Medical Image Analysis 73 (2021) 102167

Contents lists available at ScienceDirect

Medical Image Analysis

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

Computer-aided diagnosis tool for cervical cancer screening with weakly supervised localization and detection of abnormalities using adaptable and explainable classifier

Antoine Pirovano[a,b,*], Leandro G. Almeida[a], Said Ladjal[b], Isabelle Bloch[b,c], Sylvain Berlemont[a]

[a] Keen Eye, 74 rue du Faubourg Saint-Antoine, Paris 75012, France
[b] LTCI, Telecom Paris, Institut Polytechnique de Paris, 19 Place Marguerite Perey, Palaiseau 91120, France
[c] Sorbonne Université, CNRS, LIP6, Paris, France

ARTICLE INFO

Article history:
Received 12 June 2020
Revised 28 June 2021
Accepted 7 July 2021
Available online 9 July 2021

Keywords:
Classification
Cytology
Explainability
Whole slide images
Localization
Detection
Saliency
Weakly supervised learning

ABSTRACT

While pap test is the most common diagnosis methods for cervical cancer, their results are highly dependent on the ability of the cytotechnicians to detect abnormal cells on the smears using brightfield microscopy. In this paper, we propose an explainable region classifier in whole slide images that could be used by cyto-pathologists to handle efficiently these big images (100,000x100,000 pixels). We create a dataset that simulates pap smears regions and uses a loss, we call classification under regression constraint, to train an efficient region classifier (about 66.8% accuracy on severity classification, 95.2% accuracy on *normal*/*abnormal* classification and 0.870 KAPPA score). We explain how we benefit from this loss to obtain a model focused on sensitivity and, then, we show that it can be used to perform weakly supervised localization (accuracy of 80.4%) of the cell that is mostly responsible for the malignancy of regions of whole slide images. We extend our method to perform a more general detection of abnormal cells (66.1% accuracy) and ensure that at least one abnormal cell will be detected if malignancy is present. Finally, we experiment our solution on a small real clinical slide dataset, highlighting the relevance of our proposed solution, adapting it to be as easily integrated in a pathology laboratory workflow as possible, and extending it to make a slide-level prediction.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

The World Health Organization (WHO) states (WHO, 2014) that around 90% of cervical cancers could be avoided if they were detected and treated earlier. With around 500×10^3 new cases each year, screening for cervical cancer needs to be efficient and precise. Currently, pap smear (Papanicolaou and Traut, 1943) is the most commonly used diagnosis method for cervical cancer screening. It is perfomed by a visual check of squamous epithelial cells scratched at the joint section between the cervix and the uterus (the Transformation Zone), which are set inside a preservative liquid, stained using Hematoxylin and Eosin (H&E) and spread on a slide. This slide is then analyzed by a trained anapathologist or a cytotechnician, navigating the full slide through a cornucopia of cells (up to 100×10^3) in order to find potentially, high risk, pre-cancerous cytomorphologic changes (these changes can happen up to 7 years before an actual cancer develops), revealing cancerous lesions. A positive classification (classified as *abnormal*) will lead to further analysis (Wright et al., 2002). Although they are recognized as being an efficient method, these tests are highly dependent on the expertise of the cytotechnician to localize abnormal cells on smears slides using brightfield microscopy. This task, which is like looking for a needle in a haystack, leads to several drawbacks such as missed abnormalities which cause false negative cases, extra work on false positive cases and fatigue of cytotechnicians. Furthermore, about 93% of smears are categorized as normal (negative), so there is a need to prune most negative cases while keeping a sensitivity as close to 100% as possible in order to enable practitioners to focus on difficult and abnormal cases.

Also, despite the efforts to standardize the methodology to classify correctly slides and ensure reproductibility, there is still a high inter-observers variability in diagnosis (Stoler and Schiffman, 2001; Sherman et al., 2007)). It is not clearly defined how anapathologists should proceed and what they should rely on to make their

* Corresponding author at: Keen Eye, 74 rue du Faubourg Saint-Antoine, 75012 Paris, France.
E-mail address: antoine.pirovano@keeneye.tech (A. Pirovano).

https://doi.org/10.1016/j.media.2021.102167
1361-8415/© 2021 Elsevier B.V. All rights reserved.

decision. Latest guidelines (Solomon et al., 2002; Nayar and Wilbur, 2015) indicate that the size, shape and texture of the nucleus of a cell are essential discriminative features, along with the ratio between the nucleus and the cytoplasm sizes, also called Nucleo-Cytoplasmic Ratio (NCR).

With the recent emergence of deep learning methods, specifically deep Convolutional Neural Networks (CNN) succeeding on a large panel of tasks, it has been a growing area of research to use and adapt these methods to assist medical doctors in diagnosis, prognosis and medical procedures (Bar et al., 2018; Liu et al., 2017; Ronneberger et al., 2015; Naylor et al., 2019). In the case of cervical cancer screening, the Herlev dataset (Jantzen et al., 2005) enables researchers to compare their methods regarding cells classification. This dataset is composed of 917 images showing single cells, categorized using the seven labels of the WHO classification: *normal columnar, normal intermediate, normal superficial, light dysplastic, moderate dysplastic, severe dysplastic* and *carcinoma in situ*. The first three categories belong to the category of *normal* cells and the last four are *abnormal* (in order of severity, with *carcinoma in situ* hinting the presence of an actual cancer).

After reviewing related work in Section 2, we describe, in Section 3, the loss we propose to train a CNN that avoids critical mistakes. We will see that the proposed method actually performs very well in predicting cell malignancy. In Section 4, we make a step closer to Whole Slide Image (WSI) classification and medical support by taking advantage of tools developed in Section 3 to perform an adapted training on images we create and that simulate pap smears tiles. We also show that, thanks to an attribution method, we can perform weakly supervised localization of the cell reponsible for the predicted label along with weakly supervised detection of abnormal cells, which might help medical practioners to understand the outcome of our method, strenghten their confidence in the model, and decrease the time spent on each slide. In Section 5, we use a 90 WSIs dataset to apply our methods and demonstrate the clinical and practical usefulness of our work.

2. Related work

Since 2012 and the success of AlexNet (Krizhevsky et al., 2012) on Imagenet Challenge (Deng et al., 2009), deep learning has been considered as a revolution in the field of computer vision, reaching state-of-the-art performances for almost all tasks on which it has been applied, e.g. Natural Language Processing using Recurrent Neural Networks (Cho et al., 2014), Gaming using Reinforcement Learning (Mnih et al., 2015).

In image processing in particular, CNN architectures, first introduced in LeCun et al. (1989), perform really well. And, over the years, several architectures have emerged. Currently, Resnet-101 (He et al., 2016), which proposes the use of skip connections over blocks to avoid de-learning on latest blocks what has been learned on early blocks, is acknowledged to be one of the best architectures for classification and serves as the core of various derived architectures and tasks.

With this growing interest in deep learning, cervical cancer screening has been identified as a high stake subject that requires to tackle several problems: having efficient classifiers (up to cell level), define relevant features, and standardize the process that leads to a slide label and being able to analyze quickly huge images.

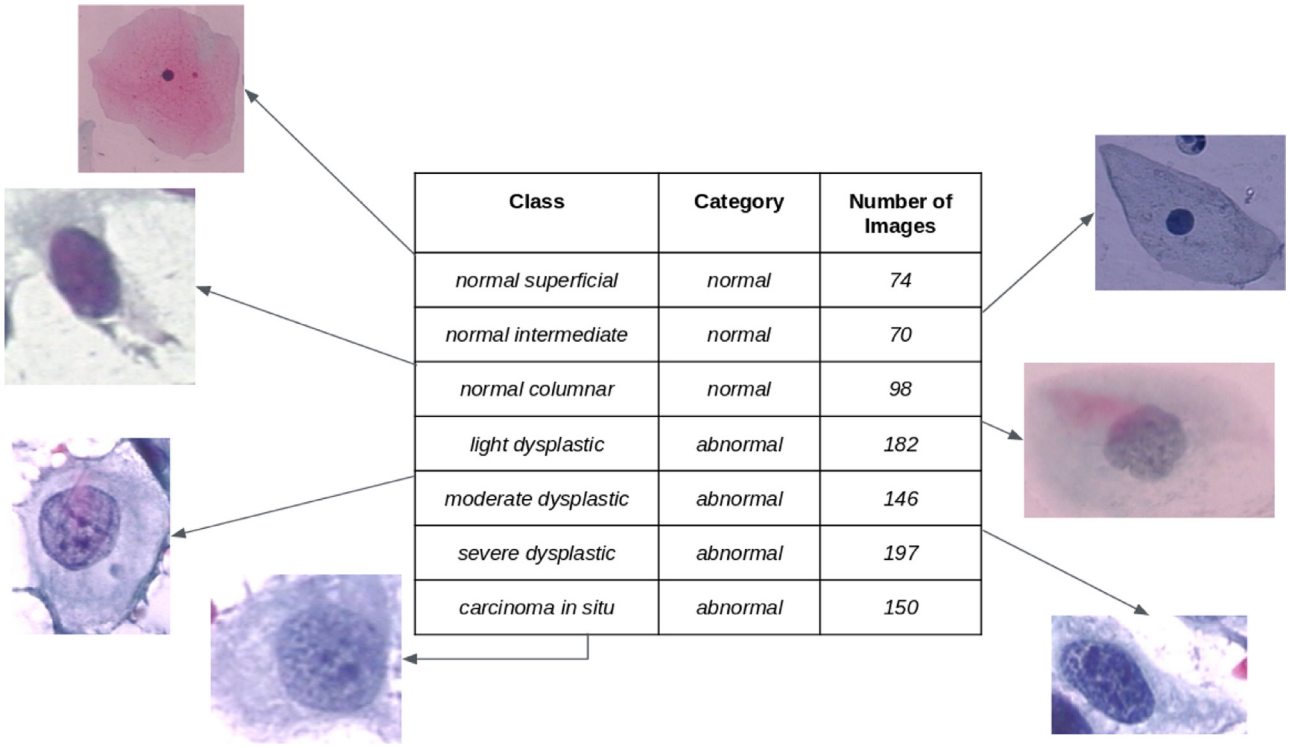
Regarding cell classification for Pap-smear analysis, most of the literature focuses on the binary “abnormal”/“normal” classification from Herlev dataset. In Bora et al. (2016) the authors used an unsupervised Feature Selection model after a CNN feature extractor to reach a F1 score of 0.90 and an accuracy of 94%. In Zhang et al. (2017), the most current deep learning methods have been used and a deep neural network (pretrained on ImageNet) has been trained on Herlev dataset categories to provide a full pipeline that reports the best performances with an accuracy of 98.3% and an Area Under the receiver operating characteristic Curve (AUC) of 0.99. In Forslid et al. (2017), a Resnet architecture was trained on Herlev dataset categories resulting in an accuracy of 84.45%. More recently, in Lin et al. (2019), the authors tackle the multi (7)-class classification challenge and propose to use, in addition to the image centered on the nuclei, cytoplasm and nuclei segmentation masks to guide the training and help the prediction. It enables them to reach an accuracy of 64.5% on the 7 classes classification task.

Regarding region (potentially containing several cells) classification, the results in Kwon et al. (2018) show an overall accuracy of 84.5% for binary abnormal/normal classification and accuracy of 76.1% for a 3 labels dataset (*Negative for Intraepithelial Lesion or Malignancy (NILM), Low-grade Intraepithelial Lesion (LSIL)* and *High-grade Intraepithelial Lesion (HSIL)*). In Harinarayanan and Nirmal (2018), a dataset of regions of Pap smears (961x961 pixels) has been labeled as “usable for diagnosis” or not. The model reaches 83.01% accuracy on the test set and the authors provide assistive maps to help pathologists by using feature maps, similarly to Grad CAM (Selvaraju et al., 2017). In Zhang et al. (2014), the authors detect and segment cytoplasm and nucleus and rely on these segmentation features to train four classifiers: “artifact” filters, “nucleus”/“artifact” classifier, “abnormal”/“normal” nucleus classifier and “abnormal” cell/hard negative classifier (each sample is going through classifiers in this order as long as it is not classified as “artifact” or “normal”). They report a system with a sensitivity of 88.1% coupled with a specificity of 100%.

Regarding WSI, mostly due to the absence of large public dataset, cytology applications are not really popular and early work were showing limitating results (Kitchener et al., 2011). In Dov et al. (2021), authors work on the classification of thyroid cytology slides with regard to The Bethesda System (TBS). They use a semi-supervised approach using 142 annotated WSIs to train a tile classifier and compute heat-maps. Then they train an aggregator that can be fed with tile label and global label. They report an AUC of 0.985 on tiles, and of 0.872 with an accuracy of 0.44 at slide-level (on the 5 classes problem that is TBS). Recently Lin et al. (2021) propose the first study on a large size dataset of pap smear WSIs and succeeded to reach a sensitivity of 0.9 aside with a specificity of 0.8.

Even if these studies show interesting results and performances, most Whole Slide Image classification methods using CNNs deal with histology slides (study of tissues to detect diseases). A WSI is the result of the digitalization of a pathology slide and is generally a high resolution image with around ten billion pixels. Classifying these images implies a high computational cost. Several works have been done to improve the efficiency/accuracy trade-off that this task requires. Camelyon-16 is the most famous dataset regarding this task, it includes 400 WSIs labeled according to the presence or not of metastases on sentinel lymph node biopsies. The most popular way to process these images is to cut them into tiles, sampling them and to work using these patches (Liu et al., 2017; Li et al., 2019; Campanella et al., 2019; Shi et al., 2020). Srinidhi et al. (2021) offer a complete review of approaches for WSI classification in histopathology. This is why, in this work, we are interested in getting closer to a pap-smear WSI classifier by working on a tile-level classifier.

Moreover, working with tiles enables to perform weakly supervised localization (Courtiol et al., 2018), i.e. highlighting which regions of these big images are responsible for the medical label. Methods that aim at understanding, after training, what the model learned to perform on a given task are called interpretability methods and are divided into three main groups: feature visualization, which consists in finding an input that maximizes the answer for



Class	Category	Number of Images
normal superficial	normal	74
normal intermediate	normal	70
normal columnar	normal	98
light dysplastic	abnormal	182
moderate dysplastic	abnormal	146
severe dysplastic	abnormal	197
carcinoma in situ	abnormal	150

Fig. 1. Herlev dataset: image examples and repartition in classes and categories.

a given neuron (or group of neurons) (Zeiler and Fergus, 2014); perturbation methods, which consist in pertubating a given input to find perturbations that impact the prediction (Fong and Vedaldi, 2017); and gradient-based methods, which rely on the fact that, for deep models, gradient (of the output with respect to the input) is a good approximation of the saliency of a model (Simoyan et al., 2013). Perturbation and gradient-based methods are part of attribution methods that aim at highlighting regions responsible for the predicted label. The main method that arose from the literature is a gradient-based method called Integrated Gradient (Sundararajan et al., 2017) that we present in detail in Section 3.3.

3. Regression constraint for explainable cervical cancer classifier

In this section we present in detail the method we developed to improve severity classification and how we use the attribution method called Integrated Gradient to prove the relevance of the training of our model.

3.1. Herlev dataset

As explained in the introduction (Section 1), the Herlev Dataset (Jantzen et al., 2005) is a cytology image set showing single cells composed of 917 images divided into two categories: “normal” vs “abnormal”.

Then “normal” cells are labeled regarding their maturity, associated with the layer of the squamous epithelium they come from. “Abnormal” (*Dysplastic*) cells are gradually classified according to their likelihood to turn into cancerous cells, based on the expertize of several cyto-technicians and doctors, while *Carcinoma in situ* are cells that actually have cancerous changes. This results in a 7 classes problem. Images in the set are encoded in 24-bit RGB with sizes ranging from 50 to 400 pixels wide.

Fig. 1 shows the distribution of images in this dataset classes and categories (note that images are scaled for a better visualization).

We turned this dataset into a “severity” focused dataset by merging all *normal* classes into one single class resulting in a 5 classes problem (*normal, light dysplastic, moderate dysplastic, severe dysplastic* and *carcinoma in situ* in order of severity).

3.2. Improving herlev severity classification using regression constraint

3.2.1. Classification approach

First, we train a ResNet-101 architecture on four independent splits of Herlev Severity dataset (4 random folds to ensure statistical significance of improvements) using multi-class cross-entropy loss that we note $\mathcal{L}_{\mathcal{CE}}(p; y_x^{cls}) = -\sum_{i=1}^{5} y_{x,i}^{cls}.log(p_i)$, where $p = (p_1, \ldots, p_5)$ are class probability neurons (resulting of soft-maxed logits neurons) and y_x^{cls} the one hot label associated with the image x (zeros array with a 1 at ground trunth class index).

3.2.2. Classification results

Performances are the following: 72.6% average overall accuracy and a very problematic confusion between *normal* and *carcinoma in situ* classes (due to *normal columnar* cells that look like *carcinoma* cells with a high NCR).

3.2.3. Regression approach

This motivates the idea of formalizing this task as a regression problem i.e. classes being represented by a regression score (1 for *normal* samples up to 5 for *carcinoma*) and using a Mean Square Error (MSE) loss $\mathcal{L}_{\mathcal{MSE}}(s; y_x^{reg}) = (s - y_x^{reg})^2$ with s the predicted score and y_x^{reg} the regression score associated with the image x.

3.2.4. Regression results

The regressor pipeline showed promising results by solving completely this challenge of differentiating *normal columnar* and

Architecture	Accuracy	Normal / Abnormal Accuracy	mean AUC	average MSE
Resnet-101 Classifier	67.6 ± 2.5	90.3 ± 1.6	0.9 ± 0.01	/
Resnet-101 Regressor	58.2 ± 5.5	89.3 ± 2.8	/	0.71 ± 0.14
Resnet-101 {Classifier + Regressor}	**72.6 ± 3.5**	**95 ± 1.8**	**0.92 ± 0.01**	**0.59 ± 0.1**

Fig. 2. Table of classification results for the three studied architectures (classifier, regressor and classifier under regression constraint) on four evaluation metrics (overall accuracy, binary accuracy, ROC-AUC and MSE).

carcinoma cells and performing with an average overall MSE of 0.71 but giving an average overall accuracy of 58.2%.

3.2.5. Classifier under regression constraint method

So, finally, we unify these two pipelines into a single one which we call "Classifier with Regression Constraint". It consists in summing the classification loss (softmax cross-entropy) with the regression loss thus strongly penalizing classification errors when the predicted class and the ground truth classes are medically distant. For that we turn classification probabilities $p = (p_1, \ldots, p_5)$ (output of the classifier) into a regression score s using a fixed fully connected layer w^r containing regression scores per class (e.g. $w^r = [1, 2, 3, 4, 5]$ as shown in Fig. 9):

$$s = RegConst(p; w^r) = \sum_{i=1}^{5} (p_i . w_i^r) \tag{1}$$

Our training loss $\mathcal{L}$ is thus:

$$\mathcal{L}(x, y_x) = \mathcal{L}_{\mathcal{CE}}(p; y_x^{cls}) + \mathcal{L}_{\mathcal{MSE}}(s; y_x^{reg}) \tag{2}$$

where x is an image, y_x the label (encoded as one hot vector y_x^{cls} for cross-entropy and as a regression score y_x^{reg} for the regression constraint).

3.2.6. Classifier under regression constraint results

The trained model jointly improves regression and classification with an average overall accuracy of 72.6% and a average overall MSE of 0.59. The binary "abnormal"/"normal" classification also benefits from this method with an average of 95% (average improvement of 4.7% compared to classification pipeline). The average accuracy of 72.6% for 5 classes accuracy classification is an average improvement of 8.1% compared to results reported in Lin et al. (2019) (see Section 2).

3.3. Interpretability, attribution and explainability

Attribution, introduced in Section 2, is a crucial task when it comes to medical applications. Indeed, since the health of patients is at stake, there is a need to strengthen the confidence of practitioners in the models, and especially to demonstrate that what is learned is relevant and relies on medical features. In order to compute attribution maps (heatmaps that highlight regions that participated to the given label), we applied the Integrated Gradient method (Sundararajan et al., 2017) to highlight on which cytomorphological features our model relies to predict the severity. This attribution method consists in interpolating the image from a baseline image (that is representative of the absence of object, e.g. a white image in the context of cervical cell classification). Given a pixel value x_i of the image x at position i in the image domain Ω, x' the baseline image (same size as x), F the model outputting a score (e.g. class probability for the classifier pipeline or severity score for the regression pipeline) given an input, and m the number of steps of the interpolation, the value $A(i)$ of the attribution map given by the Integrated Gradient method for a pixel at position i is computed as:

$$A(i) = \frac{(x_i - x_i')}{m} . \sum_{k=0}^{m} \frac{dF(x' + \frac{k}{m}.(x - x'))}{dx_i} \tag{3}$$

In order to reinforce our point, we propose a measure to quantify how much a region of an image contributes to the predicted label. Given a region $\mathcal{R}$ of an image x (subset of Ω), we denote by A_R the contribution of this region to the predicted label, which is computed as:

$$A_R = \frac{\sum_{i \in \mathcal{R}} \mid A(i) \mid}{\sum_{i \in \Omega} \mid A(i) \mid} \tag{4}$$

Note that, the *completeness* axiom defined in Sundararajan et al. (2017) ensures that, for a baseline defined as before, the attribution over the whole image (denominator) in non-zero.

We can observe that the model seems to rely more on the nucleus region for more severe classes (see Fig. 3), which is coherent since most discriminative features for severe cells are contained in the nucleus. However, we can not exclude that it could also be a simple bias introduced by the relative surface of nuclei on abnormal cells.

4. Towards whole slide image classification and medical support

In this section, we propose to apply the two methods introduced in the previous section (classification using regression constraint and attribution method using integrated gradient) to build a model able to predict a label on tiles containing several cells and to return a heatmap of the "interesting" regions for a WSI. We also benefit from these explanation maps to perform localization of the cell responsible for the predicted severity and detection of other "abnormal" cells.

4.1. Building a tile dataset based on herlev cells

To create realistic tiles, we need proper cytology background images. To this end, we use a pap smear WSI of size around

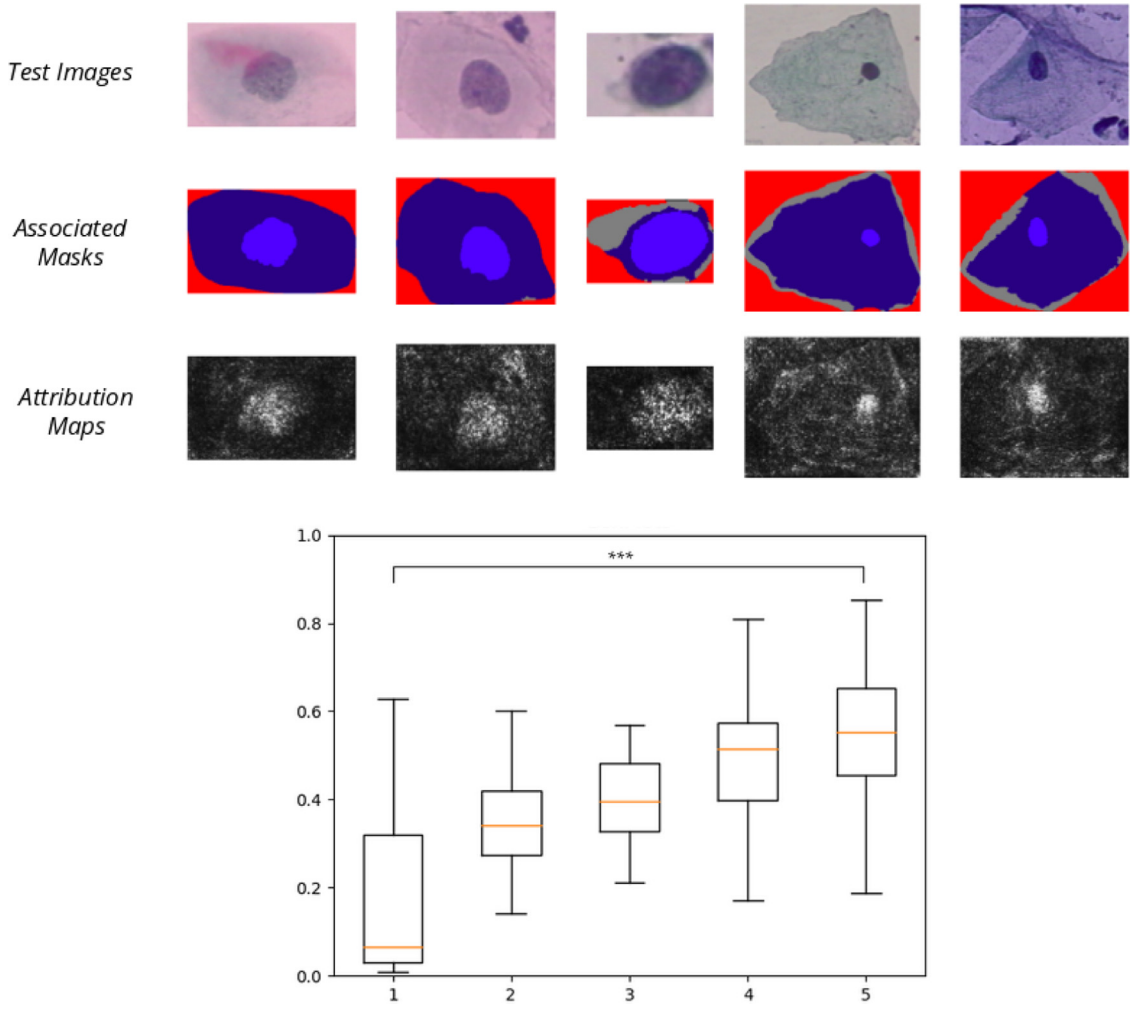


Fig. 3. Herlev images, their associated nuleus segmentation maps and attribution maps using integrated gradients on trained model (top); distribution of percentage of attribution in nucleus per class (bottom).

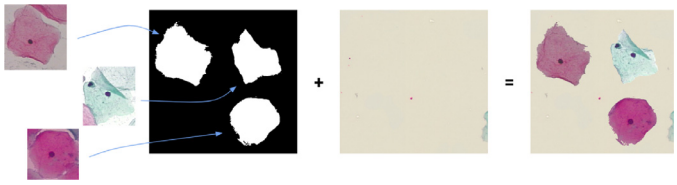


Fig. 4. Simulated tile creation process: copy paste non overlapping cells on cytology slide background tiles.

100,000x100,000 pixels, tile it (800x800 pixels non overlaping tiles), and extract "flat white" regions (by thresholding).

To create our dataset (see Fig. 4), we use the mask given by the Herlev dataset to extract only the cytoplasm and the nucleus from these images and paste it on the background images previously created (just making sure they do not overlap).

The challenge presented by what we call the "simulated" dataset of cytology tiles is to predict the maximum severity present on the tile, i.e. *normal* tiles are composed only of *normal* cells and other tiles are labelized by the degree of the most severe cell in it (see Fig. 16 (top)). Note that, in the figures, we show the ground truth boxes with a color code for clarity but these boxes are never used in the training, only global tile labels are used. We make sure that each Herlev cell is used only in one split of the simulated tile dataset.

Thus we created a dataset of 1808 images (1309 for training, 171 for validating and 328 for testing), each image containing between 1 and 15 Herlev cells. The training set contains 217 *normal* samples, 267 *light dysplastic* samples, 284 *moderate dysplastic* samples, 288 *severe dysplastic* samples and 253 *carcinoma in situ* samples, while the test set contains 60 *normal*, 74 *light dysplastic*, 77 *moderate dysplastic*, 67 *severe dysplastic* and 50 *carcinoma in situ*.

The problem of ordered classification task in known as "ordinal regression". In the following paragraphs, we start by training a classification architecture before detailing a method that is generally used to tackle these ordinal regression challenges. Finally, we apply the classification pipeline under regression constraint on the simulated tile dataset to show and validate the improvement that this method brings. We perform 5 trainings per pipeline to ensure the statistical significance of the improvements brought by the different methods considered. The improvements are assessed by three evaluation measures: overall accuracy, binary *normal*/*abnormal* accuracy and quadratic KAPPA value. Quadratic KAPPA (Brennan and Prediger, 1981) is a measure used in the context of ordinal regression problems. It consists in computing, based on the confusion matrix, a single value that takes into account the distance between classes. We define a normalized confusion matrix $M = (m_{i,j})$ such as $\sum_{i=1}^{N}(\sum_{j=1}^{N}(m_{i,j})) = 1$ (for a N classes classification problem). The expected agreement proportion P_e is $P_e = \sum_{i=1}^{N}(\sum_{j=1}^{N}(m_{i,j}).\sum_{k=1}^{N}(m_{k,i}))$ and the observed agree-

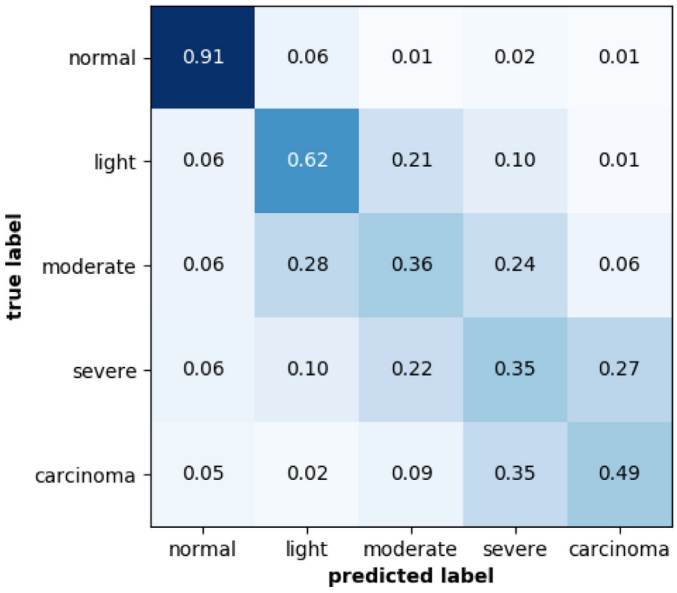


Fig. 5. Resnet-101 classifier average confusion matrix on "simulated" cytology tiles test set over 5 random folds.

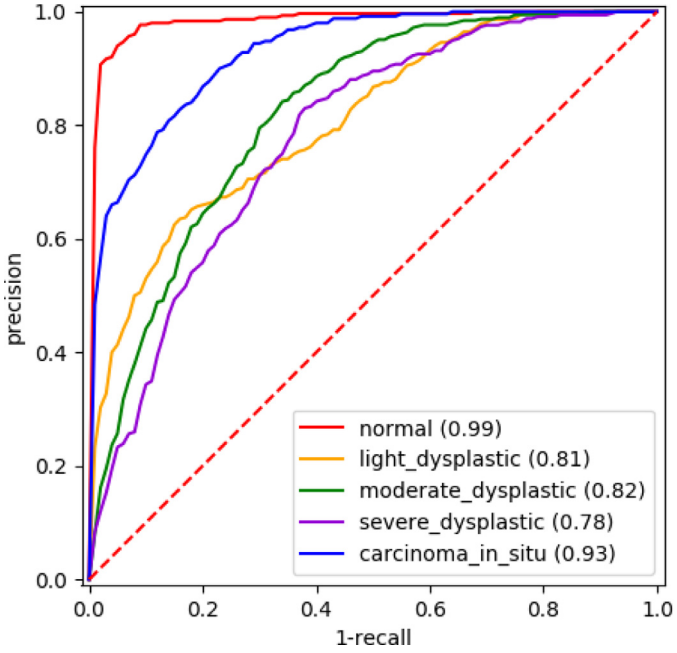


Fig. 6. Resnet-101 classifier ROC curves and on "simulated" cytology tiles test set.

ment proportion P_o is $P_o = \sum_{i=1}^{N} m_{i,i}$, KAPPA value K is then calculated as follows:

$$K = \frac{P_o - P_e}{1 - P_e} \tag{5}$$

The values of K range between -1 (worst predictor) and 1 (perfect predictor), with 0 being equivalent to a random predictor.

4.2. Baseline classification methods

In this subsection, we are going to study the performances of a standard classification and of two popular methods designed for ordinal regression.

4.2.1. Classification pipeline results

We start by training a regular (softmax cross-entropy for loss) classifier pipeline (see Fig. 15) on these simulated tiles. To deal with the size of the images (800x800 pixels), we added a 7x7 max pooling layer after the third block (inspired from "ROI Pooling" in Ren et al. (2015)). We show, in Fig. 5, that the confusion matrix computed on the 328 test images reveals an average overall accuracy of 54.6% and a binary classification accuracy of 93.6%.

We can observe in Fig. 6 the ROC curves for each class with an average mean AUC of 0.866, revealing that the network learned almost perfectly the *normal* class (AUC of 0.99) at the expense of other classes. The average quadratic KAPPA value is 0.784.

These two figures highlight that the classifier makes mistakes between *carcinoma in situ* samples and *normal* ones (this is once again due to *normal columnar* cells, and we will confirm that in the next section using attribution).

4.2.2. Ordinal regression pipeline results

In Cheng et al. (2008), the authors present their pipeline to address ordinal regression problems. Instead of training classes one against the others, the method consists in benefiting from the order of classes to train one binary classifier per class to predict whether the input sample passes the level of each class or not. For our problem it would be equivalent to train 5 classifiers. It is implemented by activating each pre-softmax neuron with a sigmoid activation thus outputting an independant score for each class (see Fig. 15). The ground truth vector is [1, 0, 0, 0, 0] for *normal* class, [1, 1, 0, 0, 0] for *light dysplastic*, and so on up to [1, 1, 1, 1, 1] for *carcinoma in situ* samples. These labels will make the model learn how to predict ordered scores (since every neuron with a lower index than the one of the ground truth class is expected to be activated). At inference, the predicted class is the last (in order of classes) class to be predicted with a score above a decided threshold (e.g. 0.5).

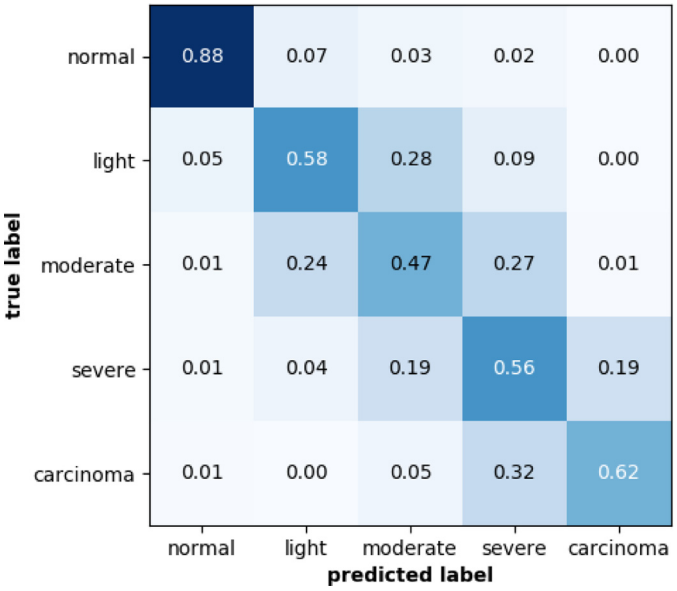


Fig. 7. Resnet-101 ordinal pipeline average confusion matrix on simulated cytology tile test set over 5 random folds.

We train a Resnet-101 with the ordinal regression pipeline on the simulated tiles dataset we created before.

Fig. 7 shows the confusion matrix obtained. We report an average overall accuracy of 61.4%, an average binary *normal* / *abnormal* accuracy of 93.7% and an average quadratic KAPPA value of 0.829 using ordinal regression pipeline.

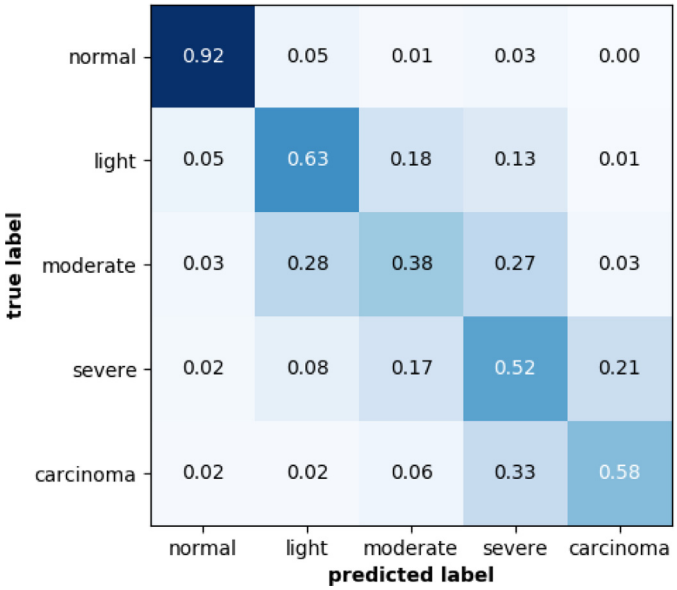


Fig. 8. Resnet-101 "Soft Labels" pipeline average confusion matrix on simulated cytology tile test set over 5 random folds.

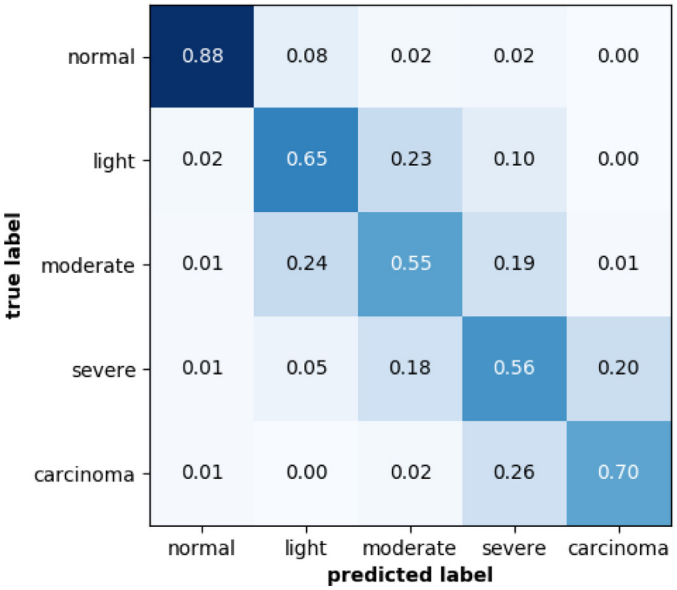


Fig. 10. Resnet-101 {classifier + regressor} average confusion matrix on simulated cytology tile test set over 5 random folds.

4.2.3. Soft labels for ordinal regression pipeline results

Another, more recent, method proposes to tackle this ordinal regression problem using "Soft Labels" (Diax and Marathe, 2019). It simply consists in changing ground truth labels to be less critical than one-hot vectors. For that, positions of classes are defined (e.g. [1, 2, 3, 4] for 4-ordered classes) and ground-truth labels are encoded as a softmax of the negative distances (absolute value of the difference of the positions) between classes. For example, instead of having [0, 0, 1, 0] as ground truth label for class 3, we define the distance vector d as [2, 1, 0, 1], thus the ground truth label is the softmax of negative distances,which is [0.0724, 0.1966, 0.5344, 0.1966] (see Fig. 15).

We train a Resnet-101 with the "Soft Labels" pipeline on the simulated tile dataset (same random 5 folds). Fig. 8 shows the confusion matrix obtained. We report an average overall accuracy of 61.5%, an average binary *normal* / *abnormal* accuracy of 94.4% and an average quadratic KAPPA value of 0.832 using "Soft Labels" pipeline. This approach statistically improves ordinal regression approach.

4.3. Proposed method: {classification + regression} pipeline results

We also trained a Resnet-101 {Classification + Regression} architecture as before on this dataset.

Fig. 9 illustrates the method explained in Eqs. (1) and (2).

4.3.1. {Classification + regression} pipeline with linear distances

First, regression constraint weights are set to be linear (e.g. [1, 2, 3, 4, 5]).

Fig. 10 shows the confusion matrix which highlights that most samples are well classified and that, once again, we avoid prediction mistakes between *normal* and *carcinoma in situ* tiles. It gives an accuracy of 66.8%. Fig. 11 confirms that the classification is really good for the *carcinoma in situ* and *normal* samples with a respective AUC of 0.96 and 0.99. The average mean AUC is 0.884.

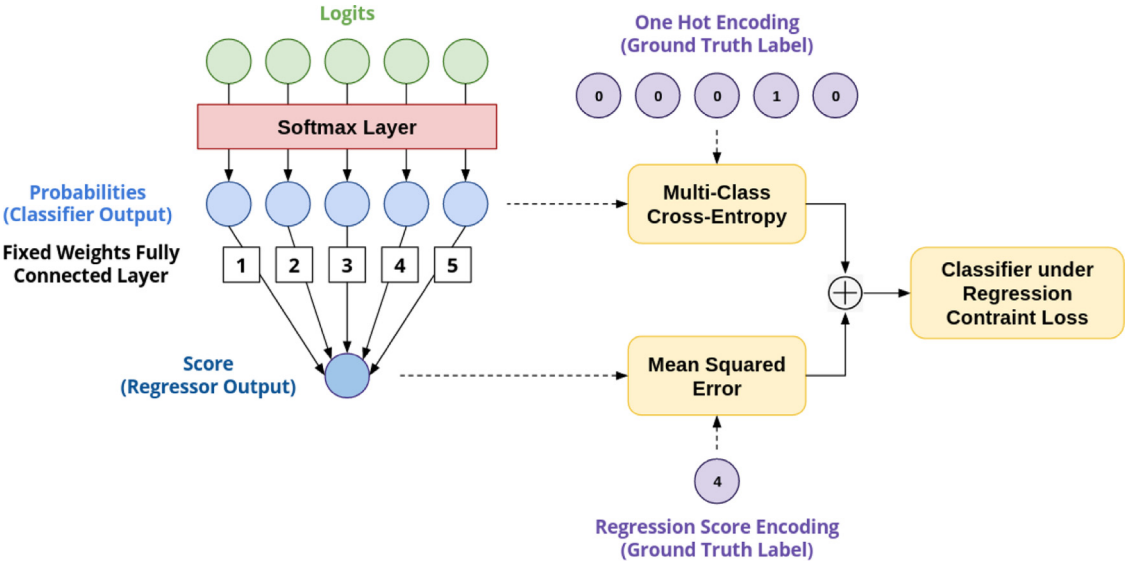


Fig. 9. Illustration of classifier with regression constraint architecture and losses.

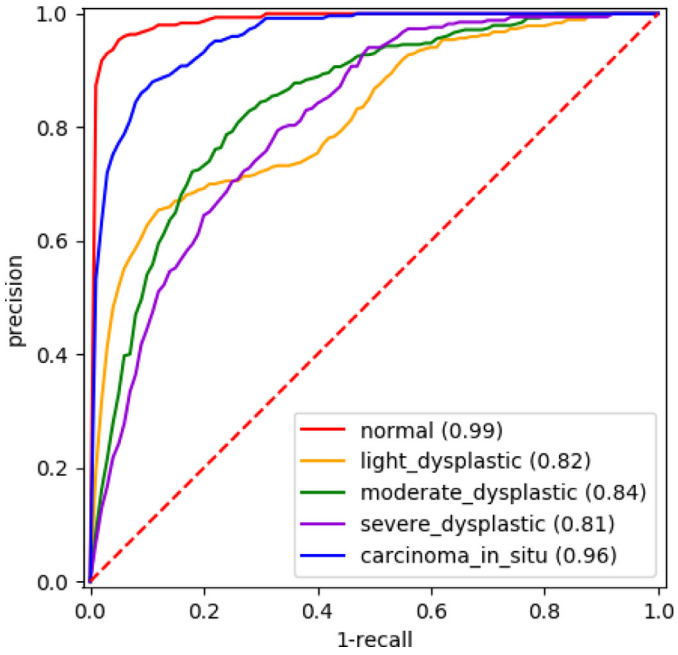


Fig. 11. Resnet-101 {classifier + regressor} average ROC curves and on simulated cytology tile test set over 5 random folds.

Interestingly, binary *normal* / *abnormal* classification also benefits from this contribution, reaching an average accuracy of 94.5%. We can also report an average classification sensitivity of 98.4% along with a specificity of 90.7%. The obtained average quadratic KAPPA value is 0.837.

We also report Positive Predicted Value (PPV or Precision) evolution with the increase of the ratio between the number of negative samples and the number of positive samples in Fig. 12. Indeed, we expect to have way more negative samples than positive samples in a real case study. As we explained, the goal is to focus on having no false negative samples to avoid missing critical cases, and according to this requirement the highest the PPV the better. We extend this discussion in Section 5, showing that we do have false positive samples but in an acceptable proportion.

4.3.2. {Classification + regression} pipeline with sensitivity focus results

As explained in Section 1, there is a need to prune "easy" normal cases that represent the vast (up to 93%) majority of cases, so medical doctors can focus on tricky abnormal cases. Nevertheless, we want to make sure that when a case is predicted as "normal" it is the right prediction, i.e. sensitivity of 100% (no False Negative) to avoid medical doctors missing an "abnormal" case.

For that we benefit from our regression constraint implementation to add more "distance" between the "normal" class and the "abnormal" ones (sensitivity focus) as follows: 1 for *normal* samples, 4 for *light dysplastic* samples, 5 for *moderate dysplastic* samples, 6 for *light dysplastic* samples and 7 for *carcinoma*. This is implemented by changing the weights for the fixed weights fully connected layer of the regression constraint (w_r becomes $[1, 4, 5, 6, 7]$). Note that this shift of 3 between the "normal" class regression score and the "light dysplastic" class regression score is purely hand-crafted.

Fig. 13 shows the confusion matrix for 5 trainings with sensitivity focus. It gives an accuracy of 66% with a sensitivity of 99.5% coupled with a specificity of 91%. As expected, this change gives a better sensitivity (highlighted in red in Fig. 13), but on the other hand the model has to make a compromise that penalizes the overall accuracy. It improves the sensitivity by 1.1%. We also re-

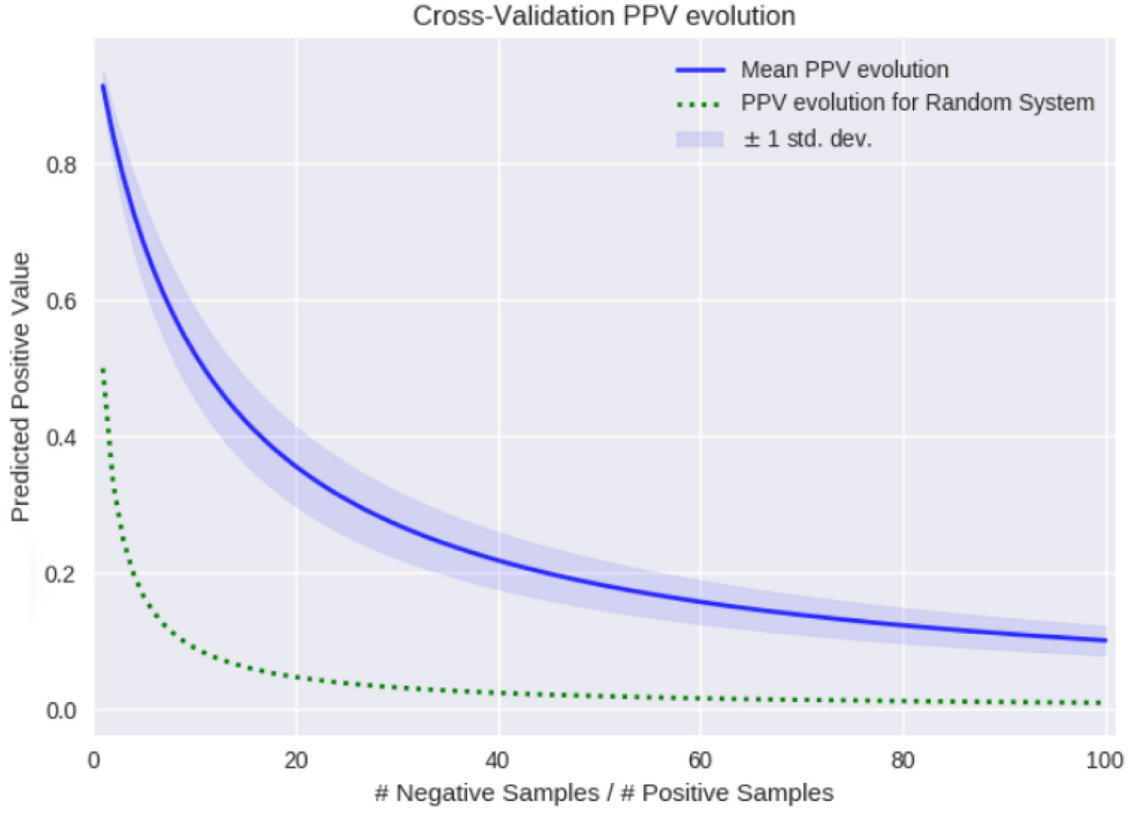


Fig. 12. PPV evolution w.r.t. ratio between negative samples and positive samples.

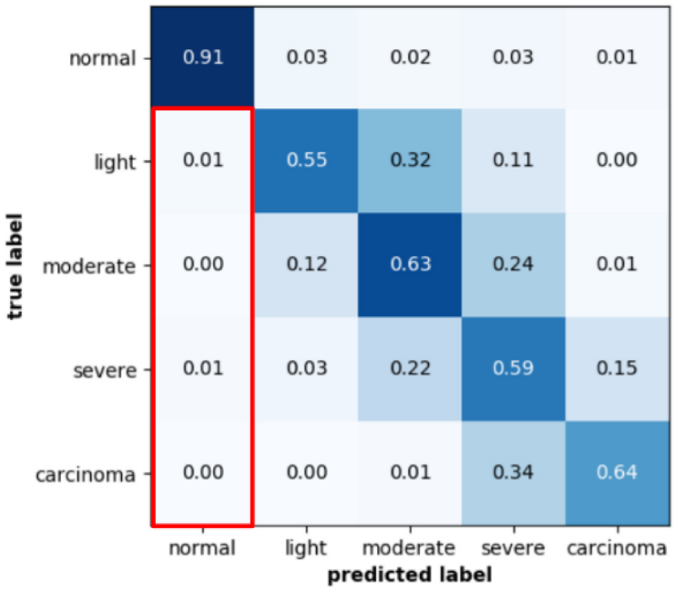


Fig. 13. Resnet-101 {classifier + regressor} with sensitivity focus average confusion matrix on simulated cytology tile test set over 5 random folds; In red, false negative probabilities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

port that the KAPPA measure also benefits from this change with a value of 0.870. It can be explained by the fact that we strengthen the regression constraint on the classifier by increasing the "distance" between the "normal" class and the "abnormal" ones, thus the regression constraint pushes severity scores towards abnormal scores, thus avoiding false negative cases and resulting in an improvement of the binary accuracy and the sensitivity.

4.3.3. Pipeline comparison

Fig. 15 illustrates pipelines to which our regression contraint method is compared and Fig. 14 shows the distribution of performances over the 5 random folds for each pipeline, i.e. the overall accuracy, binary accuracy and KAPPA value over the 5 trainings. It shows that the regression constraint really improves the general performances and particularly forces the network to learn features that are discriminative regarding the severity. Mann–Whitney *U* test (Nachar (2008)) highlights a statistical improvement from the ordinal regression pipeline to the regression constraint one regarding overall accuracy value distribution over the 5 trainings with a p-value of 0.005.

4.4. Integrated gradient for explanation maps: Weakly supervised localization and abnormality detection

Now that we have a classifier (the {Classifier + Regressor} Pipeline one) that works well on our simulated cytology tile dataset, we will check that our model relies on the right cells to make its decision by using the Integrated Gradient method (presented in Section 3.3). The baseline image used is a white (800x800) image since it is representative of the absence of objects in the cytology context. Moreover it is classified by the model as being *normal* which shows that it is a good baseline for severity attribution (since there are indeed no abnormality on it).

4.4.1. Qualitative results

Fig. 16 shows the result of the Integrated Gradient (bottom) on test images (top). Two observations are interesting to note: first, for the *normal* tile example, all cells have been identified as contributing to the predicted label and the cell that has the strongest attribution is the *normal columnar* one. This hints that the model has learned to identify these cells to avoid making the confusion with *carcinoma* cells (that also have a high NCR). Secondly, it also highlights that for *abnormal* tiles at least one of the most severe

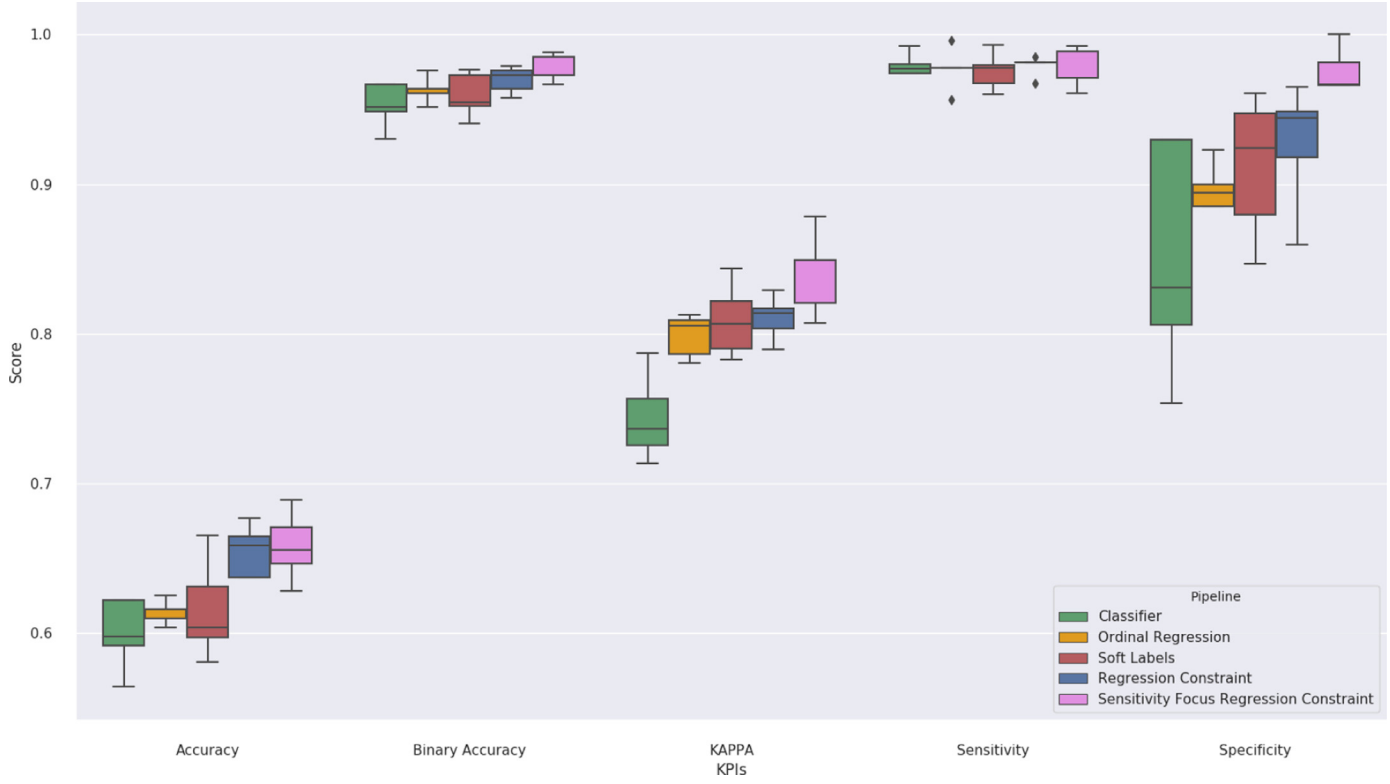


Fig. 14. Overall accuracy, binary accuracy, KAPPA, sensitivity and specificity distributions for each pipeline.

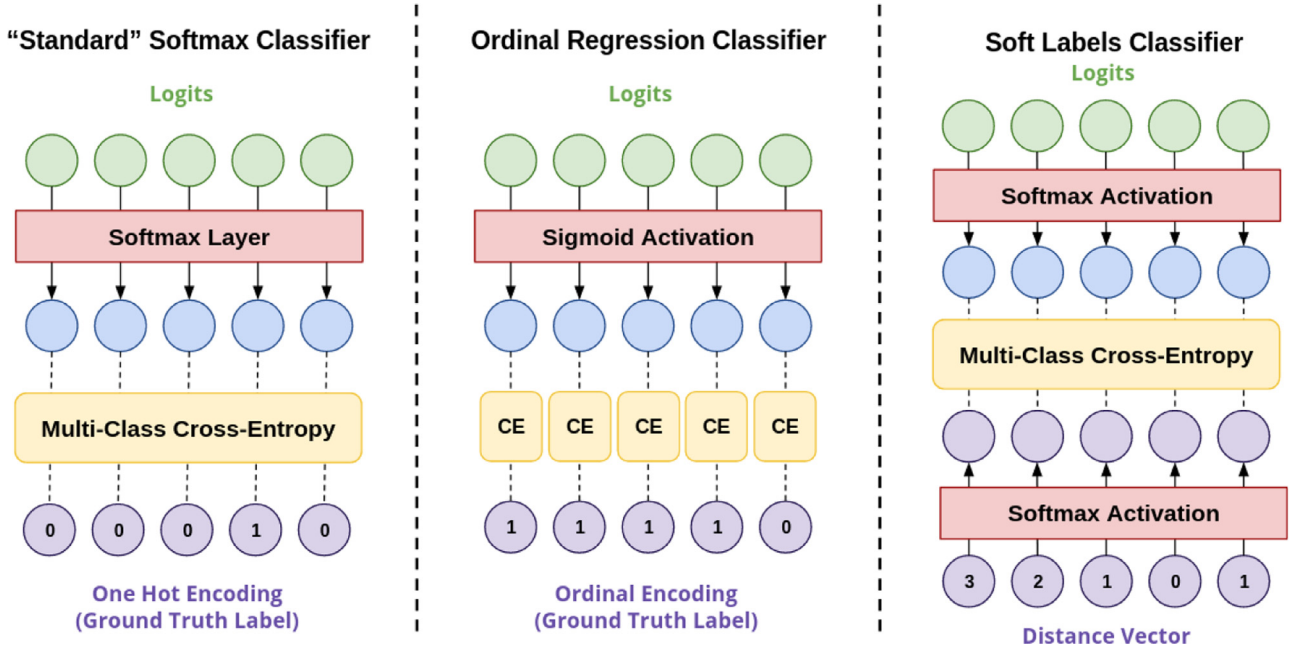


Fig. 15. Illustration of classifier, ordinal regression and Soft labels architectures and losses.

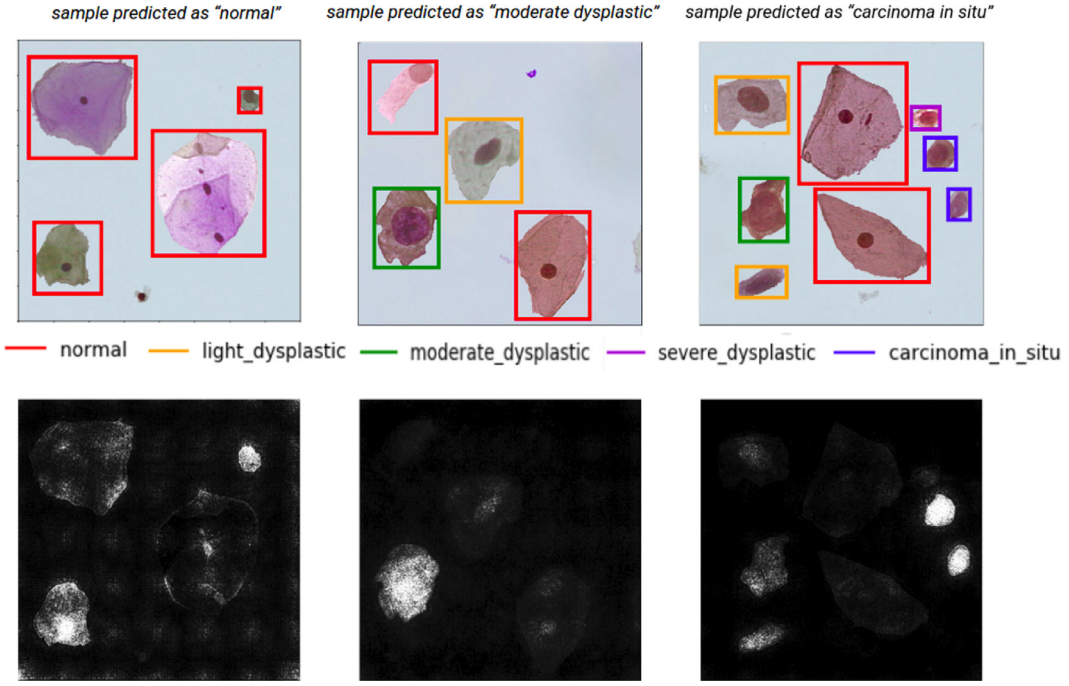


Fig. 16. Simulated tile examples (with colored ground truth cell boxes) and the associated attribution maps w.r.t. to the predicted class.

cells is clearly identified by the model as strongly contributing to the predicted label, and that cells that are *abnormal* but not the highest severity seem to contribute a bit as well. More generally, we can notice that the model learns to find some cells that are discriminative to make its prediction and some cells are just ignored.

These qualitative observations, in addition to strengthening the confidence in our model training and predictions to come, really put forward the potential for medical support through localization and more generally detection to guide diagnosis.

Interestingly, when we run the integrated gradient process on images that confused the simple classifier model (predicted *normal* for a *carcinoma in situ* sample), we can observe, in Fig. 17, that the error is due to a *normal* cell (and more precisely the *normal columnar* one at the top right of the image) while {classifier + regressor} model ignores this cell and classifies correctly this sample as being *carcinoma in situ*. This enforces the fact that the regression constraint enables to focus on these difficult cases and to drive the training towards discrimative and relevant features.

In the previous section, attribution maps have proved to be useful for the interpretability of what has been learned by the model. They also reveal the possibility to be used for explanatory localization. In the next section, we extend this approach by proposing a

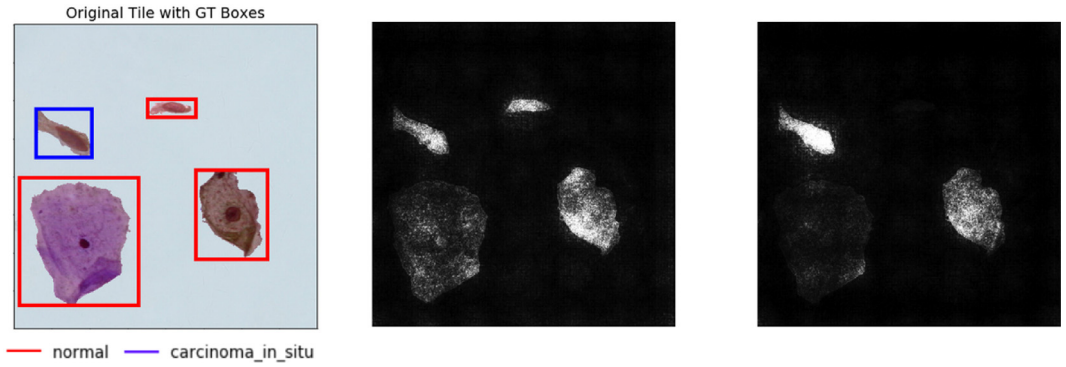


Fig. 17. Image and attribution map for a *carcinoma in situ* sample that has been classified as *normal* by classifier and as *carcinoma in situ* by {classifer + regressor}.

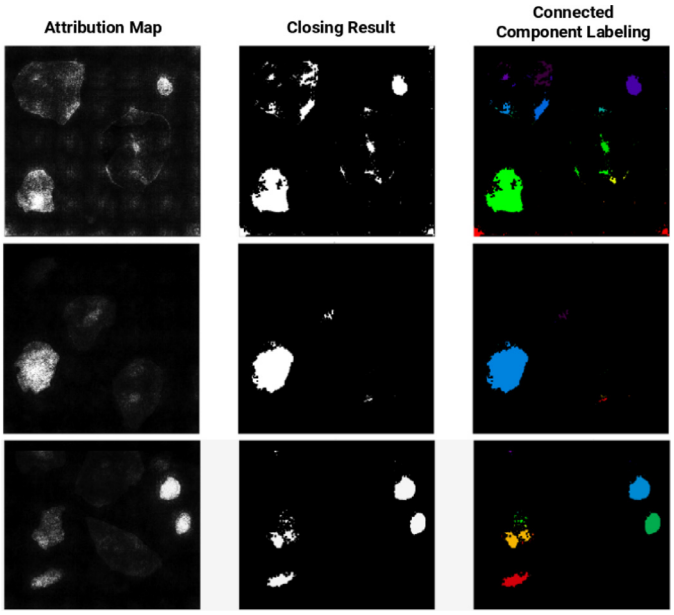


Fig. 18. Proposed three step process to localize the most contributing cells from the attribution map: binarization, closing and connected component.

method to localize and detect in a weakly supervised manner most abnormal cells in a region containing several cells.

4.4.2. Proposed method: weakly supervised localization of the most contributing cell

The previous qualitative results provide a hint for a potential localization (while no boxes were used during training).

To go from the attribution map obtained by Integrated Gradient to what we call "candidate boxes", the steps are:

1. Binarize the attribution map (e.g. 128 threshold);
2. Apply a morphological closing operation (e.g. using a 9 pixels disk structuring element);
3. Identify individual objects using connected component labeling;
4. Compute bounding boxes for each object labeled.

Results for example tiles can be seen in Fig. 18.

Thus, after obtaining all candidate boxes, we first filter out boxes that are too small (under 50 pixels) then we select the most contributing box by computing the density inside each box left. Fig. 19 shows the resulting localization boxes associated with the global label prediction.

The resulting weakly supervised localization accuracy is 80.4%.

4.4.3. Proposed method: weakly supervised abnormal cell detection

We showed that we were able to localize pretty precisely the cell that contributes the most to the predicted label. But, as explained before, the model has learned to focus on two or three cells to predict the label of the region and sometimes it seems to also use abnormal cells of lower severity degrees to predict. For example, in Fig. 16 (right) the model predicted correctly the class *carinoma in situ* and we can observe that it strongly relies on the two *carcinoma* cells on the right but also uses the three cells (and more particularly their nucleus) on the left that are *abnormal* (two *light dysplastic* and one *moderate dysplastic*) while ignoring the two cells in the middle that are indeed *normal* ones. Thus, we can enter a context of "abnormality" detection and try to find abnormal cells.

So, instead of keeping only the box with the highest density, we keep all candidate boxes (after size filtering) and point to the middle of the box.

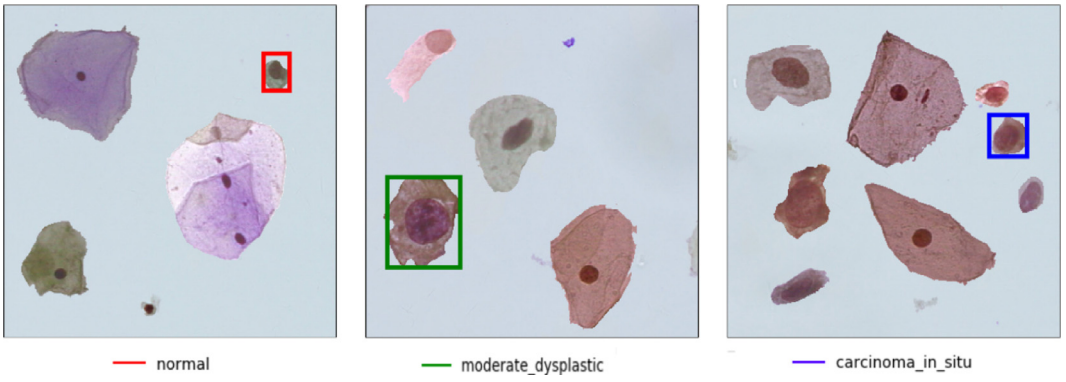


Fig. 19. Weakly supervised localization on simulated tile examples: box around most contributing cell with color associated with predicted tile label.

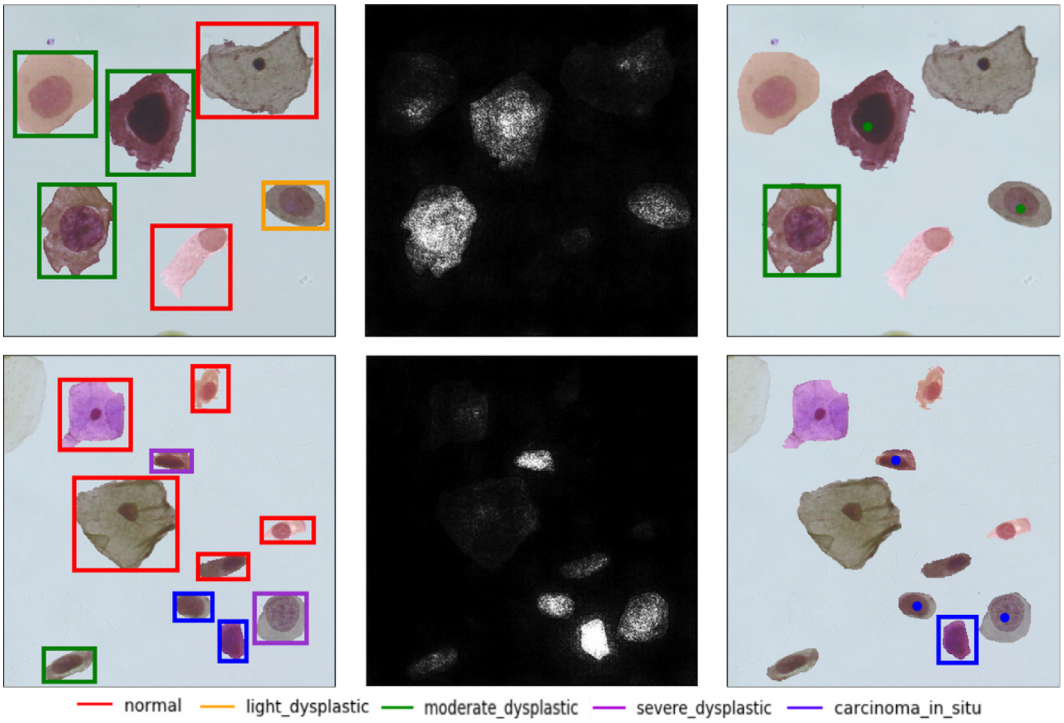


Fig. 20. Weakly supervised abnormal cell detection examples: box around most contributing cells and point annotation on other highly contributing cells.

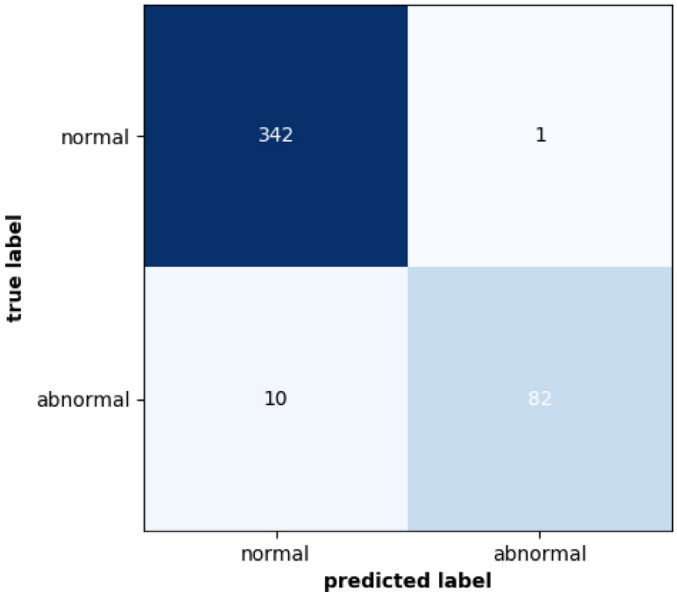


Fig. 21. Resnet-101 classifier confusion matrix on real cytology 10X tile test set.

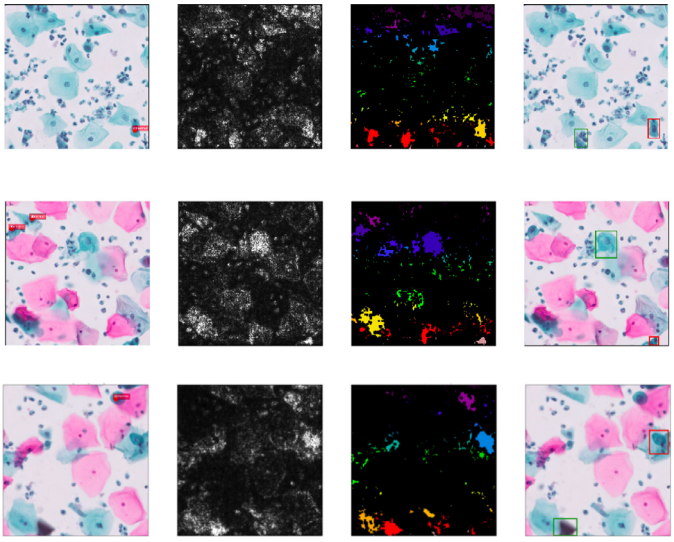


Fig. 22. Example of weakly supervised localization on real cytology 10X tiles; Images and ground truth point annotation (left); Integrated gradients results (middle); Images and proposed localization results (right), in red most contributing cell and in green other(s) highly contributing cell(s). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We count a true positive (TP) if the point is inside an *abnormal* box, false positive (FP) if it is inside a *normal* box, a true negative (TN) if a *normal* box has no point inside and a false negative (FN) if an *abnormal* box has no point inside (which is expected given the fact that the model generally uses two or three cells to predict and that a tile can have up to 12 *abnormal* cells).

Thus, we count 501 TP along with 104 FP and 433 TN for 376 FN, which gives an accuracy of 66.1%. From this confusion matrix, we also derive a sensitivity of 57.1% and a specificity of 80.6%. Fig. 20 shows some test images, their severity attribution map and the associated detection. Additionally (and maybe even more essentially), we show that in all cases where *abnormal* cells are present, we detect at least one, which ensures medical support efficiency.

5. Real clinical case study and integration in a pathologist workflow

5.1. Clinical dataset and tile classification

In this section, we discuss the performances of the proposed methods on a real clinical dataset that includes artifacts and overlapping cells. We asked an expert cytopathologist to make her

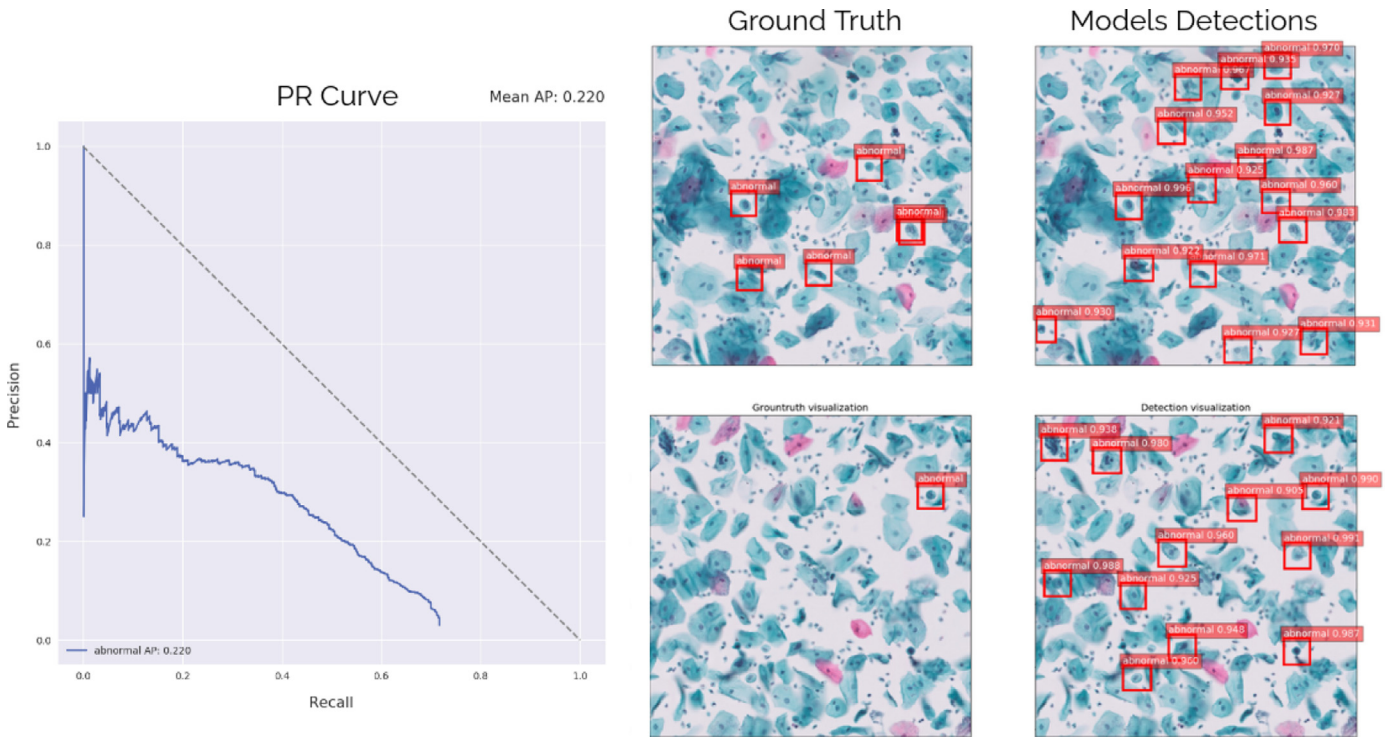


Fig. 23. Results of Faster-RCNN object detection approach for cell detection; PR Curve (left); Images and ground truth annotations (middle); Images and detection (with abnormality score above 0.9) from trained Faster-RCNN (right).

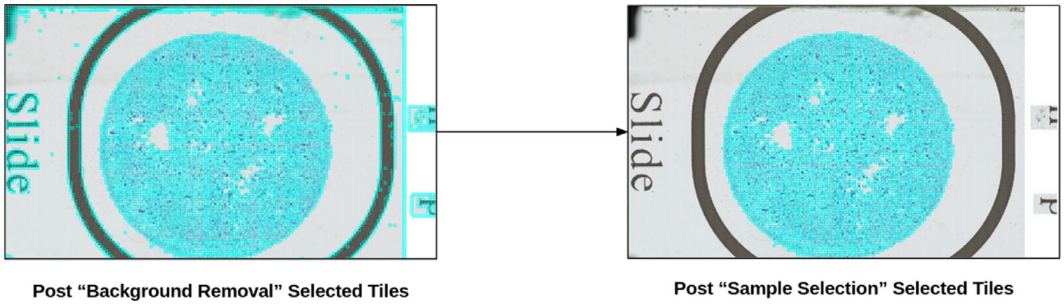


Fig. 24. Tile selection process to detect informative tiles; "background removal" result (left) and "sample selection" result (right).

diagnosis on 24 pap smears WSI and to mark some abnormal cells on abnormal slides. We extracted, by tiling where cells were marked, 568 "abnormal" images at 10X magnification, thus obtaining a binary classification dataset and more than 1900 "normal" tiles extracted from "normal" slides.

We trained the same Resnet-101 classifier architecture (using regression constraint) using 80% of these data and evaluated the performances on the 20% left (randomly splitted with regards to slides). We balance the train set regarding classes by sampling more frequently "abnormal" samples that are underrepresented in our dataset.

Fig. 21 shows the confusion matrix obtained for 10X magnification on test images. It shows an accuracy of 82%, a sensitivity of 65.3% and a specificity of 92.5%. We also report a KAPPA measure of 0.812 and an AUC of 0.991. Thus, with a reasonable ratio between negative samples and positive samples of 100 and the average number of tiles per slide around 5000, using our system we expect to have around 15 false positive tiles to review by experts on negative slides.

Using integrated gradient, we computed attribution maps and applied the post processing described in Section 4.4.2 to localize abnormal cells on "abnormal" tiles. In the case where another candidate box is 80% as dense (in terms of attribution) as the best candidate box, we also return this box as being an abnormality localization.

We report a localization accuracy of 32.8% (qualitative results obtained can be observed in Fig. 22).

This localization accuracy is quite satisfactory regarding the localization context that is pretty complicated. Indeed, there are generally around 15 cells per 10X region thus created, moreover there are artifacts as it can be observed on the third example. This localization accuracy also indicates the high number of FP detections. However to our point of view, even when the localization is wrong (see second example in Fig. 22), it still captures rather interesting cells (dark blue cell with high NCR).

This kind of supervision remains weakly-supervised even with cells annotated by the pathologist since we never use cell localization at training time and we are going to show that we are able to localize some cells. The pathologist needs only to annotate few cells (which is much less tedious than annotating all abnormal cells), and this proves sufficient for our method to predict the class of the global tiles and localize abnormality. Typically training an object detection pipeline would require such a heavy annotation and would not give much better results. We completed anno-

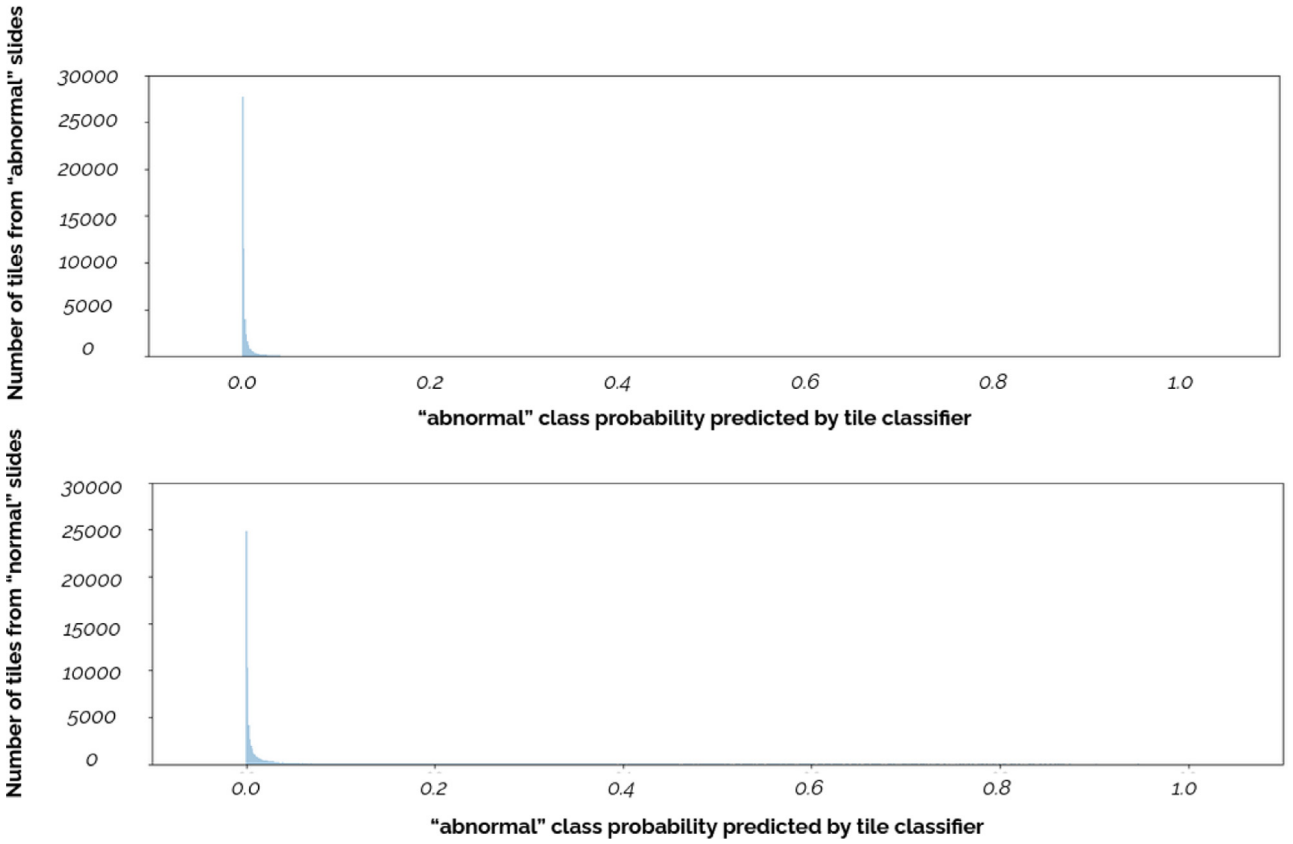


Fig. 25. Histograms w.r.t. *abnormal* tile scores for tiles from 10 *normal* slides vs 10 *abnormal* slides.

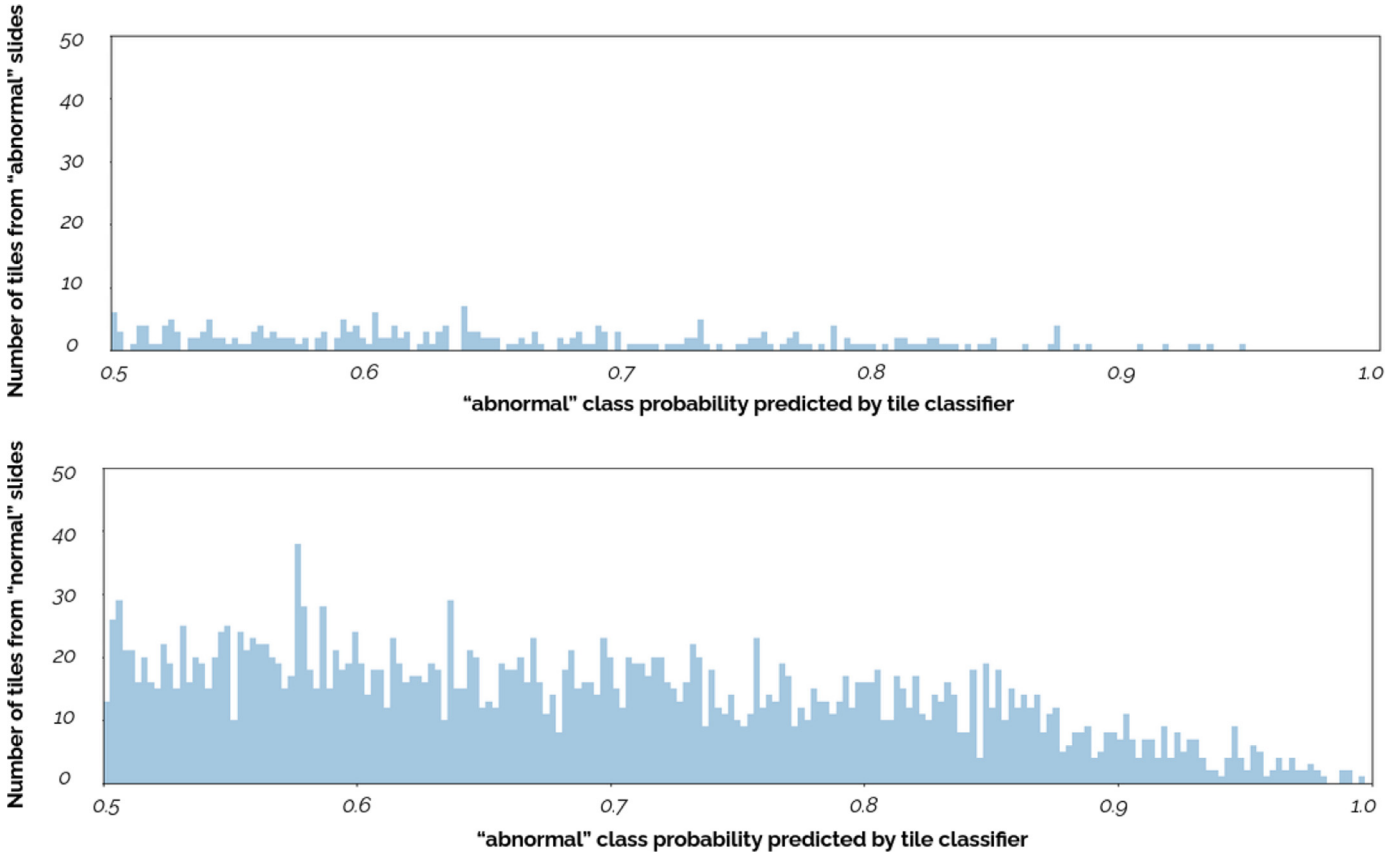


Fig. 26. Zoom (for *abnormal* class probability above 0.5) on the histograms w.r.t. *abnormal* tile scores for tiles from 10 *normal* slides vs 10 *abnormal* slides.

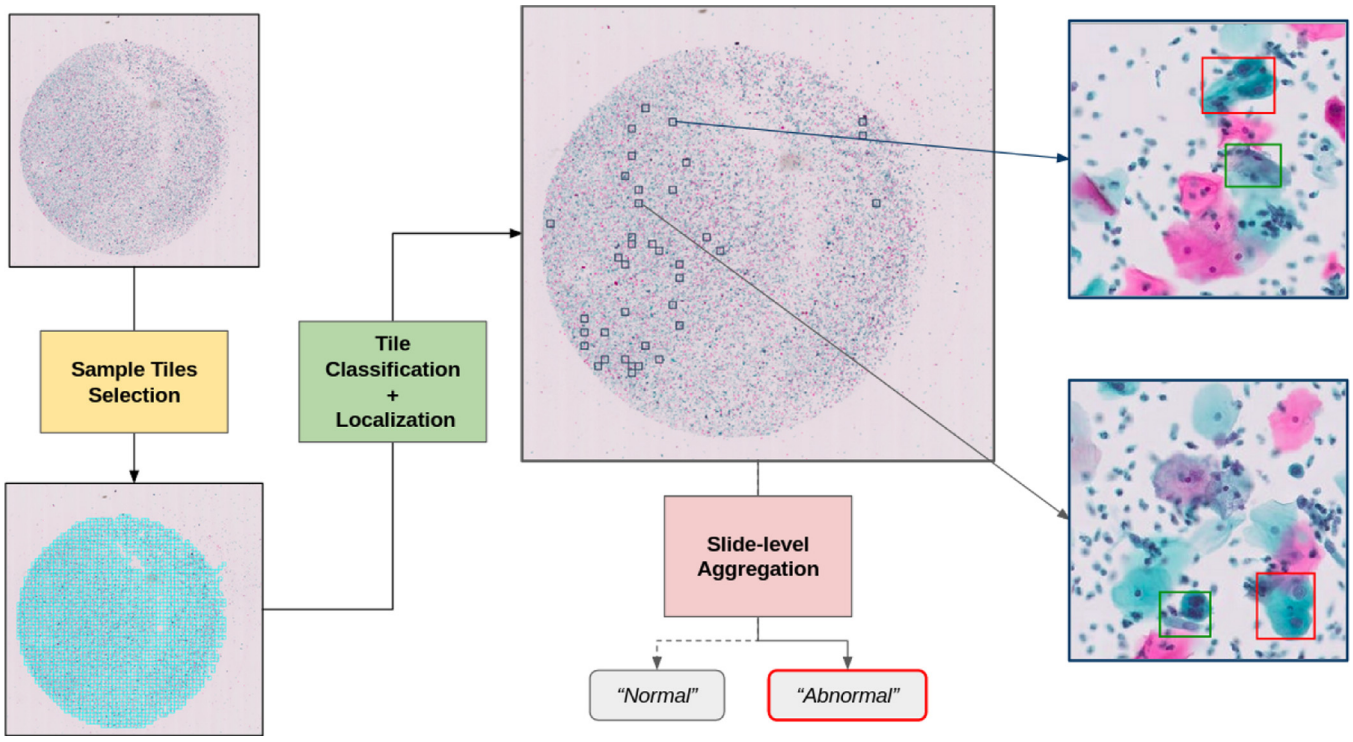


Fig. 27. Qualitative results of the proposed method for computer-aided decision: Tile Selection, tiles classification, cell localization and slide-level aggregation for proposed diagnosis.

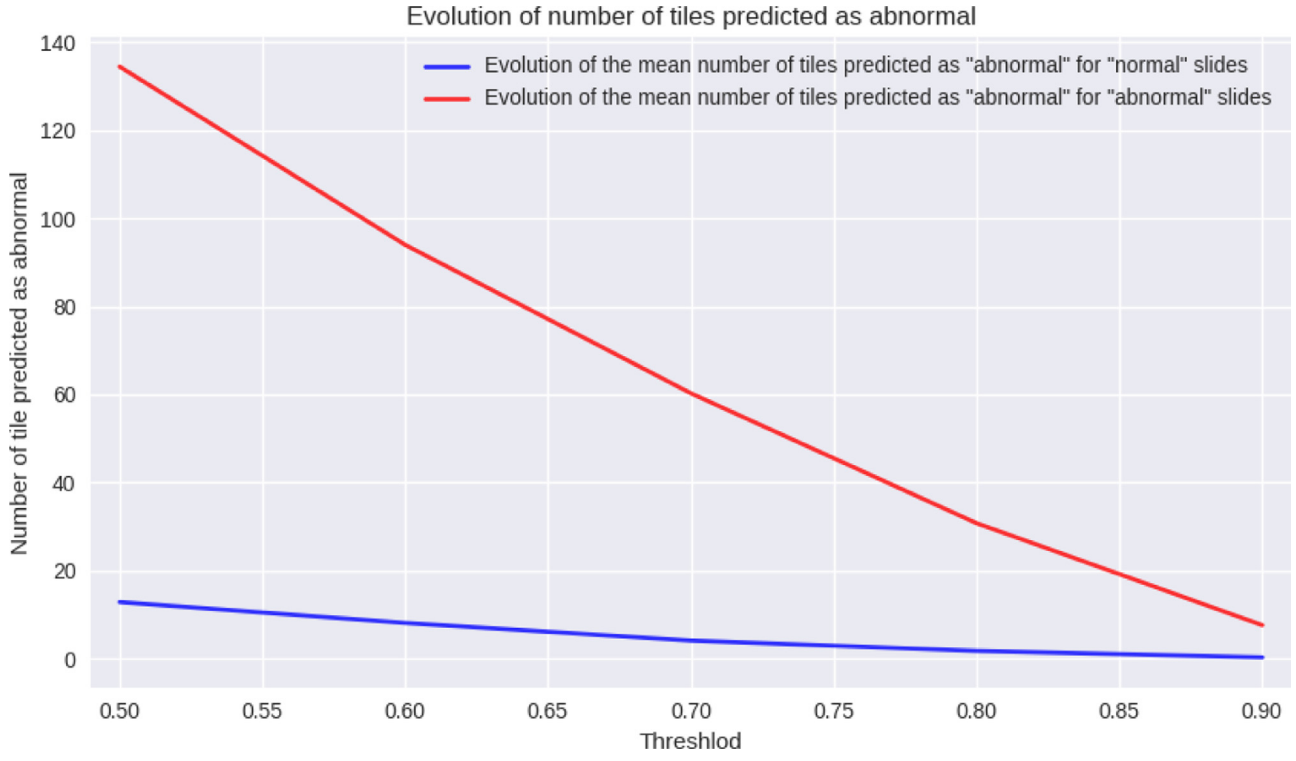


Fig. 28. Impact of tile-level decision threshold on the number of tiles selected w.r.t. slide ground truth label.

tations of potential abnormality in tiles where abnormal cells were marked, thus reaching about 3.300 annotations and 568 fully annotated tiles. We trained a Faster-RCNN (Ren et al., 2015) model for object detection and obtained an area under Precision-Recall Curve of 0.22 due to the high sensitivity that triggers a high number of FP detections. Moreover, our classification approach is twice faster than the object detection approach. Quantitative and qualitative results can be observed in Fig. 23. Both figures highlight how sensitive the model is by detecting too many cells with a high "abnormality" probability (over 0.9 on the detections showed) and how there is a compromise to make between precision and recall performances (on the Precision-Recall curve).

5.2. Proposed method: integration into pathologist workflow

To validate the clinical interest of our work, we gathered 40 new slides for which only global diagnosis is known (20 “normal” and 20 “abnormal”) and we made a prediction on each tile of the sample.

Our CAD tool starts with what we call “sample tile selection” process that aims at selecting tiles that are part of the sample and not digitalization artifact or background. It starts with a removal of all “flat” (non informative) tiles by computing the histogram of each tile and considering as background the ones that have over 95% of their histogram in a window size of 30 pixels, called “background removal”. Then, we select only neighbor tiles that form the biggest cluster, we call this “sample selection”. This process (results in Fig. 24) gives an average of number of tiles per slide of 3300 at 10X (with a minimum of 934 tiles and a maximum of 7223 tiles).

Fig. 25 shows that most tiles are classified as being *normal* (severity score between 0 and 0.5) regardless of the fact that the slide is “normal” or “abnormal”. This is expected since only some cells are abnormal on an abnormal slide. Obviously, false positive tiles are expected but we relax highly the regions to analyze before making decision, which could result in a significant gain of slide review time.

Fig. 26 shows that significantly more tiles are classified as being “abnormal” (severity score between 0.5 and 1) for “abnormal” slides which enforces the confidence in the model.

The whole computer-aided tool process and results are illustrated in Fig. 27 where we can observe that 38 regions (on more than 2700 potentially before classification) have been classified as being abnormal and that cells that made this decision have a high NCR and chromatin condensation.

For comparison, the Faster-RCNN we trained detects between 1000 and 10,000 cells per slides and there is no correlation between the number of cells detected and the label of the slides (i.e. there are no more *abnormal* cells detected on *abnormal* slides as on *normal* slides).

Thus our work allows us to reduce the amount of tiles to analyze and can guide pathologists to make their decisions on some regions instead of having to screen the complete WSI. Moreover, the localization method enables to guide the review towards discriminative cells. These contributions might avoid false negative slides by directly proposing cells of interest and could make slide review way faster by reducing the amount of data to process for a cytopathologist. In the next subsection, we extend this method by considering a simple aggregation to obtain slide-level predictions.

5.3. Proposed method: from tile-level predictions to slide-level diagnosis

We propose to study the impact of the threshold used to decide whether a tile is *abnormal* or not on the number of tiles classified as *abnormal* per slide. Fig. 28 shows the evolution of the average number of tiles selected per slide w.r.t. the slide label and the threshold on *abnormal* class probability. It confirms that statistically our method enables to select more tiles on *abnormal* slides than on *normal* slides.

Therefore, we propose to use this number of selected tiles as a predictive value for slide-label. For that, we compute accuracy and specificity w.r.t. the threshold on *abnormal* probability and the threshold on the number of selected tiles that triggers the *abnormal* label for the slide. Fig. 29 shows that the accuracy varies between 0.5 and 0.775 while specificity varies between 0.5 and 0.83.

Finally, the best configuration is to threshold at 0.1 on tile scores (that is enough to remove the vast majority of *normal* tiles) and to use a threshold of 30 tiles predicted as *abnormal* to decide that a slide is *abnormal*. This configuration gives an accuracy of 0.775 and a specificity of 0.83. We point out that, using this configuration, there are in general around 100 tiles to review on FP slides, which makes the correction by an expert fast and guided (except an outlier *normal* slide that requires more than 1000 tiles to review which would be equivalent as reviewing the whole slide).

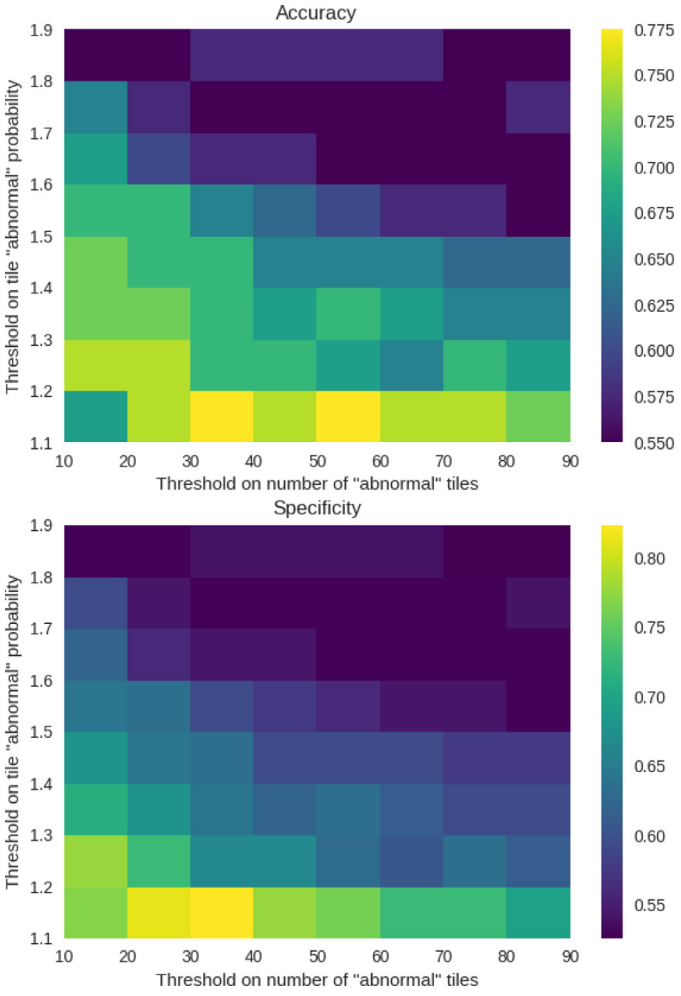


Fig. 29. Impact of threshold on tile scores and on the number of selected tiles on the slide-level prediction.

6. Discussions and conclusions

In this work we showed that our proposed method (classifier under regression constraint) can be applied to the new task of classifying tiles from cytology images in the context of cervical cancer screening. We showed, using an attribution method, that our model learned, under weak supervision, to find the cells responsible for the predicted label. We also showed that the proposed architecture outperforms a simple classifier and other state-of-the-art methods for ordinal classification in terms of overall accuracy and severity prediction.

Aiming at providing a tool that helps practitioners, we successfully tuned our model to achieve a sensitivity of 99.5% regarding *normal* tiles (almost never classify an *abnormal* tile as *normal*) while maintaining a binary accuracy of 95.2% and a good performance regarding severity stratification with a multi-class accuracy of 66%. Furthermore, we provide the user with a localization of the cause of the label up to cell level, which is an essential feature in order to gain the confidence of the practitioner in the tool, and for this tool to be integrated in the current workflow of cytopathologists. Besides, our attribution proposal can be used to detect rele-

vant cells without requiring experts to give extensive annotations at cell level. Finally, we propose to use these tile predictions to make a performant slide-level prediction.

These very encouraging results on tiles are a critical step towards an efficient and explainable Whole Slide Image classifier. The next step would be to design a system capable of aggregating in the order of 10 000 tiles while maintaining the same sensitivity, binary classification and explainability. The ingredients needed for this challenge include a reliable pruning pre-processing to alleviate the burden of testing all tiles followed by a suitable aggregation method through which explainability can be safely propagated back through each individual tile.

We will also consider refining region-based results using our state-of-the-art model trained directly on Herlev dataset that should improve the results.

Moreover, liquid-based cytology is widely used worlwide for primary indication of other cancers such as urinary or thyroid cancer screening, which makes our work even more relevant medically.

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.

CRediT authorship contribution statement

Antoine Pirovano: Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft, Visualization. **Leandro G. Almeida:** Methodology, Software, Validation, Supervision. **Said Ladjal:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Isabelle Bloch:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision, Project administration, Funding acquisition. **Sylvain Berlemont:** Validation, Resources, Supervision, Project administration, Funding acquisition.

References[1]

Bar, Y., Diamant, I., Wolf, L., Greenspan, H., 2018. Chest pathology identification using deep feature selection with non-medical training. Comput. Methods Biomech. Biomed. Eng. 6 (3), 259–263.

Bora, K., Chowdhury, M., Mahanta, L. B., Kundu, M. K., Das, A. K., 2016. Pap smear image classification using convolutional neural network. 55, pp. 1–8.

Brennan, R.L., Prediger, D.J., 1981. Coefficient kappa: some uses, misuses, and alternatives. Educ. Psychol. Meas. 41 (3), 687–699.

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Krauss Silva, V.W., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. 25, 1.

Cheng, J., Wang, Z., Pollastri, G., 2008. A neural network approach to ordinal regression. pp. 1279–1284.

Cho, K., Van Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. pp. 1724–1734.

Courtiol, P., Tramel, E.W., Sanselme, M., Wainrib, G., 2018. Classification and disease localization in histopathology using only global labels: a weakly-supervised approach. Comput. Res. Repos. (CoRR), Arvix.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., Li, F., 2009. Imagenet: a large-scale hierarchical image database. pp. 248–255.

Diax, R., Marathe, A., 2019. Soft labels for ordinal regression. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4733–4742. doi:10.1109/CVPR.2019.00487.

Dov, D., Kovalsky, S.Z., Assaad, S., Cohen, J., Elliott Range, D., Pendse, A.A., Henao, R., Carin, L., 2021. Weakly supervised instance learning for thyroid malignancy prediction from whole slide cytopathology images. Med. Image Anal. 67, 772–780. doi:10.1016/j.media.2020.101814.

Fong, R. C., Vedaldi, A., 2017. Interpretable explanations of black boxes by meaningful perturbation. pp. 3449–3457.

Forslid, G., Wieslander, H., Bengtsson, E., Wählby, C., Hirsch, J., Stark, C. R., Sadanandan, S. K., 2017. Deep convolutional neural networks for detecting cellular changes due to malignancy. pp. 82–89.

Harinarayanan, K.K., Nirmal, J., 2018. Classification driven assisted screening for cervical cancer using deep neural network.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. pp. 770–778.

Jantzen, J., Norup, J., Dounias, G., Bjerregaard, B., 2005. Pap-smear benchmark data for pattern classification. Nat. Inspired Smart Inf. Syst. 1–9.

Kitchener, H.C., Blanks, R., Dunn, G., Gunn, L., Desai, M., Albrow, R., Mather, J., Rana, D.N., Cubie, H., Moore, C., Legood, R., Gray, A., Moss, S., 2011. Automation-assisted versus manual reading of cervical cytology (mavaric): a randomised controlled trial. Lancet Oncol. 12 (1). doi:10.1016/s1470-2045(10)70264-3.

Krizhevsky, A., Sutskever, I., Hinton, G., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105.

Kwon, M., Kuko, M., Martin, V., Kim, T. H., Martin, S. E., Pourhomayoun, M., 2018. Multi-label classification of single and clustered cervical cells using deep convolutional networks.

LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D., 1989. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1 (4), 541–551.

Li, Y., Wu, J., Wu, Q., 2019. Classification of breast cancer histology images using multi-size and discriminative patches based on deep learning. IEEE Access 7, 21400–21408.

Lin, H., Chen, H., Wang, X., Wang, Q., Wang, L., Heng, P., 2021. Dual-path network with synergistic grouping loss and evidence driven risk stratification for whole slide cervical image analysis. Med. Image Anal. 69. doi:10.1016/j.media.2021.101955.

Lin, H., Hu, Y., Chen, S., Yao, J., Zhang, L., 2019. Fine-grained classification of cervical cells using morphological and appearance based convolutional neural networks. IEEE Access.

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G.E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P.Q., Corrado, G.S., Hipp, J.D., Peng, L., Stumpe, M.C., 2017. Detecting cancer metastases on gigapixel pathology images.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D., 2015. Human-level control through deep reinforcement learning. Nature 518, 529–533.

Nachar, N., 2008. He mann whitney U: a test for assessing whether two independent samples come from the same distribution. Tutor. Quant. Methods Psychol. 4 (1), 13–20.

Nayar, R., Wilbur, C.D., 2015. The pap test and bethesda 2014. Acta Cytol. 59, 121–132.

Naylor, P., Boyd, J., Laé, M., Reyal, F., Walter, T., 2019. Predicting residual cancer burden in a triple negative breast cancer cohort. pp. 933–937.

Papanicolaou, G.N., Traut, H.F., 1943. Diagnosis of uterine cancer by the vaginal smear. Yale J. Biol. Med. 15 (6), 127.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1440–1448.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. pp. 234–241.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: visual explanations from deep networks via gradient-based localization. pp. 618–626.

Sherman, M., Dasgupta, A., Schiffman, M., Nayar, R., Solomon, D., 2007. The bethesda interobserver reproducibility study (birst). Cancer Cytopathol. 111 (1), 15–25.

Shi, X., Su, H., Xing, F., Liang, Y., Qu, G., Yang, L., 2020. Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis. Med. Images Anal. 60.

Simoyan, K., Vedaldi, A., Zisserman, A., 2013. Deep inside convolutional networks: visualising image classification models and saliency maps. Comput. Res. Repos. (CoRR), Arvix.

Solomon, D., Davey, D., Kurman, R., Moriarty, A., O'Connor, D., Prey, M., Raab, S., Sherman, M., Wilbur, D., Wright, T.J., Young, N., 2002. The 2001 bethesda system terminology for reporting results of cervical cytology. J. Am. Med. Assoc. 287 (16), 2114–2119.

Srinidhi, C.L., Oxan, C., L., M.A., 2021. Deep neural network models for computational histopathology: a survey. Med. Images Anal. 67. doi:10.1016/j.media.2020.101813.

Stoler, M.H., Schiffman, M., 2001. Interobserver reproducibility of cervical cytologic and histologic interpretations realistic estimates from the ascus-lsil triage study. J. Am. Med. Assoc. 285 (11), 1500–1505.

Sundararajan, M., Taly, A., Yan, Q., 2017. Axiomatic attribution for deep networks. In: 34th International Conference on Machine Learning, 70, pp. 3319–3328.

WHO, W.H.O., 2014. Comprehensive Cervical Cancer Control: A Guide to Essential Practice, second ed..

Wright, T. C., J., Cox, T. J., Massad, S. L., L. B. Twiggs, L. B., Wilkinson, E. J., for the 2001 ASCCP-Sponsored Consensus Conference, 2002. 2001 consensus guidelines for the management of women with cervical cytological abnormalities. 287(16), pp. 2120–2129.

Zeiler, M. D., Fergus, R., 2014. Visualizing and understanding convolutional networks. pp. 818–833.

Zhang, L., Kong, H., Ting Chin, C., Liu, S., Fan, X., Wang, T., Chen, S., 2014. Automation-assisted cervical cancer screening in manual liquid-based cytology with hematoxylin and eosin staining. Cytometry Part A 85.

Zhang, L., Lu, L., Nogues, I., Summers, R.M., Liu, S., Yao, J., 2017. Deeppap: deep convolutional networks for cervical cell classification. IEEE J. Biomed. Health Inform. 21, 1633–1643.