



Advancing histopathology in Health 4.0: Enhanced cell nuclei detection using deep learning and analytic classifiers

S. Pons, E. Dura, J. Domingo *, S. Martin

Dpt. of Informatics, ETSE, University of Valencia, Avda. de la Universidad, s/n., Burjasot, 46100, Valencia, Spain

ARTICLE INFO

Keywords:
Histopathological image analysis
Deep learning cell nuclei detection
Histopathological image classifiers

ABSTRACT

This study contributes to the Health 4.0 paradigm by enhancing the precision of cell nuclei detection in histopathological images, a critical step in digital pathology. The presented approach is characterized by the combination of deep learning with traditional analytic classifiers.

Traditional methods in histopathology rely heavily on manual inspection by expert histopathologists. While deep learning has revolutionized this process by offering rapid and accurate detections, its black-box nature often results in a lack of interpretability. This can be a significant hindrance in clinical settings where understanding the rationale behind predictions is crucial for decision-making and quality assurance.

Our research addresses this gap by employing the YOLOv5 framework for initial nuclei detection, followed by an analysis phase where poorly performing cases are isolated and retrained to enhance model robustness. Furthermore, we introduce a logistic regression classifier that uses a combination of color and textural features to discriminate between satisfactorily and unsatisfactorily analyzed images. This dual approach not only improves detection accuracy but also provides insights into model performance variations, fostering a layer of interpretability absent in most deep learning applications.

By integrating these advanced analytical techniques, our work aligns with the Health 4.0 initiative's goals of leveraging digital innovations to elevate healthcare quality. This study paves the way for more transparent, efficient, and reliable digital pathology practices, underscoring the potential of smart technologies in enhancing diagnostic processes within the Health 4.0 framework.

1. Introduction

Histopathology is a medical field that investigates diseases through the microscopic examination of tissue preparations. Tissue samples, primarily obtained through biopsy procedures, are meticulously prepared as thin slices and treated with specific chemical compounds. These compounds react with the organic components found in the targeted cell types, often tumor cells. As a result, these cells, along with their nuclei, are stained with distinctive colors or varying color intensities, enabling their detection by histopathologists.

Histopathological images are captured using specialized cameras that are attached to optical microscopes. These images require high resolution to discern details at a scale of a few micrometers μm. Considering the size of a typical tissue preparation, this high resolution results in exceptionally large images, reaching up to 100.000 pixels per side. These images are referred to as Whole Slide Images (WSI) and can be classified as Gigapixel images due to their enormous size. Each of these Gigapixel images may contain several thousand cells within it.

The detection and classification of abnormal cells, especially tumor cells, are typically carried out by physicians relying on their expertise. Additionally, the identification of cell nuclei is essential for studying their condition and spatial arrangement. In an ideal scenario, all cells should be analyzed to determine the quantity of those of interest and their spatial distribution. However, this is a labor-intensive task for which both time and the availability of trained specialists are limited resources. Typically, the approach is to select a representative region, often based on the specialist's judgment, and conduct a thorough analysis of that chosen area.

Furthermore, interpreting the characteristics of each cell poses challenges, even for well-trained professionals. Variations in staining intensity and color can occur due to random factors during the sample preparation process. Achieving a high degree of consensus among different histopathologists can be a challenging task.

The goal of this work is to improve the detection of cell nuclei in histopathological images done by deep learning models taking into account the outcome of a previous detection model. The method evaluates for which images the system works poorly and modifies the used model to improve its performance. Done that, a procedure to decide which

* Corresponding author.
E-mail address: Juan.Domingo@uv.es (J. Domingo).

https://doi.org/10.1016/j.csi.2024.103889
Received 27 March 2024; Received in revised form 18 June 2024; Accepted 11 July 2024
Available online 14 July 2024

model must be applied to new cases is devised, in the form of a classical logistic regression classifier whose input features are selected. As a side effect, at least certain degree of interpretability is gained. Partial interpretability is not obtained by direct analysis of the inner workings of the neural network, but by examining its outcomes. By doing so, the intention is to understand which interpretable features can account for the differences between two groups of results: the images with good performance and the ones with poor detection rates.

1.1. Relevant work

As stated in Section 1, all steps of histopathological image analysis are labor-intensive. Some steps require a high degree of previous experience and are prone to error due to inter-observer variability, as pointed out in [1]. The different steps, which we now briefly mention, are treated more extensively with appropriate references up to the date of publication in the survey of Hayakawa [2].

The first problem is the reduction of color variability due to the use of different staining protocols, differences in the capture devices (scanners) and variations in thickness of the tissue sections. Even not being the main subject of this paper, the interested reader can find a comprehensive survey in [3].

The second problem is the detection of cell nuclei, their separation and their classification. These tasks have been from long ago the focus of computer-aided techniques. Initially, these techniques relied on classical image processing algorithms, primarily involving color analysis and color normalization (as summarized in Roy et al.'s review [4]). Shape detection was also a component of these methods, with Das et al.'s review [5] specifically focusing on breast cancer. While these algorithms were reasonably effective in terms of detection quality, they have largely been replaced by color analysis, detection, and classification techniques based on deep learning. This shift can be observed in works such as those by Runz et al. [6] and Khened et al. [7]. A new and interesting approach from Huang at al.[8] allows not only simple detection but multi-class detection (i.e.: nuclei of different types) using selected areas of the image to train a sub-network.

Finally, the transition to deep learning model is carefully summarized in the recent review of Basu et al. [9] that explains with detail the different learning architectures and models currently used. The motivation behind this transition to the deep learning lies in the noticeable improvement in result quality, typically measured in terms of sensitivity and accuracy, as shown in [10]. Nevertheless, the combination of visual attributes and appearance of the objects to be detected with usual convolutional networks can be also a suitable approach, as shown in [11] in the wider context of general object detection. In this case, visual features are found by using algorithms specifically tailored to detect them.

Another significant factor is that deep learning can be applied by individuals without specialized knowledge in image processing or strong programming skills. However, it is important to acknowledge that deep learning also comes with significant drawbacks, notably its substantial computational requirements and the lack of interpretability. This means that it is challenging to provide human-understandable explanations for the factors, whether they are low, middle, or high-level features within an image or any sub-image, that determine the degree of success in the task at hand (in this case, nuclei detection). At least some degree of explainability can be gained by visualization techniques like heat maps [12]. Other possibility is the use of the so-called clustering explanations, i.e.: presenting the user the input images forming a group with the most representative images of each category found by the deep learning model so that the user intuitively finds visual resemblances. An example showing 13 classes of tissues with lesions is [13]. Finally, there are saliency techniques, which help identify which weights of a neural network are more relevant in producing a given result. However, these techniques are primarily applied to convolutional networks and can become challenging to interpret when convolutions are applied successively to prior convolutions (as discussed in [14]). Furthermore, saliency techniques are either inapplicable or not yet well-developed for the final, fully-connected non-linear layers that are commonly used in the final stages of deep learning models.

Finally, a third problem is the detection and/or classification of whole portions of the image as different tissues or different stages in the evolution of a tissue. A recent example of the use of deep learning for this task is [15]. This problem is specially prone to the application of transfer-learning techniques, i.e.: the reuse of the weights learned by a previous model as starting point for the learning of a new model. A relevant example in histopathology is [16], in which patches are extracted from Whole Slide Images and fed into a convolutional neural network (CNN) for feature extraction; discriminatory patches are selected and then fed to Efficient-Net architecture pre-trained on the ImageNet dataset.

The conclusion is that, even previous classical methods exist and is worth to use them for auxiliary tasks, the central detection task is currently primarily delegated to deep learning architectures.

2. Materials and methods

This work is primarily methodological in nature. Its main objective is to propose a procedure that can be applied in current cell nuclei detection cases, and as such, it needs to be validated using a well-established set of data that can serve as a benchmark. To achieve this, two collections of manually labeled histopathological images, provided and documented in [17,18], will be used.

The first dataset used in this project is a publicly available dataset [17] consisting of 100 histopathological images of colon cancer, each measuring 500×500 pixels, at 20x magnification. These images represent sections of colon tissue affected by cancer and were obtained using microscopy techniques. This dataset will be referred as DB1 from now on.

The second dataset used in this project contains publicly accessible breast cancer histopathology images [18]. This dataset, consists of 50 images with a size of $512\ x\ 512$ pixels at 40x magnification. This dataset will referred as DB2.

These collections, DB1 and DB2, will be referred to as the "original image sets". Examples of images for both datasets, DB1 and DB2, can be found in Fig. 1. The nuclei are displayed in the images as darker purple spots with irregular shape. The fact of having chosen images of different magnifications is intentional, since this is one of the characteristics that affect most the recognition ability of deep learning systems, and we want to test the adaptability of these systems to different situations.

The overall approach can be outlined as follows:

- Utilize the labeled images from the original image sets, partition them into suitable sets for training, validation, and testing, and employ them to train a deep learning model using the YOLOv5 [19] framework. The trained model will be referred to as "original model" from now on.
- Make a statistical analysis to determine if the performance of the trained model significantly varies across different images. Preliminary findings suggest that there are indeed variations in performance, and visual inspection indicates that issues related to sample preparation or staining may have a substantial influence.
- Extract from the original image datasets those images where the model exhibited the lowest performance and train a new model using these images. Again, divide this set into training, testing, and validation groups. Data augmentation will be necessary since the sets with low performance extracted from each of the original image datasets are relatively small. This newly trained network will be referred to as the "specific model".
- Evaluate the performance of the specific model on the images for which it was specifically trained (along with their augmented set). As will be demonstrated later, the results are expected to be better than those obtained with the original model, as the new model is more specialized.

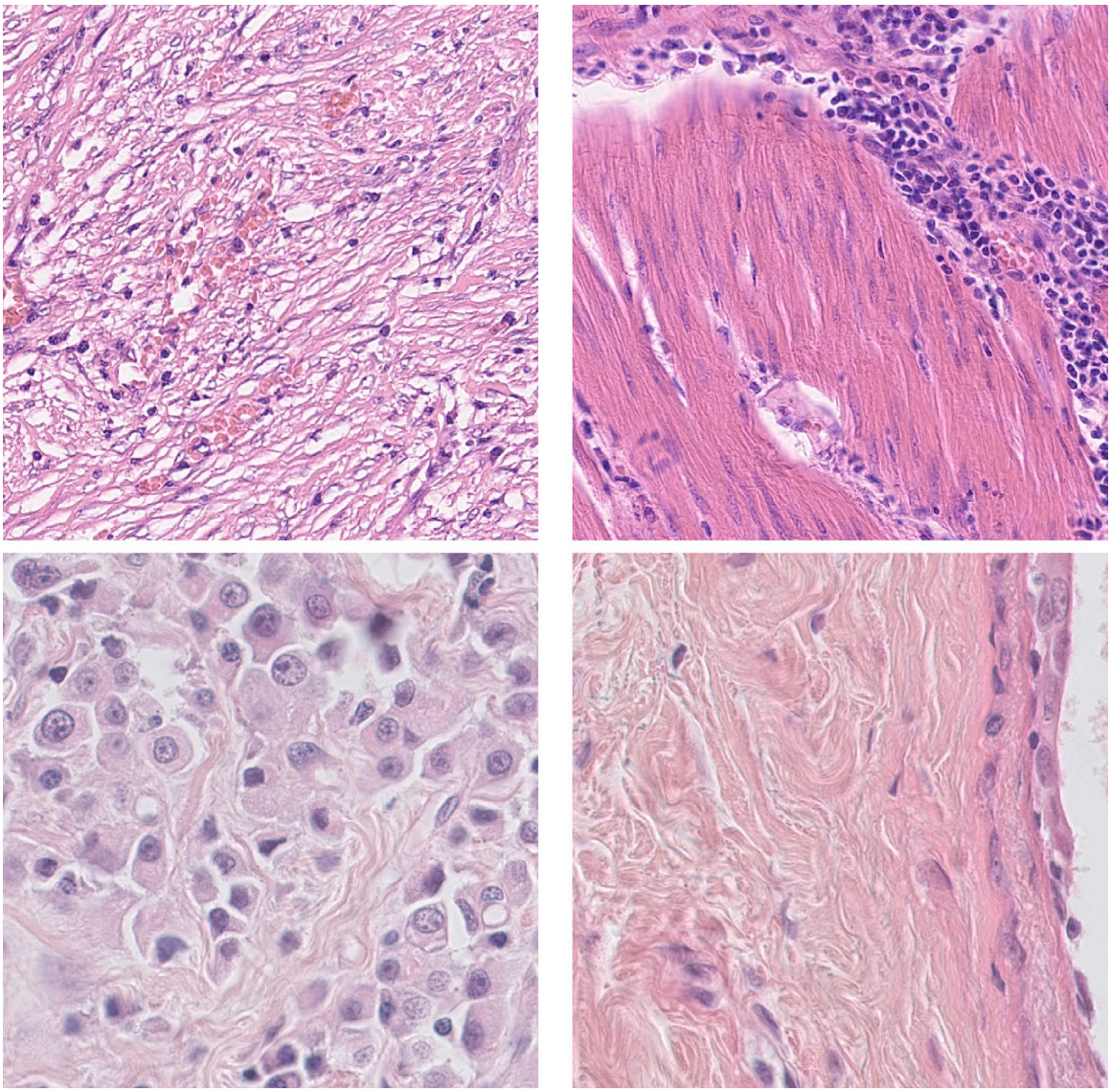


Fig. 1. Typical example of images found for dataset 1 (DB1) in top row and for dataset 2 (DB2) in bottom row. The nuclei are displayed in the images as darker purple spots with irregular shape. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- Extract color and texture descriptors from all the images, and train a classical classifier, in this case, a logistic regression model, using these descriptors. The goal is to classify the original set into two classes: images for which the original model worked sufficiently well and images for which it did not. This will be accomplished using various combinations of color and texture descriptors.

With respect to data augmentation, the procedure was done including the following transformations: random flipping, image rotation between −10 and 10 degrees and image shearing with shearing parameter between −5 and 5. The python package *Albumentation* [20] was used, which can be obtained from https://albumentations.ai/. In addition to its flexibility, this package offers the advantage of being able to transform the ground truth, too, which was necessary for our particular case.

The core concept here is that when presented with a new image for analysis, the logistic model can be applied to predict whether the original or the specific deep learning model should be used. Statistically, this approach is likely to yield better results than relying solely on a single model as the new image is more similar to those images used to train the model that is to be applied to it, either original or specific.

As a collateral but important characteristic, this approach allows us to identify which color and texture descriptors are effective in differentiating normal from low-performance (i.e., more challenging) images. This can give interesting clues to the histopathologists about the most common factors that degrade the quality of the images and suggest ways to detect them automatically.

The following sections will delve into these points in more detail:

Section 2.1 will describe the dataset of images used. Section 2.2 will provide insights into the training of the original model using YOLOv5. Subsequently, Section 3.2 will outline the image descriptors that were tried with the ones ultimately selected to build the logistic regression model. In Section 3, quantitative results from the previous steps will be presented. Finally, Section 4 will summarize the findings and offer suggestions for future work.

2.1. Original images

Two collections of images are used for the experiments. The first one, referred as DB1 previously is a set of 100 images from colorectal adenocarcinoma images, consisting of more than 20,000 annotated nuclei which were collected by Sirinukunwattana et al. and documented

and used in [17]. The second collection, referred as DB2 previously, is composed by 50 images of breast cancer with a total of 4022 annotated cells; it was collected from Naylor et al. reviewing other collections and extending them and was documented and used in [18]. These two sets are publicly available (see URLs in [21] for colorectal adenocarcinoma and in [22] for breast cancer).

2.2. Analysis with YOLOv5

In this study, the YOLOv5 framework, a state-of-the-art object detection technique originally introduced by Redmon [23], is employed to detect cell nuclei. This algorithm utilizes a single convolutional neural network (CNN) to process the entire image, partitioning it into grids and assigning each grid the responsibility of detecting and predicting objects within its defined region. YOLOv5 is available in four distinct versions [24], differing in model size and accuracy, ranging from the compact YOLOv5s to the more extensive YOLOv5x. For this research, we have chosen to use the YOLOv5x model as our benchmark model for training, is the largest and most accurate variant in the YOLOv5 family. It has more layers and parameters to capture fine-grained details as small object detection. Therefore is more suitable for our case as accuracy is prioritized over speed.

3. Results and discussion

3.1. Performance results and analysis

For all experiments carried out with YOLOv5, 600 epochs and 16 batches were used. To evaluate the performance of the models trained four metrics, the accuracy, precision, recall and F1-score, were calculated as:

$$\text{Accuracy} = \frac{TP}{TP+FP+TN+FN} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP+FP} \tag{3}$$

$$\text{F1} - \text{score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

Where TP represent the number of false positives (i.e.: nuclei detected by the model which were marked as such in the manually annotated ground truth), FP the number of false positives, FN the number of false negatives and TN the number of true negatives (which is taken as 0 in this case, since most pixels are not centers of nuclei and therefore are not marked neither by the model not by the histopathologist).

For both datasets, DB1 and DB2, a k-fold cross validation technique was used for evaluating the quality of the models based on test data. Being $k = 10$ for DB1 and $k = 7$ for DB2. The purpose of this, is to provide a more robust and reliable estimate of the model's performance. By repeatedly training and testing the model on different subsets of the data, it generates multiple values of the performance metrics. The average of these values gives a more accurate estimate of how well the model is likely to perform on unseen data. It also avoids bias in the results.

For the first dataset DB1, and for each of the k folding sets, 65 images were randomly selected for training and 25 for validation. The remaining 10 images were used for testing. Likewise for the second dataset DB2: 30 images were randomly selected for training and 13 for validation, leaving the remaining 7 for testing.

In Table 1 the average values for the k tests of the four metrics are displayed for both datasets respectively. It can be appreciated that the overall results are better for DB1 than for DB2.

However to delver deeper into these results a box-whisker plot with the interquartile range (IQR) is used to determine if an image deviate from the expected or common behavior. Such image will be considered as a outlier. Box plots with their IQRs are calculated tested for the four metrics. They are as shown in Figs. 2, 3, 4 and 5 for both datasets, DB1 and DB2.

Table 1
Metrics results for each dataset DB1 and DB2: average Accuracy, average Precision, average Recall and average F1-score.

	Accuracy	Precision	Recall	F1-score
DB1	0.73	0.76	0.96	0.84
DB2	0.66	0.67	0.96	0.79

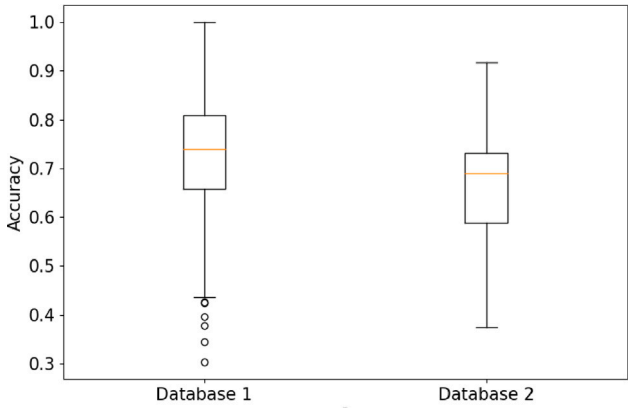


Fig. 2. Box plot diagrams for DB1 and DB2 for Accuracy metric.

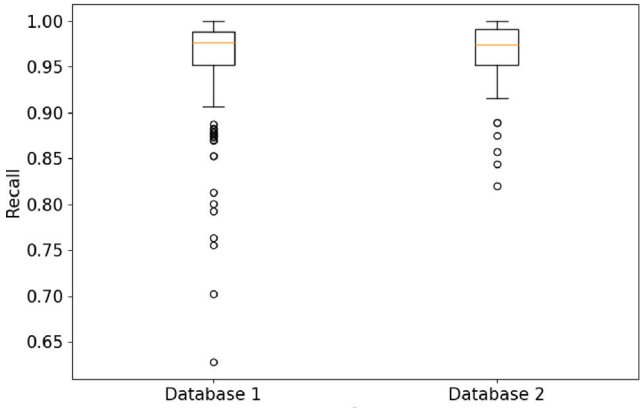


Fig. 3. Box plot diagrams for DB1 and DB2 for Recall metric.

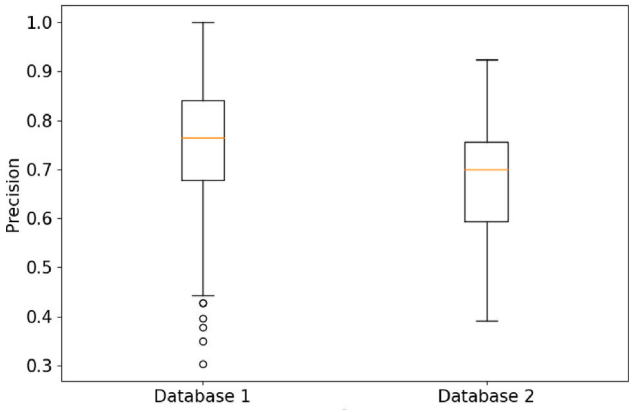


Fig. 4. Box plot diagrams for DB1 and DB2 for Precision metric.

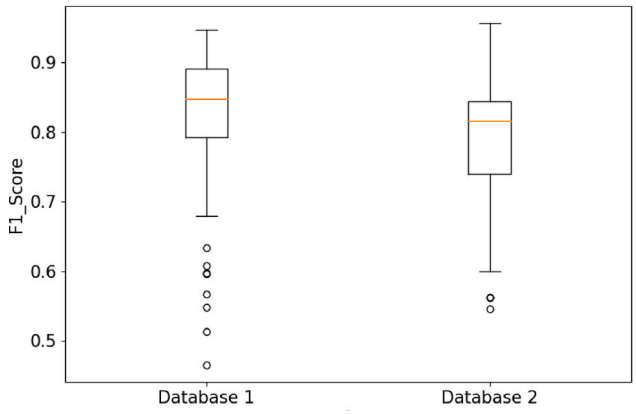


Fig. 5. Box plot diagrams for DB1 and DB2 for F1-score metric.

Table 2
Interquartile range(IQR) for each metric of DB1 and DB2.

	Accuracy	Precision	Recall	F1-score
DB1	0.15	0.16	0.04	0.10
DB2	0.14	0.16	0.04	0.10

The final criterion to declare an image as outlier is that it falls significantly below the lower whisker (i.e.: below 1.5 times the IQR) in the box plot for one or more of the four analyzed metrics. This is the most restrictive criterion (i.e.: a single failure is enough to include the image in the outlier category) so the union of all cases marked as circles in any of the box plots in Figs. 2 to 5 will constitute the set of images with low model performance.

It can be observed that the first dataset has a higher number of outliers for all metrics results. This higher incidence indicates greater variability or dispersion in the obtained results. These outliers represent unusual cases that significantly differ from the general patterns of the results. The presence of more outliers can be attributed to various factors such as the diversity of cell characteristics, image quality, the presence of noise, or the complexity of the samples. These factors are manifest as changes in color and texture, which are further analyzed in Section 3.2. These deviations can pose special challenges for the model and require to train a specific model to improve its results as it will be seen in Section 3.2.

On the other side, it has to be noted that the value of the interquartile range is almost the same in both datasets as can be seen in Table 2. Despite the fact that DB1 has more outliers, the similarity of the interquartile range indicates that the variability of the results, excluding the outliers, is comparable in both datasets. This may suggest similar cell detection characteristics in both datasets, which could be an important finding in the analysis of the results.

In DB1, 19 out of its 100 images were outliers for at least one of the four metrics. In DB2, 6 out of its 50 images were outliers with the same criterion. This indicates that, apart of using simply the mean values or even mean plus typical deviations, a finer analysis must be done to gain insight of which images specifically lower the performance of the model.

Examples of both types of images (normal and outliers) for both datasets are shown in Section 3.2.

3.2. Statistical analysis and classical classifiers

The findings given in Section 3.1, even being globally acceptable in terms of accuracy, precision and recall, reveal the presence of low-performance cases (outliers) as shown in Figs. 2, 3, 4 and 5.

These cases(outliers) far below the average performance have a two-fold effect: firstly a new case resembling these outliers is likely to yield similarly poor results. Secondly they have contributed to the network training process, which degrades the overall performance of the model, even for the good cases.

Table 3
Value of metrics for DB1 with "original model" and "specific model" respectively for images with lower performance: average accuracy, average precision, average Recall and F1-score.

	Accuracy	Precision	Recall	F1-score
DB1 (orig.model)	0.59	0.62	0.94	0.73
DB1 (spec.model)	0.67	0.71	0.94	0.80

Table 4
Value of metrics for DB1 with "original model" and "specific model" respectively for images with higher performance: average accuracy, average precision, average Recall and F1-score.

	Accuracy	Precision	Recall	F1-score
DB1 (orig.model)	0.73	0.75	0.97	0.84
DB1 (spec.model)	0.72	0.75	0.92	0.83

Table 5
Value of metrics for DB2 with "original model" and "specific model" respectively for images with lower performance: average accuracy, average precision, average Recall and F1-score.

	Accuracy	Precision	Recall	F1-score
DB2 (orig.model)	0.60	0.63	0.94	0.74
DB2 (spec.model)	0.73	0.74	0.98	0.83

Table 6
Value of metrics for DB2 with "original model" and "specific model" respectively for images with higher performance: average accuracy, average precision, average Recall and F1-score.

	Accuracy	Precision	Recall	F1-score
DB2 (orig.model)	0.71	0.73	0.96	0.83
DB2 (spec.model)	0.67	0.73	0.88	0.80

Hence, our strategy is based on dividing the original data set into two subsets: one containing the images with acceptable performance and the other containing the images (outliers) with lower performance.

Then, a new deep learning model (formerly called the "specific model") is trained using YOLOv5, with the subset containing the images with lower performance. This new model retained the same network structure and size as in the original model. However, there is a difference in the number of low-performance images which was much smaller than the original sample size. Specifically, this category includes 19 out of the 100 images in DB1 and 6 out of the 50 images in DB2 as was indicated in Section 3.1. This is insufficient for appropriate training requiring the use of data augmentation techniques. To address this lack of samples, 100 images were generated from the 19 in DB1 and 50 from the 6 in DB2 using data augmentation. This choice aimed to keep the same sizes of each original sample for which these type of models exhibited reasonable results.

Both, the new trained "specific model" and the "original model" are now applied to both subsets of images for each dataset: those with higher performance from the original experiment and those with lower values. Results for these two models applied to each subset of each dataset can be found in Tables 3, 4, 5 and 6.

The results in Tables 3 and 5 show that the model trained specifically with the lower-performance subset outperform the original model

when it is applied to its corresponding type of images. Also, Tables 4 and 6 show that this model slightly degrades its performance when applied to the upper-performance subset. This outcome is not surprising since more homogeneous training data lead to more accurate models.

The next step involves delving into why certain images perform better in order to provide some insights useful for the histopathologists about which causes provoke some images being poorly classified. This will be achieved by training a simple classical classifier, specifically a logistic regression model. This model will take features related to color and texture as input and effectively segregate both subsets.

The choice of the features to classify the images is based on previous experience and guidelines valid for the general field of image description and image recovery. A good reference to the general context is [25]. Out final selection has been:

- As color features, the values of the bins of the normalized color histogram of the Hue coordinate, independently of the intensities and saturations. This histogram divided the full Hue coordinate range into 10 bins.
- As texture features the following ones were tested:
  - The Gabor energy features, described in [26]. This method calculates the integral of the convolution of the image with a Gabor filter, which is a Gaussian filter multiplied by a 2D-sinus function. Its parameters are the Gaussian support size (which was taken as 9 pixels) and its σ value (taken as 0.5). Gaussian filters are scaled (in our case, by factors 1, 2 and 3) and sinus function is rotated through several orientations (in our case, 0°, 45°, 90° and 135°). This gives a total of 12 descriptors (product of number of magnifications and number of orientations).
  - The Gaussian Markov Random Field described in [27]. This method assumes that the neighborhood of each pixel can be modeled as a Markov random field, i.e. there is a linear dependency between the value at each point and its neighbor values and the residuals of the fitting to the model are approximately Gaussian. Descriptors are the estimators of the model parameters. The only parameter required is the size of the local neighborhood to adjust the model, which was taken as 4 pixels.
  - The values of the morphological granulometry distribution function with a horizontal linear structuring element. A detailed description of the granulometry can be found in [28]. As a simple definition, a granulometry is the proportion of the area of an image that remains after a morphological opening with a given structuring element of certain size. It intuitively represents the area of the shapes that fully contain the structuring element. As the element gets larger the shapes are progressively eliminated so a granulometry starts at 1 (full image remains) for 0-size structuring elements and ends at 0 (image completely vanishes) for a size of the structuring element bigger than the biggest shape present in the image. For this experiment the element starts at size 2 pixels and reaches 20 pixels in intervals of 2.
  - The coefficients of the expansion of the former function in a b-spline basis of functions. Since the former descriptor is a function, it is common to smooth it by projection on a function basis, being the most common the polynomial b-splines. The number of basis functions was taken as 5.
  - The values of the morphological granulometry with a vertical linear structuring element. This is as before by changing only the structuring element. It is a common practise when using linear structuring elements to take always at least two perpendicular orientations to take into account differently oriented shapes.
  - The coefficients of the expansion of the former function in a b-spline basis of functions, for the same reason as before.
  - The local spatial size distribution, as described in [29]. This is an improvement of the simple granulometry that calculates its correlation with a version of itself displaced to all points of a neighborhood of each point. It is more robust and rotationally invariant. The structuring elements were taken as before in size and limits and the neighborhood was taken as a disk with radius 4 pixels.

This makes a total of eight sets of descriptors, comprising the color descriptor and seven texture descriptors. Each of these descriptors can be either included or excluded, leading to a collection of $2^8 = 256$ possibilities. For both sets of images, specifically colon and breast cancer images, a logistic classifier was trained for each of these 256 possible combinations.

Due to the relatively small sample size (100 images for colorectal cancer and 50 for breast cancer), a bootstrap approach was employed. In this approach, a logistic classifier is trained using all images except one, and then applied to the remaining image. The classifier's performance is assessed by counting the number of correctly classified remaining images. Despite the computational intensity of this task, given the substantial number of models to be trained and tested, the procedure remains entirely feasible. Each model is fitted using a function available in the R programming language, taking just a few milliseconds for completion.

From the pool of 256 possibilities, we selected those that achieved perfect classification results, meaning that all remaining images in the bootstrap were correctly assigned for both sets of images. It is worth to note that color descriptors were included in all of these selected combinations, although color alone did not yield perfect classification. Finally, from the possibilities with perfect classification, the combination with fewer descriptors is selected. This combination comprised the color histogram, Gabor energies, granulometry values using a vertical linear structuring element, and corresponding values obtained with a horizontal structuring element.

Graphical examples with the nuclei detected in terms of TP, FP and FN are shown in Figs. 6, 7, 8 and 9.

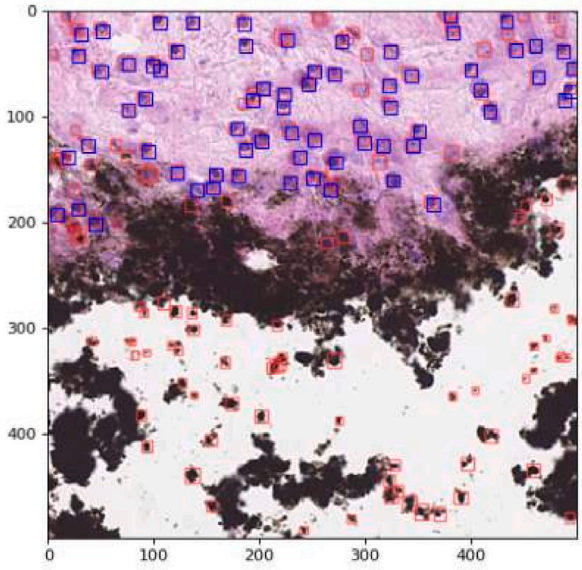
Fig. 6 which corresponds to DB1 (colon cancer) shows a typical example of an over-stained area that confuses the network, which interprets isolated spots outside the tissue as cell nuclei whereas Fig. 7 is one example of a normal case in which most nuclei are correctly detected. In this second case the texture of the image is rather uniform and the color is also similar to what it is expected for a slide appropriately prepared and stained.

Fig. 8 taken from DB2 (breast cancer) is again an example of poor detection with many false positives and false negatives but in this case the errors are mostly caused by the lack of sufficient contrast between the nuclei and the rest of the cell. Conversely, Fig. 9 shows nuclei highly contrasted against the background and therefore easily detected which puts this image in the group of those appropriate for the original model.

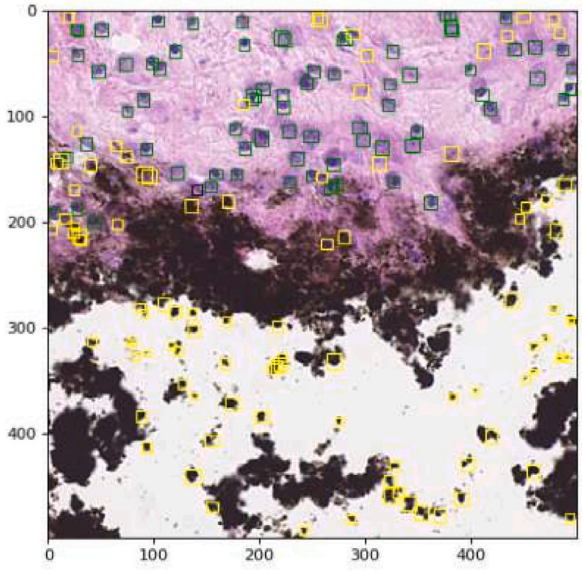
The choice of low level features found by the systematic tests done with all possible combinations of features seems now sensible looking at the examples: sometimes as in the case of Fig. 6 both, texture and color are the most relevant factors. Also, and for this particular type of images, texture is not oriented with micro-borders in any privileged direction, so granulometries with vertical and horizontal linear structuring elements are both important. In the second case of Fig. 8 the reason for the poor performance is the lack of contrast, captured by the features related with color distribution.

4. Conclusions

Digital histopathology has reached a substantial degree of maturity. Automatic aids are now common to help the work of histopathologists. Tools for automatic detection and classification are highly valuable in diagnosis and prognosis but their blind application is discouraged, as in any other field of medicine, due to the harmful outcomes they may



Ground truth data (blue square) vs detection from YOLOv5 (red square)



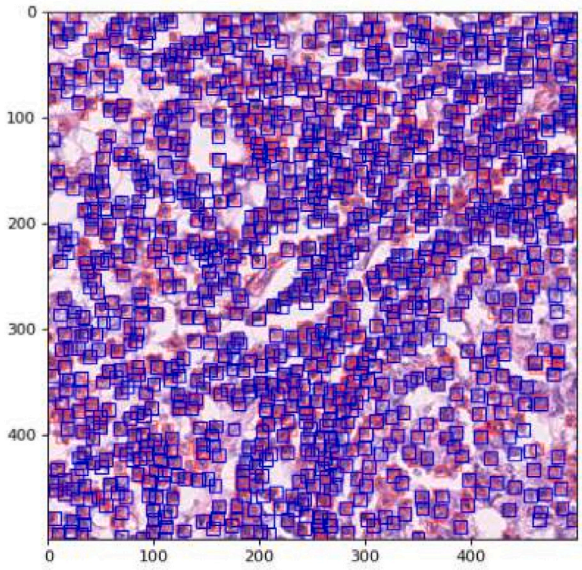
True positives (green squares), False positives (yellow squares) and False negatives (black squares)

Fig. 6. Example of image from the DB1 (breast cancer) where a low performance was obtained by the original model trained with YOLOv5. On the top (a) the ground truth data vs the detected by the original model is displayed. On the bottom (b) a detailed detection is given in terms of True positives, False positives and False negatives.

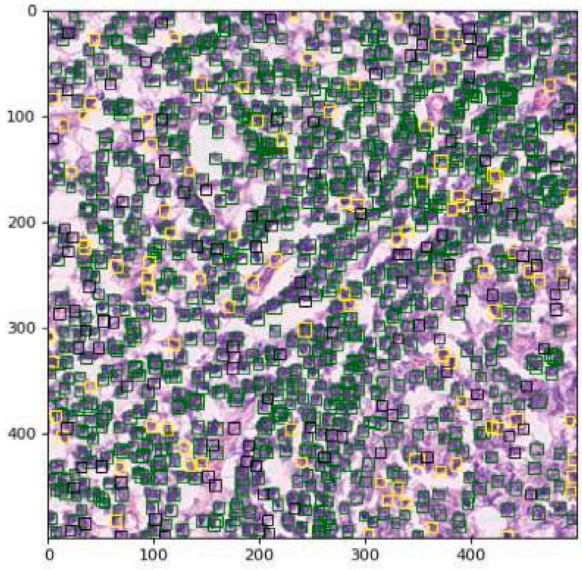
provoke. The declared objective of Health 4.0, to provide patients with better, more value-added, and more cost-effective healthcare services while also improving the industry's efficacy and efficiency cannot be obtained without tools that not only automate the work but also provide means to assess of the work done. Since final supervision of a specialist is always necessary, providing tools to help doctors in routine tasks like previous image classification is a must.

This work contributes to such task by providing a method to tune a previously trained deep learning model: the outcome of the model shows in which specific types of images is not performing well and guides the choice of the sample to train a better model, specifically tuned to improve the cases of poor performance. This fact improves the behavior of the deep learning technique on the training set, but it is also used to train a classical logistic classifier that discriminates which model should be used for a new input image, thus establishing a fast initial screening. Moreover, by visual observation of which new cases are classified as more prone to worse results, the histopathologists can determine the reasons to such behavior (bad staining techniques, staining errors in a particular batch, ...) and propose corrective actions for the future.

Also as further work we will consider the possibility of using not only two but several models. This can be done either by dividing the dataset into two groups, use a logistic regression and repeat the process several times learning a sort of nested models, each of which refines the former one or by choosing a number of groups and training a multinomial classifier in a single step.



Ground Truth Data (blue square) vs detection from YOLOv5 (red square)



True positives (green squares), False positives (yellow squares) and False negatives (black squares)

Fig. 7. Example of image from DB1 (breast cancer) where a good performance was obtained by the original model trained with YOLOv5. On the top (a) the ground truth data vs the detected by the original model is displayed. On the bottom (b) a detailed detection is given in terms of True positives, False positives and False negatives.

Also, it would be interesting to evaluate several user's explanation, in this case pathologists, for the classification made by the model as worse and better performance images. This could be done through interviews, questionnaires...etc. It would improve the trustworthiness of the system.

CRediT authorship contribution statement

S. Pons: Writing – original draft, Validation, Software, Investigation, Conceptualization. **E. Dura:** Writing – review & editing, Investigation, Formal analysis, Conceptualization. **J. Domingo:** Writing – review & editing, Software, Methodology, Funding acquisition, Conceptualization. **S. Martin:** Validation, Investigation, Data curation.

Declaration of competing interest

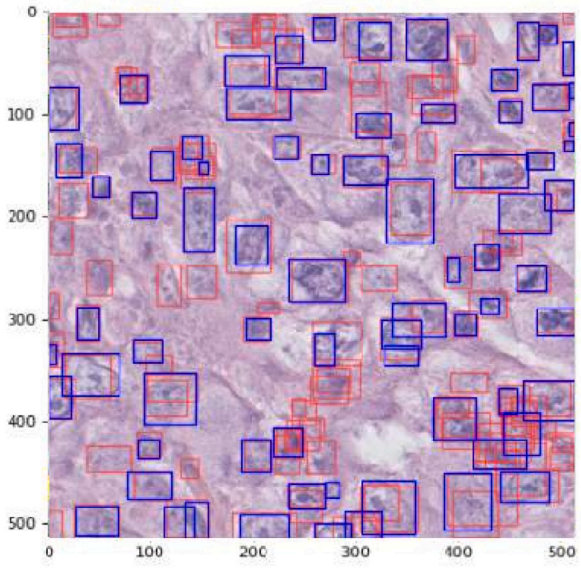
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

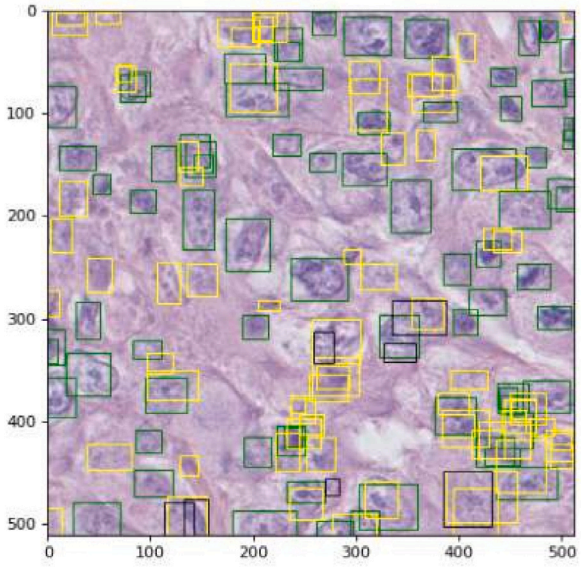
No data was used for the research described in the article.

Acknowledgments

This publication is part of the I+D+i project PGC type B with reference PID2020-117114GB-I00 funded by the Spanish Ministry of Science, Innovation and Universities, MCIN/AEI/10.13039/501100011033/.

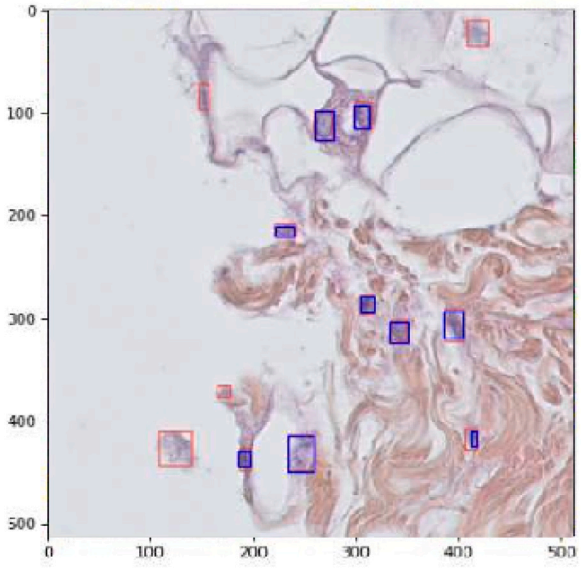


Ground Truth Data (blue square) vs detection from YOLOv5 (red square)

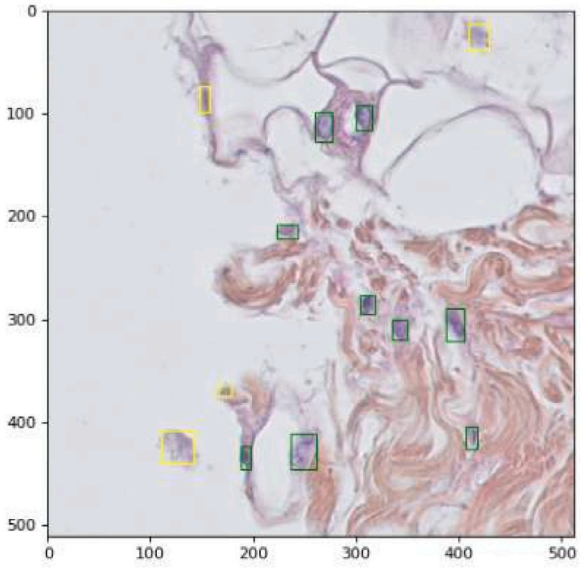


True positives (green squares), False positives (yellow squares) and False negatives (black squares)

Fig. 8. Example of image from DB2 (colon cancer) where a low performance was obtained by the original model trained with YOLOv5. On the top (a) the ground truth data vs the detected by the original model is displayed. On the bottom (b) a detailed detection is given in terms of True positives, False positives and False negatives.



Ground Truth Data (blue square) vs detection from YOLOv5 (red square)



True positives (green squares), False positive (yellow squares) and False negatives (black squares)

Fig. 9. Example of image from DB2 (colon cancer) where a good performance was obtained by the original model trained with YOLOv5. On the top (a) the ground truth data vs the detected by the original model is displayed. On the bottom (b) a detailed detection is given in terms of True positives, False positives and False negatives.

References

[1] C. Kang, C. Lee, H. Song, M. Ma, S. Pereira, Variability matters: Evaluating inter-rater variability in histopathology for robust cell detection, in: L. Karlinsky, T. Michaeli, K. Nishino (Eds.), Computer Vision – ECCV 2022 Workshops, Springer Nature Switzerland, Cham, 2023, pp. 552–565.

[2] T. Hayakawa, V.B.S. Prasath, H. Kawanaka, B.J. Aronow, S. Tsuruoka, Computational nuclei segmentation methods in digital pathology: A survey, Arch. Comput. Methods Eng. 28 (1) (2021) 1–13, http://dx.doi.org/10.1007/s11831-019-09366-4.

[3] T.A. Azevedo Tosta, P.R. de Faria, L.A. Neves, M.Z. do Nascimento, Computational normalization of H&E-stained histological images: Progress, challenges and future potential, Artif. Intell. Med. 95 (2019) 118–132, http://dx.doi.org/10.1016/j.artmed.2018.10.004, URL https://www.sciencedirect.com/science/article/pii/S093336571830424X.

[4] S. Roy, A. kumar Jain, S. Lal, J. Kini, A study about color normalization methods for histopathology images, Micron 114 (2018) 42–61, http://dx.doi.org/10.1016/j.micron.2018.07.005, URL https://www.sciencedirect.com/science/article/pii/S0968432818300982.

[5] A. Das, M.S. Nair, S.D. Peter, Computer-aided histopathological image analysis techniques for automated nuclear atypia scoring of breast cancer: a review, J. Dig. Imag. 33 (5) (2020) 1091–1121, http://dx.doi.org/10.1007/s10278-019-00295-z.

[6] M. Runz, D. Rusche, S. Schmidt, M.R. Weihrauch, J. Hesser, C.-A. Weis, Normalization of HE-stained histological images using cycle consistent generative adversarial networks, Diagnost. Pathol. 16 (1) (2021) 71, http://dx.doi.org/10.1186/s13000-021-01126-y.

[7] M. Khened, A. Kori, H. Rajkumar, G. Krishnamurthi, B. Srinivasan, A generalized deep learning framework for whole-slide image segmentation and analysis, Sci. Rep. 11 (1) (2021) 11579, http://dx.doi.org/10.1038/s41598-021-90444-8.

[8] J. Huang, H. Li, X. Wan, G. Li, Affine-consistent transformer for multi-class cell nuclei detection, in: 2023 IEEE/CVF International Conference on Computer Vision, ICCV, IEEE Computer Society, Los Alamitos, CA, USA, 2023, pp. 21327–21336, http://dx.doi.org/10.1109/ICCV51070.2023.01955, URL https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.01955.

[9] A. Basu, P. Senapati, M. Deb, R. Rai, K.G. Dhal, A survey on recent trends in deep learning for nucleus segmentation from histopathology images, Evol. Syst. 15 (1) (2024) 203–248, http://dx.doi.org/10.1007/s12530-023-09491-3.

[10] H. Chen, X. Qi, L. Yu, Q. Dou, J. Qin, P.-A. Heng, DCAN: Deep contour-aware networks for object instance segmentation from histology images, Med. Image Anal. 36 (2017) 135–146, http://dx.doi.org/10.1016/j.media.2016.11.004, URL https://www.sciencedirect.com/science/article/pii/S1361841516302043.

[11] M.J. Khan, A. Zafar, V. Tumanian, D. Yue, G. Li, Object detection boosting using object attributes in detect and describe framework, in: 2019 IEEE 31st International Conference on Tools with Artificial Intelligence, ICTAI, 2019, pp. 886–893, http://dx.doi.org/10.1109/ICTAI.2019.00126.

[12] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, Int. J. Comput. Vis. 128 (2) (2020) 336–359, http://dx.doi.org/10.1007/s11263-019-01228-7.

[13] K. Faust, Q. Xie, D. Han, K. Goyle, Z. Volynskaya, U. Djuric, P. Diamandis, Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction, BMC Bioinformat. 19 (1) (2018) 173, http://dx.doi.org/10.1186/s12859-018-2184-4.

[14] Q. Zhang, Y.N. Wu, S.-C. Zhu, Interpretable convolutional neural networks, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8827–8836, http://dx.doi.org/10.1109/CVPR.2018.00920.

[15] Y. Xue, J. Ye, Q. Zhou, L.R. Long, S. Antani, Z. Xue, C. Cornwell, R. Zaino, K.C. Cheng, X. Huang, Selective synthetic augmentation with histogan for improved histopathology image classification, Med. Image Anal. 67 (2021) 101816, http://dx.doi.org/10.1016/j.media.2020.101816, URL https://www.sciencedirect.com/science/article/pii/S1361841520301808.

[16] N. Ahmad, S. Asghar, S.A. Gillani, Transfer learning-assisted multi-resolution breast cancer histopathological images classification, Vis. Comput. 38 (8) (2022) 2751–2770, http://dx.doi.org/10.1007/s00371-021-02153-y.

[17] K. Sirinukunwattana, S.E.A. Raza, Y.-W. Tsang, D.R.J. Snead, I.A. Cree, N.M. Rajpoot, Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images, IEEE Trans. Med. Imaging 35 (5) (2016) 1196–1206, http://dx.doi.org/10.1109/TMI.2016.2525803.

[18] P. Naylor, M. Laé, F. Reyal, T. Walter, Segmentation of nuclei in histopathology images by deep regression of the distance map, IEEE Trans. Med. Imaging 38 (2) (2019) 448–459, http://dx.doi.org/10.1109/TMI.2018.2865709.

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

[20] A. Buslaev, V.I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, A.A. Kalinin, Albumentations: Fast and flexible image augmentations, Information 11 (2) (2020) http://dx.doi.org/10.3390/info11020125, URL https://www.mdpi.com/2078-2489/11/2/125.

[21] K. Sirinukunwattana, S.E.A. Raza, Y.-W. Tsang, D.R.J. Snead, I.A. Cree, N.M. Rajpoot, Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images, 2016, URL https://warwick.ac.uk/fac/cross_fac/tia/data/crchistolabelednucleihe.

[22] N.P. Jack, W. Thomas, L. Marick, R. Fabien, Segmentation of nuclei in histopathology images by deep regression of the distance map, 2018, http://dx.doi.org/10.5281/zenodo.1175282, We would like to thank Ligue Nationnal contre le Cancer for funding my PhD..

[23] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 779–788, http://dx.doi.org/10.1109/CVPR.2016.91.

[24] G. Jocher, Ultralytics/yolov5: v3.1 - bug fixes and performance improvements, 2020, http://dx.doi.org/10.5281/zenodo.4154370, https://github.com/ultralytics/yolov5.

[25] Y. Rubner, C. Tomasi, Perceptual Metrics for Image Database Navigation, in: The Kluwer international series in engineering and computer science. Robotics: vision, manipulation and sensors, Springer US, 2013, URL https://link.springer.com/book/10.1007/978-1-4757-3343-3.

[26] I. Fogel, D. Sagi, Gabor filters as texture discriminator, Biol. Cybernet. 61 (2) (1989) 103–113, http://dx.doi.org/10.1007/BF00204594.

[27] R. Chellappa, S. Chatterjee, Classification of textures using Gaussian Markov random fields, IEEE Trans. Acoust. Speech Signal Process. 33 (4) (1985) 959–963, http://dx.doi.org/10.1109/TASSP.1985.1164641.

[28] P. Soille, Morphological Image Analysis, second ed., Springer Berlin, Heidelberg, 2002, http://dx.doi.org/10.1007/978-3-662-05088-0.

[29] G. Ayala, J. Domingo, Spatial size distributions: Applications to shape and texture analysis, IEEE Trans. Pattern Anal. Mach. Intell. 23 (12) (2001) 1430–1442, http://dx.doi.org/10.1109/34.977566.