A Spatial Regulated Patch-Wise Approach for Cervical Dysplasia Diagnosis

Ying Zhang*, Yifang Yin*, Zhenguang Liu*, Roger Zimmermann*

*School of Computing, National University of Singapore
*School of Computer Science, Northwestern Polytechnical University
School of Computer and Information Engineering, Zhejiang Gongshang University
{dcszyi, idsyin, dcsrz}@nus.edu.sg, liuzhenguang2008@gmail.com

Abstract
Cervical dysplasia diagnosis via visual inspection is a challenging problem. Recent approaches use deep learning techniques to extract features and require the downsampling of high-resolution cervical screening images to smaller sizes for training. Such a reduction may result in the loss of visual details that appear weakly and locally within a cervical image. To overcome this challenge, our work divides an image into patches and then represents it from patch features. We aggregate patch patterns into an image feature in a weighted manner by considering the patch–image relationship. The weights are visualized as a heatmap to explain where the diagnosis results come from. We further introduce a spatial regulator to guide the classifier to focus on the cervix region and to adjust the weight distribution, without requiring any manual annotations of the cervix region. A novel iterative algorithm is designed to refine the regulator, which is able to capture the variations in cervix center locations and shapes. Experiments on an 18-year real-world dataset indicate a minimal of 3.47%, 4.59%, 8.54% improvements over the state-of-the-art in accuracy, F1, and recall measures, respectively.

Introduction
Cervical cancer ranks fourth among the most frequent cancers in women and approximately 90% of its deaths occur in less developed countries (WHO 2018). However, statistics show that cervical cancer is more than 90% treatable if it is detected at an early stage (Gotlieb et al. 2017). An abnormality might be identified by cervical intraepithelial neoplasia (CIN) which is the precancerous change and abnormal growth of squamous cells on the surface of the cervix (Kumar V 2007). There exist a few screening methods to diagnose cervical dysplasia including but not limited to, a Pap smear test, an HPV test, and visual examinations. The first two are conducted in a laboratory setting and require professional medical devices as well as highly trained experts. Thus, they might not be easily deployed in less developed regions or countries where deaths from cervical cancer occur more often. Consequently, performing cost-effective and non-invasive visual screening shows great potential in the medical field (Do et al. 2018; Feng and Zhou 2016). In a visual inspection, a non-physician takes colposcopic photographs of the cervix after the application of 5% acetic acid (VIA) to the cervix epithelium and submits them to a physician for further interpretation if certain visual characteristics appear. In this paper, the cervical images refer to the colposcopic photographs taken by the VIA approach.

Naturally, visual inspection lends itself to a conversion of the cervical dysplasia diagnosis problem into a binary image classification, solvable by computer. The CIN grades 0 and 1 are labeled as normal while CIN 2+ identifies abnormal cases. The cervical images have a few challenging characteristics. (1) They are usually of high resolution. (2) In a majority of cases the images contains a large amount of irrelevant background as the non-cervix tissues or medical instruments (Gordon et al. 2006). (3) The transformation zone, the most common area on the cervix for abnormal cells to develop, takes up a relatively small fraction of the full image. (4) The correlation between the visual clues of abnormal tissues and CIN grades is relatively weak (Xu et al. 2017). Thus, it is not very effective to directly import the full images into modern classification models as the heavy compression might lead to a loss of the weak visual details of the abnormal tissue (Zhou, Zhang, and Wu 2018) and the small part of cervical lesion results in information bias (Xue, Ng, and Qiao 2020). To alleviate some of these challenges, nearly all conventional approaches manually label a tight bounding box of the cervical region or even the transformation zone and crop it for further examination (Xu et al. 2017; Hu et al. 2019; Liu et al. 2013). However, such manual labeling is not always available and its quality strongly depends on the subjective experience of annotators (Xue, Ng, and Qiao 2020). Different from such approaches, this paper proposes a method that does not require such manual annotations but instead automatically focuses on the proper region of interest (ROI) for detection.

The proposed framework is illustrated in Fig. 1. Each image is divided into sub-regions (patches), from which the image-level features will eventually be reassembled and trained together with the labels in a supervised manner. The model leverages three sets of information during the training: (1) The patch visual contents, which is used to learn the distinguishable local patterns of abnormal tissues. (2) The patch–image relationship, which encapsulates that patches should contribute to various image features differently. This
Figure 1: The workflow of our proposed explainable cervical dysplasia diagnosis model. We divide a cervical image into non-overlapping patches and jointly leverage the patch visual contents, image-patch relationship and basic domain knowledge to eventually output an image-level binary label as well as an interpretation map. Patch features are individually extracted and are aggregated to an image-level feature via patch-weights learning by considering the image-patch relationship. Basic domain knowledge is incorporated to further adjust the weights spatial distribution via an temporal iterative algorithm.

is inspired by the fact that patches in the normal cases should be cancer-free and thus their patterns should trigger the detection weakly. (3) Basic domain knowledge, the fact that cervical colposcopic images are always taken from the same viewing direction indicates a correlation between certain visual contents and their spatial location. Hence we bring in an additional spatial regulation to guide the weights-learning so that the model can focus more on the foreground regions through a novel iterative algorithm. Finally, all information sets are integrated in the training to output not only a binary label but also an interpretation map. In summary, our main contributions include:

• A patch-wise solution for cervical dysplasia diagnosis is proposed. Patch features capture local visual details and are essential for medical applications where lesions generally cover only a small percentage of the high-resolution screening images. Patch patterns are aggregated into image-features in a weighted manner. Such weights can be visualized as a heatmap to make the diagnosis more explainable.

• A novel spatial regulator is introduced to guide the classifier focus on the cervix region. The training requires no manual ROI annotations. Such a hassle free solution is valuable for professional fields with a limited number of experts. The regulator is refined by a novel image-specific iterative algorithm to capture the data variations.

• We evaluate the approach on an 18-year real-world dataset including 490 non-redundant sessions and observe at least 3.47%, 4.59%, 8.54% improvements over existing approaches in accuracy, F1, and recall scores, respectively.

Related Work
We focus on the existing literature that uses the visual information for detection since other modalities are not always available. The early approaches focus on feature engineering and various color or texture based features are proposed (Li et al. 2007; Kim and Huang 2013; Song et al. 2014). A more recent work (Xu et al. 2017) proposes a combination of the pyramid histogram in LAB color space, HOG and LBP (PLAP-PLAB-PHOG) to further surpass the previous studies. With the quick development of deep learning, the latest studies explore end-to-end feature-learning and observe a better performance. For example, Xu et al. (Xu et al. 2017) use the CaffeNet to outperform existing hand-crafted features. Other convolutional neural networks have been employed such as the AlexNet (Xu et al. 2016), LeNet (Vasudha and Juneja 2018) and Faster RCNN (Hu et al. 2019). Sato et al. (Sato et al. 2018) designed their own network structure and obtain a validation accuracy of around 50% on an in-house dataset. All the above solutions require a pre-processing stage to manually annotate the small cervix region in the raw screening images and the cropped region is used for analysis. However such manual labeling requires expert knowledge and is very time-consuming so that many solutions cannot be applied to the general un-annotated cervical screening images. On the other hand, most of these techniques (Xu et al. 2017, 2016; Vasudha and Juneja 2018; Sato et al. 2018) output a classification label without explaining how the decisions come from, which restrict them to be applied in hospitals due to trust problem (Gu et al. 2020). In contrast, our approach generates not only a binary label but also an interpretation map and the whole process requires no manual annotation for the cervix-region.

Methodology
In the proposed approach we divide each high-resolution image into same-sized patches which capture more local information, and aim to predict the image-level label (abnormal or normal) from such patch-collections. Mathematically, we denote a set of $n$ images as $\{I_1, I_2, \ldots, I_n\}$ and their binary labels as $\{Y_1, Y_2, \ldots, Y_n\}$. For each label, $Y_i \in \{0, 1\}$
where 1 indicates an abnormal (pre-cancer stage) case and 0 otherwise. For an image \( I_i \) containing \( k \) patches, we use \( X_i = \{x^j_i\}, j = 1 \sim k \) to denote its patch-collection. Each patch occupies a set of non-overlapping image pixels and \( \mathcal{L}_i \) denotes all patch-locations for the image \( I_i \).

**Baseline Patch Feature Learning & Aggregation** An intuitive approach for the patch-wise learning is to extract the individual patch patterns, aggregate them as an image-level features, and pair with an image-level label for binary classification. Patch pattern could be a hand-crafted feature or learned through an end-to-end feature representation model. We adopt the latter one to minimize the manual efforts by using a CNN-based network. Concretely, each image patch goes through a stacked of convolutional layers, followed by non-linear activation functions such as ReLu. Denote the parameters in the non-linear transformation as \((w, \theta)\) and the patch features as \( g(x^j_i | w, \theta) \), then the image feature \( f(I_i) \) is created by aggregating all its patch features as in Eqn:\[1\]

\[
f(I_i | w, \theta, X_i) = \bigoplus_{x^j_i \in X_i} g(x^j_i | w, \theta)
\]

where \( \bigoplus \) represents an aggregation operation such as concatenation or average. The network architectures could be adjusted for differently sized patches with details in the experiment section. The image-level features are imported to a classifier to output an image-level prediction probability and the binary label is threshold-determined easily. We use a multilayer perceptron network with two hidden layers and the training objective is to minimize the binary cross-entropy loss across \( n \) training examples as in Eqn:\[2\]

\[
L_c = \sum_{i=1}^{n} -(\log(P_i) \cdot Y_i + (1 - \log(P_i)) \cdot (1 - Y_i))
\]

where \( P_i \) is the predicted probability that an image \( I_i \) is likely to be cervical cancerous.

**Patch Weights Learning** The above formulation ignores the relationship between image label and patch labels. Actually, for a given cervical screening image, it will be a normal case if and only if all its composed patches are free from abnormal tissue. Assume that we have the ground-truth binary label \( y^j_i \in \{0, 1\} \) for each individual patch \( x^j_i \), then the label-relationship between the patches and their corresponding image is represented by the following equation:

\[
Y_i = \begin{cases} 
0, & \text{iff } \sum y^j_i = 0, \\
1, & \text{otherwise}. 
\end{cases}
\]

Practically, we do not have the ground truth binary labels for patches so that a patch-level classification is impossible. But such a relationship is still of value to improve the classification as the patch patterns found in the normal (negative) case images should not (or not that much) trigger abnormal (positive) labels whereas the opposite is not true. So we relax the patch labels from a hard binary value (0 or 1) to a soft continuous value (between 0 and 1) and convert the patch-image relationship from the label-level in Eqn:\[3\] to the feature-level as in Eqn:\[4\]. Specifically, we assume that patches contribute to its corresponding image feature with different weights.

\[
f(I_i | w, \theta, X_i) = \bigoplus_{x^j_i \in X_i} a^j_i \cdot g(x^j_i | w, \theta)
\]

where \( a^j_i \) denotes the weight for patch \( x^j_i \). This new image-level feature is used to achieve the objective in Eqn:\[2\].

**Domain-Knowledge Driven Spatial Regulation** So far the patch-weights are learnt solely from the patch visual contents but actually some basic domain knowledge could further facilitate a more reasonable patch-weight distribution. In real-world situations, all cervical colposcopic screening images are photographed from a vagina (bottom)-uterus(up) direction. As observed from Fig:\[3\] when viewed in such an upwards way, the cervix region looks quite like a circular blob-alike region and its center is near the cervix-os, which is the opening in the lower part of the cervix between the uterus and the vagina \[Fayed 2019]\]. Such a blob-alike region covers the transformation zone where cluster most of the abnormal cervical cancer tissues if exist.

This indicates a correlation between some visual contents and their spatial locations in an image. We use such correlation to further regulate the weight distribution so that the training can focus more on the patches within this blob-alike region. Specifically, we introduce a 2D spatial regulator, denoted by \( S \), to indicate the location of the cervical region and propose to learn the key patches in a manner such that they have a good spatial match with the cervix region. As such, another spatial loss is further introduced as below:

\[
L_{spatial}(S) = \sum_{i=1}^{n} \sum_{m \in \mathcal{L}_i} D(a_m, s_m)
\]
where $\mathcal{L}_i$ records all patch locations in an image and $D(\cdot, \cdot)$ calculates the dis-similarity between the weight and the spatial regulator by subtracting their dot-product sum by Eqn. (6).

$$D(a, s) = 1 - \frac{a - a_{\min}}{a_{\max} - a_{\min}} \cdot \frac{s - s_{\min}}{s_{\max} - s_{\min}}$$  \hspace{1cm} (6)

The overall loss is finalized as a combination of the classification error and the spatial regulation loss as in Eqn. (7).

$$L = \lambda \cdot L_{\text{spatial}} + L_c$$  \hspace{1cm} (7)

where $\lambda$ is an adjustable weight. In this work, we test our hypothesis by formulating the spatial regulator as a 2D Gaussian distribution with parameters $(\mu, \sigma)$. Here, $\mu$ defines the offset along the horizontal $(h)$ and vertical $(v)$ dimensions and $\sigma$ is the covariance matrix to define the Gaussian shape.

$$\mu = \begin{bmatrix} \mu_h \\ \mu_v \end{bmatrix}$$  \hspace{1cm} (8)

$$\sigma = \begin{bmatrix} \sigma_{hh} & \sigma_{hv} \\ \sigma_{vh} & \sigma_{vv} \end{bmatrix}$$  \hspace{1cm} (9)

Based on these two parameters, we can calculate the density value $s_m$ for any location $m \in \mathcal{L}_i$ by Eqn. (10).

$$s_m(\mu, \sigma) = \frac{1}{2\pi |\sigma|^{1/2}} \exp\left(-\frac{1}{2}(s_m - \mu)^T \sigma^{-1}(s_m - \mu)\right)$$  \hspace{1cm} (10)

However, such density parameters $(\mu, \theta)$ are unknown and we will explain in next section how to estimate them accurately for each cervical image.

**Image-Specific Iterative Spatial Regulator Estimation**

To estimate the spatial regulator’s parameter, a naive approach is to choose a pre-defined value according to the prior knowledge and fix it the same for all images. However, such an one-off setting lacks compatibility with various aspects of different cervical images. For example, the cervical region might not always be placed at the image center, or the cervix shape might not be a perfect circle. Thus, instead of using a single and fixed regulator, we design an iterative estimation for each individual image. In particular, for a given epoch at time $t$ and for each individual image, we estimate the image-specific Gaussian parameters from its own patch-weights collections $\{a_m\}$, $m \in \mathcal{L}_i$ via Maximun Likelihood. Subsequently the image-specific spatial regulator $\hat{S}_t^i$ is updated based on the estimated Gaussian parameters and will be used to calculate the spatial loss in the next round at time $t + 1$. The overall training spatial loss is refined to the following form:

$$L^{t+1} = \lambda \cdot L_{\text{spatial}}(\hat{S}_t^i) + L_c$$  \hspace{1cm} (11)

where $\hat{S}_t^i = \{\hat{S}_t^i(I_1), \hat{S}_t^i(I_2), ..., \hat{S}_t^i(I_n)\}$ contains a collection of refined image-specific spatial regulators.

**Algorithm 1:** Iterative spatial regulator update.

**Result:** Regulators $S$

Initialization: $\hat{S}_0 = G_0$;

for each training epoch $t$ do

for each training image $I_j$ do

\{ $f(x_i^t)$ \} ← Patch feature extraction;

\{ $a_i^t$ \} ← Patch weight learning;

$(\mu, \sigma)$ ← Parameters estimation from patch weights;

if $(\mu, \sigma)$ is valid then

$\hat{S}_t^i(I_j)$ ← Update the regulator;

end end

end

Alg.1 summarizes the major steps for the iterative procedure. We initialize each training image’s regulator $G_0$ to the same values by setting $\mu_h = w_0, \mu_v = h_0, \sigma_{hh} = w_0, \sigma_{vv} = h_0, \sigma_{hv} = 0$, and $\sigma_{vh} = 0$, where $w_0$ and $h_0$ are half of the image width and height, respectively. For each image, after patch feature extraction, we learn the patch weights, based on which the parameters of its regulator are estimated. The regulator will be updated if the estimated parameters pass validation by checking if the center location is within the image size. Via such training, the algorithm focuses on the region-of-interests on-the-fly without increasing additional computation complexity.

**Experiments**

**Dataset** We evaluate the approach on a real-world database from the U.S. National Cancer Institute (NCI) (Institute 2020). The dataset is accessible based on request and under constrained agreement. This dataset is from a longitudinal study in Costa Rica: Proyecto Epidemiologico Guanacaste 2020). We evaluate the approach on a real-world database from the U.S. National Cancer Institute (NCI) (Institute 2020).
parts. Each session may consist of more than one images which share similar visual contents, so the train/test split is based on sessions for fairness. All images have resolutions around 2400 x 1600. Evaluations includes balanced-accuracy, F1, precision, recall and AUC scores of ROC curve and equal error rate (EER).

**Parameters**  
By default, the image is divided into 42x42 patches. The feature representation CNN has three convolutional layers with 12, 24, 48 filters and the size of filters is 3 x 3, followed by ReLu activation and max-pooling. The weights-learning module contains three dense layers of 800, 512, 128 nodes. The aggregated feature is passed to a dense layer activated by Sigmoid. The default parameters are Xavier initialized for all layers. By default, the learning rate is $0.1$ over a 150-epoch training with early-stopping.

![Figure 4: Two images from the same session are very similar. Our training and testing data contains mutually-exclusion sessions.](image)

**Baselines**  
We compare our model with the following state-of-the arts: 1) **Xu’17** (Xu et al. 2017): it is one of the most active groups working in this field and they use a CaffeNet based transfer learning model to solve the problem which surpassed the best reported hand-crafted feature (around 1% improvement) on the same dataset. We follow their parameter settings and achieve similar results as reported in this paper. 2) **Vasuda’18** (Vasudha and Juneja 2018): it uses the LeNet-based transfer learning but its training and testing data contain the same-session’s images so their reported results contains some bias. 3) **VGG16, InceptionNet**. Both above methods are based on transfer-learning so we further test some more recent backbone networks including VGG16 and InceptionNet. VGG16 is much larger than InceptionNet so we freeze different blocks of VGG16 and report the results accordingly. 4) **Hu’19+AN1500** (Hu et al. 2019): We compare a simplified version of this paper that uses RCNN to crop cervix region (ROI) before classification. This model requires additionally experts-labeled ROI ground truth (not available for public) so we cannot fully reproduce it. But as our images are captured in a very controlled environment and the image centers mostly correspond to the ROI, we annotate the center $k \times k$ pixels as ROI labels instead. To choose a proper $k$, we test the traditional transfer-learning models by using the center $k \times k$ as input for $k = 900, 1200, 1500$ and observe the size 1500 give the best results. So we choose $k = 1500$ as the annotations to train the RCNN accordingly. Note that this paper (Hu et al. 2019) uses an in-house training dataset 2.4 times patient numbers as ours.

**Quantitative Results**  
Table 1 reports the performance on multiple metrics where a few observations are made:

- The proposed method surpasses existing solutions in most scenarios where recall has the largest minimal-increment of 8.54%, followed by F1 of 4.59% and accuracy of 3.47%. In rare-cancer detection, the recall is a very important measurement to avoid missing of a cancerous case. We visualize the approaches’ ranking in Fig. 5a.

<table>
<thead>
<tr>
<th>Model</th>
<th>Annotation</th>
<th>Accuracy</th>
<th>F1</th>
<th>Precision</th>
<th>Recall</th>
<th>AUC ROC</th>
<th>1-EER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vasudha’18</td>
<td>No</td>
<td>0.7415</td>
<td>0.6923</td>
<td>0.7297</td>
<td>0.6585</td>
<td>0.8143</td>
<td>0.7895</td>
</tr>
<tr>
<td>Xu’17</td>
<td>No</td>
<td>0.7657</td>
<td>0.7326</td>
<td>0.7</td>
<td>0.7683</td>
<td>0.8049</td>
<td>0.7632</td>
</tr>
<tr>
<td>VGG16,FZ1*</td>
<td>No</td>
<td>0.7267</td>
<td>0.6752</td>
<td>0.7067</td>
<td>0.6463</td>
<td>0.775</td>
<td>0.6491</td>
</tr>
<tr>
<td>VGG16,FZ2</td>
<td>No</td>
<td>0.7145</td>
<td>0.6581</td>
<td>0.6986</td>
<td>0.622</td>
<td>0.7926</td>
<td>0.7193</td>
</tr>
<tr>
<td>VGG16,FZ3</td>
<td>No</td>
<td>0.7082</td>
<td>0.6627</td>
<td>0.6548</td>
<td>0.6707</td>
<td>0.7939</td>
<td>0.7368</td>
</tr>
<tr>
<td>VGG16,FZ4</td>
<td>No</td>
<td>0.7318</td>
<td>0.6871</td>
<td>0.6914</td>
<td>0.6829</td>
<td>0.7822</td>
<td>0.7018</td>
</tr>
<tr>
<td>InceptionNet</td>
<td>No</td>
<td>0.6049</td>
<td>0.4458</td>
<td>0.6512</td>
<td>0.3415</td>
<td>0.6825</td>
<td>0.6491</td>
</tr>
<tr>
<td>Hu’19 + AN1500**</td>
<td>Yes</td>
<td>0.5764</td>
<td>0.3455</td>
<td>0.6786</td>
<td>0.2317</td>
<td>0.8145</td>
<td>0.7807</td>
</tr>
<tr>
<td>Vasuda’18 + Crop1500`</td>
<td>Yes</td>
<td>0.7737</td>
<td>0.7362</td>
<td>0.7407</td>
<td>0.7317</td>
<td>0.8214</td>
<td>0.7807</td>
</tr>
<tr>
<td>Xu’17 + Crop1500</td>
<td>Yes</td>
<td>0.7581</td>
<td>0.7134</td>
<td>0.7467</td>
<td>0.6829</td>
<td>0.8194</td>
<td>0.7895</td>
</tr>
<tr>
<td>VGG16,FZ2 + Crop1500</td>
<td>Yes</td>
<td>0.7518</td>
<td>0.7143</td>
<td>0.6977</td>
<td>0.7317</td>
<td>0.8299</td>
<td>0.7456</td>
</tr>
<tr>
<td>Proposed</td>
<td>No</td>
<td>0.8084</td>
<td>0.7821</td>
<td>0.7216</td>
<td>0.8537</td>
<td>0.8354</td>
<td>0.7719</td>
</tr>
</tbody>
</table>

*FZ$n$: freeze the lower $n$ blocks of the VGG16. **AN$n$: annotate center $n \times n$ pixels as cervix bounding box.

*Crop$n$: crop the center $n \times n$ pixels as the model inputs.

Table 1: Performance Comparison across Different Approaches.
global-features in our problem and this explains the effectiveness of our patch-wise solutions.

**ROC Curve**  
Fig. 5b compares the ROC curves among all annotation-free approaches. The closer the curve is to the upper-left corner, the better that approach works. Our model works the best but there is still room for improvements.

**Score Distribution & Feature Visualization**  
Fig. 5c shows the predicted score distribution where the normal (blue) cases are skewed-left while the abnormal (green) ones are skewed right. Such a two-end skewed-distribution suggests an effective classification, which is further validated by the 2D T-SNE plot in Fig. 5d. Points from each category are clustered together while the two clusters are separated.

**CIN Grades & Temporal Factor**  
The binary classes (cancer and non-cancer) come from five CIN grades labeled from 0 (normal) to 4 (cancerous), we also analyze the prediction statistics for each category in Fig. 5e. The medium values (the red line in each box), generally go higher along the grade level, indicating a potential possibility for fine-level categorization. Lastly, we discuss the temporal factor influence. Each image is labeled within 1-year from its image-taken date so we divide this 1-year window into four quarters and visualize the average probability in each quarter in Fig. 5f. We have not observed obvious trending and this matches the fact that abnormal cells have the potential to progress to cancer, but may also regress to normal or remain unchanged (Wang et al. 2013).

**Interpretability**  
Other than the improvements under standard metrics, our approach has a better interpretability by outputing an additional heatmap from the patch-weights. The first two columns Fig. 6 illustrates a few examples. Column (a) is the screening images and column (b) is the interpretation maps where the bright color indicates the areas where features are mainly learnt from. We can see that they mostly cover the transformation zone within the cervix region. This is an area of changing cells, and it is the most common place on the cervix for abnormal cells to develop.

**Impact of Spatial Regulator**  
We remove the spatial regulation module and show the interpretation maps in the column (c) in Fig. 6. The maps are significantly varied from the original maps in column (b) by highlighting mostly the outer background parts which are incorrect. Thus, we surmise that an explainable result might be very important when using computational solutions for cervical dysplasia diagnosis as the doctor can validate the predictions more easily.

**Impact of Iterative Algorithm**  
To illustrate the effectiveness of our iterative regulator refinement, we visualize the estimated Gaussian over the temporal dimension in the last
four columns of Fig. [4] The yellow color corresponds to the Gaussian center. The updates are reflected in e.g., the cervix shape, the center location, the width-height-ratio and the rotation perspectives. 1) Location-wise, the highlighted areas are gradually shifted to the cervical central area for each individual image through the training. 2) Size-wise, the spread of the Gaussian is gradually converged to a small region. At the very beginning, the Gaussian occupies almost the whole image and it is similar for all examples. After a few steps, the Gaussian focus only on the potential ROI which takes a small ratio over the image. 3) shape-wise, the Gaussian rotates differently for different cases.

**Table 2: Performance vs. Patch Size.**

<table>
<thead>
<tr>
<th>Patch Size</th>
<th>Accuracy</th>
<th>F1</th>
<th>Precision</th>
<th>Recall</th>
<th>ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5xDefault</td>
<td>0.734</td>
<td>0.687</td>
<td>0.705</td>
<td>0.670</td>
<td>0.789</td>
</tr>
<tr>
<td>Default</td>
<td>0.808</td>
<td>0.782</td>
<td>0.721</td>
<td>0.853</td>
<td>0.835</td>
</tr>
<tr>
<td>2xDefault</td>
<td>0.791</td>
<td>0.764</td>
<td>0.708</td>
<td>0.829</td>
<td>0.821</td>
</tr>
<tr>
<td>4xDefault</td>
<td>0.736</td>
<td>0.701</td>
<td>0.663</td>
<td>0.743</td>
<td>0.765</td>
</tr>
</tbody>
</table>

**Impact of Patch Size.** Results using different patch sizes (n x Default size) are reported in Table 2. A larger patch size focuses more on the global clues and results a relatively lower performance. This matches the second observations in our quantitative results discussion where a shallow network works better. Another plausible reason might be that a larger patch size will reduce the total number of training patches and this further affect the accuracy eventually.

Figure 6: (a) Input cervical images and each row is an example. (b) The final maps where the bright color indicates the features from those regions are more likely to trigger the prediction labels. Our highlights well correspond to the cervix transformation zones. (c) The maps using our model but without spatial regulation and the highlight regions focus more on the background regions. (d-g) These four columns show the refined spatial regulator learnt at epochs 1, 32, 48 and the final one. Via the proposed iterative algorithm, we can see that the regulator gradually changes in locations, sizes and shapes until a good fit to the cervix.

**Conclusion and Future Work**

This work proposes a patch-wise solution for cervical cancer image classification. Compared to the majority of solutions in this field which use high-resolution images directly, it has the advantage of retaining local visual details and this is important for medical applications. During the patch feature aggregation, the approach also learns the patch contribution weights, from which an interpretation map is created to indicate which regions mostly trigger the prediction. The framework further integrates basic domain knowledge and introduces an adjustable spatial regulator to control where the classifier should focus on. We have designed a novel iterative training to capture the cervix-image data diversity on-the-fly, so that the variations in terms of center location, size, and rotation are automatically refined for individual images, making the approach more flexible and dynamic. Extensive experiments have been performed to evaluate the approach with significant improvements observed.

In future, we will explore to incorporate some stopping criteria, so that the approach may finalize at a stage to cover a more complete view of the transformation zone. Another direction is to study the effect of various kernels so that the approach can cater to the cervix shape more accurately.

**Acknowledgement**

This research was supported by the MSIT, Korea, under the ITRC support program (IITP-2020-2020-0-01789) supervised by the IITP. We thank David Levitz and Yonit Zall from MobileODT Ltd. for advise and suggestions.
References


