Computer-Aided Diagnosis tool for Cervical Cancer Screening with Weakly Supervised Localization and Detection of Abnormalities using Adaptable and Explainable Classifier

Antoine Pirovano\textsuperscript{a,b}, Leandro G. Almeida\textsuperscript{a}, Said Ladjal\textsuperscript{b}, Isabelle Bloch\textsuperscript{b,c}, Sylvain Berlemont\textsuperscript{a}

\textsuperscript{a}Keen Eye, 74 rue du Faubourg Saint-Antoine, 75012 Paris, France  
\textsuperscript{b}LTCI, Telecom Paris, Institut Polytechnique de Paris, 19 Place Marguerite Perey 91120 Palaiseau, France  
\textsuperscript{c}Sorbonne Universit\'e, CNRS, LIP6, Paris, France

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 $500 \times 10^3$ new cases each year, screening for cervical cancer needs to be efficient and precise. Currently, pap smear (Papanicolaou and Traut (1943)) are the most commonly used diagnosis method for cervical cancer screening. They are 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 $100 \times 10^3$) 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 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 reproducibility, there is still a high inter-observers variability in diagnosis (Stoler and Sherman (2007); Sherman et al. (2007))). It is not clearly defined how anapathologists should proceed and what they should rely on to make their 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 responsible for the predicted label along with weakly supervised detection of abnormal cells, which might help medical practitioners to understand the outcome of our method, strengthen 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 (Forstlid 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 (NLM), 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 limiting 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)) proposes 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 a given neuron (or group of neurons) (Zeiler and Fergus (2014)); perturbation methods, which consist in perturbing 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 details 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 expertise 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 $L_{CE}(p; y^{cls}) = -\sum_{i=1}^{N} y_{i}^{cls} \log(p_{i})$, where $p = (p_{1}, \ldots, p_{5})$ are class probability neurons (resulting of softmaxed logits neurons) and $y^{cls}$ the one hot label associated with the image $x$ (zeros array with a 1 at ground truth 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 \( L_{\text{MSE}}(s; y^s_x) = (s - y^s_x)^2 \) with \( s \) the predicted score and \( y^s_x \) 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 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 \( w' \) containing regression scores per class (e.g. \( w' = [1, 2, 3, 4, 5] \) as shown in Fig. 9):

\[
    s = \text{RegConst}(p; w') = \sum_{i=1}^{5} (p_i w'_i)
\]

Our training loss \( L \) is thus:

\[
    L(x, y_s) = L_{\text{CE}}(p; y^s_x) + L_{\text{MSE}}(s; y^s_x)
\]

where \( x \) is an image, \( y_s \) the label (encoded as one hot vector \( y^s_x \) for cross-entropy and as a regression score \( y^s_x \) 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 an 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 \( \Omega \), \( x' \) 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:

\[
    A(i) = \frac{(x_i - x'_i)}{m} \sum_{k=0}^{m} \frac{dF(x' + \frac{k}{m}(x - x'))}{dx_i}
\]

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

\[
    A_R = \frac{\sum_{i \in R} |A(i)|}{\sum_{i \in \Omega} |A(i)|}
\]

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.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Accuracy</th>
<th>Normal / Abnormal Accuracy</th>
<th>mean AUC</th>
<th>average MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResNet-101 Classifier</td>
<td>67.6 ± 2.5</td>
<td>90.3 ± 1.6</td>
<td>0.9 ± 0.01</td>
<td>/</td>
</tr>
<tr>
<td>ResNet-101 Regressor</td>
<td>58.2 ± 5.5</td>
<td>89.3 ± 2.8</td>
<td>/</td>
<td>0.71 ± 0.14</td>
</tr>
<tr>
<td>ResNet-101 (Classifier + Regression)</td>
<td>72.6 ± 3.5</td>
<td>95 ± 1.8</td>
<td>0.92 ± 0.01</td>
<td>0.59 ± 0.1</td>
</tr>
</tbody>
</table>

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).

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 100,000x100,000 pixels, tile it (800x800 pixels non-overlapping 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 labeled 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 $\sum_{j=1}^{N}(\sum_{i=1}^{N}m_{i,j}) = 1$ (for a N classes classification problem). The expected agreement proportion $P_e$ is $P_e = \sum_{i=1}^{N}(\sum_{j=1}^{N}m_{i,j})/N^2$ and the observed agreement proportion $P_o$ is $P_o = \sum_{i=1}^{N}m_{i,i}$. KAPPA value $K$ is then calculated as follows:

$$K = \frac{P_o - P_e}{1 - P_e} \quad (5)$$

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 independent 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).

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.

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 Eq. (1) and (2).
4.3.1. \textit{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. 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. \textit{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 report 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 constraint 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 \textit{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 \textit{normal} tile example, all cells have been identified as contributing to the predicted label and the cell that has the strongest attribution is the \textit{normal columnar} one. This hints that the model has learned to identify these cells to avoid making the confusion with \textit{carcinoma} cells (that also have a high NCR). Secondly, it also highlights that for \textit{abnormal} tiles at least one of the most severe cells is clearly identified by the model as strongly contributing to the predicted label, and that cells that are \textit{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 \textit{normal} for a \textit{carcinoma in situ} sample), we can observe, in Fig. 17, that the error is due to a \textit{normal} cell (and more precisely the \textit{normal columnar} one at the top right of the image) while \{classifier + regressor\} model ignores this cell and classifies correctly this sample as being \textit{carcinoma in situ}. This enforces the fact that the regression constraint enables to focus on these difficult cases and to drive the training towards discriminative 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 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.

Result 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 \textit{carcinoma in situ} and we can observe that it strongly relies on the two \textit{carcinoma} cells on the right but also uses the three cells (and more particularly their nucleus) on the left that are \textit{abnormal} (two \textit{light dysplastic} and one \textit{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.

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 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 1,900 “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 Figure 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 annotations of potential abnormality in tiles where abnormal cells were marked, thus reaching about 3,300 annotations and 568 fully annotated tiles. We trained a

---

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...
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.

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. Figure 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. Figure 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).

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 relevant 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 preprocessing 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 regions-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 worldwide for primary indication of other cancers such as urinary or thyroid cancer screening which makes our work even more relevant medically.

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

References
