



**HAL**  
open science

# Computer-aided diagnosis methods for cervical cancer screening on liquid-based Pap smears using convolutional neural networks: design, optimization and interpretability

Antoine Pirovano

► **To cite this version:**

Antoine Pirovano. Computer-aided diagnosis methods for cervical cancer screening on liquid-based Pap smears using convolutional neural networks: design, optimization and interpretability. Artificial Intelligence [cs.AI]. Institut Polytechnique de Paris, 2021. English. NNT: 2021IPPAT011. tel-03335365

**HAL Id: tel-03335365**

**<https://theses.hal.science/tel-03335365v1>**

Submitted on 6 Sep 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



INSTITUT  
POLYTECHNIQUE  
DE PARIS

NNT : 2021IPPAT011

Thèse de doctorat



# Computer-aided diagnosis methods for cervical cancer screening on liquid-based Pap smears using Convolutional Neural Networks: design, optimization and interpretability.

Thèse de doctorat de l'Institut Polytechnique de Paris  
préparée à Télécom Paris

École Doctorale de l'Institut Polytechnique de Paris (ED 626)  
Spécialité de doctorat : Signal, Images, Automatique et Robotique

Soutenance présentée et soutenue à Palaiseau (91477), le 07/05/2021, par

**ANTOINE PIROVANO**

Composition du Jury :

Charles Kervrann Directeur de recherche, Inria Rennes	Président
Valery Naranjo Ornedo Professeur, Institute for Research and Innovation in Bioengineering, Universitat Politècnica de València	Rapporteuse
Thomas Walter Directeur de recherche, Centre for Computational Biology (CBIO), MINES ParisTech	Rapporteur
Henning Müller Professeur, University of Applied Sciences and Arts Western Switzerland	Examineur
Laetitia Lacoste-Collin Praticienne hospitalière, Medipath	Examinatrice
Saïd Ladjal Maître de conférence, Télécom Paris (LTCl)	Co-directeur de thèse
Isabelle Bloch Professeure, Télécom Paris (LTCl)	Directrice de thèse
Sylvain Berlemont CEO, Keen Eye	Invité



# Acknowledgments

I'd like to start by thanking the members of the jury (Charles Kervrann, Valery Naranjo, Thomas Walter, Henning Muller and Laetitia Collin) that accepted to dive into my work and evaluate it.

To me, these acknowledgments words might be the most important part of this manuscript as for the past three years, and overall through my whole life, I have been surrounded by kind, caring and inspiring people that made me feel happy and thus made this work possible. Most of the following words will be written in French.

J'aimerais commencer par dire que je suis une personne privilégiée et qu'il me semble important de mettre le modeste succès de ces travaux en perspective dans ce contexte qui avantage bien trop les personnes comme moi.

Cela étant dit, j'ai en plus bénéficié d'un environnement particulièrement doux, attentionné et dynamique qui a fait de ce doctorat une aventure humaine incroyable durant laquelle, au-delà de l'aspect technique, j'ai pu grandir et m'épanouir pleinement. C'est pourquoi il est important pour moi de prendre le temps de mettre en avant les personnes qui ont participé à ce voyage et de leur exprimer toute ma reconnaissance.

Saïd and Isabelle, merci pour l'accompagnement, vos conseils et votre exigence mais aussi pour votre soutien et votre confiance en mon travail. Vous m'avez fait me sentir à l'aise tout en me guidant quand j'en avais besoin. Je n'aurais pas pu imaginer meilleur encadrement.

Sylvain, il est difficile de te dire à quel point j'admire profondément tout ce que tu as créé et fais à Kene Eye, ta disponibilité et la façon dont tu gères cette incroyable entreprise. J'ai eu la chance d'observer Keen Eye grandir et ça m'a tant appris puis finalement maintenant même donné l'envie d'entreprendre à mon tour. J'ai aussi une pensée pour Leandro qui m'a accompagné sur la première partie de ce doctorat.

L'équipe DS. La team vous le savez, je vous aime vraiment très (trop?) fort. Je suis super fier de nous et de la façon dont nous avons grandi ensemble et construit, basé sur la confiance, une équipe forte qui cherche sans arrêt à s'améliorer et dans laquelle tout le monde peut trouver une place où il/elle se sent bien. Hippo, merci pour ton soutien et la prise en main du management de l'équipe que tu as si bien géré. Thomas, je rigole rien qu'en écrivant ton nom en me rappelant toi, en boule, en nage, dans le coin de la meeting room pleine... T'es juste le meilleur. PK, j'admire ta gentillesse et la façon dont tu t'intéresses à tout et dont tu es toujours partant pour une douce poelade. Yan, je suis toujours aussi admiratif déjà de tes compétences et de ta pertinence technique, mais aussi de ta qualité de communicant, tu m'as tellement aidé et a toujours été super patient avec moi (same pour Thomas au passage). Louis, merci pour toutes les fois où tu m'as fait sourire et rire, je suis super impressionné par ta rigueur d'un côté et de l'autre côté des pertes de contrôles totales que tu sais offrir. Mélanie, merci pour toutes les fois où tu m'as porté et encouragé lors de nos projets, c'est

un vrai plaisir de travailler avec une personne si motivée, engagée et intelligente. Il me tarde d'assister à ta soutenance dans deux ans et demi, tu vas tout defoncer.

## KeenEye

Figure 1: Keen Eye logo.

Plus généralement, je tiens à remercier toutes les personnes qui composent Keen Eye ou qui y ont croisé mon chemin. J'avais toujours pensé que travailler en entreprise c'était rentrer dans un standard et ne partager que du professionnel. Vous m'avez montré à quel point c'est faux et permis de m'épanouir dans cette incroyable activité. Merci aussi à Yaelle, Jacques, Zelma et Philippe d'avoir partagé leurs expertises, de s'être intéressés à mes travaux et d'avoir répondu à mes questions.

Aussi, j'ai eu la chance d'être intégré dans une super équipe de doctorants à Télécom Paris. En particulier, je tiens à remercier Nicolas, Bastien, Clément, Mateus et Emanuele pour leur gentillesse et disponibilité qui m'ont permis de me sentir si bien à Télécom.

Si professionnellement cette période a été intense émotionnellement et pleine de super expériences, personnellement c'est encore plus vrai.

Je veux ici commencer par remercier ma famille, mon papa Alain, ma maman Nelly, mes soeurs Mathilde, Coline et Adèle et mes quatre grands-parents Raymond, Jeannine, Claude et Michelle pour tout leur amour et l'éducation qu'ils m'ont donné.

Je tiens aussi à remercier mes amis, les personnes avec qui j'ai grandi, qui font de moi la personne que je suis et qui constituent encore aujourd'hui une partie importante des gens qui me rendent heureux: Antoine Momo, Adrian, Paul, Enzo, Francois, Fanny.

Ma vie aujourd'hui est à Paris et cette vie s'est construite en grande partie grâce au sport et plus particulièrement aux personnes que j'y ai rencontré. Thomas et Laura, mes coachs et amis, merci pour tout, vos conseils, votre générosité, votre accompagnement, votre gentillesse, votre humour. Vous voir construire une famille ensemble ces deux dernières années est vraiment une source de bonheur incroyable surtout avec l'arrivée de la petite Alma. Hassan et Antoine, je suis tellement heureux d'avoir trouvé des amis comme vous, vous m'inspirez tellement par votre gentillesse et votre intelligence. Avec vous je sais que si j'ai un doute ou une question, j'ai un endroit où trouver des réponses et de la bienveillance.

Je veux aussi remercier mes colocataires qui m'ont créé un cocon dans lequel j'ai pu m'épanouir et travailler malgré la période compliquée, Nora, Juliette, Cintia, Fabien, Thiago, Gaetano, Clément et Chloe.

Et je vais finir par remercier Daniela, ma petite amie que j'aime très fort et qui me donne confiance pour la suite.

Vous avez tous contribué à la réussite de ce doctorat. Je vous aime !

Antoine



# Publications

## Papers in conference proceedings

A. Pirovano, L. G. Almeida, S. Ladjal, “Regression Constraint for an Explainable Cervical Cancer Classifier”, in Groupe d’Etudes du Traitement du Signal et des Images (GRETSI), Aug. 2019. Available: <https://arxiv.org/abs/1908.02650>

A. Pirovano, H. Heuberger, S. Berlemont, S. Ladjal, I. Bloch, “Improving Interpretability for Computer-aided Diagnosis tools on Whole Slide Imaging with Multiple Instance Learning and Gradient-based Explanations”, In: Cardoso J. et al. (eds) Interpretable and Annotation-Efficient Learning for Medical Image Computing. IMIMIC 2020, MIL3ID 2020, LABELS 2020. Lecture Notes in Computer Science, vol 12446. Springer, Cham. [https://doi.org/10.1007/978-3-030-61166-8\\_5](https://doi.org/10.1007/978-3-030-61166-8_5)

## Papers in international journals

A. Pirovano, H. Heuberger, S. Berlemont, S. Ladjal, I. Bloch, “Automatic Feature Selection for Improved Interpretability on Whole Slide Imaging”, in Machine Learning and Knowledge Extraction (MDPI), vol. 3(1), pp. 243-262; DOI:10.3390/make3010012. 2021.

A. Pirovano, L. G. Almeida, S. Berlemont, I. Bloch, S. Ladjal, “Computer-Aided Diagnosis tool for Cervical Cancer Screening with Weakly Supervised Localization and Detection of Abnormalities using Adaptable and Explainable Classifier”, in Medical Image Analysis, vol. 73. DOI:10.1016/j.media.2021.102167. 2021.



# Résumé long en français

Le cancer du col de l'utérus est le deuxième cancer le plus important pour les femmes après le cancer du sein. En 2012, le nombre de cas recensés dépasse 500,000 à travers le monde, dont la moitié se sont révélés mortels.

Jusqu'à maintenant, le dépistage primaire du cancer du col de l'utérus est réalisé par l'inspection visuelle de cellules, prélevées par frottis vaginal, par des cytopathologistes utilisant la microscopie en fond clair dans des laboratoires de pathologie. Chaque lame peut contenir jusqu'à 100.000 cellules. En France, environ 5 millions de dépistage sont réalisés chaque année et environ 90% mènent à un diagnostic négatifs (i.e. pas de changements précancéreux détectés).

En terme d'ordre de grandeur cette consiste à chercher quelques balles de ping-pong sur une vingtaine de terrain de foot en sachant que statistiquement ces balles ne peuvent être trouvées que sur deux de ces terrains, ce qui fait de ces analyses au microscope une tâche extrêmement fastidieuses et coûteuses en temps pour le cyto-techniciens et peut nécessiter l'avis conjoint de plusieurs experts. Ce processus impacte la capacité à traiter cette immense quantité de cas et à éviter les faux négatifs qui sont la cause principale des retards de traitements médicaux. Le manque d'automatisation et de traçabilité des dépistage deviennent ainsi de plus en plus critique à mesure que le nombre d'experts diminue.

En ce sens, l'intégration d'outils numériques dans les laboratoires de pathologie devient une réelle problématique de santé publique et la voie privilégiée pour l'amélioration de ces laboratoires.

Depuis 2012, l'apprentissage profond a révolutionné le domaine de la vision par ordinateur, en particulier grâce aux réseaux de neurones à convolutions qui se sont montrés fructueux sur un large panel d'applications parmi lesquelles plusieurs en imagerie bio-médicale. Parallèlement, le processus de digitalisation de lames entières a ouvert l'opportunité pour de nouveaux outils et de nouvelles méthodes de diagnostic assisté par ordinateur.

Dans cette thèse, après avoir motivé le besoin médical et introduit l'état de l'art en terme de méthodes d'apprentissage profond pour le traitement de l'image et en particulier pour le traitement de lames entières, nous présentons nos contribution au domaine de la vision par ordinateur traitant le dépistage du cancer du col de l'utérus dans un contexte de cytologie en milieu liquide.

Notre première contribution consiste à proposer une méthode simple de régularisation pour l'entraînement de modèles dans le contexte d'une classification ordinaire (i.e. classes suivant un ordre). Nous démontrons l'avantage de notre méthode pour la classification de cellules utérines en utilisant sur le jeu de données Herlev sur lequel nous définissons un nouvel état-de-l'art (66.8% de précision pour la classification de sévérité, 95.2% pour la classification binaire entre les classes "normal" et "anormale" et un score KAPPA de 0.87). De plus, nous proposons

de nous appuyer sur des explications basées sur le gradient pour réaliser une localisation faiblement supervisée (précision de 80.4%) et plus généralement une détection d’anormalité. Pour cela, nous créons un jeu de donnée qui simule des régions de lames contenant plusieurs cellules. Finalement, nous montrons comment nous intégrons ces méthodes pour créer un outil assisté par ordinateur qui pourrait être utilisé afin de réduire la charge de travail des cytopathologistes en proposant un diagnostic a priori sur la lame et en identifiant des cellules d’intérêt afin de guider la revue de la lame par l’expert.

La seconde contribution se concentre sur la classification de lames entières et l’interprétabilité de ces approches. Nous formalisons le design commun des architectures de classification de lames entières s’appuyant sur un contexte de “multiple instance learning” et proposons une approche d’interprétabilité par morceaux s’inscrivant dans ce formalisme. Cette approche repose sur les méthodes d’explicabilité basées sur le gradient, la visualisation de caractéristiques et le contexte de “multiple instance learning”. A travers cette méthode, nous sommes capable d’expliquer aux experts sur quoi repose les décisions prise par l’algorithme. Nous étendons ce travail en proposant une nouvelle façon de calculer des cartes de chaleurs pouvant expliquer les décisions et guider l’expert dans sa revue. Deux études quantitatives, nous ermettent de valider la méthode et de prouver que nous améliorons la qualité des cartes de chaleur (de plus de 29% pour la mesure d’AUC). Finalement, nous appliquons ces méthodes pour le dépistage du cancer du col de l’utérus en utilisant un détecteur d’ “anormalité” qui guide l’entraînement pour l’échantillonnages de régions d’intérêt.

Finalement, nous concluons en discutant des perspectives que ces travaux ouvrent.



# Contents

<b>List of Abbreviations</b>	<b>xiii</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Medical context</b>	<b>5</b>
2.1 Cervical cancer screening . . . . .	5
2.1.1 Cervical cancer: what is it? . . . . .	5
2.1.2 Cervical cancer screening . . . . .	7
2.1.3 Liquid-based cytology slides content . . . . .	8
2.1.4 Management of abnormal cases, evolutions and further medical exams	14
2.1.5 Conclusion: advantages and limitations of liquid-based cytology Pap tests	16
2.2 Whole Slide Imaging process: towards digital pathology next generation tools	17
2.2.1 Camelyon-16: Breast Cancer detection on Biopsies . . . . .	18
2.2.2 The Cancer Genome Atlas (TCGA) . . . . .	19
2.2.3 Herlev dataset . . . . .	19
2.2.4 Conclusion . . . . .	20
<b>3 Deep learning: An opportunity for efficient computer-aided diagnosis tools</b>	<b>21</b>
3.1 Classification . . . . .	22
3.1.1 Metrics . . . . .	22
3.1.2 Feature extractor architectures . . . . .	24
3.1.3 Cervical cancer usage . . . . .	26
3.2 Localization and object detection . . . . .	27
3.2.1 Architectures . . . . .	28
3.2.2 Cervical cancer usage . . . . .	31

---

3.3	Whole slide image classification . . . . .	33
3.3.1	Methods and architectures . . . . .	33
3.3.2	Other methods for WSI classification . . . . .	39
3.3.3	Cervical cancer and cytology applications . . . . .	40
3.4	Explanations and interpretability . . . . .	41
3.4.1	Methods for interpretability . . . . .	41
3.4.2	Evaluation and quantification of interpretability . . . . .	43
3.4.3	Medical usage . . . . .	43
3.5	Conclusion and discussion . . . . .	44
<b>4</b>	<b>Cell and region classification for cytology whole slide imaging computer-aided diagnosis tools: proposed methods and experiments</b>	<b>45</b>
4.1	Cell-level classification . . . . .	46
4.1.1	Herlev severity . . . . .	46
4.1.2	Backbone comparison . . . . .	46
4.1.3	Feature fine-tuning . . . . .	48
4.1.4	Regression approach . . . . .	49
4.1.5	Classifier under regression constraint . . . . .	50
4.1.6	Interpretability . . . . .	52
4.2	Simulated tiles classification . . . . .	53
4.2.1	Simulated dataset . . . . .	54
4.2.2	Classification . . . . .	55
4.2.3	Interpretability . . . . .	62
4.2.4	Weakly supervised localization . . . . .	64
4.2.5	Weakly supervised abnormal cell detection . . . . .	65
4.3	Liquid-based cytology whole slide image classification . . . . .	66
4.3.1	Pre-processing . . . . .	68

4.3.2	Integration in a computer-aided diagnosis pipeline . . . . .	69
4.3.3	From tile-level predictions to slide-level diagnosis . . . . .	71
4.4	Conclusion . . . . .	72
<b>5</b>	<b>WSI classification: proposed methods and experiments</b>	<b>75</b>
5.1	Multiple instance learning approach for whole slide classification . . . . .	76
5.1.1	CHOWDER model . . . . .	76
5.1.2	Attention-based model . . . . .	77
5.2	Improving interpretability . . . . .	77
5.2.1	Formalization and tile scores . . . . .	77
5.2.2	Proposed method: Using gradient-based explanations . . . . .	78
5.2.3	Feature identification on trained CHOWDER and attention-based models	80
5.2.4	Tile-level explanations . . . . .	83
5.2.5	Feature-based heat-maps . . . . .	84
5.2.6	Measure of interpretability through heat-maps relevance . . . . .	84
5.2.7	Analysis of the number of features . . . . .	86
5.2.8	Colocalization filtering . . . . .	88
5.2.9	Application to the SFP Challenge . . . . .	89
5.3	LBC slides classification . . . . .	93
5.3.1	Medipath dataset . . . . .	93
5.3.2	MIL classical approach . . . . .	93
5.3.3	“Abnormality” detector sampling approach . . . . .	95
5.4	Conclusion . . . . .	100
<b>6</b>	<b>Conclusion and Perspectives</b>	<b>103</b>
<b>A</b>	<b>From Machine Learning to Deep Learning</b>	<b>109</b>
A.1	Machine Learning and image classification before 2012 . . . . .	109

---

A.1.1	Hand-crafted feature extraction . . . . .	109
A.1.2	ML classification methods . . . . .	111
A.1.3	2012 LSVRC Edition . . . . .	112
<b>B</b>	<b>Introduction to Convolutional Neural Networks and their training strategies</b>	<b>115</b>
B.1	Convolution layers . . . . .	115
B.2	Supervised learning . . . . .	118
B.3	Training strategies to reduce overfitting: regularization and transfer learning	120
B.4	Other kinds of learning . . . . .	124
	<b>Bibliography</b>	<b>139</b>

# List of Abbreviations

<b>AI</b>	Artificial Intelligence
<b>AGUS</b>	Atypical Glandular cells of Undetermined Significance
<b>AIS</b>	Adenocarcinoma In Situ
<b>ANN</b>	Artificial Neural Network
<b>ASC-H</b>	Atypical Squamous Cells that cannot exclude High-grade lesion
<b>ASCUS</b>	Atypical Squamous Cells of Undetermined Significance
<b>AUC</b>	Area Under the Curve
<b>CAD</b>	Computer-Aided Diagnosis
<b>CIN</b>	Cervical Intraepithelial Neoplasia
<b>CNN</b>	Convolutional Neural Network
<b>DL</b>	Deep Learning
<b>FPN</b>	Feature Pyramid Network
<b>GPU</b>	Graphical Processor Unit
<b>H&amp;E</b>	Hematoxylin & Eosin
<b>HPV</b>	Human Papilloma Virus
<b>HSIL</b>	High-grade Squamous Intraepithelial Lesion
<b>IOU</b>	Intersection Over Union
<b>LBC</b>	Liquid Based Cytology
<b>LSIL</b>	Low-grade Squamous Intraepithelial Lesion
<b>LSVRC</b>	Large-Scale Visual Recognition Challenge
<b>MIL</b>	Multiple Instance Learning
<b>ML</b>	Machine Learning
<b>MSE</b>	Mean Squared Error
<b>NCR</b>	Nucleo-Cytoplasmic Ratio
<b>NILM</b>	Negative for Intraepithelial Lesion or Malignancy

<b>ROI</b>	Region Of Interest
<b>ROC</b>	Receiver Operating Characteristic
<b>ReLU</b>	Rectified Linear Unit
<b>RPN</b>	Region Proposal Network
<b>SEC</b>	Squamous Epithelial Cells
<b>SCC</b>	Squamous Cells Carcinoma
<b>SGD</b>	Stochastic Gradient Descent
<b>SIFT</b>	Scale-Invariant Feature Transform
<b>SN</b>	Stain Normalization
<b>SVM</b>	Support Vector Machine
<b>WHO</b>	World Health Organization
<b>WSI</b>	Whole Slide Imaging

# Introduction

In the early 1940's, Dr. Papanicolaou discovered that a visual inspection of cells sampled at the entrance of the cervix could give evidence about the potential development of a cancer. Moreover, the World Health Organization (WHO) states that 90% of cervical cancer cases could be avoided if detected earlier. However, with more than 500,000 new cases every year and about 250,000 deaths, cervical cancer still is a major worldwide healthcare issue.

Today, the screening of cervical cancer is performed by highly trained cytopathologists assisted by cytotechnicians. These experts are screening microscopy slides containing the sampled cells in a drop of preservative liquid, this is called Liquid-Based Cytology (LBC). The slide that can be observed in Figure 1.1 highlights the difficulty of this task. Indeed, a single slide can contain up to 100,000 cells, and the diagnosis may rely on a few cells only. Moreover, on the one hand, most of the time there is no abnormality to be found but, on the other hand, when abnormalities are to be found (around 10% of cases) it is critical not to miss them. This requires time and expertise. These conditions make the task of screening efficiently cervical cancer a real challenge.

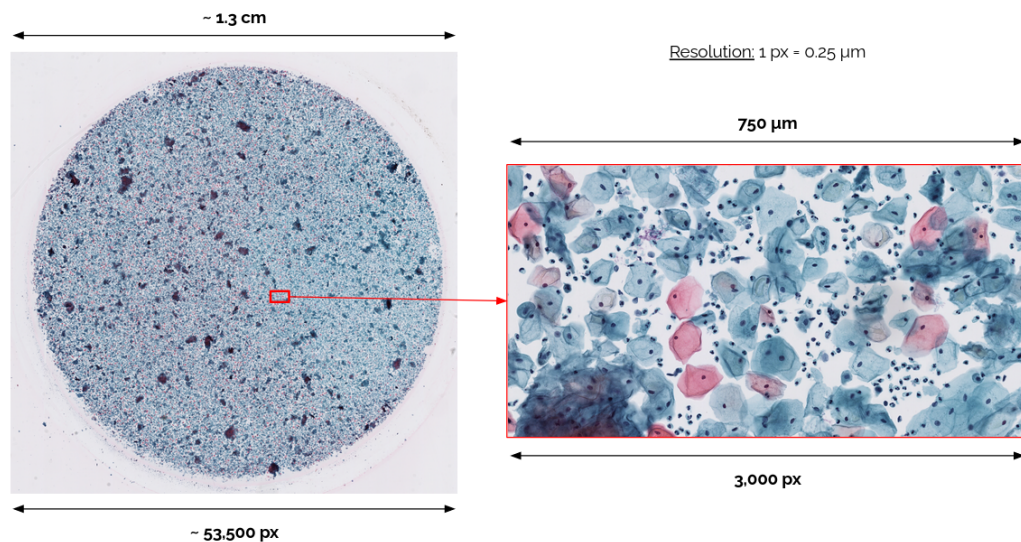


Figure 1.1: Illustration of a liquid-based cytology slide for cervical cancer screening

In the mean time, the emergence of computer science gave birth to a new field, called digital pathology, that regroups methods that bring together medical applications and computer science (scanners, viewers, data management ...). Recently, with the success of Machine

Learning (ML) methods in a large panel of fields and the improvement of whole slide imaging process, digital pathology is about to be revolutionized. ML defines the subset of Artificial Intelligence (AI) methods that are able to learn to perform a task through a training with examples. For images, Convolutional Neural Networks (CNN), that were initially inspired by the human visual cortex, are particularly popular, due to their incredible performances through a wide range of tasks. Whole Slide Image (WSI) are microscopy slides digitized at a really high resolution that enabled the creation of virtual microscopy. Thus, there is an opportunity for AI-based Computer-Aided Diagnosis (CAD) tools that can help cytopathologists navigating through this enormous quantity of information. This defines the context of the Ph.D. and raises the questions it will answer. Such as how to train an efficient cell classifier ? Or how to build an integrable and explicative CAD tool for cervical cancer screening ?

Thus the objective of this thesis is the development of an accurate, fast and explainable CAD tool to help cytopathologists in their daily routine.

This manuscript presents the work I did during my Ph.D. under the supervision of Saïd Ladjal and Isabelle Bloch, professors at Telecom Paris. The proposed methods have also been developed in close collaboration with the Data Science team of Keen Eye (Hippolyte Heuberger, Louis Jeay, Paul Klein, Thomas Le Meur, Melanie Lubrano and Yan Petit), Leandro G. Almeida (former CSO) and Sylvain Berlemont (CEO).

In Chapter 2, we detail the medical context with a particular focus on the type of cells and of malignancy that can be found on cervical LBC slides, and more generally the specificities of cervical cancer screening.

In Chapter 3, we introduce Deep Learning (DL) methods on which we, and more generally the computer vision community, rely on. We present the most popular feature extractors and their application for different tasks (image classification, object detection and WSI classification) and different cancers (breast, prostate, lung ...) that get close to our problem and will enable to appreciate our contributions in the two next chapters.

In Chapter 4, we propose a method to perform an efficient and medically relevant classification of images of single cells. The method, that we call regression constraint, enables to efficiently and simply introduce the notion of distances between classes in the training of a model. In the second part of this chapter, we study the direct application of this method for the classification of regions that may contain up to a dozen of cells (on a simulated dataset), and we extensively use an attribution method to perform weakly supervised localization. We finally extend this work on real slides, and integrate these methods in a pipeline that could be used to reduce the workflow of cytopathologists.

In Chapter 5, we are mainly interested in the interpretability, i.e. explaining how a trained model makes a specific decision and what has been learned at training time. We start by questioning the concept of explicative heat-maps as currently defined in the literature, and defining the current framework of most popular WSI classification architectures. In this common design, we propose a piece-wise interpretability method that enables to identify features that have been learned as contributing to describe the “tumor” class in a public



dataset called Camelyon-16 (breast cancer biopsy slides). We further use these features to compute new explicative heat-maps, and we demonstrate, through two measures, that they improve the interpretability. We validate this approach on another dataset that contains cervical cancer biopsy slides which enable to suspect the limitations for a direct usage on LBC slides. In the end, we verify this hypothesis by working on a 393 “abnormal” cervical LBC slides and propose to relax the learning context through weak “abnormality” detection, which enables us to reach acceptable performances, and we use our interpretability work to highlight the relevance of what has been learned.

Finally, in addition to being a conclusion to the manuscript and our work, Chapter 6 opens discussions and offers perspectives about future works that could be done to extend the path we started to trace toward CAD tools for cytopathologists routine.

Our contributions consist of both new methods in computer vision and DL, and applications to the field of digital pathology.



# Medical context

## Sommaire

<b>2.1</b>	<b>Cervical cancer screening . . . . .</b>	<b>5</b>
2.1.1	Cervical cancer: what is it? . . . . .	5
2.1.2	Cervical cancer screening . . . . .	7
2.1.3	Liquid-based cytology slides content . . . . .	8
2.1.4	Management of abnormal cases, evolutions and further medical exams . .	14
2.1.5	Conclusion: advantages and limitations of liquid-based cytology Pap tests	16
<b>2.2</b>	<b>Whole Slide Imaging process: towards digital pathology next generation tools . . . . .</b>	<b>17</b>
2.2.1	Camelyon-16: Breast Cancer detection on Biopsies . . . . .	18
2.2.2	The Cancer Genome Atlas (TCGA) . . . . .	19
2.2.3	Herlev dataset . . . . .	19
2.2.4	Conclusion . . . . .	20

## 2.1 Cervical cancer screening

According to the World Health Organization (WHO) [WHO 2014], cervical cancer is the second most important cancer for women after breast cancer. In this section, we present stakes related to cervical cancer, how is it generally detected in the first place and why Artificial Intelligence (AI) gives great promises to improve the performance of cervical cancer screening.

### 2.1.1 Cervical cancer: what is it?

According to the National Cancer Institute, a cancer is defined as an abnormal and uncontrolled cells division that can leads to a tumor and invades nearby tissues.

The cervix (see Figure 2.1) is an organ that ensures, during pregnancy, that the embryo stays in the uterus and is protected from bacterias.

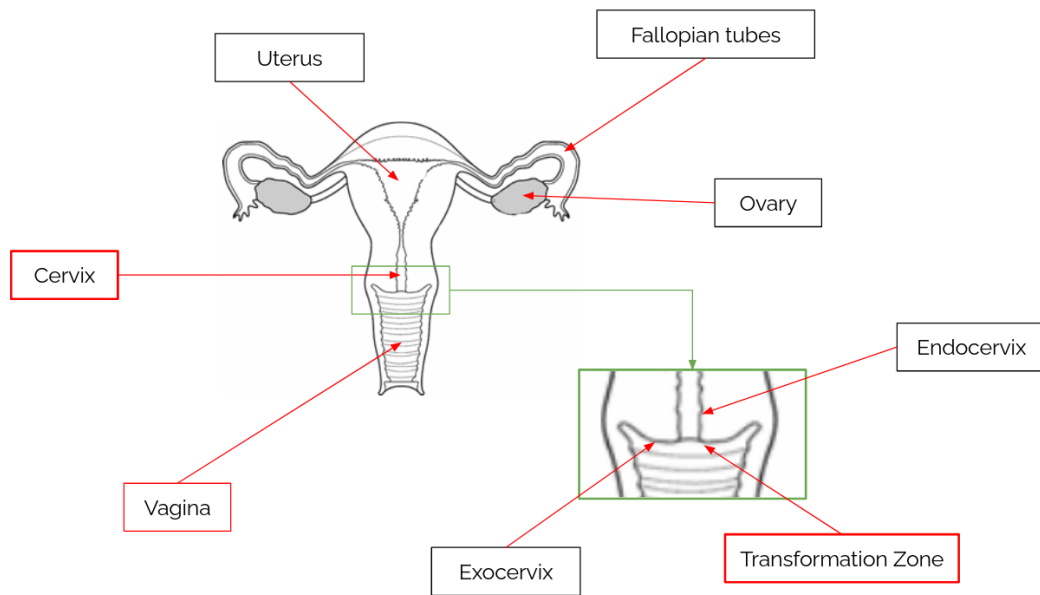


Figure 2.1: The cervix organ.

Cervical cancer is a type of cancer that develops inside the epithelium (outer layer of the skin) of the cervix. Most of cervical cancers cases (about 90%) are developing in the exocervix and are called Squamous Cells Carcinoma (SCC), the other 10% are Adenocarcinoma that take place in glandular cells that form the endocervix. A very small percentage of cases combine both (adenosquamous carcinoma). Generally, a cervical cancer starts from the transformation zone (or junction zone) where endocervix and exocervix join, it is a zone where a lot of cellular changes take place.

This cancer is mainly due to an infection by a virus called Human Papilloma Virus (HPV) which counts over 200 genotypes that can lead to precancerous lesions called Cervical Intraepithelial Neoplasia (CIN).

After being infected by HPV, it generally takes between 7 and 10 years for the infection to turn into an actual cancer.

The main symptom of cervical cancer is unusual bleedings e.g. outside of periods or after menopause. Treatments generally consist of surgery (e.g. ablation) or radiation therapy.

Cervical cancer is the second deadliest cancer after breast cancer with over 500 000 new cases detected each year and over 250 000 deaths. WHO also states that around 90% of cervical cancers could be avoided if they were detected and treated earlier.

Today, over 50 millions Pap tests (see Section 2.1.2) are made each year worldwide, among which above 95% are classified as negative (e.g. 3.8% cases are abnormal in [Maraqa, Lataifeh, and Otay 2017]).

### 2.1.2 Cervical cancer screening

Cervical cancer is generally first detected using either a HPV test or a cytology analysis. Both exams require a gynecologist to make a smear, i.e. sampling cells from the cervix using a swab. HPV tests consist in detecting high-risk HPV phenotype DNA by molecular biology. The second exam relies on the visual screening from a cytopathologist to detect abnormal precancerous changes on microscopy slides. This second method is the base and the core of our work.

Note that cytology defines the study of cells and relies on sampling methods such as smears, which differs from histology that defines the study of tissues which implies performing a biopsy. From the point of view of image analysis the main difference is that tissues are more structured, textured and cells organization can be observed while in LBC we have to deal with an ensemble of cells freely floating in liquid.

First introduced by Aurel Babes [Babes 1928] in 1927, visual inspection of cells through a microscope for cervical cancer detection has been improved and popularized by Georgios Papanicolaou [Papanicolaou and Traut 1943] in the early 1940's through sampling, staining and interpretation methods [Diamantis, Magiorkinis, and Androutsos 2010]. Indeed, Dr. Papanicolaou proposed to detect precancerous changes that can be observed in the morphology and staining of cells in the early stages of the infection, thus giving his name to the method: Pap test or Pap smear.

A Pap test consists first in collecting samples from the cervix by scratching the epithelium at the transformation zone (see Figure 2.1). The collected cells are stained using Pap staining (see Figure 2.2), which uses three stains [Marshall 1983]:

1. Hematoxylin that stains cell nuclei in blueish colors;
2. Orange-G that stains keratin which is a protein that is secreted to protect epithelium from external aggressions;
3. Eosin that stains cell cytoplasm in pinkish colors.

Secondly, a Pap test consists in analyzing visually the cells that have been sampled and colored. At first, conventional Pap smears, where cells were directly put down and spread on the microscopy slides, were highly used. But more recently Liquid-based Pap smears, where an additional step consisting in placing cells in a preservative liquid to remove parasite objects such as mucus, have proved to offer a better interpretation for cervical cancer screening [Karimi-Zarchi et al. 2013; Qureshi et al. 2017; Singh, Anjum, and Qureshi 2018].

Then, cytology experts observe these slides, under a microscope, looking for atypical cells. Next, we present cells that can be observed on Pap smear Liquid-Based Cytology (LBC) slides and their interpretation w.r.t. the medical decision.

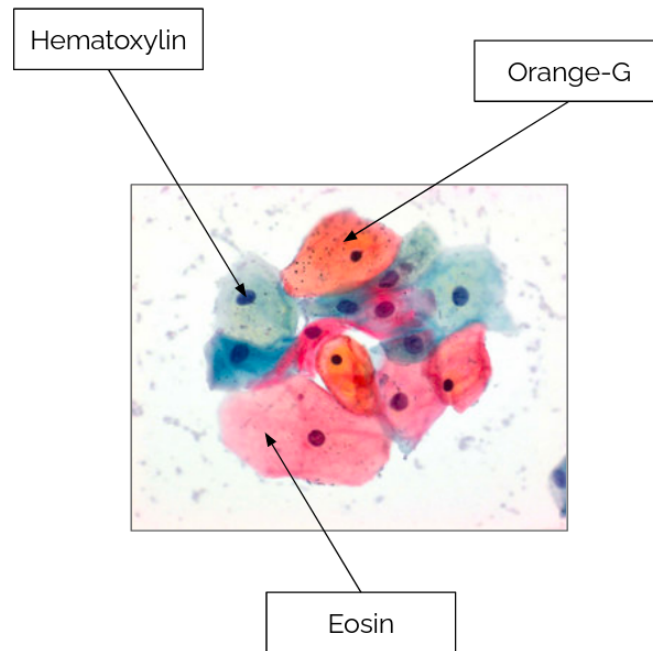


Figure 2.2: Pap stain illustration; image from <https://www.sigmaaldrich.com/catalog/product/mm/115925>.

### 2.1.3 Liquid-based cytology slides content

#### 2.1.3.1 Content of “normal” slides

As explained before, cells are sampled at the junction between endocervix (inner cervix) and ectocervix (outer cervix), and respectively contain glandular cells and squamous epithelial cells.

Generally, most cells observed on a Pap test are Squamous Epithelial Cells (SEC). These cells are present in the external layer of the epidermis that covers hollow organs such as the cervix. There are three kinds of SEC: Parabasal, Intermediate and Superficial cells (see Figure 2.3).

Parabasal cells are immature squamous cells, they are the smallest epithelial cells and can be found as single cells or in groups (see Figure 2.4). Their characteristics (cytomorphology) are the following: they are round or oval shaped with a high nucleus over cytoplasm ratio (NCR) and a dense cyanophilic (dark blue) cytoplasm.

Intermediate cells are semi-mature squamous cells that are smaller and less angular than superficial cells. They can be recognized thanks to their cyanophilic cytoplasm (stained in blue, due to Hematoxylin), their border tends to fold and they have vesicular nuclei (deeply stained membrane and pale center) with reticular chromatin (chromatin that forms some kind of network). Navicular cells are a kind of benign intermediate cells that are filled with

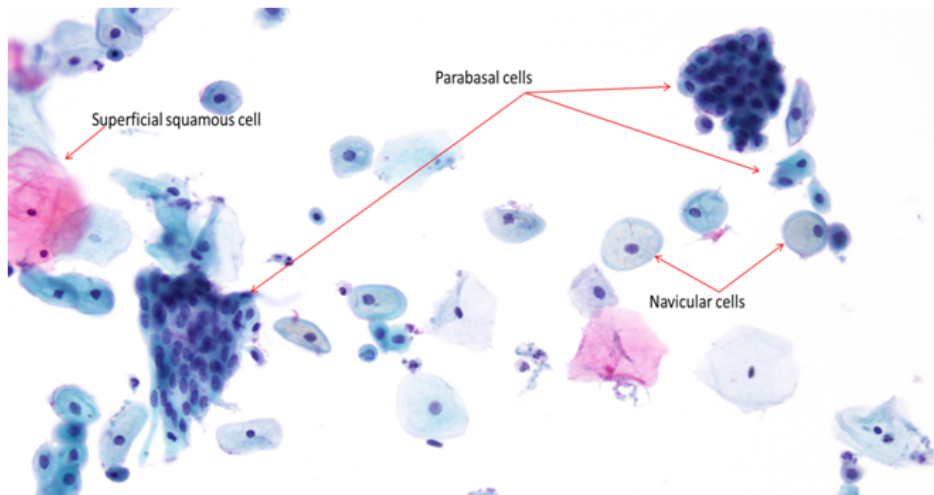


Figure 2.3: Illustration of Squamous Epithelial Cells (SEC).

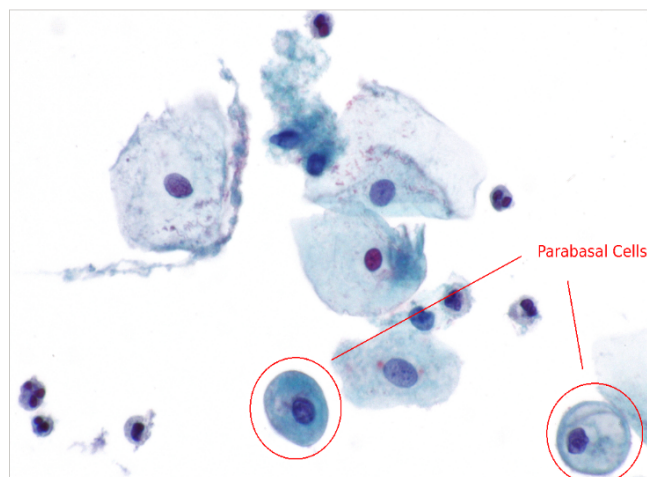


Figure 2.4: Parabasal cells.

glycogen, giving it a central halo with a yellow hue. Figure 2.5 illustrates intermediate and navicular cells characteristics.

Superficial cells are mature squamous cells. They are the biggest cells found on Pap smears, are polygonal (angular borders) and have small and dark nuclei with pyknotic appearance (condensation of chromatin, hint of cells death). They also have keratohylin granules. Their cytoplasm can also be kind of transparent. Figure 2.6 shows superficial cells and keratohylin granules.

The other types of cells are glandular columnar cells that are elongated stick-shaped small cells that pave the inner cervix (see Figure 2.7).

Also bacterias and artifacts organisms can be found on Pap smears. For example, bacterias might be abundant over a Pap test slide. They look like small really dark nuclei and are totally

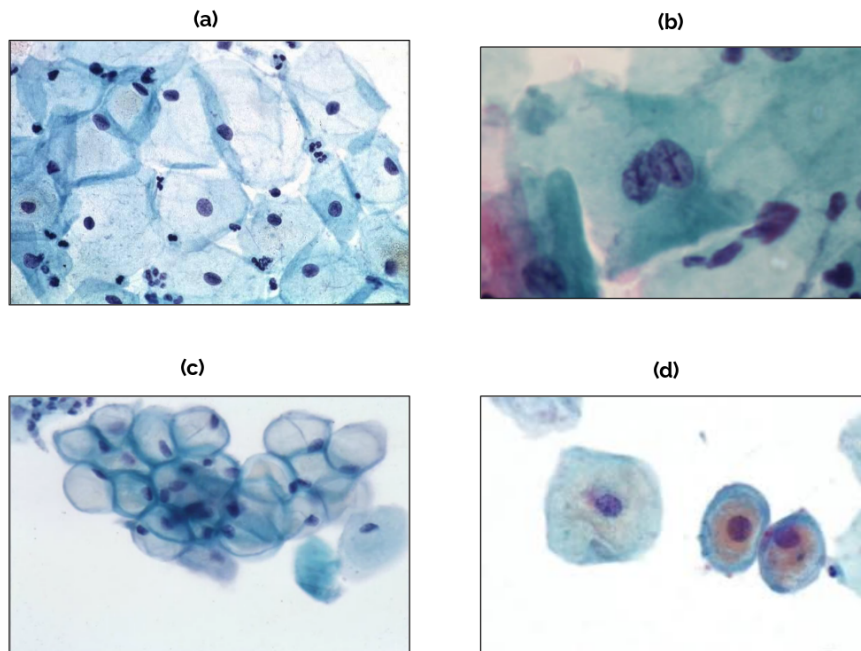


Figure 2.5: Intermediate cells and their characteristics. (a) A group of navicular cells; (b) An example of reticular chromatin nucleus; (c) A group of navicular cells; (d) An example of glycogen cytoplasm.

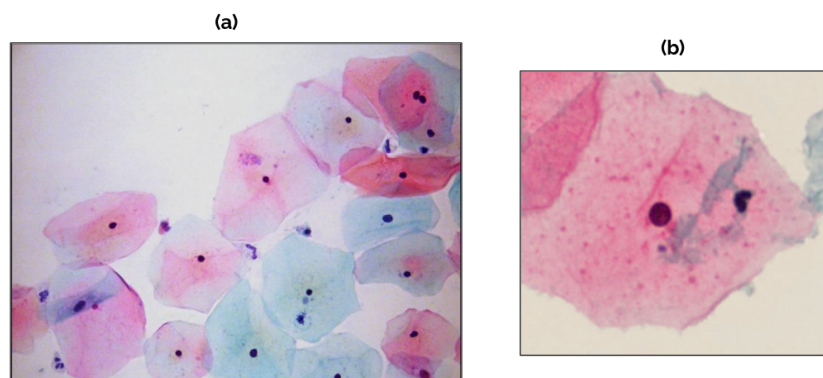


Figure 2.6: Superficial cells (a) and keratohylin granules (b).

benign. There can also be blood cells such as neutrophils (white blood cells) and erythrocytes (red blood cells). Examples of these artifacts are shown in Figure 2.8

Finding only these kind of cells and organisms will lead to a “Negative for Intraepithelial Lesion or Malignancy” (NILM) classification, which means that no pre-cancerous changes have been detected.



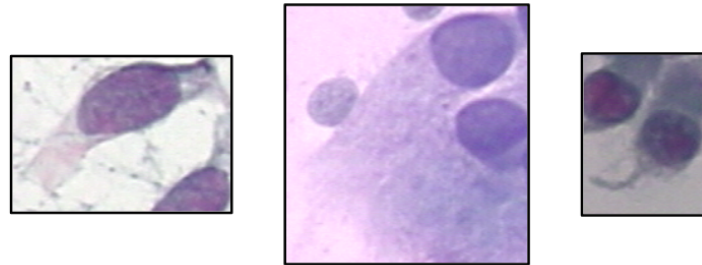


Figure 2.7: Illustration of columnar cells (from Herlev dataset [Jantzen et al. 2005], see Section 2.2.3)

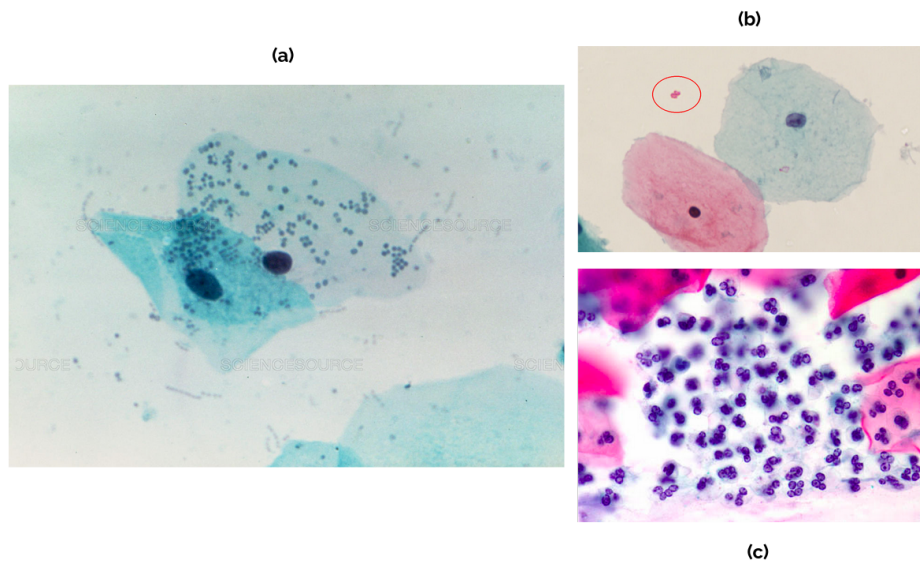


Figure 2.8: Pap smear artifacts. (a) Bacterias on superficial cells (from <https://www.sciencesource.com/>); (b) Erythrocyte; (c) Neutrophilis (from <http://pathology.jhu.edu/cytopath/masterclass/general/1gen16b.htm>).

### 2.1.3.2 “Abnormal” Slides

Now that we have seen what is expected on a “normal” or negative Pap test, we can try to understand what are the atypia that cytopathologists are looking for when they perform reviews of slides.


There are three classification systems to classify an abnormal Pap smear slide: WHO [Ri-otton et al. 1973], Richart [Wilbanks et al. 1968] and Bethesda [Solomon et al. 2002]. The correspondence table that links these systems can be found in Figure 2.9. WHO and Richart were created for histology exams while Bethesda was created, in 1988, for cytology exams

and adapted until 2014 in order to offer guidelines and to standardize cytology slides interpretation and results, i.e. improve intra-observer reproducibility [Stoler and Schiffman 2001; Sherman et al. 2007].

**(a)**

Cervical Cancer Screening Classifications		
Histology exam		Cytology exam
World Healthcare Organization	Richart	Bethesda Classification
Negative	Negative	NILM
Squamous atypia	Squamous atypia	ASC-US
Mild dysplasia	CIN 1	LSIL
Moderate dysplasia	CIN 2	HSIL
Severe dysplasia	CIN 3	
Carcinoma in situ		
Invasive cervical cancer	Invasive cervical cancer	Invasive cervical cancer



**(b)**

Figure 2.9: (a) Pap smear classification systems (from [Riotton et al. 1973, [Wilbanks et al. 1968] and [Solomon et al. 2002]]); (b) Illustration of Bethesda system abnormality grades (from <https://www.incytediagnostics.com/about/news-and-publications/asc-us-vs-asc-h-what-is-the-difference/>).

Thus, the guidelines [Solomon et al. 2002] and [Nayar and Wilbur 2015] arose from a medical consensus in 2001 and then in 2014 to standardize and define cytomorphological features that are discriminative for Pap smear cells classification regarding malignancy, terminology used to report Pap test results, and management of abnormal Pap tests. Mainly, the four characteristics used to determine whether a cell is abnormal or not are:

1. The cytoplasm color: Pap smears are stained using Hematoxylin (pink) and Eosin (blue) stainings and mature cells will have their cytoplasm mostly stained by Hematoxylin, so changes are expected on these cells;
2. The nucleus texture: condensation of the chromatin in the nucleus and vanishing borders are major features for abnormal cells;

3. The nucleus shape: a concave and round shape is expected for a normal cell;
4. The Nucleo-Cytoplasmic Ratio (NCR), ratio between the nucleus size and the cytoplasm size: the higher the more abnormal (except for one type of normal cells);

Note that from here on we will only deal with squamous cells atypia that represent the vast majority of cervical cancer cases.

**ASC-US** ASCUS are Atypical Squamous Cells of Undetermined Significance (see Figure 2.10), meaning that this grade is used to classify cells (generally superficial or intermediate cells) that appear mildly abnormal but the cause of changes is unclear. They appear with a nuclear enlargement that makes the nucleus twice to three times bigger than normal ones. Regarding the texture, there are several types of ASCUS: they may be hyperchromatic with fine chromatin and smooth nuclear membrane. Multinucleation or a mildly irregular nuclei membrane and/or an increase in the chromatin granularity with enlarged nuclei is also a sign of ASCUS. Regarding the cytoplasm, ASCUS are sometimes recognized by orangeophilia (which is keratinized cytoplasm which makes the cytoplasm appear orange).

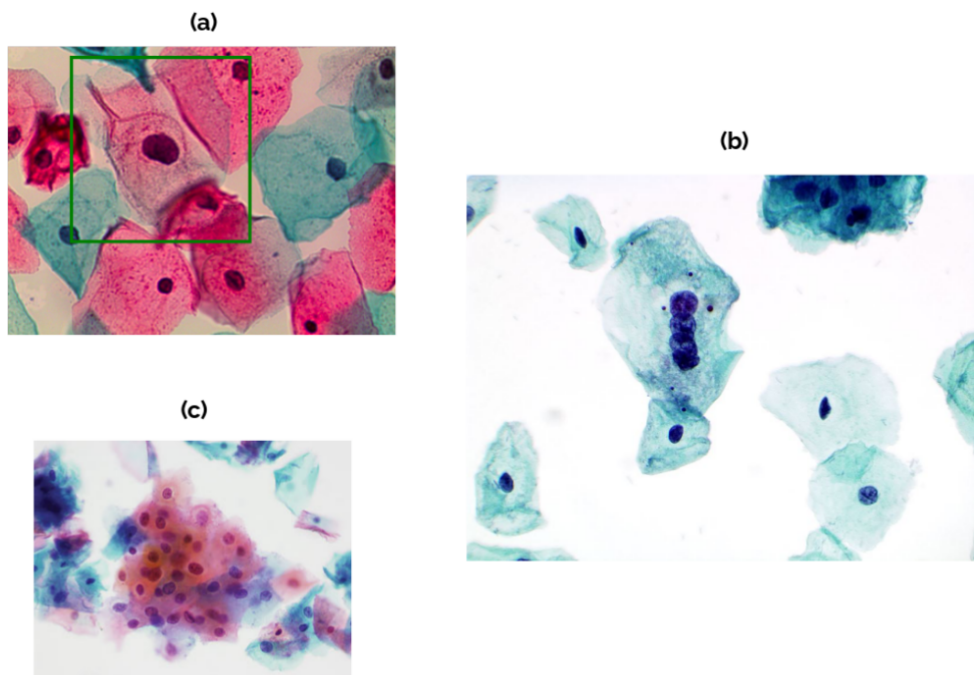


Figure 2.10: Atypical Squamous Cells of Undetermined Significance grade. (a) ASC-US cell example; (b) Multinucleation; (c) Orangeophilia cytoplasm.

**LSIL** LSIL are Low-grade Squamous Intraepithelial Lesions (see Figure 2.11). What differentiates them from ASCUS is mainly the enlargement of nuclei that is even more important (about 3 or 4 times bigger than normal), the chromatin that appears more granular and

the presence of koilocyte (squamous epithelial cells that contain an acentric, hyperchromatic nucleus displaced by a perinuclear vacuole).

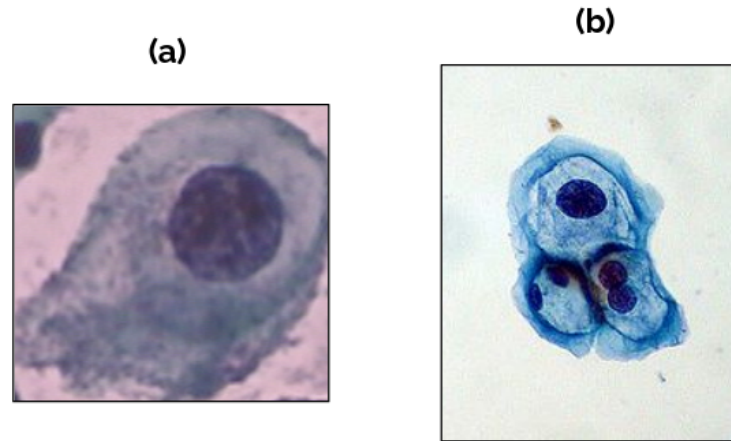


Figure 2.11: Low-grade Squamous Intraepithelial Lesions grade. (a) LSIL cell example; (b) Koilocytes.

**ASC-H** ASC-H (see Figure 2.12) are Atypical Squamous Cells that cannot enable to exclude High-grade intraepithelial lesion. In [Chivukula and Shidham 2006], the authors explain that this label is given to cells that “exhibit some equivocal features suggestive of but not sufficient to call “HSIL”, and that cytomorphological criteria associated with ASC-H class are wide and can easily be confused with LSIL and High-grade Squamous Intraepithelial Lesion (HSIL)”. In [Hata et al. 2019], the authors deeply study this class and conclude that “the presence of small dysplastic cells displaying marked hyperchromasia, thickening of nuclear contour, and prominent nucleoli” are most discriminative cytomorphological features for ASC-H.

**HSIL and above** HSIL are High-grade Squamous Intraepithelial Lesion (see Figure 2.13). They can be identified with their highly enlarged nuclei and their reduced cytoplasm that implies a NCR above 50%. The nucleus reveals important irregularities and granularities (such as crowding) and appears most often to be hyperchromatic, but can also appear to be hypochromatic.

#### 2.1.4 Management of abnormal cases, evolutions and further medical exams

These grades are associated with a risk of evolving to an actual cancer and chances to observe a natural regression of the lesion. Figure 2.14 shows these probabilities for each grade. We can, for example, observe that more than 70% of ASCUS cases will lead to a regression of

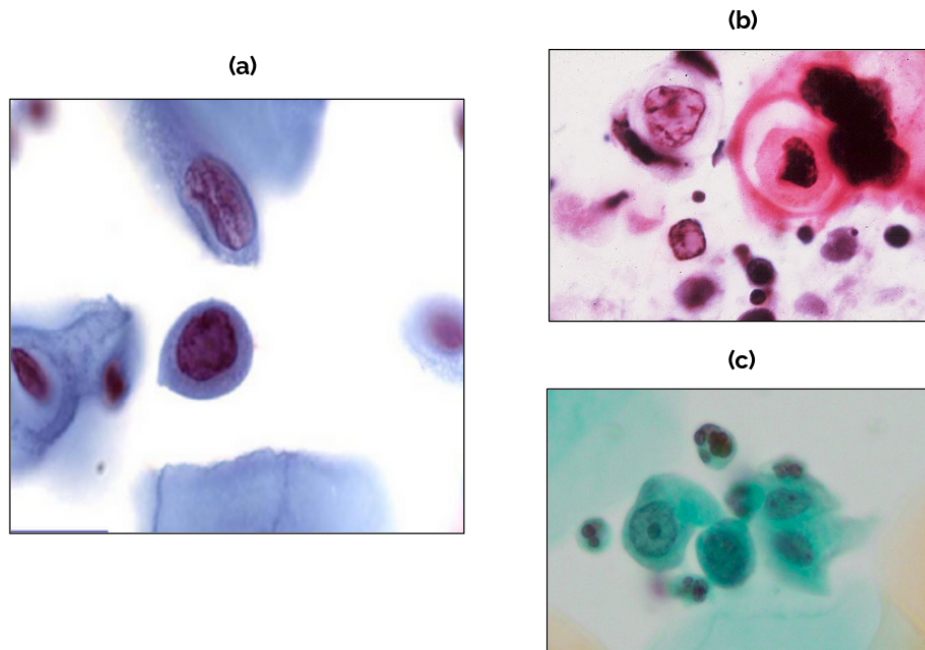


Figure 2.12: Atypical Squamous Cells that cannot exclude high-grade intraepithelial lesion grade. (a) ASC-H cell example; (b) Marked hyperchromasia (from <http://pathology.jhu.edu/cytopath/masterclass/general/maligcri/1genp26.htm>); (c) Prominent nucleoli (from [https://screening.iarc.fr/atlascyto\\_detail.php?flag=0&lang=2&Id=cyto7756&cat=F1a2](https://screening.iarc.fr/atlascyto_detail.php?flag=0&lang=2&Id=cyto7756&cat=F1a2)).

lesions, around 50% for LSIL while only around 30% of HSIL cases have chances to regress. More generally, the more severe the diagnosis is, the more chances there are to observe an evolution towards an invasive cancer and the less chances there are to observe a regression.

In that regard, [Wright et al. 2002] present the medical consensus that came out of a conference gathering 121 experts in cervical cancer screening. The recommendations for women with abnormal Pap tests are summarized in Figure 2.15.

We can observe that cytology test is an efficient primary test and its results indicate the following exam to do and its timing.

Colposcopy consists in inspecting cervix, using a binocular magnifier, looking for physical lesions. Histological biopsies consist in removing a tissue sample from the cervix for further microscopy analysis, that enables to grade the abnormality more precisely and to detect potential infiltrations.



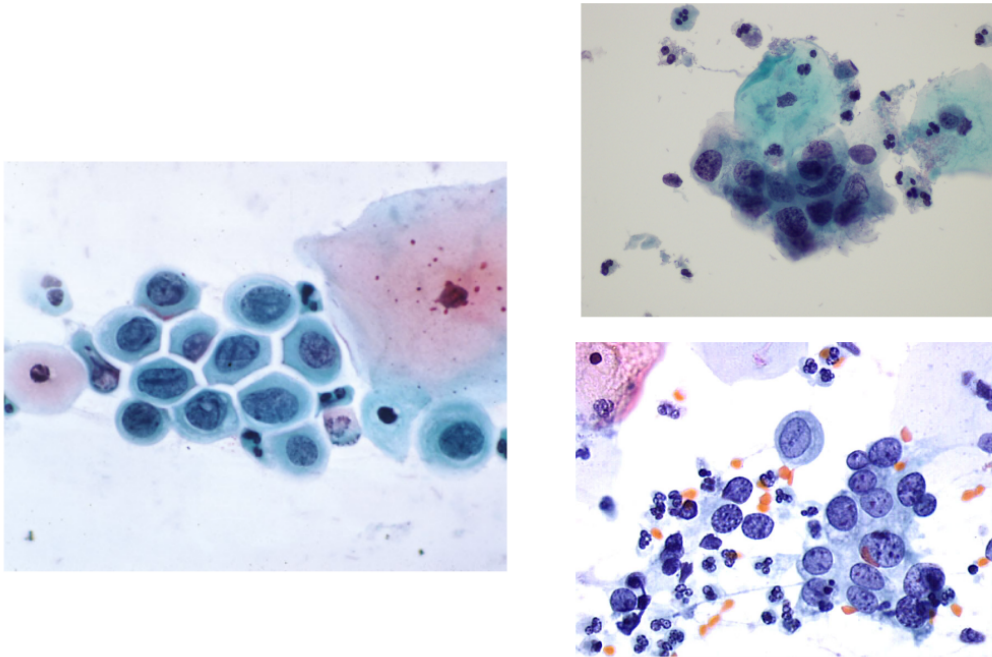


Figure 2.13: High-grade Squamous Intraepithelial Lesion grade examples (from <https://www.eurocytology.eu/en/course/1297> and [https://screening.iarc.fr/atlascyto\\_detail.php?flag=0&lang=1&Id=cyto7719&cat=F1c4](https://screening.iarc.fr/atlascyto_detail.php?flag=0&lang=1&Id=cyto7719&cat=F1c4)).

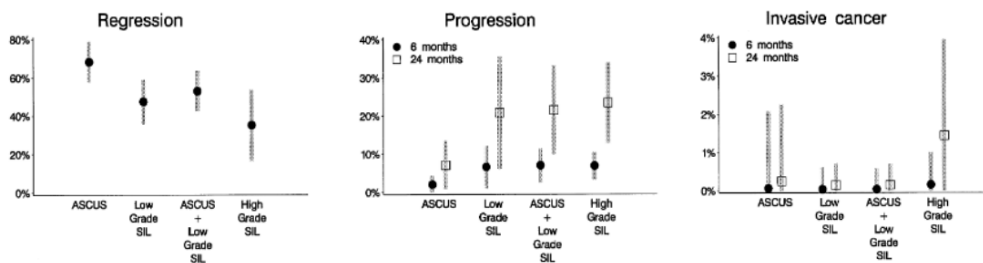


Figure 2.14: Regression (left), progression (middle) and invasive cancer (right) rates w.r.t. LBC Pap smear diagnosis.

### 2.1.5 Conclusion: advantages and limitations of liquid-based cytology Pap tests

[Schwartz 2002] offers a great review and study about the efficiency and limitations of LBC Pap tests in Switzerland. First, it highlights how important a regular and frequent Pap test is in order to avoid invasive cancer [Janerich, Hadjimichael, and Schwarz 1995]. However, this method has some limitations. Indeed, if over-grading is pretty rare (generally a great specificity is measured), the sensitivity of Pap tests is generally estimated between 50% and 60% due to an important number of false negative cases. Studies show that around 20% of women with pre-cancerous lesions or cancer have had a negative cytology exam in the

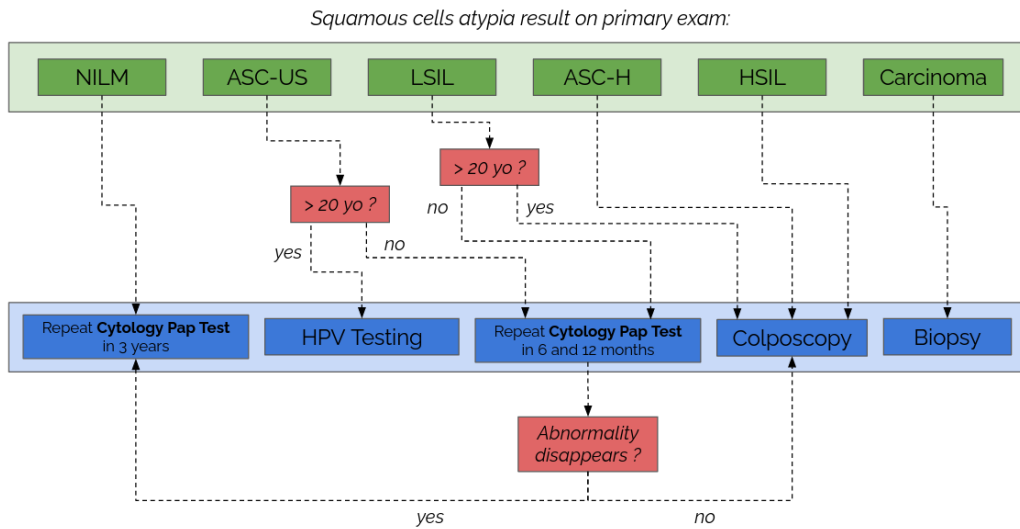


Figure 2.15: Management of abnormal cases.

past two years [Nanda, McCrory, and Myers 2000, Morell et al. 1982, Gay, Donaldson, and Goellner 1985, Kristensen et al. 1991, Joste, Crum, and Cibas 1995].

Even if promises offered by Pap tests are of interest (proved to be successful in developed countries, good specificity, cost effective ...), there are limitations that are inherent to Pap tests such as sampling limitations i.e. if cells have not been sampled in the transformation zone. Also cells of interest may be removed during the slide preparation, and cells of interest can be hidden under other cells. Some of these limitations have been tackled by the introduction of LBC exams. However, the heavy process and sparsity of cells of interest induce a lot of work and fatigue, while most of the time there is “nothing” to find.

To tackle the low sensitivity, two processes are applied: quality control process in pathology laboratories, that consists in re-screening a certain percentage of slides classified as NILM and repeat regularly Pap tests (at least every 3 years).

Moreover, in spite of efforts that have been made recently with the Bethesda consortium, Pap test screening is completely rater dependent.

In that sense, a Computer-Aided Diagnosis (CAD) tool with a sensitivity of 100% (or close) would enable to reduce the workflow of cytopathologists almost regardless of the associated specificity.

## 2.2 Whole Slide Imaging process: towards digital pathology next generation tools

Whole Slide Imaging [Farahani, Parwani, and Pantanowitz 2015, Nishat et al. 2017, Kumar, Gupta, and Gupta 2020] is a process that enables to create Whole Slide Images (WSI) i.e.

digitized microscopy slides that can then be visualized and used for medical applications through a “virtual microscope”. The digitization is performed by a scanner (e.g. NanoZoomer from Hamamatsu or Ultra-Fast Scanner from Philips) that outputs digital files in adapted format (e.g. NDPI or PhilipsTIFF). Most scanning processes are tile-based, i.e. given a zoom level (or magnification level) the scanner iteratively scans patches of fixed size. It may also include “pyramid” information, i.e. digitized slide at lower magnification which enable to simulate a microscope more efficiently and precisely (see Figure 2.16).

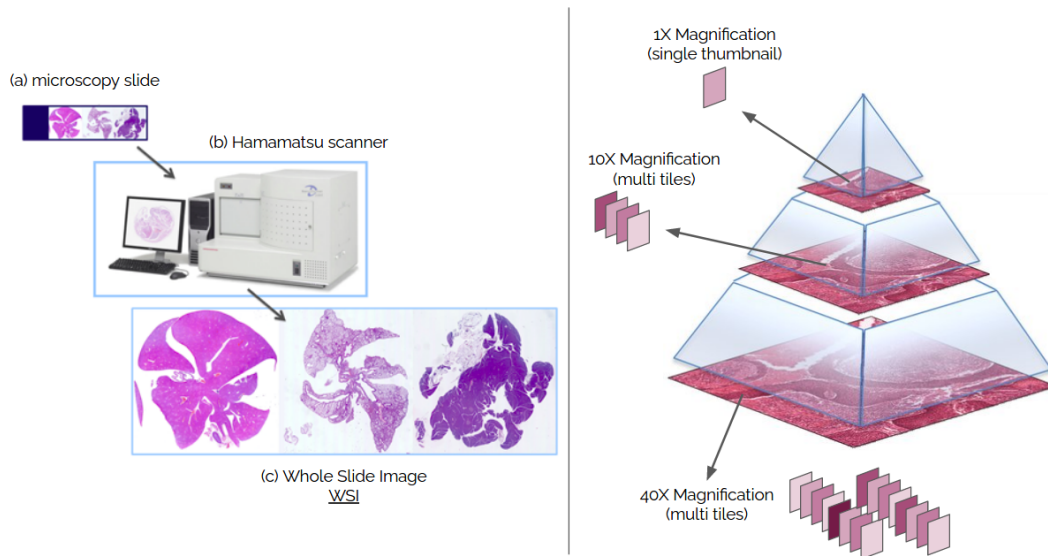


Figure 2.16: Whole Slide Imaging; From a microscopy slide to a WSI through a scanner (left; from <https://www.rhem.cnrs.fr/index.php/nos-services-en-ligne/numerisation-lames>); “Pyramid” image organization.

Moreover public libraries (such as OpenSlide or ASAP) have been developed to enable computer science researchers to work with these images and, in the mean time, several public WSI dataset have been published and strengthen the link between computer vision and medical applications. Most popular public WSI datasets contain histology slides, and the size of dataset we have for LBC use case is small. Thus, in this thesis, we will rely on these larger histology datasets to develop and validate some of our methods before applying to our dataset of interest.

### 2.2.1 Camelyon-16: Breast Cancer detection on Biopsies

Camelyon-16 [Ehteshami Bejnordi et al. 2017] is the most popular dataset containing WSIs. It was introduced during the IEEE International Symposium on Biomedical Imaging conference in 2016 to develop, evaluate and compare classification algorithms. The task (see Figure 2.17) consists in classifying histological slides (from biopsies) between two classes: “normal” slides that contain only “normal” tissue, and “tumor” slides that contains both “normal” tissue and “tumor” tissue called “metastases”. The dataset contains 345 WSIs divided into 209 “normal” cases and 136 “tumor” cases digitized at 40X magnification.



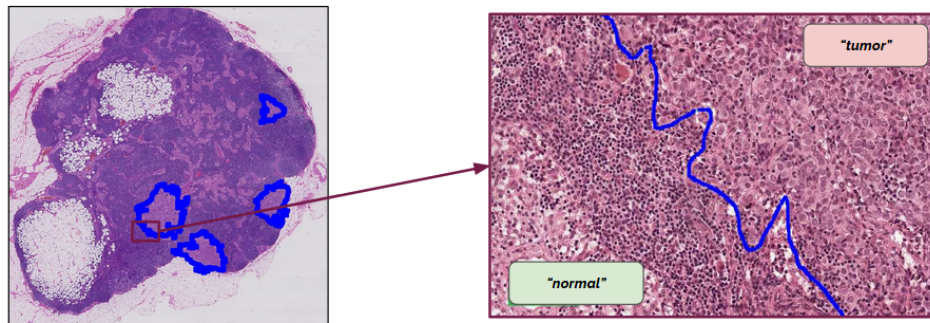


Figure 2.17: Camelyon-16 dataset. Thumbnail (right) and about 5X zoom on tumorous region (left)

### 2.2.2 The Cancer Genome Atlas (TCGA)

Another popular dataset of interest is TCGA [Tomczak, Czerwińska, and Wiznerowicz 2015] (stands for “The Cancer Genome Atlas”). It contains data (WSI, genomic ...) of about 33 cancer types (such as lung cancer, cervical cancer ...) through more than 11,000 cases. It aims at accelerating cancer research and discoveries. For example, it contains more than 1580 slides for lung cancer divided into three classes (“normal”, “Lung squamous cells carcinoma” and “Lung adenocarcinoma”) which gave birth to an important number of researches (e.g. [Chen, Chen, and Yu 2021] and other methods this work compares to).

### 2.2.3 Herlev dataset

There is no such WSI dataset available for cervical cancer in cytology context.

In this context, Herlev dataset [Jantzen et al. 2005] proposes to tackle single cells classification. This dataset is composed of 917 images showing single cells (between 50 and 400 pixels large), 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). It additionally gives the segmentation masks for nucleus, cytoplasm and background.

Note that two other datasets exist for cervical cancer in cytology context ([ISBI 2015] and [Ahmady Phoulady and Mouton 2018]), but they consist in segmenting cytoplasm and detecting nuclei which is a challenging task but do not interest us here.

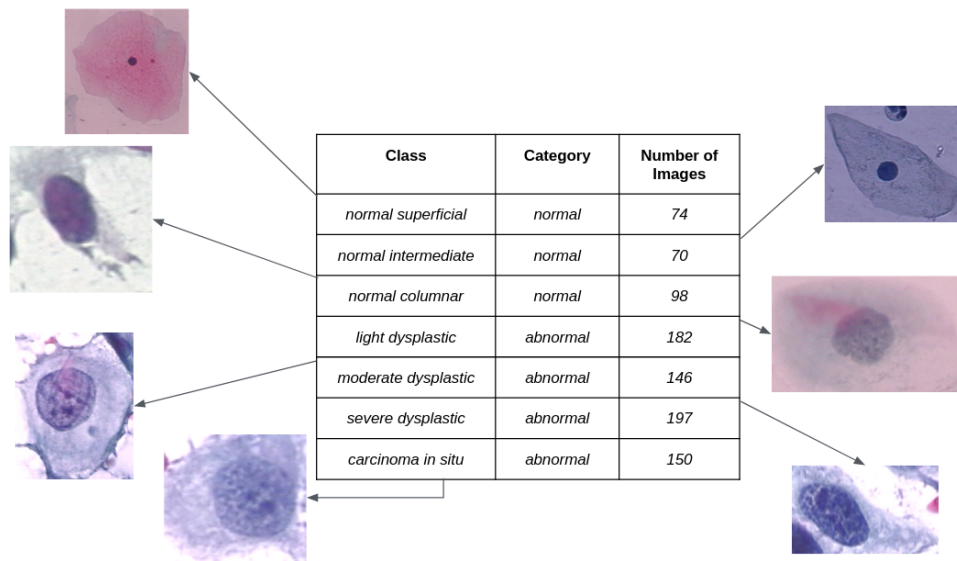


Figure 2.18: Herlev dataset illustration.

#### 2.2.4 Conclusion

These WSIs are considered to be the front door to the next CAD tools [Bera, Schalper, and Rimm 2019]. But, as it has been introduced here, the context of LBC is a complicated one with slides that can contain hundred of thousands of cells with really precise and complex characteristics to learn and detect in order to be efficient. Thus in the next chapter we present methods that are the most promising ones: Deep Learning (DL) methods.

# Deep learning: An opportunity for efficient computer-aided diagnosis tools

---

## Sommaire

---

<b>3.1</b>	<b>Classification</b>	<b>22</b>
3.1.1	Metrics	22
3.1.2	Feature extractor architectures	24
3.1.3	Cervical cancer usage	26
<b>3.2</b>	<b>Localization and object detection</b>	<b>27</b>
3.2.1	Architectures	28
3.2.2	Cervical cancer usage	31
<b>3.3</b>	<b>Whole slide image classification</b>	<b>33</b>
3.3.1	Methods and architectures	33
3.3.2	Other methods for WSI classification	39
3.3.3	Cervical cancer and cytology applications	40
<b>3.4</b>	<b>Explanations and interpretability</b>	<b>41</b>
3.4.1	Methods for interpretability	41
3.4.2	Evaluation and quantification of interpretability	43
3.4.3	Medical usage	43
<b>3.5</b>	<b>Conclusion and discussion</b>	<b>44</b>

---

Over the past decades, two technologies have emerged: Deep Learning (DL), a subset of Artificial Intelligence (AI) methods that enable deep architectures to learn complex features and perform on a wide range of tasks; and Whole Slide Image (WSI), microscopy slides digitized at high resolution enabling digital and virtual microscopes to be developed. Mixed together, they revolutionized the field of digital pathology that mainly consists in improving the pathologist workflow using digital information [Zarella et al. 2019].

In this chapter, we introduce different DL<sup>1</sup> feature extractors and their specificities along with their applications for image classification, objects detection and WSI classification. Each

---

<sup>1</sup>Deep Learning

section ends with a presentation of applications to medical problems close to cervical cancer screening. In the end, we go through approaches that reveal what is learned by these models, which are called interpretability methods.

The background on DL is recalled in the appendices. In Appendix A, we go through the methods that reached the Large-Scale Visual Recognition Challenge (LSVRC) state-of-the-art performances, starting with AlexNet [Krizhevsky, Sutskever, and Hinton 2012] which revolutionized the field of computer vision in 2012. In Appendix B, AlexNet architecture is detailed, and common strategies for training are summarized.

## 3.1 Classification

Image classification is a task that consists in associating a class (or a score) to an image.

### 3.1.1 Metrics

Most of the the time, the metric with which the performances of the trained model will be measured (on the test set) defines the final activation and the loss used for training. For example, softmax activation and categorical cross-entropy loss are particularly efficient for accuracy and Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) metrics (see Figure 3.1). These metrics are the most popular ones, even if some other metrics can be encountered [Chicco and Jurman 2020].

Accuracy is defined, from the confusion matrix, as the proportion of images well classified i.e. the number of image well classified divided by the total number of images. In the context of binary classification and medical applications, sensitivity and specificity are also often reported. Sensitivity is the proportion of images well classified when they belong to the positive class, and specificity is the proportion of images well classified when they belong to the negative class (see equations in Figure 3.1). ROC curves are computed as the sensitivity (or true positive rate) in function of the recall (or 1 - specificity) using different decision thresholds and AUC is the the area under this curve.

When it comes to continuous scoring, other loss functions, activation functions and metrics might be used. A standard implementation for continuous scoring is to use a single neuron with a linear activation function and a Mean Square Error (MSE) loss function. For a predicted score  $y$  and a ground truth score  $y'$ , the MSE loss  $L_{MSE}$  is computed as:

$$L_{MSE} = (y - y')^2.$$

This approach can be particularly interesting in the medical field where gradual labels (also referred to as “ordinal regression”) are often used, for example when predicting the severity or malignancy of a disease. In this case, a popular metric is called Cohen’s KAPPA measure.

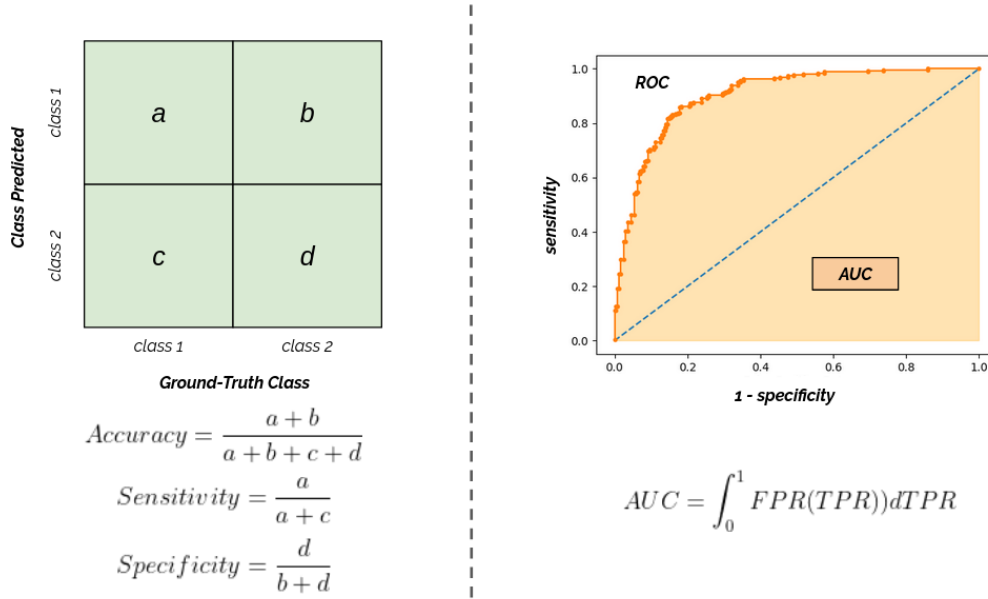


Figure 3.1: Confusion matrix and ROC Curve.

Quadratic Cohen’s KAPPA [Brennan and Prediger 1981] is a measure used in the context of ordinal regression problems (ordered classes). 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  $M$ , with coefficients  $m_{i,j}$  in column  $j$  and line  $i$ , such that  $\sum_{i=1}^N (\sum_{j=1}^N (m_{i,j})) = 1$  for a  $N$  classes classification problem. The expected agreement proportion  $P_e$  is  $P_e = \sum_{i=1}^N (\sum_{j=1}^N (m_{i,j}) \cdot \sum_{k=1}^N (m_{k,i}))$  and the observed agreement proportion is  $P_o = \sum_{i=1}^N m_{i,i}$ . KAPPA value  $K$  is then calculated as follows:

$$K = \frac{P_o - P_e}{1 - P_e}.$$

The value ranges between -1 (worst predictor) and 1 (perfect predictor) with 0 corresponding to a random predictor.

Using this measure or MSE and AUC, architectures or training strategies can be designed to perform on ordinal regression task. [Cheng, Wang, and Pollastri 2008] and [Diaz and Marathe 2019] are the most popular methods and will be further detailed, respectively in Section 4.2.2.2 and Section 4.2.2.3, where they will be used as baselines.

A popular way to compare different methods for a task is to perform what is called a  $K$  cross-validation study. It consists in creating  $K$  different random splits (training/validation/testing sets), and performing training and testing to ensure that the method is not dependent on the split. It also enables to perform statistical tests between performance distributions to show that a method is statistically better than another one, e.g. a Mann Whitney U Test [Nachar 2008].

### 3.1.2 Feature extractor architectures

In this section we go through most popular CNN-based feature extractors that all come from edition of LSVRC [Singh 2016] following 2012 edition.

AlexNet [Krizhevsky, Sutskever, and Hinton 2012] (see Appendix A) revealed the power of deep learning CNN feature extractors. In 2013 edition, ZFNet [Zeiler and Fergus 2014], a slightly modified AlexNet architecture, improved top-5 classification error rate from about 16% to about 12%. Architecturally, the changes consist in replacing first 11x11 (stride 4) convolutional layers by a 7x7 convolutional layer with stride 2. The authors claim that these changes enable the model to capture more details. Interestingly, these changes were motivated by the visualization of filters proposed in their work. We will go more into details on this in Section 3.4 about interpretability of models.

This same year, VGG (Visual Geometric Group) architectures [Simonyan and Zisserman 2015] were proposed. They rely on successive small 3x3 convolutional layers. The authors claimed that two 3x3 consecutive convolutional layers have the same receptive field as one 5x5 convolutional layer but are lighter computationally speaking. It takes images of size 224x224(x3) as input and alternates 3x3 convolutional layers with increasing number of filters (or depth) and 2x2 (stride 2) max pooling layers with two or three fully connected layers in the end. For example, the most popular implementation called VGG-16 is made of two blocks of two convolutional layers followed by three blocks of three convolutional layers (respective depths being of 64, 128, 256, 512, 512), each block being intercut by max-pooling layers thus outputting a 25088-descriptor that is fed to two consecutive 4096-fully connected layers and a final 1000-fully connected layer. It is shown to be a very efficient implementation reaching a performance of 6.8% top-5 error rate in 2014 which places it in second place behind another well known model called GoogLeNet with 6.7% top-5 error rate.

GoogLeNet architecture [Szegedy et al. 2015] relies on a module called Inception (see Figure 3.2). This module is motivated by the idea of mixing multi-scale information, and thus consists of three parallel neural networks with different filter sizes and a pooling layer, using 1x1 convolution for dimension reduction and concatenating the resulting feature maps. GoogLeNet is made of three first convolutional layers of filter sizes 7x7, 3x3 and 1x1, followed by 9 successive inception modules and a final 7x7 pooling outputting a 1024-descriptor fed into two fully connected networks. An interesting contribution of this work is also the two auxiliary classification branches placed after the third and the sixth inception module that oblige early layers to learn relevant features. The authors also extended their work a year later and improved performances on LSVRC (5.6% top-5 error rate) by adapting inception modules [Szegedy et al. 2016].

In 2015, by adding skip connections (see Figure 3.3) to deep VGG-like networks, an ensemble of ResNet [He et al. 2016] reached 3.57% top-5 error rate. Skip connections (or connection shortcuts) consist in adding (summing or concatenating) to a block output the input so the information extracted by earlier blocks can not be lost in further blocks.

The architecture called DenseNet [Huang et al. 2017a] extends this idea with dense blocks

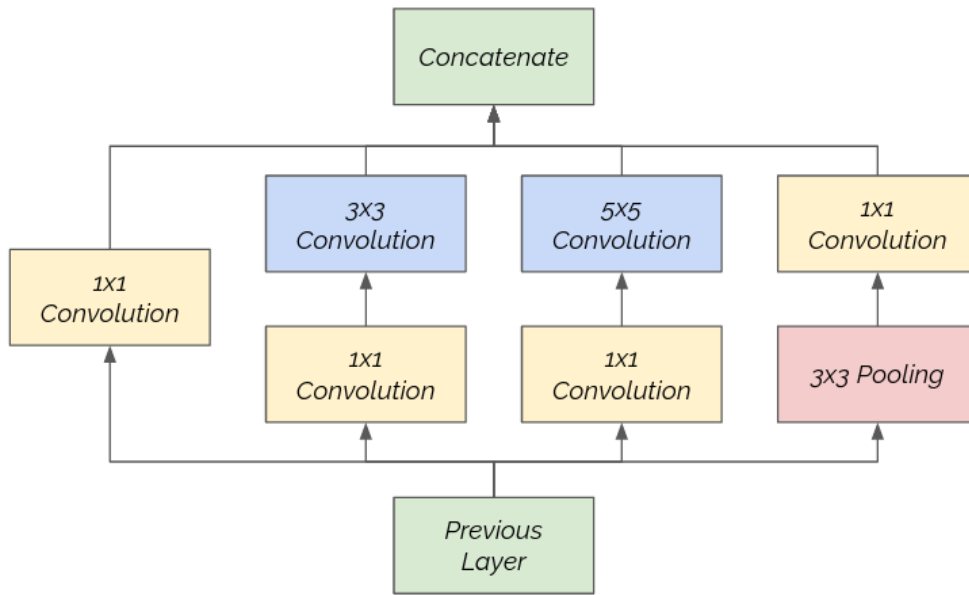


Figure 3.2: Inception module.

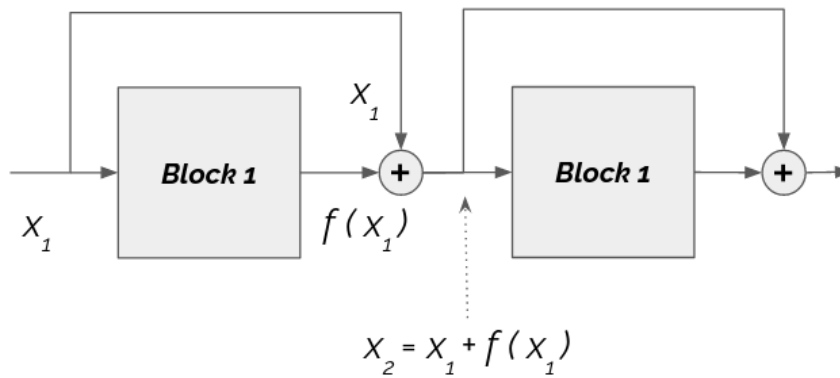


Figure 3.3: Skip connections.

where every layer receives (by concatenation) outputs from all previous layers in the block and passes forward its output to all layers that follow. In addition to making models lighter

with generally twice as less parameters than in a DenseNet model to obtain equivalent performances, this improves the obtained results over equivalent the single ResNet model from 22.4% top-1 error rate to 22.1%.

The latest breakthrough in feature extractor comes from EfficientNet architectures [Hoon Tan et al. 2019] that come from a compound model scaling that consists in optimizing, through a grid search on architectural hyper-parameters (width, depths, resolution ...) and under a constraint on the total number of parameters, a baseline architecture to perform best. In this work the authors propose to create a baseline architecture called EfficientNet-B0 with neural architectural search [Tan et al. 2019] based on light MobileNet feature extractor [Howard et al. 2017] and then propose nine (up to EfficientNet-B8) models scaled from the baseline architecture. In the end, for example, Efficient-B7 reaches a top-1 error rate of 15.6% on LSVRC with only 66 million parameters which is about as much as a ResNet-152 architecture that performs at 22.3%.

Finally, the current state-of-the-art method for LSVRC comes from [Xie et al. 2020b] and is an EfficientNet-B8 model which is trained benefiting from additional adversarial examples and using auxiliary batch normalization [Ioffe and Szegedy 2015] during training not to bias original images batch norm layers parameters. The performances are a top-1 error rate of 14.5%.

This paragraph introduced most popular and efficient Convolutional Neural Networks (CNN) architectures and detailed their specificities and promises. It enables us to understand the tools that we will further extensively use in our work and that are widely used for image classification in medical applications as we will show in next paragraphs. We will also see how they are directly integrated in pipelines for more challenging tasks such as object detection or WSI classification.

### 3.1.3 Cervical cancer usage

Medical applications of the methods described so far quickly emerged for different indications or image modalities. Among these applications, cervical cancer related applications also were studied notably thanks to the publication of Herlev dataset (described earlier in Section 2.2.3) which pushed for cell classification researches.

Regarding cervical cells classification, most of the literature focuses on the “abnormal”/“normal” (from now on it will be referred to as “binary”) 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 AUC of 0.99. A similar approach has been evaluated in [Taha et al. 2017] with an AlexNet architecture followed by a Support Vector Machine (SVM) and reached 99.51% recall, 99.5% precision and 99.19% accuracy on the binary problem. In my opinion one limitation of both previous methods lies



in their preprocessing, which consists in padding input images with black or white pixels to fit the input size expected by the pretrained model they use. Indeed, we know that “abnormal” cells are significantly smaller than “normal” cells, which implies more padding for “abnormal” cells, thus padding can be learned as a discriminative feature for the problem. 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 (*NILM*, *LSIL* and *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%. [Hyeon et al. 2017] propose to classify patches (about 8,300 per binary class) extracted by medical experts using a VGG-16 model pretrained on ImageNet to extract features and train a SVM. They obtain a F1 score of 0.78 regarding binary classification.

Finally, [Rahaman et al. 2020] offer a great review of methods dealing with the classification of cells from cytopathology slides using most of the time Herlev dataset [Jantzen et al. 2005] and sometimes ISBI dataset [ISBI 2015] (both described in Section 2.2) or in-house datasets. Interestingly, as we explained before, Nucleo-Cytoplasmic Ratio (NCR) is a critical characteristic for severity classification. So earliest works were consisting in a two stage algorithm: first detecting and segmenting nucleus and cytoplasm, and then a second stage to classify the cell using segmentation results.

The main limitation of these work is that they highly tackle the binary “normal”/“abnormal” classification of cells and do not enter medical guidelines defined previously in Chapter 2.

## 3.2 Localization and object detection

Localization is defined as classifying an image and additionally proposing a bounding box of the region of the image that is responsible for the predicted label. Object detection consists in predicting a bounding box for each object that can be observed, and the class associated to each box. Figure 3.4 illustrates these principles.

Both tasks are of interest because, in the context of Computer-Aided Diagnosis (CAD)

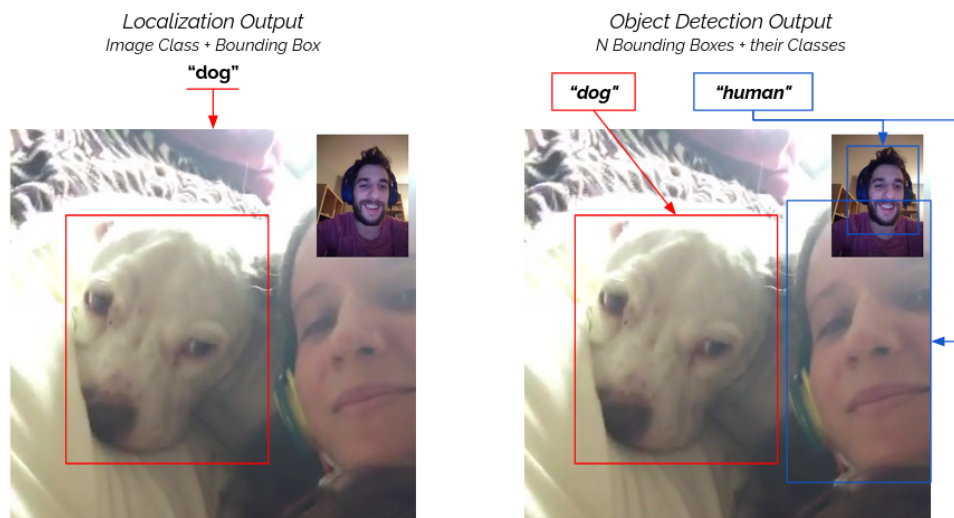


Figure 3.4: Localization and object detection.

tools, localizing regions (e.g. cells) of interest in images can help pathologists to make their reviews more quickly by being guided by these localizations.

The most popular datasets for object detection are Pascal VOC [Everingham, Van Gool, and Williams 2010] and COCO [Lin et al. 2014] which, respectively, contain 20 and 80 classes of objects.

The main metric used, in localization and object detection tasks, is the mean Average Precision (mAP) that consists in computing the mean over all classes of the average precision. It is the same as AUC but the positive prediction is defined with an additional Intersection Over Union IOU thresholding applied to the bounding boxes and the ground truth boxes. Given two regions (e.g. here)  $A$  and  $B$ , IOU is computed as the ratio between the size of their intersection  $A \cap B$  and the size of their union  $A \cup B$ .

### 3.2.1 Architectures

Over the years two types of architectures for localization and object detection were developed: two-stage architectures and single-stage architectures. They offer a trade-off between mAP and speed: indeed [Huang et al. 2017b] compare most popular object detection architectures and highlight that single-stage detectors are faster (due to lower computational complexity) but less efficient in term of mAP. We present here the principles of these approaches and their most popular implementations.

### 3.2.1.1 Two-stages architectures

This section deals with object detection using Region-based Convolutional Neural Networks (RCNNs). We propose a review of major RCNN papers with a specific detailed review of Faster-RCNN architecture.

Overfeat [Sermanet et al. 2014] technique started to bring an answer to object detection by proposing to use a classification network that takes, as input, crops from the original image and regions that are collateral and are predicted as the same class. The problem is that this process is really expensive (computationally speaking) to predict on every region. Moreover, a priori, the scale at which objects are expected is not known.

**RCNN** RCNN method [Girshick et al. 2014] tackles computational limitations exposed before by cropping about 2000 Regions Of Interest (ROI) using a region proposal algorithm (based on edge boxes [Zitnick and Dollar 2014]) to reduce the number of candidate boxes that are to classify. Then, every region is resized to fit the expected input size of a CNN (e.g. 224x224 for a VGG-16) that outputs low dimension descriptors. Finally, these descriptors are given to two distinct branches. The first branch has several SVMs, one per class that outputs whether the associated class is present in this crop or not. The second one is a bounding box regression branch that outputs the correction to apply to the ROI. This approach reached a new state-of-the-art performance on Pascal VOC 2007 improving from about 50% to around 66%. The main drawbacks of RCNN is that there is no training computation shared between CNN features extraction, classification SVMs and bounding box regression branches and that it is pretty slow (about 50 seconds per image).

**Fast RCNN** Fast RCNN architecture [Girshick 2015] uses the same region proposal and CNN feature extraction ideas as RCNN, except that the region proposal algorithm is used to extract regions from the intermediate feature maps of the CNN feature extractor instead of the original image. This makes the whole architecture trainable, and the computation is thus shared among the layers. The varying size of ROIs proposed by the region proposal algorithm implies an additional (max-)pooling layer w.r.t. a grid that matches the fully connected layer expected input size. This approach mainly improves performances of RCNN regarding testing time since it takes about 2 seconds to make a prediction (25 times faster than RCNN) while reaching about 66.9% on PASCAL VOC.

The real bottleneck is now the region proposal algorithm that counts for more than 80% of the computation time.

**Faster RCNN** Faster-RCNN [Ren et al. 2015] proposes to use a Region Proposal Network (RPN) i.e. a CNN trained to propose ROIs. The two advantages from this approach are that features are shared with the region proposal algorithm thus it is trained to perform on the specific task, and the region proposal becomes almost cost-free since most of the computation is already done by the CNN feature extractor.

RPN consists of a fully convolutional network that outputs ROIs by predicting bounding boxes and an objectness score (i.e. probability to contain an object) for each of them. First, there is a  $n \times n$  (originally  $n = 3$ ) convolutional layer over the last feature maps with same padding and a depth of 256. Then, there are two  $1 \times 1$  convolutional layers. One is a “class” layer predicting  $2 \times k$  values ( $k$  binary “object” / “not object” classification) and the other one is a “regression” layer predicting  $4 \times k$  ( $k$  bounding box regression parametrized scores). This  $k$  parameter is generally set to 9 and defines the number of anchors used at each position in the feature maps. Anchors are virtual boxes mapped on the original image considering positions on the feature map. For example, for  $k = 9$ , there are 3 scales (e.g. 64 pixels, 128 pixels and 256 pixels) combined with 3 shapes ( $[1,1]$ ,  $[1,2]$ ,  $[2,1]$ ), which gives 9 anchors. With regards to their IOU with the ground truth bounding boxes, some anchors are paired or not with a ground truth bounding box to train the RPN. The main advantage of Faster RCNN is that it is trainable end-to-end and thus all components are adapted to the task of interest.

Performances achieved by Faster RCNN are around 0.2 second (10 times faster than Fast RCNN, RP algorithm adds only 0.1 second) at test time and a mAP of 73.9% on PASCAL VOC dataset.

Figure 3.5 illustrates all the RCNN architectures presented.

These two-stages approaches enable a good control of the design that can guide the training of models. For example, [Eggert et al. 2017] deal with the question of small object detection when using a Faster RCNN architecture. They propose a theoretical approach to understand what is the minimum size of an object to be detectable, before looking independently at this issue for the RPN and for the classifier. In the end, they propose a practical solution using two RPNs at two different levels of the CNN and stabilize results for datasets with small objects. Feature Pyramid Networks (FPN) [Lin et al. 2017] also tackle this object size issue by combining highest-level feature maps up-scaled with early feature maps, and using the combination to predict at different scales. This improved state-of-the-art results on COCO by applying FPN to Faster-RCNN (using ResNet-101 CNN feature extractor) with a mAP (threshold on IOU at 0.5) from 55.7% to 59.1%.

### 3.2.1.2 Single-stage architectures

In [Redmon et al. 2016] YOLO (You Look Only Once), an architecture able to perform object detection in only 22 ms and still performs at 63.4% of mAP on Pascal VOC, is proposed. Instead of a detection based on region proposal, that implies time consumption by going back to the original image for each region, the idea is to divide the image directly according to a regular grid (e.g.  $7 \times 7$ ) and to have two parallel branches: one that predicts a class probability for each grid cell, and one that predicts bounding box regression based on grid cells with an associated objectness score (from ground truth bounding boxes). These outputs are combined to obtain final predictions. The same authors improved these performances with additional contributions with YOLO v2 [Redmon and Farhadi 2016] (15 ms at test time and a 76.8% mAP) and Yolo v3 [Redmon and Farhadi 2018] (45 ms at test time and a 83.6%

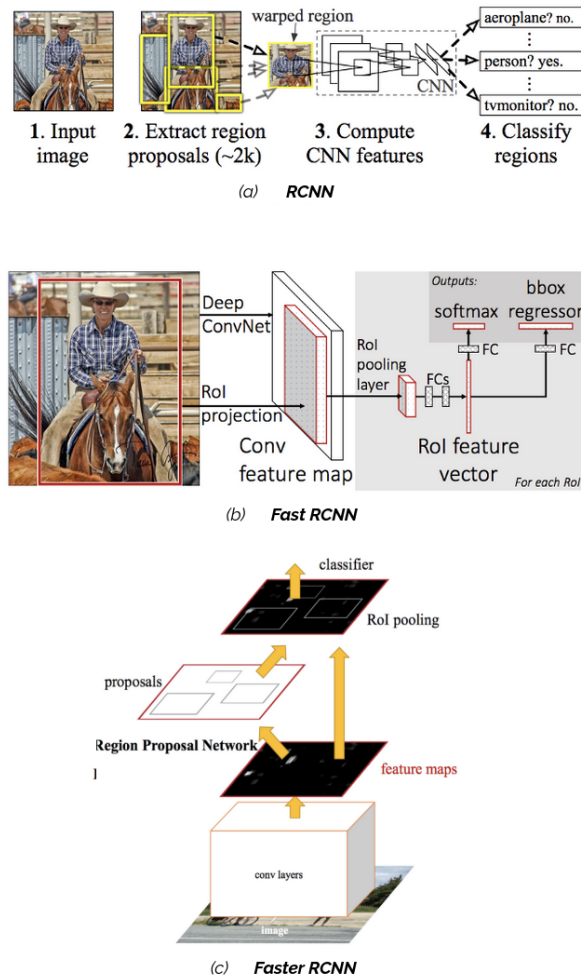


Figure 3.5: Region-based convolutional neural networks. (a) RCNN; (b) Fast-RCNN; (c) Faster-RCNN.

mAP). A similar approach called SSD (for Single Shot Detector) [Liu et al. 2016] achieves a mAP of 72.1% on PASCAL VOC with only 17 ms per image by using FPN on YOLO-like architecture. The main contribution that enabled single stage object detectors to reach two-stages performances is the focal loss [Lin et al. 2020], inspired by the class balancing, that enables to give more importance to poorly classified samples, and to avoid impacting what has already been well learned (mainly the easy “background” samples) during the training.

### 3.2.2 Cervical cancer usage

The general concept of objects detection might be the very fact of what defines the medical expert work on a daily basis. Indeed, a large panel of medical tasks consists in a visual inspection of an organ through imaging methods such as MRI, X-ray, echography or microscopy.

Thus, application of methods presented in the previous section have been popular over the past few years for several medical applications.

Close to our application, [Hu et al. 2019] propose to use a faster-RCNN to detect lesion on cervicography (or colposcopy), and reach an AUC of 0.91 on CIN2+ positive or negative binary classification.

Regarding the kind of images, some interesting works have been made on histopathology slides to detect nuclei or cells. For example, VOCA [Xie et al. 2019] performs multi-task training by predicting, using the same network and three branches, three different outputs (confidence score, localization vector and weight of contribution) for each pixel, that are then combined to give nuclei localizations and thus improve state-of-the-art results on a colorectal cancer dataset.

The closest and only work (to the best of my knowledge) that uses an object detection pipeline on Pap smear slides is [Meiquan et al. 2018]. The dataset is made of 500 whole slides (50 negatives / 450 positives) digitized at 20X. The test set is composed of the 50 negative samples and 50 random positive samples while the training set uses the 400 other positive slides. Training slides are annotated with 5 labels: “ASC-US”, “LSIL”, “HSIL”, “Endocervical Cells” (EC) and “Metaplastic Squamous Cells” (MSC). Each slide is divided w.r.t. a non overlapping grid with patches of 1024x600 pixels and every patch that has at least a target cell is kept. From the 5,721 training patches, 500 are used to build a validation set, the rest for actual training. The statistics of the annotation are the following: 21.2% (1962) are ASC-US, 9.3% (860) are LSIL, 10.2% (939) are HSIL, 38.8% (3589) are EC and 20.5% (1896) are MSC, for a total of 9246 annotations. For testing slides, the foreground area is extracted (using a thresholding on the Z channel from XYZ color space) and patches of 1024x60 pixels of foreground regions are extracted (which removes about 10% of patches). The trained model is a Faster-RCNN with a ResNet-101 backbone (9 anchors with shape 128x128, 256x256 and 512x512 and scale 1:1, 1:2 and 2:1). The results at cell level of this trained network on the validation set are a precision/recall of 0.74/0.52 for ASC-US, 0.83/0.5 for LSIL and 0.87/0.44 for HSIL. At slide level, on the test set, the accuracy is 0.78 for a classification into “positive” and “negative” cells, and the accuracy is 0.7 when it comes to classify among “negative”, “ASC-US”, “LSIL” and “HSIL”. It is really unclear how to go from cell based predictions to slide based diagnosis. The authors say that “the model detects the five types of target cells at first, and then counts the number of cells in each category, finally, generates a diagnosis”. We can only speculate that the most represented class is associated with the slide. It takes about 5 minutes to classify a whole new slide. The precision of positive cells is 0.91 but these performances drop to around 0.8 when it comes to differentiate the different types of positive cells.

Even if object detection methods match the workflow of cytopathologists and if [Meiquan et al. 2018] proved the interest of such approaches, the main and non negligible drawback of such methods is that there is a need for extensive annotations that require a lot of expertise and a lot of time. Moreover, it is not clearly defined how to go from cell detection to slide global label. This motivates a more diagnosis-oriented method: WSI classification.

## 3.3 Whole slide image classification

The emergence of WSI, along with deep learning methods, represents a real opportunity for the development of efficient CAD tools to help pathologists. Indeed, over the last 3 years, notably due to the datasets of WSIs publicly available (presented before [Ehteshami Bejnordi et al. 2017, Tomczak, Czerwińska, and Wiznerowicz 2015]), deep learning pipelines for WSI classification have been developed and compared.

The task of WSI classification consists in associating a label (e.g. proposition of a diagnosis) to a WSI input. It differs from regular classification task presented before due to the large size of these images as input. Indeed, due to their very large size (generally around 10 giga pixels per slide), these images cannot be fed directly into regular classification pipelines.

In spite of this challenge, recent research led to really efficient and elegant solutions.

### 3.3.1 Methods and architectures

Early methods rely on patch classification, which implies annotations from medical experts and a tiling process. Tiling consists in defining a grid with regard to which patches (called tiles) will be extracted. These tiles are then used by regular classification pipelines.

#### 3.3.1.1 Patch-based classification methods

For example, in [Liu et al. 2017], a model (Inception) is trained to classify patches (of size 299x299 pixels) as containing tumor or not (“normal” patch). The dataset is created from 270 slides from which a large number of patches (between 10.000 to 400.000) are extracted and are assigned the label “tumor” if there is a tumor inside the 128x128 region in the middle of the patch. At inference, a sliding window is used and each tile is classified, which enables to output a probability heat map to have a tumor in the whole slide. Dataset balancing is performed by balancing the sampling towards “tumor” patches (on a “tumor” slide there are between 20 and 150.000 “tumor” patches). The slide classification is evaluated with AUC-ROC (0.98 reached) and the tumor detection is evaluated with FROC (Free-response Receiver Operating Characteristic, ROC-AUC adapted to object detection, 0.885 reached). Interestingly, it is claimed that pre-training the model on ImageNet does not improve the results (but ensures a faster convergence of the model).

[Iizuka, Kanavati, and Kato 2020] address the problem of classifying colon and stomach WSI (from Hiroshima University Hospital and Fukuoka Haradoi Hospital) into three classes: “non-neoplastic” (or “normal”), “adenocarcinoma” and “adenoma”. Their approach consists in extracting 512x512 pixels tiles from pathologists annotations, and training a tile classifier. WSI label is predicted through either max-pooling probability over tiles on the whole slide, or through a Recurrent Neural Network (RNN) that recursively takes tile descriptors. An AUC of 0.99 and 0.97 is reached for, respectively, “adenoma” and “adenocarcinoma” on the

stomach application and 0.99 and 0.96 for the colon application. Interestingly, the authors say that “the average annotation time per WSI was about 10 minutes” which shows one of the limitations of such approaches, since it makes them unscalable for application on a larger dataset.

[Sing et al. 2019] are interested in classifying normal tissues in different groups w.r.t. organs and types of tissue. They rely on 1690 WSIs from rats biopsies annotated by experts with 46 classes. Three architectures and five magnifications are compared. The top model reaches an accuracy of 83.4%. An interpretability study is performed using UMAP projection method [McInnes and Healy 2018], and highlights the relevance of what has been learned by exposing semantic clusters in visualizations (we will introduce interpretability methods further in this chapter in Section 3.4).

In [Hoffman et al. 2014], the authors propose a method that consists in extracting a set of (461) features using maximum relevance method [Peng, Long, and Ding 2005] on 512x512 selected patches that represent diseases and training a SVM on 600 manually annotated tiles. Thus they reach about 97% and 98% accuracy on TCGA dataset annotated for respectively ovarian serous cystad-enocarcinoma and renal clear cell carcinoma. They also measure the Pearson correlation between predicted and ground-truth regions for each indication, and reach over 0.3 score for most types of tissue.

Today the state-of-the-art method on Camelyon-16 dataset (described in Section 2.2) is [Lee and Paeng 2018]. Their approach is a two-stage strategy where, first, 224x224 pixels “tumor” tiles are extracted from tumorous regions in “tumor” slides (annotations available) and “normal” tiles are extracted from “normal” slides to train a tile classifier. This model is used to compute heat-maps with “tumor” class probability and 11 hand-crafted features are extracted from it (e.g. largest tumorous region area or maximum confidence probability in WSI). The descriptor that is composed of these features activation is then used to train a random forest to predict the label of the slide. An AUC of 0.98 is reached and a KAPPA of 0.92 for the pN-stage classification.

All of these works are highly interesting, however, as for object detectors, they use extensive annotations, which makes them unscalable to larger datasets or requires the presence of a medical expert for every application.

### 3.3.1.2 Methods based on Multiple Instance Learning (MIL)

Here, we are interested in WSI classification architectures that use only the global label (e.g. diagnosis) to train and require no intermediate information such as cell labeling or tissue segmentation (which are time-consuming annotations). The training is regularized by introducing prior knowledge by design in the architectures which, in addition, makes the result interpretable (see Section 3.4).

Most popular approaches performing this rely on a context of Multiple Instance Learning (MIL) (framed by [Maron and Lozano-Pérez 1997]), i.e. slides are represented by bags of



tiles. Positive bags must contain at least one positive tile and negative bags contain only negative ones (see Figure 3.6). The work in [Durand, Thome, and Cord 2016] particularly contributed to popularize this MIL context by proposing an architecture called WELDON (that stands for “WEakly supervised Learning of Deep cOnvolutional neural Networks”) and reaching state-of-the-art performances on PASCAL-VOC. Their method consists in cutting images in regions (tiling) with a regular grid and computing for each region a descriptor using a backbone (i.e. CNN for feature extraction), then for each feature of the descriptors the top 3 and bottom 3 scores are summed which gives an image single value descriptor. This image descriptor is then used to classify the image.

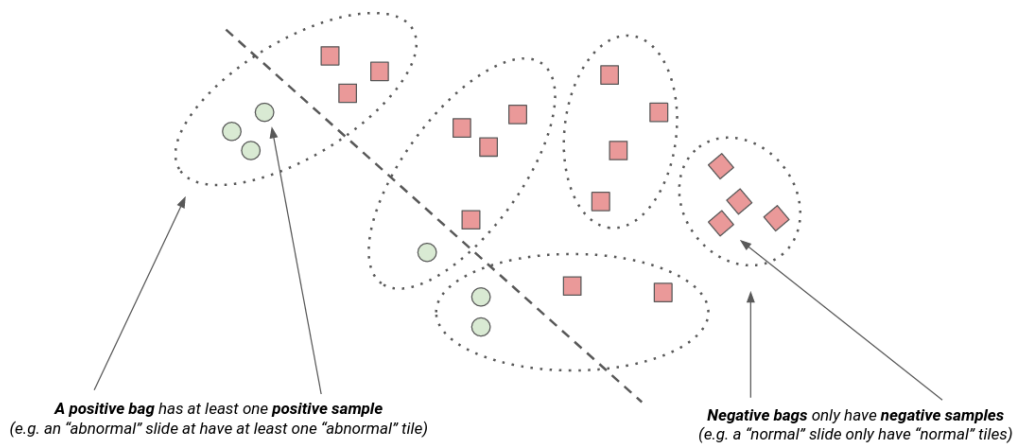


Figure 3.6: Multiple Instance Learning (MIL) context.

However, the MIL context for Camelyon-16, and in general for WSIs datasets, is more complicated than for natural images mainly for two reasons. First, there are up to 10.000 tiles of size 224x224 pixels at 20X magnification, while for natural images bags are made of only some dozens of regions. Secondly, tumorous regions can be as small as 100x100 pixels (localized disease) while in natural images most objects of interest are visually predominant.

Taking this into account, [Courtiol et al. 2018] propose CHOWDER (that stands for “Classification of HistOpathology with Weak supervision via Deep fEature aggRegation”), an extension of WELDON solution for WSI classification. Their approach mainly includes the “relaxing” of this MIL context by pre-computing tiles descriptors using a pretrained network on ImageNet. Thus slides are represented by bags of descriptors (of size 2048) instead of bags of tiles (of size 224x224 pixels). Also they add a 1x1 convolution layer to turn every tile descriptor into a single tile score. Scores are then aggregated using a min-max layer, that keeps the top- $R$  and bottom- $R$  scores (e.g. empirically  $R = 5$  gives the best results), to give a slide descriptor (of size  $2xR$ ) which is given to a two layers fully connected network, with respectively 200 and 100 hidden neurons, that outputs the predicted score. This approach reaches an AUC of 0.858 on Camelyon-16 and 0.915 on TCGA-lung (subset of TCGA dataset related to lung cancer, see more details about datasets content in Section 2.2). Additionally, the authors report that they use tile scores to compute a tumorous heat-map over WSI that

can be used to perform segmentation of tumorous regions with an area under the free-response ROC curve of 0.31 (which is good given the absence of tile-level annotation).

In general, WSI classification methods relying on MIL use preprocessing steps to encode as efficiently as possible tiles from WSIs using tissue detection, tiling and normalization methods and then use three trainable blocks to make the decision. This process is illustrated in Figure 3.7.

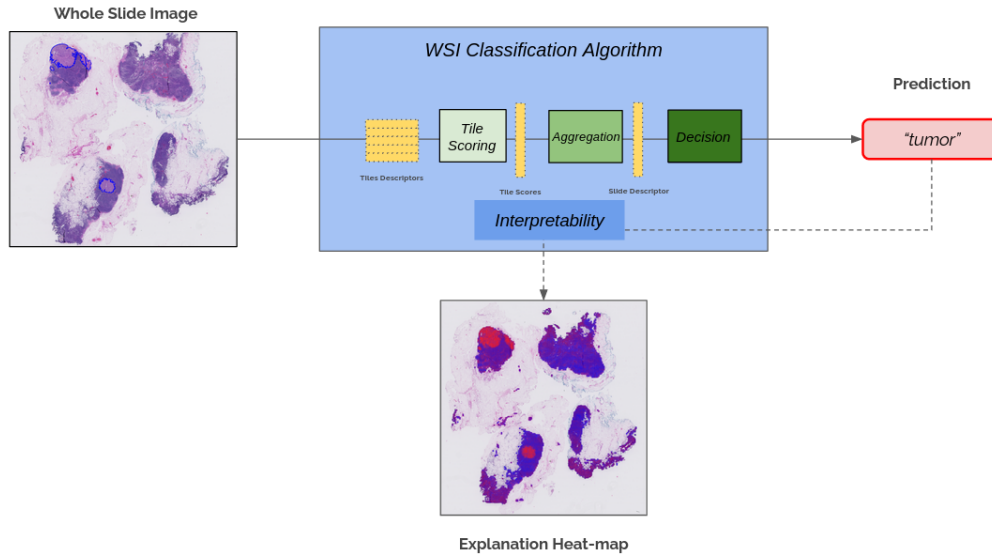


Figure 3.7: WSI classification relying on MIL.

**Preprocessing** The first step of preprocessing generally consists in detecting samples on the WSI since there are a lot of non-informative tiles that are just white background or artifacts (see Figure 3.8).

The most popular method for tissue detection relies on Otsu thresholding [Otsu 1979] and involves defining a threshold on the percentage of foreground pixels for a tile to be selected. Tissue detection through color space transformation and thresholding is also widely used (e.g. thresholding on RGB values in [Coudray, Ocampo, and Sakellaropoulos 2018] or thresholding on the saturation channel of HSV color space for [Lu et al. 2020]). Tissue detection could also be performed using a semantic segmentation pipeline such as U-Net [Ronneberger, Fischer, and Brox 2015] as in [Ianni et al. 2020].

Once the tissue is detected, another step that consists of Stain Normalization (SN) is generally carried out. The motivation behind this step is to improve the transferability to other datasets that come from a different hospital and that might use a different scanner or staining to create their WSIs (Figure 3.9). This question has been studied in [Ciompi et al. 2017] where the authors show that the color normalization of [Ehteshami Bejnordi et al. 2016] improves the training process and the generalization on another test set. A rectal cancer

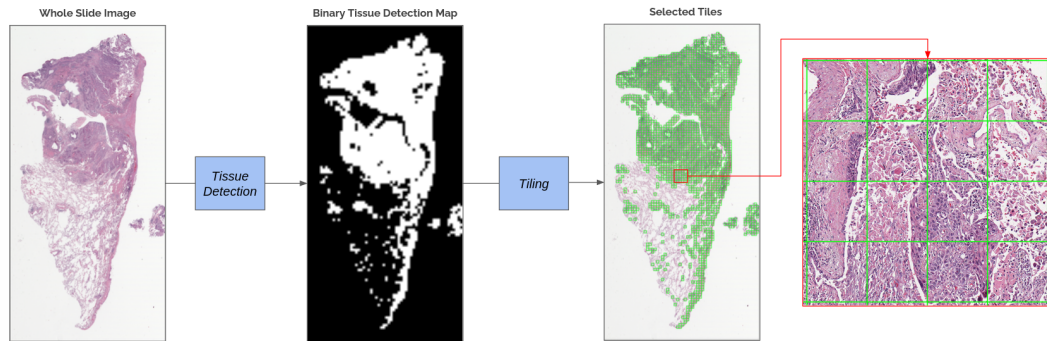


Figure 3.8: Tissue detection process.

dataset (74 slides) is used for training and a colorectal cancer dataset (10 slides) is used as a test set to measure the transferability of the learning. Indeed, training without SN gives an accuracy of 75.55% if SN is performed on the testing set and 50.96% if not, and training using SN gives an accuracy of 79.66% with SN on the testing set while only 45.65% is reached without it.

Most SN techniques consist of color deconvolution [Zhou, Hammond, and Parker 1996] to work in a color space where channels represent concentrations of stains. Several automatic stain vector computations for color deconvolution have been proposed, among which [Macenko et al. 2009] and [Khan et al. 2014] are the most popular ones. The color transfer method proposed in [Reinhard et al. 2001] is also popular due to its simplicity/efficiency trade-off. More advanced methods using state-of-the-art deep learning generative models, called Generative Adversarial Networks (GAN) [Goodfellow et al. 2014], were recently applied with more specifically cycle-GAN approach [Zhu et al. 2017], to perform stain standardization [Bel et al. 2019].

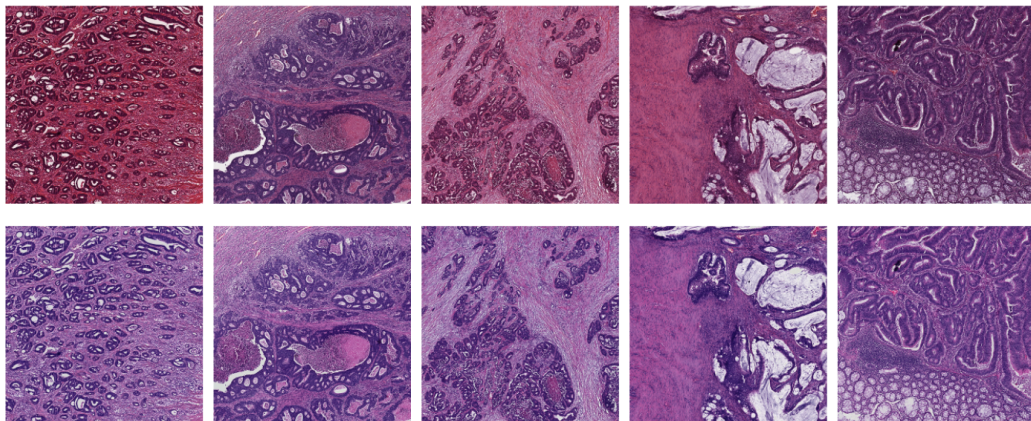


Figure 3.9: Stain normalization illustration (from [Ciompi et al. 2017]).

**Architectures and applications** We already presented CHOWDER approach, and in general WSI classification architectures follow the same idea. For example, [Ilse, Tomczak, and Welling 2018] propose to use an attention module (a two layers fully connected network with 128 and 1 hidden neurons and a softmax layer) to compute competitive and normalized (sum to 1) tile scores from tiles descriptors. Then, the slide descriptor is computed as the weighted (by tile scores) sum of tile descriptors. They report an AUC of 0.775 for a breast cancer dataset and 0.968 for a colon cancer dataset. They also highlight the relevance of attention module tile scoring by visual improvement of heat-maps.

More recently, [Campanella et al. 2019] propose a two steps training to, first, adapt the backbone part (and thus the relevance of tile descriptors computed with respect to the task), while still using only global slide-level label. Then training the WSI classification part relies on a RNN (Recurrent Neural Network [Raffel and Ellis 2015]) aggregator. They reach an AUC of 0.991 for prostate cancer classification and 0.93 on Camelyon-16.

All these architectures have the great advantage of being efficient on a large range of tasks. Some works also aim at adapting these to induce more a priori knowledge to tackle a specific task. For example [Li et al. 2019] propose a two stage training for prostate cancer classification that enables to perform a multi-scale approach that is closer to what pathologists experience. They reach an accuracy of 84.3% on a 3-classes dataset and are able to propose precise heat-maps.

On the same application as Camelyon-16 (breast cancer) but on a different dataset (BioImaging 2015 Breast Histology Classification Challenge, 2048x1536 pixels images), [Li, Wu, and Wu 2019] propose to use a four stage pipeline by tiling with tile size of 512 (big tiles) and 128 (small tiles) since diagnosis of some classes can be made at a relatively low magnification level (e.g. 5X) and others may require a higher level of magnification (e.g. 40X). Two feature extractor CNNs are trained on each magnification (associating slide diagnosis as the tile label). Small patches are clustered according to the phenotype (using thumbnail) and only patches for most discriminative clusters are kept. Thus the slide is represented by descriptors from 512 tiles and selected 128 tiles which are then aggregated using the root of degree  $p$  of the sum of vectors to the power  $p$  (called P-norm pooling) as aggregator to obtain a single slide descriptor. Finally a 4-classes SVM is trained to predict the slide label. With this approach (that in addition requires training 5 different CNNs) an accuracy of 88.89% is reached.

In general, as presented in Figure 3.7, these MIL pipelines rely on an encoding step of tiles using a CNN and then three distinct learning blocks which consist in associating a score with each tile, aggregating according to tile scores and classify the slide using the aggregated vector or value. [Lu et al. 2020] and [Li, Li, and Eliceiri 2020] improve the baselines by relaxing attention-based scoring, by respectively using a clustering layer and using contrastive learning [Chen et al. 2020a] and multi-scale embedding. [Ianni et al. 2020] propose to train their model using regularization based on dropout (inspired by [Gal and Ghahramani 2016]).

In [Coudray, Ocampo, and Sakellaropoulos 2018], TCGA-lung dataset is used with its three classes “normal”, “LUAD” (adenocarcinoma tumor), “LUSC” (squamous cells carci-

noma tumor). The 1634 slides were separated into training, validation and testing sets and tiled using a 512x512 pixels non-overlapping grid. Each tile is associated with the slide ground-truth label. An Inception-v3 model is trained and a slide score is computed using either the proportion of tiles predicted per class or the average probability over tiles of a slide. A “normal” vs “tumor” classification AUC of 0.99 and 0.993 is reached, and 0.97 for the three-classes problem. The authors also work towards what is called “discoveries” which consist in predicting from the image new clinical or medical information. Indeed, they use the same approach to try to predict 10 gene mutations that are given by TCGA database, and among these 6 gene mutations could be predicted with an AUC above 0.73 and up to 0.85.

This idea of “discovering” new links between WSI and clinical data through deep learning has been very popular lately. For example for treatment response using average pooling as aggregator [Naylor et al. 2019], or in [Coudray, Ocampo, and Sakellaropoulos 2018] where the authors present and compare [Fu et al. 2020] and [Kather et al. 2020] approaches that worked on predicting genetic information. [Naik, Madani, and Esteva 2020] train a model to predict, from Hematoxylin & Eosin (H&E) stained WSIs estrogen receptor status which is usually determined using other immunohistochemistry slides that are more expensive to prepare.

In the context of MIL, tile sampling is also a way to relax the complexity of the leaning and to regularize the training. For instance, [Xie et al. 2020a] take an approach close to [Li et al. 2019] but compute centroids of clusters and use the ensemble of tiles that are the closest to centroids to classify the slide. [Combalia and Vilaplana 2018] propose Monte-Carlo like sampling that selects tiles to constitute the next training bag and thus enable to train also the feature extractor by reducing the size of the input.

Finally, [Srinidhi, Oxan, and L. 2021] offer a great and up-to-date review of state-of-the-art methods for computational histopathology.

Recently, the gap between methods using annotations and methods using only global labels for WSI classification has been closed by [Dehaene et al. 2020]. The authors explain that the only difference is the fine-tuning of feature extractor backbone to be task specific. The solution they propose is to use Moco v2 [Chen et al. 2020b], a contrastive unsupervised learning method, to train the feature extractor in a self-supervised manner. Thus they improve CHOWDER from a mean AUC of 82.3 (when using feature extractor from ImageNet) to 0.983.

### 3.3.2 Other methods for WSI classification

Even if MIL methods are the most popular ones and are about to gain even more interest with [Dehaene et al. 2020] publication, other promising methods have been developed recently such as Neural Image Compression-based methods [Tellez et al. 2019] that propose to keep the spatial organization while it is lost in MIL context; thus [Tellez et al. 2020] adapt their first work to a multi-task learning and reach state-of-the-art performances on TUPAC16 [Veta et al. 2019] (baseline of 0.617) a dataset whose task is to predict tumor proliferation on breast biopsies with a Spearman correlation 0.632. [Pinckaers, Ginneken, and Litjens 2019] propose

a streaming-based method that reaches a Spearman correlation 0.570 on TUPAC16, and that consists in benefiting from the locality of the majority of components that constitute convolutional networks to adapt the forward and backward path to be performed directly on tiles. [Cheng et al. 2020] propose to predict segmentation maps through a teacher-student approach using self-similarity. [Barker et al. 2016] use a clustering-based method to have only a coarse representation of tissues. [Shi et al. 2020] use intermediate targets for features through semi-supervision. These methods open a new manner to tackle WSI classification that could be merged with MIL pipeline to improve the results.

### 3.3.3 Cervical cancer and cytology applications

As said before, most applications of WSI classification have been made and thought for histopathology slides.

Thus, an application of tile-based approach has been made to differentiate successfully adenocarcinoma and Squamous Cells Carcinoma (SCC) in [Idlahcen, Himmi, and Mahmoudi 2020] through the extraction of 300 tiles from each class. In November 2020, the Société Française de Pathologie (French Pathology Society) released a dataset of more than 1500 slides from cervical biopsies in four classes. This will open new opportunities for computer scientists to contribute to research in this field.

But for cytology, very few works have been published on the subject of WSI classification.

[Sornapudi et al. 2019] propose a tile-based approaches working on 25 cervical Liquid-Based Cytology (LBC) slides (19 “abnormal” and 6 “normal”) and combining it with Herlev dataset. Their approach consists in using annotations from cytopathologists to extract around 4,120 tiles/cells and then train a classifier using Herlev images and 2,800 extracted and annotated cells for labeling. They reach an accuracy of 0.888, a sensitivity of 0.882 and a recall of 0.897 using a VGG-16 architecture.

[Dov et al. 2021] are interested in classifying thyroid cytology slides according 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 reach an AUC on tiles of 0.985 and at slide level they have a mean AUC of 0.872 slide and an accuracy of 0.44 (on the 5 classes problem that is TBS).

This overview shows that, in spite of the critical importance of cervical cancer screening, using WSI classification pipeline is not popular mainly for two reasons. First, no dataset exists and, secondly, this task is way more challenging than histopathology slides classification, and there is a need to relax the MIL context by adding annotations.

## 3.4 Explanations and interpretability

Interpretability (the ability to provide explanations that are relevant and interpretable by experts in the field), also referred to as explainability here, is of critical interest in general but even more when it comes to medical applications.

The main reasons why interpretability is crucial for medical applications are:

1. For routine tools where useful features are well known and are subject to a consensus among experts (which is the case for cervical cancer screening), it is important to show that trained models rely on the relevant features to make the decision in order to gain the confidence of practitioners.
2. A good explainability would enable to assist more efficiently medical doctors in their slide reviews by identifying discriminative regions.
3. The ability to train using only slide level supervision opens a new field we call discovery. It consists in predicting, based on easier access data, outputs that generally require heavy processes, for example, predicting a the response of a patient to a treatment based on biopsy imaging (while generally the only solution to have this information is to try and wait). In order to be able to guide experts towards new discoveries, the need for reliable interpretability is obviously high.

In this section, we summarize the most popular interpretability methods, explain how they can be used to retrieve information that were not given at training time (weak supervision), present how to measure the interpretability and finally expose some medical cases using interpretability.

### 3.4.1 Methods for interpretability

While interpretability for deep learning CNN models is still at its beginning, some methods arise from the literature. “Feature Visualization” has been proposed in [Zeiler and Fergus 2014] and extensively developed in [Olah, Mordvintsev, and Schubert 2017]. It consists in visualizing in the most interpretable manner features associated with a single neuron or a group of neurons. It can be used to understand the general training of a model (see Figure 3.10).

For example, the question of transferring features learned from natural images (ImageNet) to medical images has only recently been deeply investigated [Raghu et al. 2019] while widely used and yet surprisingly good. It has also been used to measure how robust a learned feature is [Couteaux et al. 2019].

Another type of explainability methods is attribution methods. These methods measure, for each component of the input (e.g. pixels), its contribution to the prediction. They are



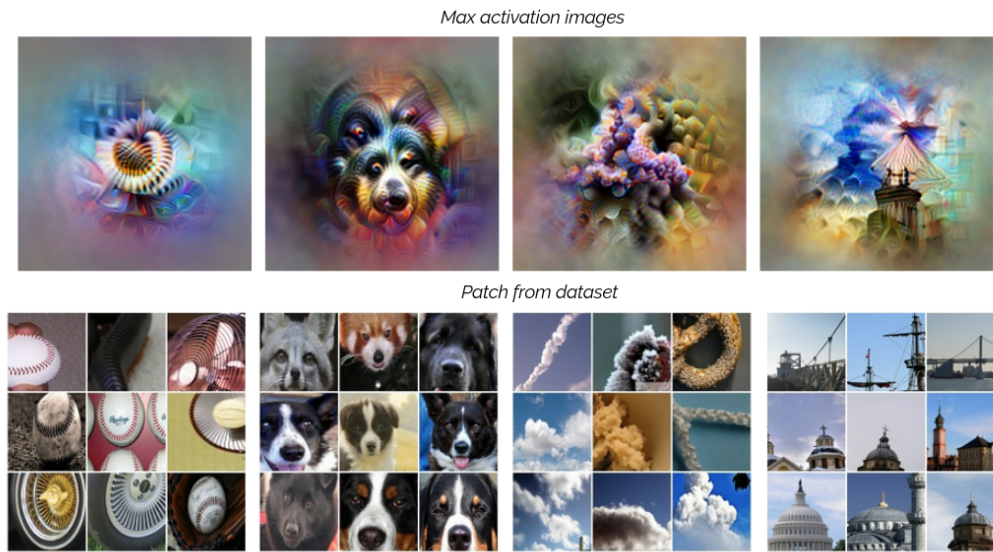


Figure 3.10: Feature Visualization examples (from [Olah, Mordvintsev, and Schubert 2017])

performed either through perturbation [Fong and Vedaldi 2017] or gradient computation (i.e. measure of the gradient of the output with respect to the input). This second group of methods is gaining more and more attention.

In [Simonyan, Vedaldi, and Zisserman 2014], the authors show that the gradient is a good approximation of the saliency of a model and even put forward a possibility to perform weakly supervised localization. This work opened a new way of accessing explanations in deep neural networks and motivated a lot of interesting researches [Sundararajan, Taly, and Yan 2017; Smilkov et al. 2017; Srinivas and Fleuret 2019; Goh et al. 2020] (see Figure 3.11).

Grad-CAM is another gradient-based attribution method [Selvaraju et al. 2017] that comes from Class Activation Mapping (CAM) [Zhou et al. 2016]. It consists in computing a weighted average of feature maps at a given depth of the feature extractor, where weights are computed using gradients of the predicted class output with respect to these feature maps.

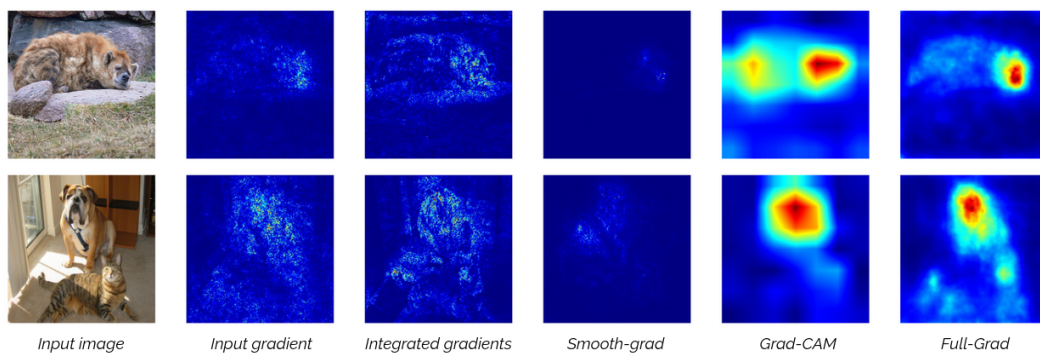


Figure 3.11: Gradient-based explanations examples (from [Srinivas and Fleuret 2019]).



Another manner to approach interpretability and highlight what has been learned to differentiate classes is through dimension reduction. t-SNE [Maaten and Hinton 2008] and UMAP [McInnes and Healy 2018] methods offer efficient dimension reduction for deep learning methods that enable to visualize how data are embedded and separated in a 2 or 3 dimensional space with an iterative and probabilistic approach.

More recently, [Olah et al. 2020] introduce circuits that bring interpretability to the next level. It takes neurons and by visualizing features associated to them (based on the previous work from these authors [Olah, Mordvintsev, and Schubert 2017]), they imagine a set of tests to validate what is interpreted. It also shows how a complex and deep CNN can be divided into smaller networks (called circuits) that perform specific tasks.

### 3.4.2 Evaluation and quantification of interpretability

As the quality of explanations improved, the importance of quantifying these improvements grew, which pushed researches to question interpretability methods and to compare them by measuring their performances.

For example, RemOve And Retrain (ROAR) [Hooker et al. 2018] method consists in removing contributing items (features, pixels ...) identified by an interpretability method from training samples in order to evaluate the completeness and relevance of this method by measuring the impact of such ablations on the learning and the performance of the model. Tests (cascade randomization, data randomization ...) are designed in [Adebayo et al. 2018] to show whether interpretability methods are sensitive to model parameters or input. In [Nie, Zhang, and Patel 2018], it is shown that some popular interpretability methods are doing a simple partial image recovery that makes them model or class sensitive and this is supported by adversarial examples.

[Kindermans et al. 2019] propose a property that an attribution method should have, called “shift invariance”, and show that some popular methods do not respect it. Note that defining properties required for interpretability methods has motivated a lot of works such as “completeness” property and “implementation invariance” property in [Sundararajan, Taly, and Yan 2017], “weak dependence” property in [Srinivas and Fleuret 2019] or “sensitive-n” in [Ancona et al. 2017].

### 3.4.3 Medical usage

Due to their growing importance, interpretability methods have been extensively used in medical applications especially to access further and more complex information than what the model is trained for.

For example in [Courtiol et al. 2018], as we said (and illustrated in Figure 3.7), interpretability is induced by design in the training and it works well with a FROC (see Section 3.3.1.1 for definition) of 0.318 on Camelyon-16 detection while it is fully weakly supervised

regarding detection. The same weakly-supervised localization is performed in [Sundararajan, Taly, and Yan 2017] for diabetic retinopathy detection (while only image label is given at training time) using their Integrated Gradients interpretability method.

In [Campanella et al. 2019], t-SNE tile visualization is proposed to show the relevance of what is learned by the model by highlighting how it separates sub-types of tissues that were never explicitly given to the model.

[Schutte et al. 2020] use a generative model (called StyleGAN [Karras, Laine, and Aila 2019]) to generate images that highlight which changes in image would be responsible for a change in prediction.

More advanced works even propose to improve performances relying on interpretability and guiding the fine-tuning of models by constraint on dimension reduction and clustering method called “projective latent interventions” [Hinterreiter, Streit, and Kainz 2020].

Even if interpretability for medical application seems to grow and be efficient enough to gain the confidence of medical experts, there are still some limitation in particular to guide pathologists for their analysis of slides and to guide them towards new discoveries. Indeed, the 0.318 FROC metric reached by [Courtiol et al. 2018] is interesting but is far from fully supervised state-of-the-art methods performing at 0.807.

### 3.5 Conclusion and discussion

In this chapter, we presented the most popular CNN feature extractors and their performances on ImageNet challenge. Object detection pipelines have been detailed and we showed that, even if they perform well in general, their need for extensive annotations becomes a limitation in their development. Thus, we introduced WSI classification mostly in a MIL context and its instantiations on histology public datasets. We also explained why these approaches are interpretable by design. Thus, we further detail importance and promises of interpretability methods while highlighting how their performances can be measured.

From this literature overview, we can identify that crucial components for an efficient automatic Pap smear WSI classification system are: an efficient tool for cell classification that goes beyond the binary “normal” vs “abnormal” classification, an interpretability scheme, and a whole slide image classifier adapted to cytology use case.

# Cell and region classification for cytology whole slide imaging computer-aided diagnosis tools: proposed methods and experiments

---

## Sommaire

---

<b>4.1</b>	<b>Cell-level classification</b>	<b>46</b>
4.1.1	Herlev severity	46
4.1.2	Backbone comparison	46
4.1.3	Feature fine-tuning	48
4.1.4	Regression approach	49
4.1.5	Classifier under regression constraint	50
4.1.6	Interpretability	52
<b>4.2</b>	<b>Simulated tiles classification</b>	<b>53</b>
4.2.1	Simulated dataset	54
4.2.2	Classification	55
4.2.3	Interpretability	62
4.2.4	Weakly supervised localization	64
4.2.5	Weakly supervised abnormal cell detection	65
<b>4.3</b>	<b>Liquid-based cytology whole slide image classification</b>	<b>66</b>
4.3.1	Pre-processing	68
4.3.2	Integration in a computer-aided diagnosis pipeline	69
4.3.3	From tile-level predictions to slide-level diagnosis	71
<b>4.4</b>	<b>Conclusion</b>	<b>72</b>

---

While Pap tests are 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 chapter, 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 use a loss function, 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 benefit from this loss function 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 easy to integrate in a pathology laboratory workflow as possible, and extending it to make a slide-level prediction.

## 4.1 Cell-level classification

In this section we are interested in classifying cell images. As introduced in Chapter 2, in order to classify a slide, pathologists need to go through a cornucopia of cells (up to  $100 \times 10^3$ ). Thus it makes sense to start by addressing the problem of automatic squamous cell classification.

### 4.1.1 Herlev severity

The Herlev Dataset ([Jantzen et al. 2005]; see Figure 2.18) is a cytology image set composed of 917 images gathered in 7 classes: normal columnar, normal intermediate, normal superficial, light dysplastic, moderate dysplastic, severe dysplastic, and carcinoma in situ. The three first classes belong to the category of normal cells and the last four are abnormal ones (in order of severity, carcinoma in situ hinting at the presence of an actual cancer). Images are between 50 and 400 pixels wide. Here, we merged normal images into a single class in order to study the medical severity or malignancy only, thus building a 5 classes dataset, we call Herlev severity consisting of: normal (242), light dysplastic (182), moderate dysplastic (146), severe dysplastic (197) and carcinoma in situ (150).

### 4.1.2 Backbone comparison

We started by comparing two backbones (or feature extractor). For that we trained the last fully connected layer of an Inception v3 and a ResNet-101 architecture (pre-trained on ImageNet) on Herlev Severity using multi-class cross-entropy loss that we note

$$\mathcal{L}_{CE}(p; y_x^{cls}) = - \sum_{i=1}^5 y_{x,i}^{cls} \cdot \log(p_i),$$

where  $p = (p_1, \dots, p_5)$  are class probability (neurons resulting of softmaxed logits neurons)

and  $y_x^{cls}$  the one hot label associated with the image  $x$  (zeros array with a 1 at ground truth class index).

We used a 5 random folds (splits of the datasets) and obtained average confusion matrices and KPIs distributions that can respectively be seen in Figure 4.1 and Figure 4.2.

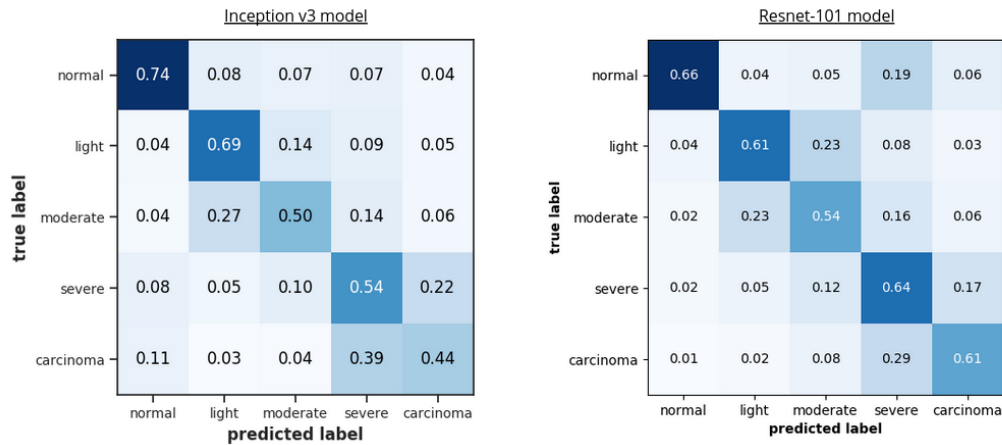


Figure 4.1: Average confusion matrix for ResNet-101 and Inception-v3 backbones over 5 random folds

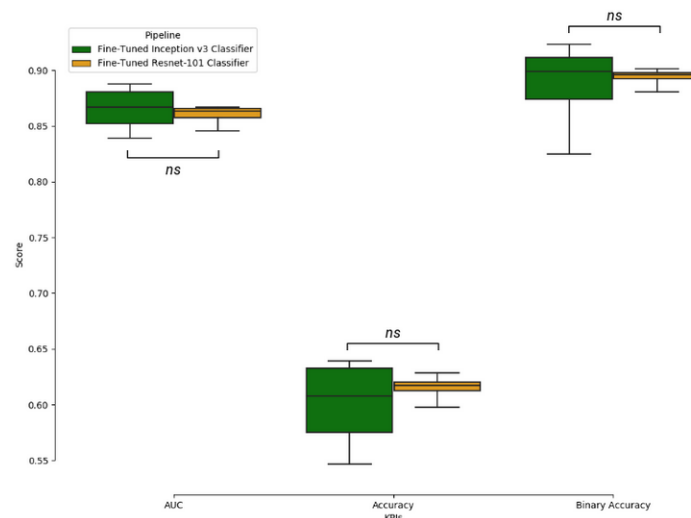


Figure 4.2: Distribution of performances (Accuracy, Area Under the Curve (AUC) and Binary Accuracy) for ResNet-101 (in yellow) and Inception-v3 (in green) backbones over 5 random folds.

Both models seem to perform the same with an average accuracy of 60% for the Inception v3 model and 61.5% for the ResNet-101 model. The result we report confirms the fact that these two feature extractors are not statistically significantly different ( $p > 0.1$ , using Mann-Whitney U Test). Nevertheless, we chose to continue with a ResNet-101 backbone because it seems to be more stable over trainings, i.e. invariant to split.

### 4.1.3 Feature fine-tuning

Then, we fully retrained the ResNet-101 (pretrained on ImageNet) to measure the impact of features fine-tuning on performances. Figure 4.4 shows the distribution of performances of fine-tuned models and compare them to the “frozen” ImageNet ResNet-101 feature extractor.

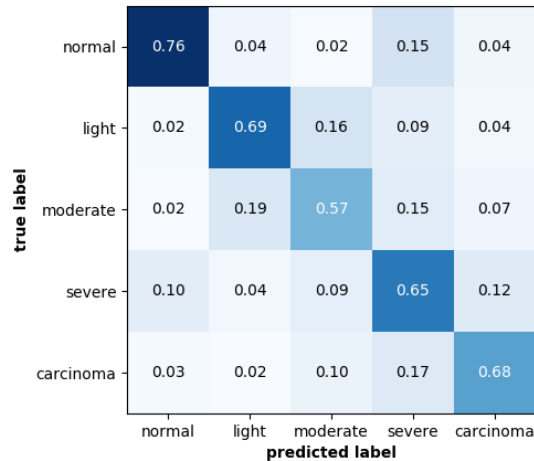


Figure 4.3: Average confusion matrix for ResNet-101 with fine-tuned features over 5 random folds.

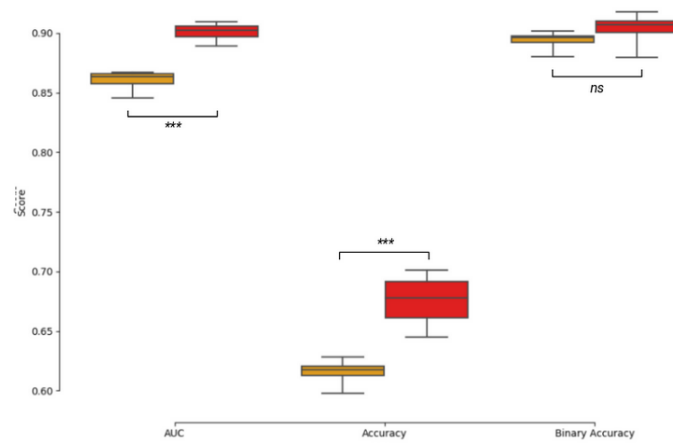


Figure 4.4: Distribution of performances (Accuracy, AUC and Binary Accuracy) for ResNet-101 with ImageNet features (in yellow) vs ResNet-101 with fine-tuned features (in red) over 5 random folds.

We report highly significant improvements regarding overall accuracy with a gain of 6.1% and an increase of 4% in mean AUC.

However in Figure 4.3 we can see that the model tends to misclassify images from the “normal” class and most severe classes (“severe dysplastic” and “carcinoma in situ”). This

was already reported in [Zhang et al. 2017] and identified to be due to the visual similarities between normal columnar and carcinoma in situ cells. Obviously, missing a potential highly abnormal diagnosis is to be avoided. Similarly, due to the fact that 93% of Pap smears are normal during routine diagnosis, misclassifying normal cells would require an additional action by the attending cytotechnicians.

#### 4.1.4 Regression approach

Since the World Health Organization (WHO) classification used in the Herlev set has an order of severity, the classification task can be interpreted as a regression problem. The regression loss will push the network to clearly differentiate normal samples from most malignant ones. We relabel Herlev samples using a score from 1 (for normal ones) to 5 (for carcinoma ones) and use a Mean Squared Error (MSE) as loss function to optimize:

$$\mathcal{L}_{MSE}(s; y_x^{reg}) = (s - y_x^{reg})^2$$

with  $s$  the predicted score and  $y_x^{reg}$  the regression score associated with the image  $x$ .

Thus, we retrain the same ResNet-101 architecture replacing the softmax layer with a fully connected layer.

Figure 4.5 and Figure 4.6 show respectively the average confusion matrix and the distribution of scores per class performed by these regressor models for the same 5 random folds.

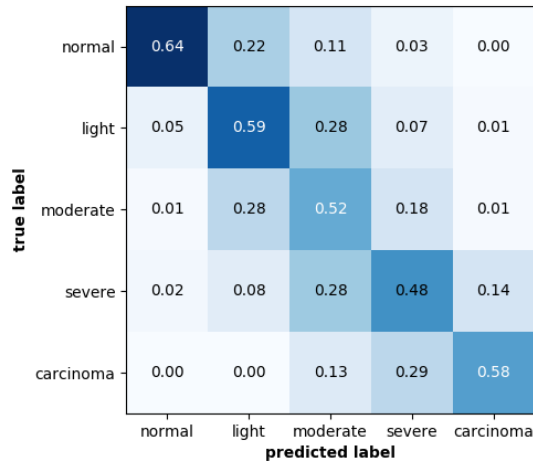


Figure 4.5: Average confusion matrix for ResNet-101 Regressor over 5 random folds.

Most importantly, we can see that models do not mis-classify any normal samples or carcinoma in situ samples with each other. A further point to note from the confusion matrix deriving from this distribution, these models do more mis-classifications than the categorical model, with an accuracy of 58.2%. However these misclassifications are less severe in the scope

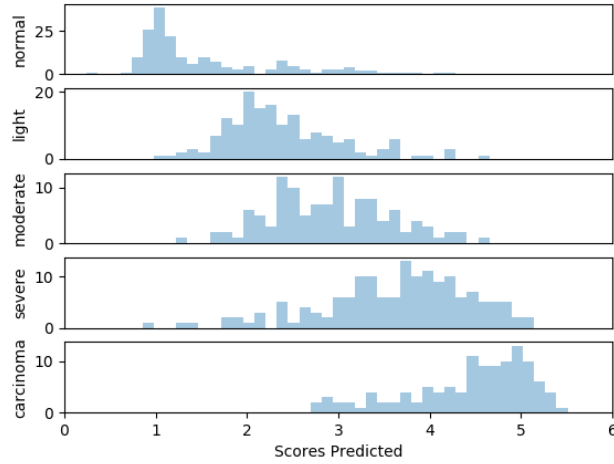


Figure 4.6: Distribution of scores predicted by ResNet-101 Regressor on test sets over 5 random folds.

due to their relative prognosis distance. This can be more easily displayed by the overall MSE of 0.707 over the test set.

#### 4.1.5 Classifier under regression constraint

While the regression loss was more adapted than a classification (cross entropy) loss to the severity task, it nonetheless did not improve the performances per class. In this section we combine the strength of both approaches into a single architecture.

We unify these two pipelines, in order to combine the strength of both approaches, into a single architecture 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  $p = (p_1, \dots, p_5)$  (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 Figure 4.19):

$$s = \text{RegConst}(p; w^r) = \sum_{i=1}^5 (p_i \cdot w_i^r) \quad (4.1)$$

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

$$\mathcal{L}(x, y_x) = \mathcal{L}_{\mathcal{CE}}(p; y_x^{cls}) + \mathcal{L}_{\mathcal{MSE}}(s; y_x^{reg}) \quad (4.2)$$

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).



In Figure 4.7, Figure 4.8 and Figure 4.9, we can see that our ResNet-101 classifier under regression constraint makes less misclassifications than the classifier and has lower MSE than the regressor (see Figure 4.6). Thus, we have an architecture performing on classification task (mean accuracy of 69.9% and mean AUC of 0.922) and on scoring severity task (average MSE of 0.654). What is particularly appreciated here is that the “extreme” classes (“normal” and “carcinoma in situ”) have the best AUC (respectively 0.98 and 0.94).

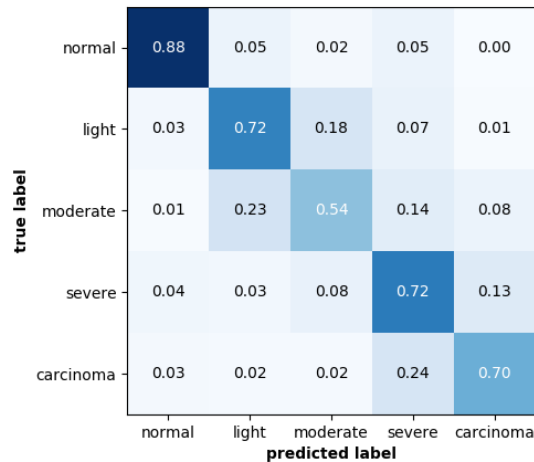


Figure 4.7: Average confusion matrix for ResNet-101 classifier under regression constraint over 5 random folds.

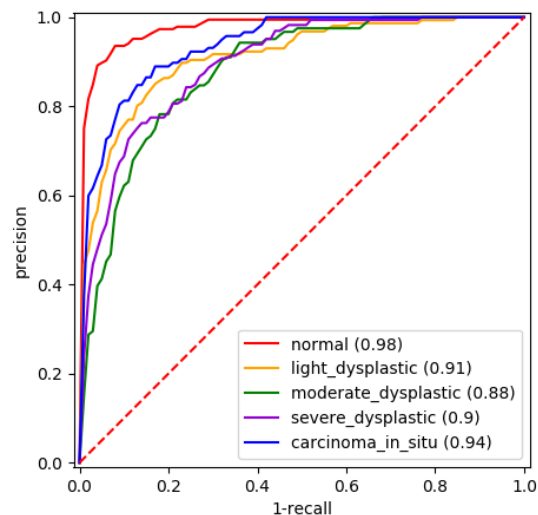


Figure 4.8: ROC and AUC obtained by ResNet-101 classifier under regression constraint on test sets over 5 random folds.

Figure 4.10 sums up the performances over the 5 random folds of the five architectures we experimented: “frozen”Inception-v3, “frozen” ResNet-101, fully retrained ResNet-101,

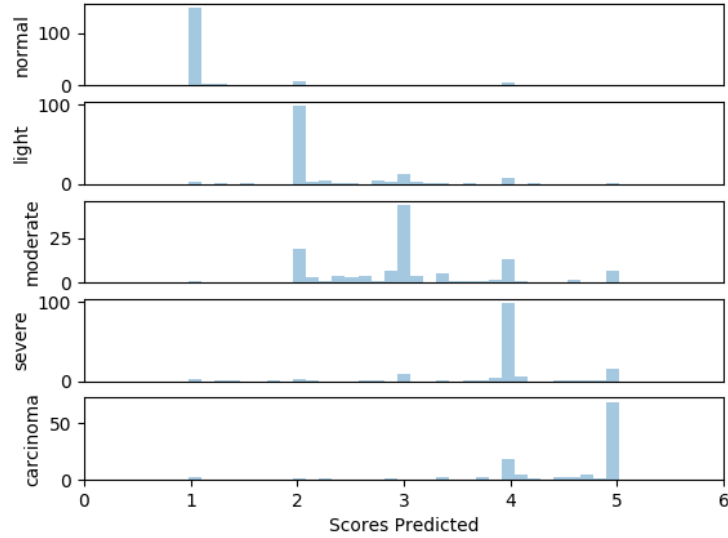


Figure 4.9: Distribution of scores predicted by ResNet-101 classifier under regression constraint on test sets over 5 random folds.

ResNet-101 regressor and ResNet-101 classifier under regression constraint. It shows that overall our proposed classifier under regression constraint is statistically better than other methods.

#### 4.1.6 Interpretability

Attribution, introduced in the previous chapter, 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 (heat-maps that highlight regions that participated to the given label), we applied the Integrated Gradient method [Sundararajan, Taly, and Yan 2017] to highlight on which cyto-morphological 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:

$$A(i) = \frac{(x_i - x'_i)}{m} \cdot \sum_{k=0}^{m-1} \frac{dF(x' + \frac{k}{m} \cdot (x - x'))}{dx_i} \quad (4.3)$$

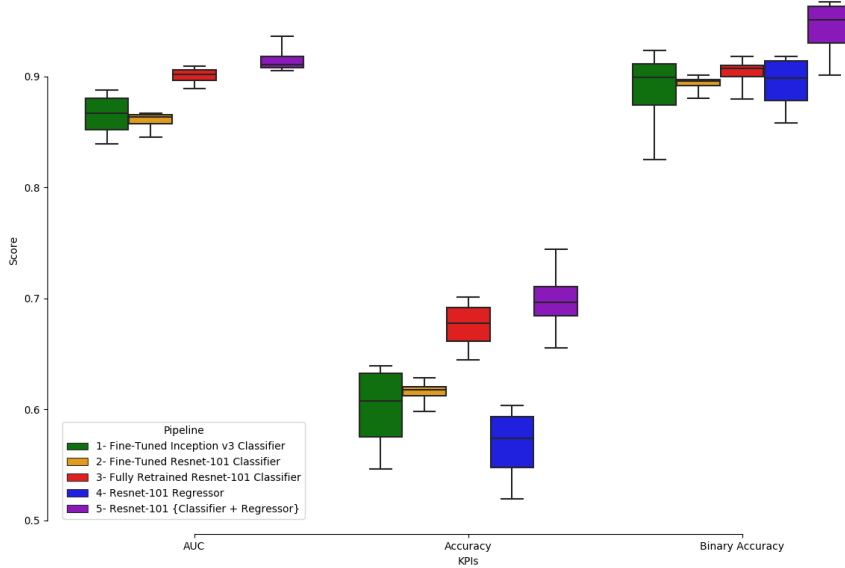


Figure 4.10: Distribution of performances (Accuracy, AUC and Binary Accuracy) for the 5 architectures experimented over 5 random folds.

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  $\Omega$ ), we denote by  $A_{\mathcal{R}}$  the contribution of this region to the predicted label, which is computed as:

$$A_{\mathcal{R}} = \frac{\sum_{i \in \mathcal{R}} |A(i)|}{\sum_{i \in \Omega} |A(i)|} \quad (4.4)$$

Note that the *completeness* axiom defined in [Sundararajan, Taly, and Yan 2017] ensures that, for a baseline defined as before, the attribution over the whole image (denominator) is non-zero.

We can observe that the model seems to rely more on the nucleus region for more severe classes (see Figure 4.11), 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.

## 4.2 Simulated tiles classification

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)

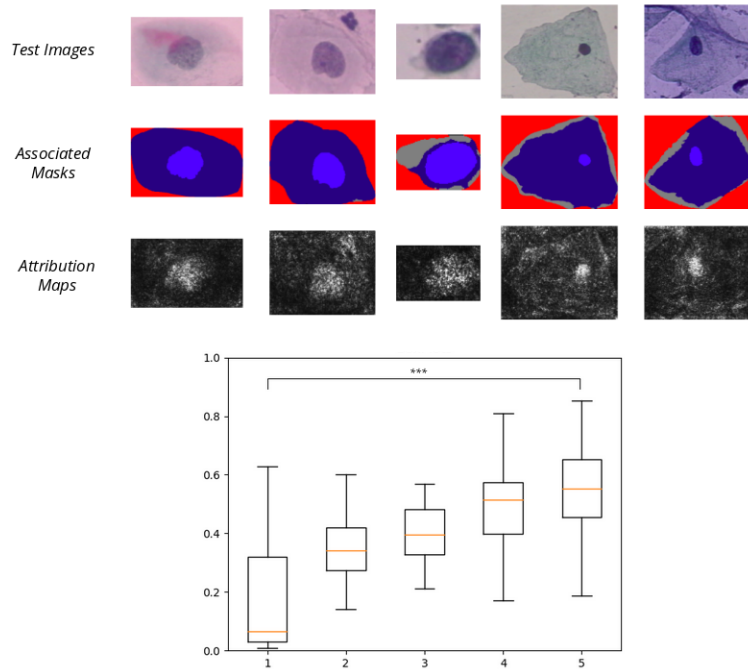


Figure 4.11: Herlev images, their associated nucleus segmentation maps and attribution maps using integrated gradients on trained model (top); distribution of percentage of attribution in nucleus per class (bottom).

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 Whole Slide Image (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. This approach has the advantage of getting us closer to Whole Slide Image (WSI) classification by working on patches instead of individual cells (indeed cyto-pathologists do not analyze every cell individually).

#### 4.2.1 Simulated dataset

To create realistic tiles, we need proper cytology background images. 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 Figure 4.12), 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). We separate single cells into 3 sets (training, validation and test) and create the “simulated” cytology tiles sets using only single cells from the corresponding set. Cells are selected randomly and placed at a random position on the tile that does not overlap other cells. Overlap is not considered in order to avoid a cell to hide an informative part of another cell that would create adversarial samples.

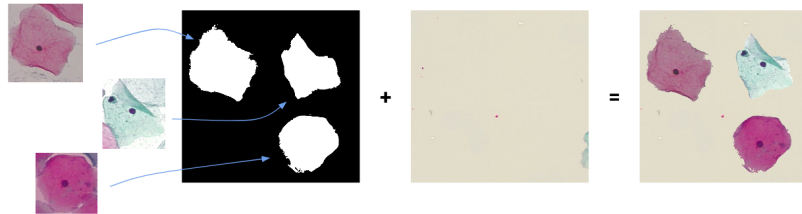


Figure 4.12: Simulated tile creation process.

The challenge presented by what we call the “simulated” cytology tiles dataset 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 Figure 4.13 (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 image labels are used). We make sure that each Herlev cell is used only in one split of the “simulated” tiles dataset.

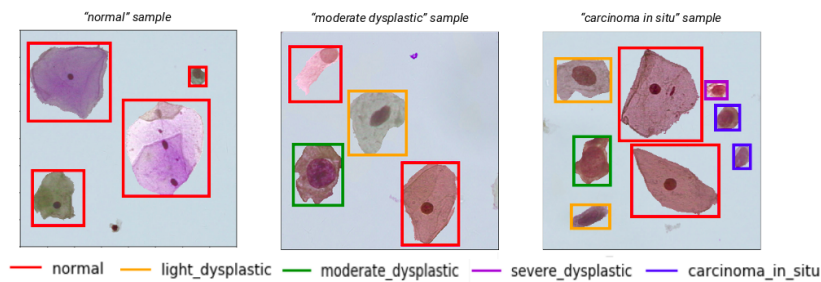


Figure 4.13: Simulated tile labels.

We created a 1808 images dataset (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*.

### 4.2.2 Classification

The problem of ordered classification task is known as “ordinal regression”. In the following paragraphs, we start by training a classification architecture before detailing two methods that are generally used to tackle these ordinal regression challenges. Finally, we apply the classification pipeline under regression constraint on the “Simulated” tiles dataset to show and validate the improvement that this method brings. We perform 5 trainings per pipeline to ensure the statistical significance of the proposed improvements by comparing three evaluation measures: overall accuracy, binary normal/abnormal accuracy and quadratic KAPPA value.

#### 4.2.2.1 Softmax cross-entropy approach

We start by training a regular (softmax cross-entropy for loss) classifier pipeline 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 Figure 4.14, 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 Figure 4.15 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.

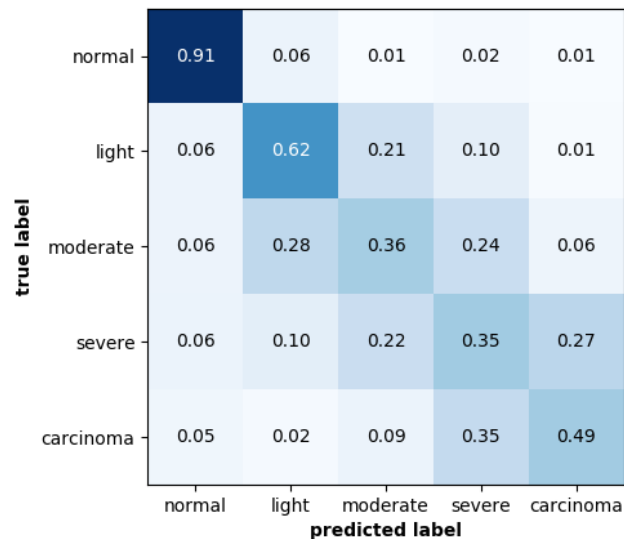


Figure 4.14: ResNet-101 classifier confusion matrix on “simulated” cytology tiles test set.

These two figures highlight that the classifier makes mistakes between *carcinoma in situ* samples and *normal* ones. Using Integrated Gradient attribution method (see Section 4.16) we show that this is once again due to *normal columnar* cells.

#### 4.2.2.2 Ordinal regression approach

In [Cheng, Wang, and Pollastri 2008], the authors present their pipeline to address ordinal regression problems. Instead of training classes one against the others, it 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 Figure 4.26). 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,

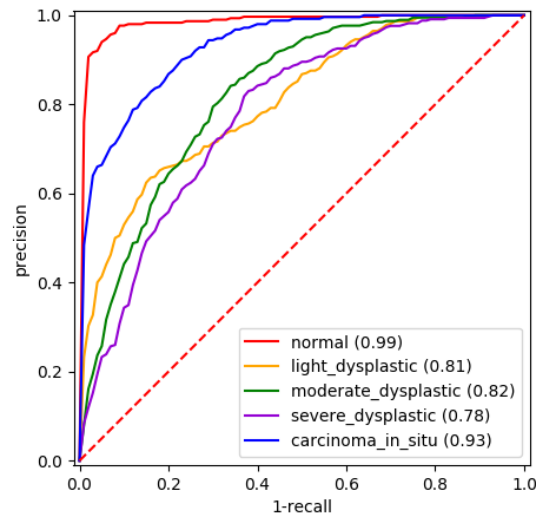


Figure 4.15: ResNet-101 classifier ROC curves and on “simulated” cytology tiles test set.

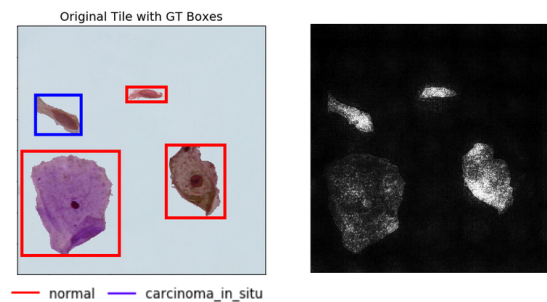


Figure 4.16: *carcinoma in situ* sample predicted as *normal* by ResNet-101 classifier and its attribution map.

1, 1, 1, 1] for *carcinoma in situ* samples.

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

Figure 4.17 shows the obtained confusion matrix. We report an average overall accuracy of 62.2%, an average binary *normal* / *abnormal* accuracy of 93.7% and an average quadratic KAPPA value of 0.83 using ordinal regression pipeline.

#### 4.2.2.3 Soft Labels for Ordinal Regression Pipeline Results

Another, more recent, method proposes to tackle this ordinal regression problem using “Soft Labels” [Diaz and Marathe 2019]. It simply consists in changing the 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

	normal	light	moderate	severe	carcinoma
normal	0.88	0.07	0.03	0.02	0.00
light	0.05	0.58	0.28	0.09	0.00
moderate	0.01	0.24	0.47	0.27	0.01
severe	0.01	0.04	0.19	0.56	0.19
carcinoma	0.01	0.00	0.05	0.32	0.62

Figure 4.17: ResNet-101 ordinal pipeline confusion matrix on “simulated” cytology tiles test set.

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. As an example, instead of having  $[0, 0, 1, 0]$  for class 3, we have a vector  $d$ , containing the opposite of distances, equal to  $[-2, -1, 0, -1]$  and then the ground truth label is  $[0.0724, 0.1966, 0.5344, 0.1966]$  (see Figure 4.26).

We train a ResNet-101 with the “Soft Labels” pipeline on the simulated tile dataset (same random 5 folds). Figure 4.18 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 the “Soft Labels” pipeline. This approach statistically improves the ordinal regression approach.

#### 4.2.2.4 Classifier under regression constraint

We consider again our classification under regression constraint method for the problem of classification of tiles. Figure 4.19 illustrates the method explained in Eq. 4.1 and 4.2. Note that the regression constraint weights are set to be linear (e.g.  $[1, 2, 3, 4, 5]$ ).

Figure 4.20 shows the confusion matrix which highlights that most samples are well classified and that, once again, as for Herlev cells, we avoid predictions mistakes between *normal* and *carcinoma in situ* tiles. It yields an accuracy of 66.8%. Figure 4.21 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



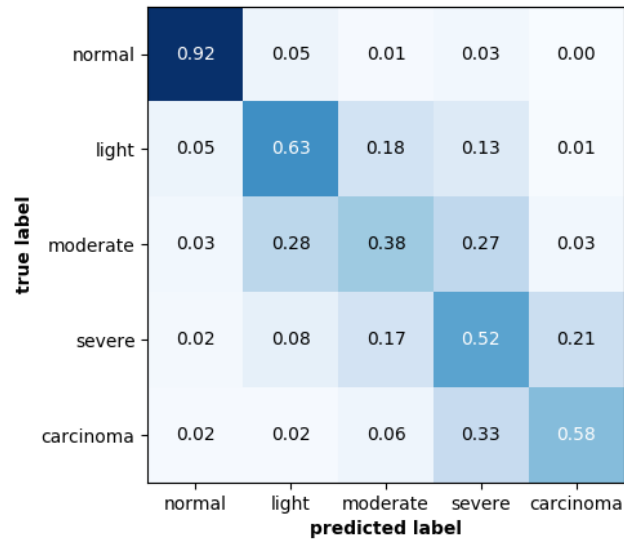


Figure 4.18: ResNet-101 “Soft Labels” pipeline average confusion matrix on simulated cytology tile test set over 5 random folds.

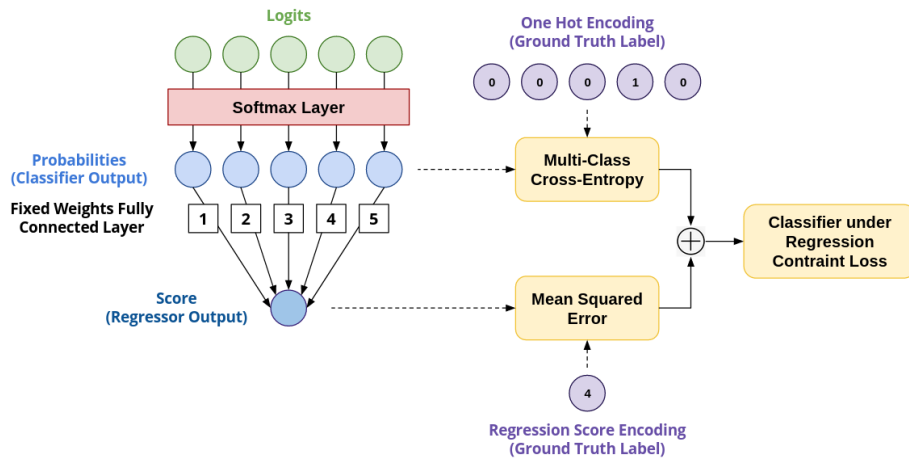


Figure 4.19: Illustration of classifier with regression constraint architecture and losses.

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 Figure 4.22. Indeed, we expect to have many more negative samples than positive samples in a real cases. 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

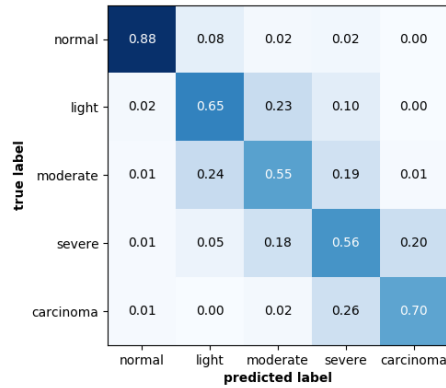


Figure 4.20: ResNet-101 {classifier + regressor} confusion matrix on “simulated” cytology tiles test set.

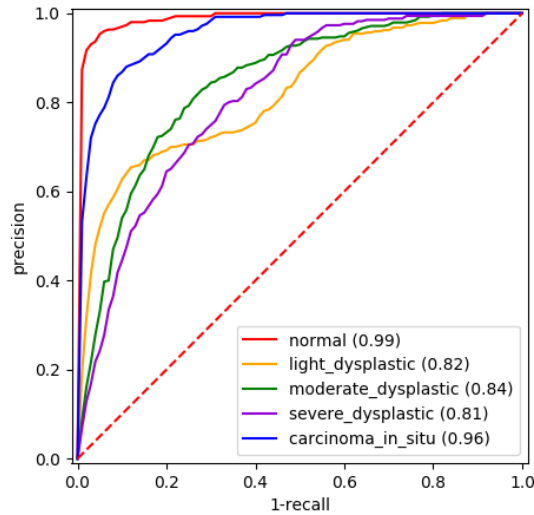


Figure 4.21: ResNet-101 {classifier + regressor} ROC curves and on “simulated” cytology tiles test set.

extend this discussion in Section 4.3, showing that we do have false positive samples but in an acceptable proportion.

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 Figure 4.23, that the error is due to a *normal* cell (and more precisely the *normal columnar* one at the top right of the image) while the {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 discriminative and relevant features.

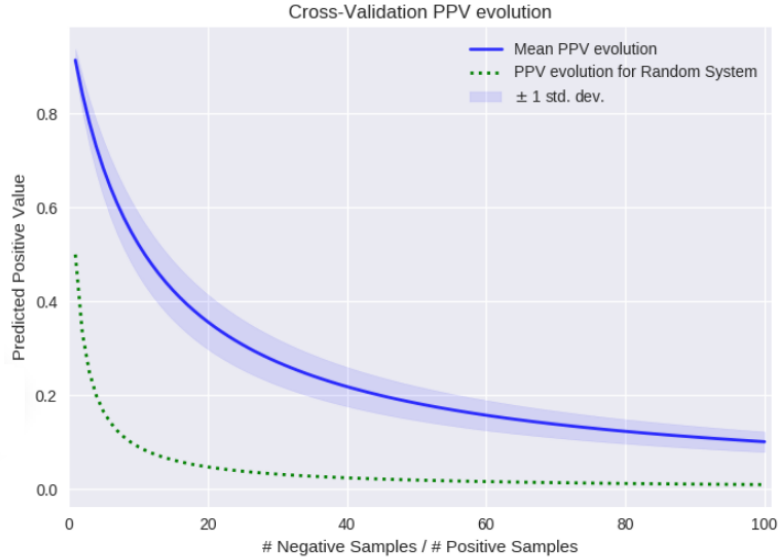


Figure 4.22: PPV evolution w.r.t. ratio between negative samples and positive samples.

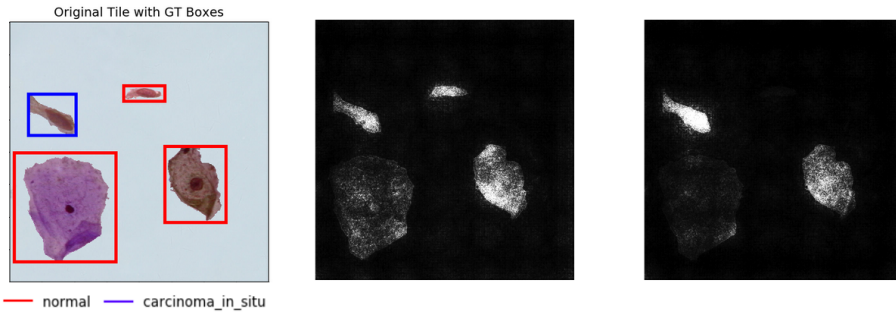


Figure 4.23: Image (left) and attribution map for a *carcinoma in situ* sample that has been classified as *normal* by classifier (middle) and as *carcinoma in situ* by {classifier + regressor} (right).

#### 4.2.2.5 Classifier under regression constraint with sensitivity focus

As explained before, 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.

Figure 4.24 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 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.

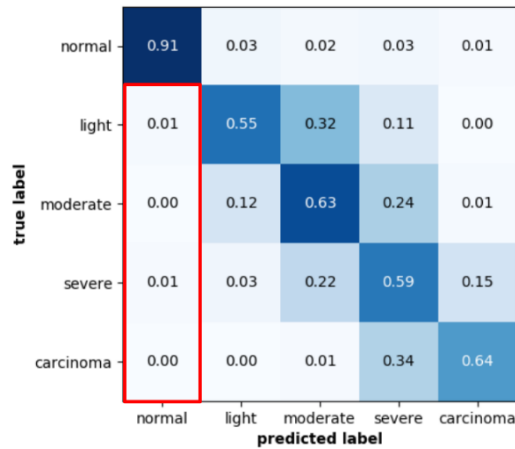


Figure 4.24: ResNet-101 {classifier + regressor} with sensitivity focus confusion matrix on “simulated” cytology tiles test set.

#### 4.2.2.6 Comparison of pipelines

Figure 4.26 illustrates pipelines to which our regression constraint method is compared and Figure 4.25 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] shows 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.

### 4.2.3 Interpretability

Now that we have a classifier (the {Classifier + Regressor} Pipeline one) that works well on our “Simulated” cytology tiles dataset, we will check that our model relies on the right

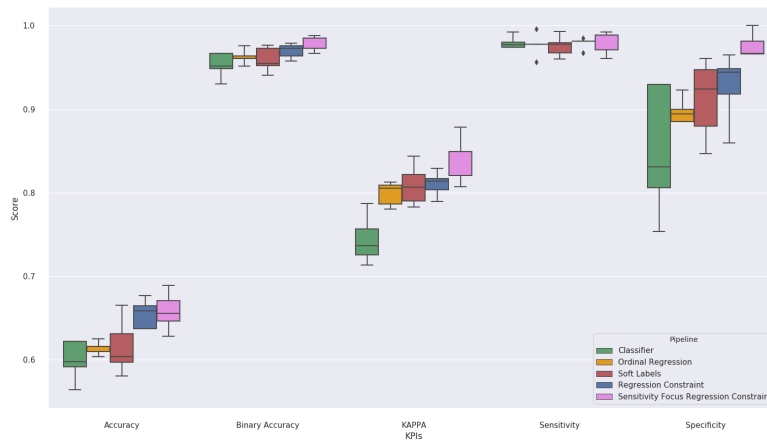


Figure 4.25: Overall accuracy, binary accuracy, KAPPA, sensitivity and specificity distributions for each pipeline.

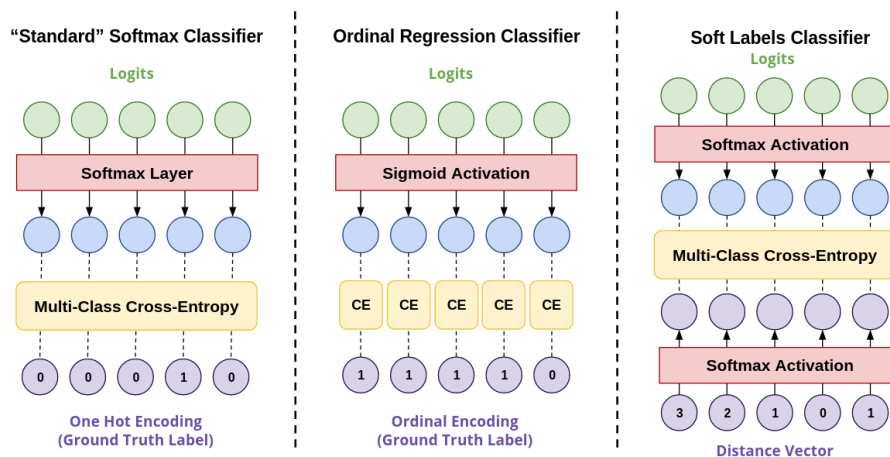


Figure 4.26: Illustration of classifier, ordinal regression and Soft labels architectures and losses.

cell(s) to make its decision by using the Integrated Gradient method presented previously. 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* so it is a good baseline for severity attribution.

Figure 4.27 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 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.

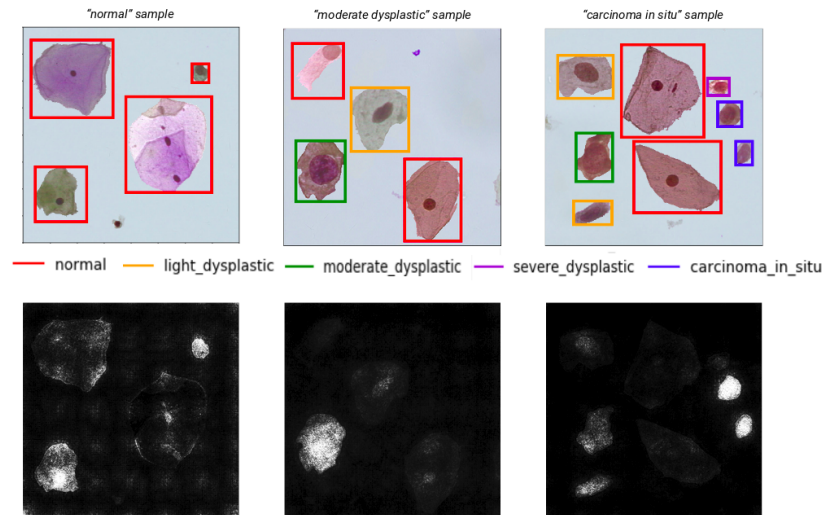


Figure 4.27: Simulated tile examples (with colored ground truth cell boxes) and the associated attribution maps w.r.t. to the predicted class.

#### 4.2.4 Weakly supervised localization

In the previous paragraph, attribution maps have proved to be useful for the interpretability of what has been learned by the model. They also hint the possibility to be used for explanatory localization. In this 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.

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 Figure 4.28.

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. Figure 4.29 shows the resulting localization boxes associated with the global label prediction.

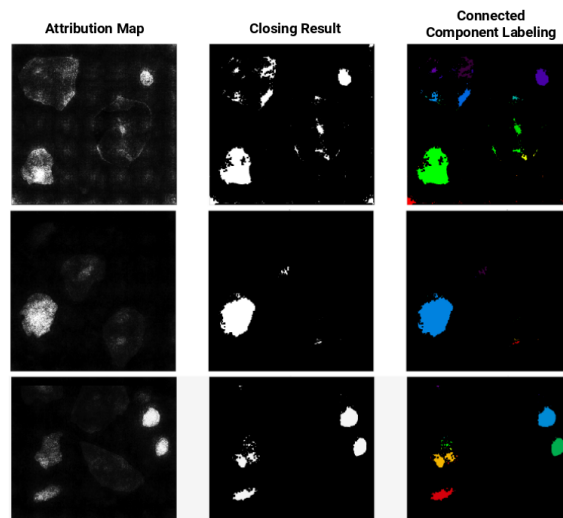


Figure 4.28: Process to localize most contributing cell from attribution map.

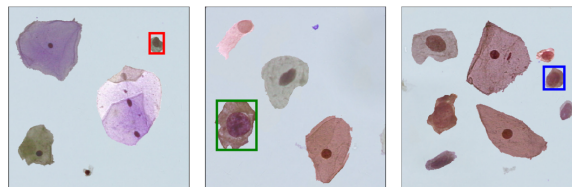


Figure 4.29: Weakly supervised localization on “simulated” tiles examples.

The resulting weakly supervised localization accuracy is 80.4%.

#### 4.2.5 Weakly supervised abnormal cell detection

We showed that we were able to localize 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 to predict. For example, in Figure 4.27 (right) the model predicted correctly the class *carcinoma 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.

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%. Deriving from this confusion matrix, we also report a sensitivity of 57.1% and a specificity of 80.6%. Fig 4.30 shows some test images, their severity attribution map and the detection associated. Additionally (and maybe even more essentially), we claim that in all cases where *abnormal* cells are present, we detect at least one which ensures medical support efficiency.

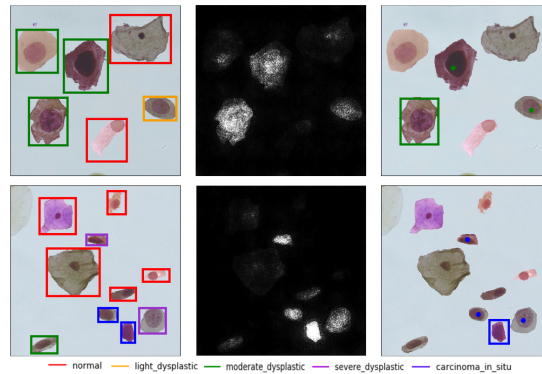


Figure 4.30: Weakly supervised abnormal cells detection examples.

### 4.3 Liquid-based cytology whole slide image 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 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 split with regards to slides). We balance the train set regarding classes by sampling more frequently “abnormal” samples that are under-represented in our dataset.

Note that we considered using “simulated” cytology tiles to increase the size of the training set but the non-overlapping of cells in these tiles simplifies the decision and would not transfer



to tiles extracted from slides which would result in a loss of performances.

Figure 4.31 shows the confusion matrix obtained for 10X magnification on test images. It shows an accuracy of 97.4%, a sensitivity of 89.1% and a specificity of 99.7%. We also report a KAPPA measure of 0.812 and an AUC of 0.991.

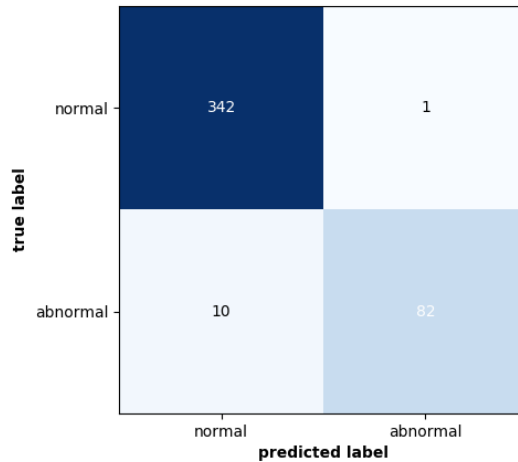


Figure 4.31: ResNet-101 classifier confusion matrix on real cytology 10X tiles test set.

Using integrated gradient, we computed attribution maps and applied the post-processing described previously 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 Figure 4.32).

This localization accuracy is quite satisfactory regarding the localization which is pretty complicated. Indeed, there are generally around 15 cells per 10X tile, 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, from our point of view, even when the localization is wrong (see second example in Figure 4.32), it still captures rather interesting cells (dark blue cell with high NCR). Note that this localization accuracy could be improved using a Herlev cell classifier to validate the “abnormality” of identified cells.

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 much heavier 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

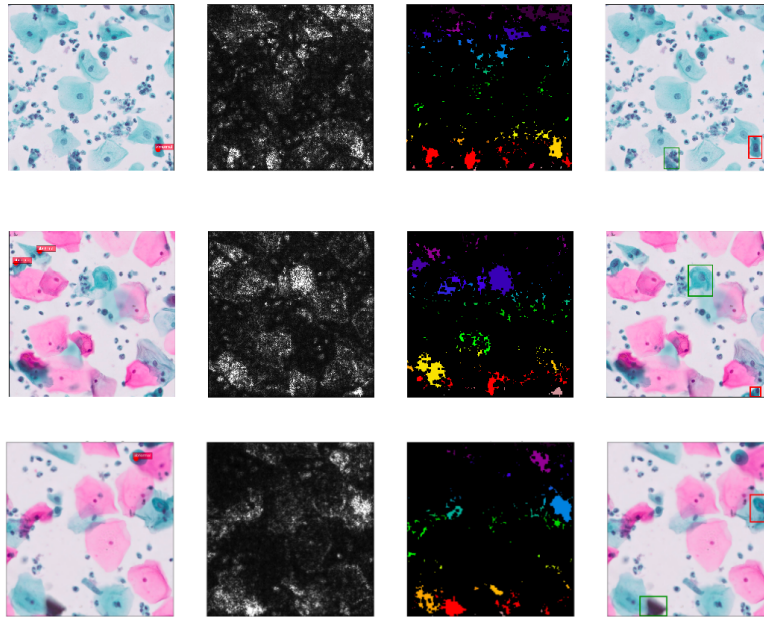


Figure 4.32: Example of weakly supervised localization on real cytology 10X tiles; Images and ground truth annotations (left); Integrated gradients results (middle); Images and localization results (right).

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 Figure 4.33. 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 that there is a compromise to make between precision and recall performances (on the Precision-Recall curve).

### 4.3.1 Pre-processing

To validate the clinical usefulness of our work, we gathered 40 new slides for which only the 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 tiles 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 neighbors tiles that form the biggest cluster. We call this “sample selection”. This process (results in Figure 4.34) 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).

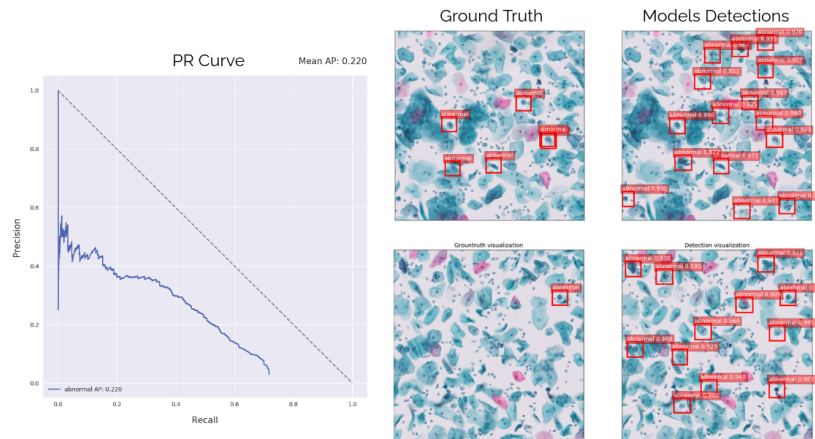


Figure 4.33: Results of Faster-RCNN object detection approach for cell detection; PR Curve (left); Images and ground truth annotations (middle); Images and detection (with abnormality score above 0.9) from trained Faster-RCNN (right).

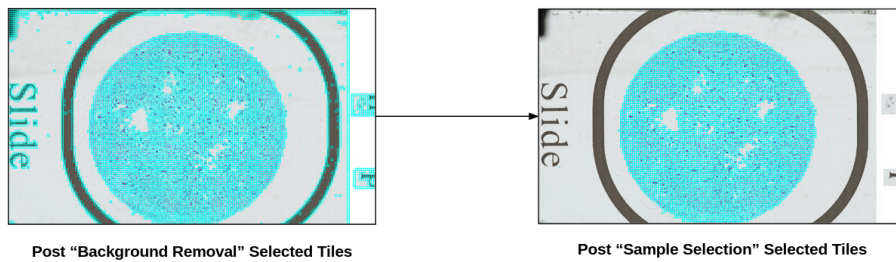


Figure 4.34: Tiles selected after sample tile selection process; respectively “background removal” and “sample selection”.

### 4.3.2 Integration in a computer-aided diagnosis pipeline

Figure 4.35 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.

Figure 4.36 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 Figure 4.37.

We can observe that 38 regions (on more than 2700 potentially before classification) have been classified as being abnormal and that cells that led to this decision have a high NCR and chromatin condensation.

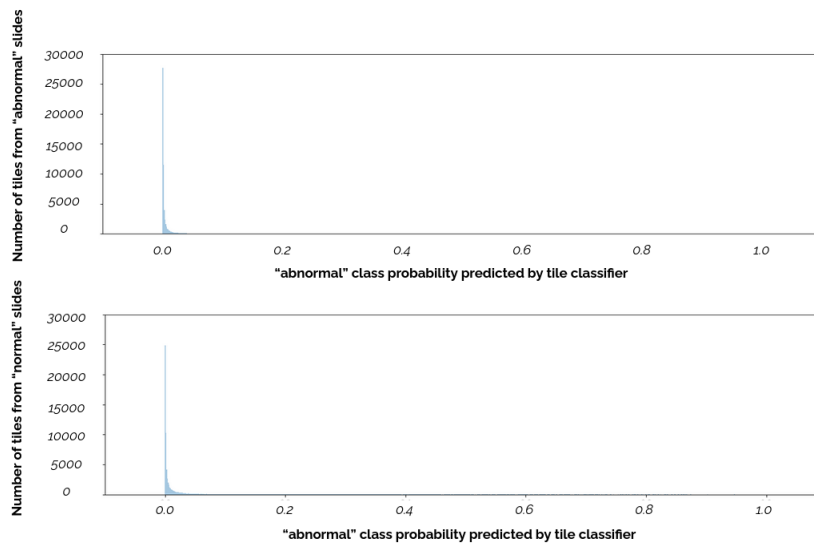


Figure 4.35: Histograms w.r.t. *abnormal* tile scores for tiles from 10 *normal* slides vs 10 *abnormal* slides.

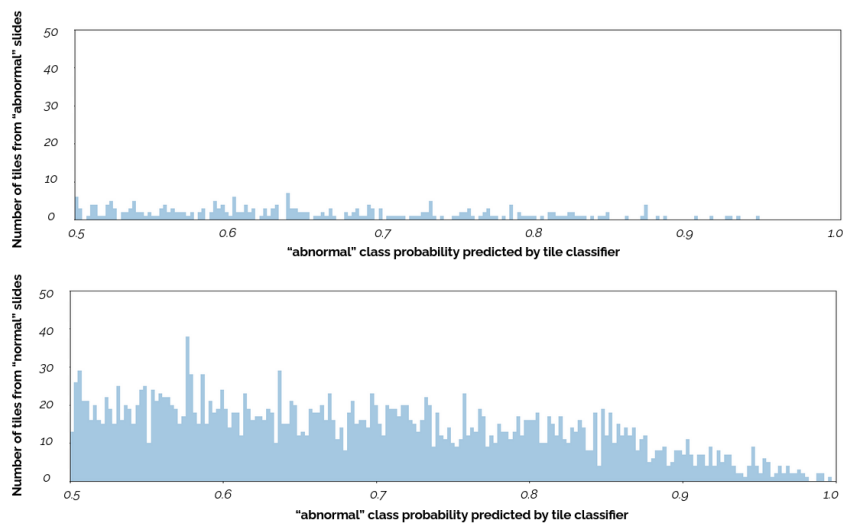


Figure 4.36: Zoom (for *abnormal* class probability above 0.5) on the histograms w.r.t. *abnormal* tile scores for tiles from 10 *normal* slides vs 10 *abnormal* slides.

For comparison, the Faster-RCNN we trained detects between 1000 and 10000 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 than 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.

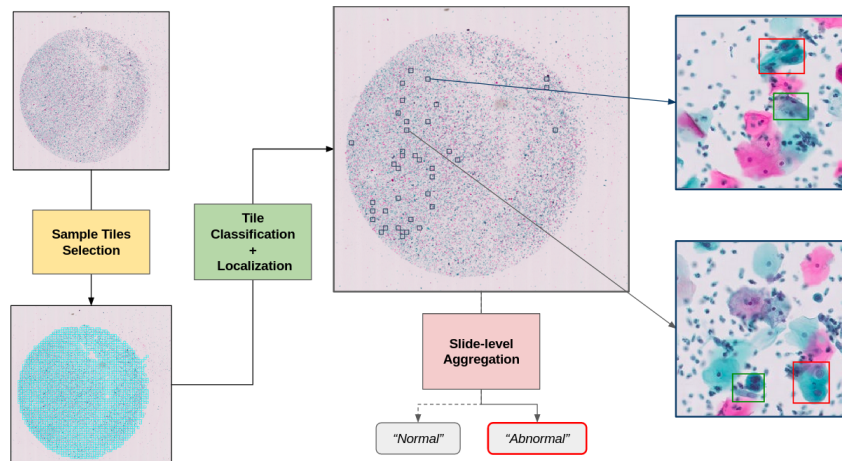


Figure 4.37: Complete pipeline and qualitative results of the proposed method for computer-aided decision.

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 much 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.

### 4.3.3 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 4.38 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 4.39 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 77.5%, a specificity of 82.3%, and a sensitivity of 73.9%. 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).

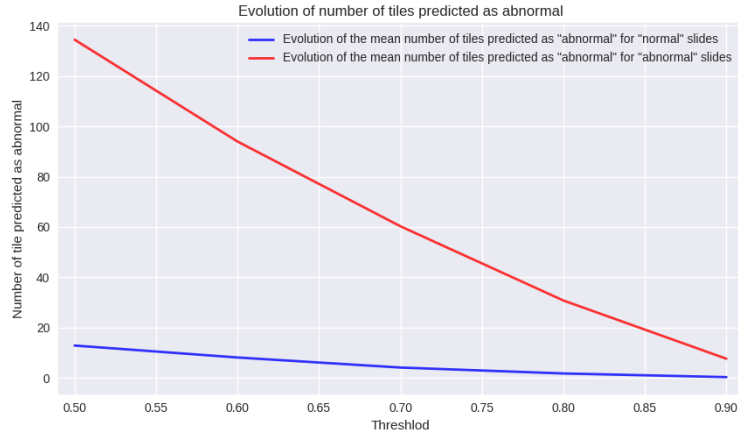


Figure 4.38: Impact of tile-level decision threshold on the number of tiles selected w.r.t. slide ground truth label.

## 4.4 Conclusion

In this chapter, we showed that our proposed method (classifier under regression constraint) can be extended to the new task of classifying tiles from cytology images. 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 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 to have 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 an efficient 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 will 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 to each individual tile.

Moreover, LBC is widely used worldwide for primary indication such as urinary or thyroid cancer screening which makes our work even more relevant medically and extendable.

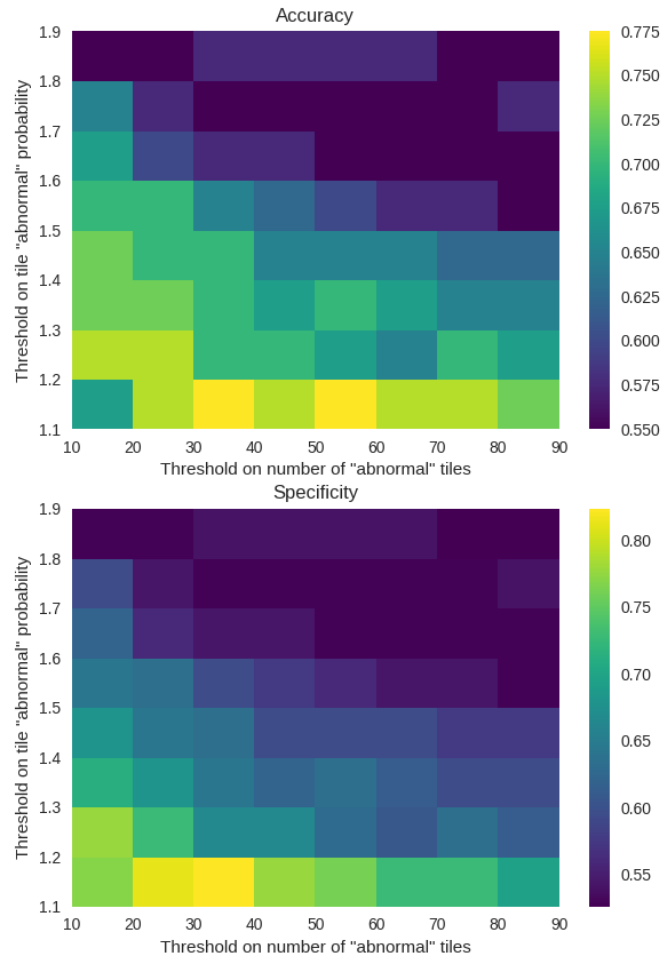


Figure 4.39: Impact of threshold on tile scores and on the number of selected tiles on the slide-level prediction.





# WSI classification: proposed methods and experiments

---

## Sommaire

<b>5.1</b>	<b>Multiple instance learning approach for whole slide classification . . .</b>	<b>76</b>
5.1.1	CHOWDER model . . . . .	76
5.1.2	Attention-based model . . . . .	77
<b>5.2</b>	<b>Improving interpretability . . . . .</b>	<b>77</b>
5.2.1	Formalization and tile scores . . . . .	77
5.2.2	Proposed method: Using gradient-based explanations . . . . .	78
5.2.3	Feature identification on trained CHOWDER and attention-based models	80
5.2.4	Tile-level explanations . . . . .	83
5.2.5	Feature-based heat-maps . . . . .	84
5.2.6	Measure of interpretability through heat-maps relevance . . . . .	84
5.2.7	Analysis of the number of features . . . . .	86
5.2.8	Colocalization filtering . . . . .	88
5.2.9	Application to the SFP Challenge . . . . .	89
<b>5.3</b>	<b>LBC slides classification . . . . .</b>	<b>93</b>
5.3.1	Medipath dataset . . . . .	93
5.3.2	MIL classical approach . . . . .	93
5.3.3	“Abnormality” detector sampling approach . . . . .	95
<b>5.4</b>	<b>Conclusion . . . . .</b>	<b>100</b>

---

As presented in Chapter 3, Multiple Instance Learning (MIL) methods proved efficient for Whole Slide Image (WSI) classification using only slide-level labels. In this chapter, after presenting two of the most popular Multiple Instance Learning (MIL) classification approaches (which we will use a lot in our experiments), we question the concept of interpretability in these architectures and propose a method that improves slide-level heat-maps by identifying features that have been learned as contributing to predictions. This approach is validated on Camelyon-16 dataset. Finally we highlight the limitations for application on Liquid-Based Cytology (LBC) datasets and provide a weakly-supervised solution that relies on the annotation of a few cells only.

## 5.1 Multiple instance learning approach for whole slide classification

As we said before, MIL context is the most popular approach for WSIs classification using only global labels (which avoids requiring medical experts to spend important amount of time drawing annotations on slides). This idea was first introduced and validated in [Courtiol et al. 2018] and [Ilse, Tomczak, and Welling 2018]. Both approaches are illustrated in Figure 5.1.

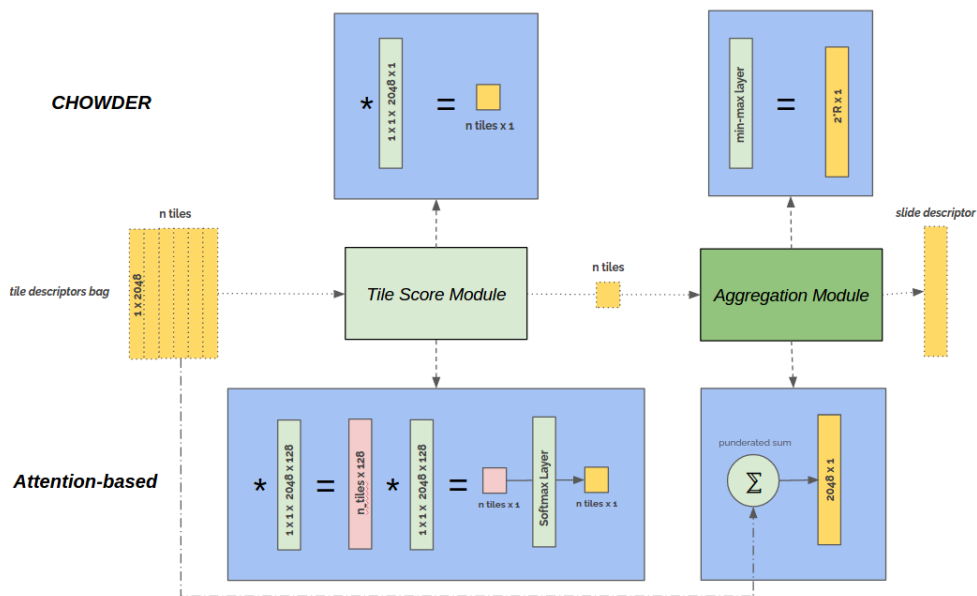


Figure 5.1: CHOWDER [Courtiol et al. 2018] and Attention-based [Ilse, Tomczak, and Welling 2018] approaches.

### 5.1.1 CHOWDER model

CHOWDER stands for Classification of HistOpathology with Weak supervision via Deep fEature aggRegation [Courtiol et al. 2018]. It consists of preprocessing steps (tissue detection, color normalization and tile-descriptors computation) that were presented in Chapter 3 to create the bag of descriptors. Then a  $1 \times 1$  convolution layer turns each tile-descriptor into a single tile score thus creating a bag of tile scores. These scores are then aggregated using a min-max layer, that keeps the top- $R$  and bottom- $R$  scores (empirically  $R = 5$  gives the best results), to give a slide descriptor (of size  $2 \times R$ ) that is fed into a two-layers fully connected network that proposes the diagnosis.

### 5.1.2 Attention-based model

The Attention-based whole slide classifier [Ilse, Tomczak, and Welling 2018] uses an attention module (two  $1 \times 1$  convolution layers with respectively 128 and 1 channels, and a softmax layer) to compute competitive and normalized (sum to 1) tile scores from tile descriptors. Then, the slide descriptor is computed as the weighted (by tile scores) sum of tile descriptors and, as in CHOWDER, given to a two-layers fully connected network that predicts the classes probabilities.

## 5.2 Improving interpretability

As introduced in Chapter 3, the great advantage of these architectures, in addition to being trainable and efficient with very few supervision, is that they are thought to mimic the workflow of pathologists, which makes the result interpretable. However explanations are relying on a single “medical” score which might limit the interpretability regarding complex tissue structures that can be found on these slides.

### 5.2.1 Formalization and tile scores

This inspiration from pathologist’s workflow to classify a slide (analyzing the whole slide at a high magnification level, identifying informative regions and making a decision based on these regions) makes most pipelines entering a common framework.

We propose to formalize this common design here.

Let  $i$  be the slide index. The slide is divided into tiles w.r.t. a non overlapping grid after a tissue detection relying on Otsu segmentation [Otsu 1979] (see Figure 5.2), and. Thus we obtain a bag of tiles. Let  $j$  be the tile index for each slide.

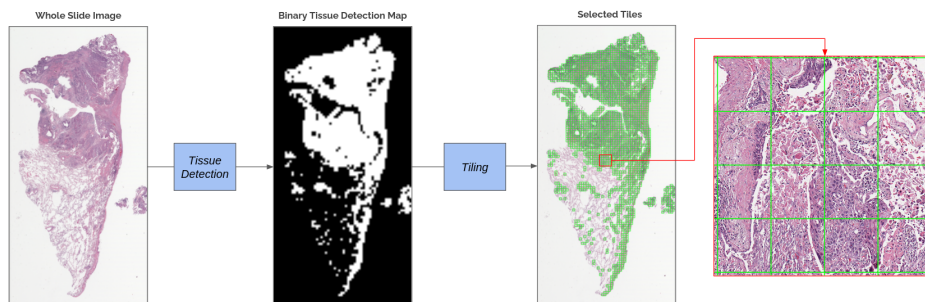


Figure 5.2: Illustration of tissue detection and tiling processes.

There are four distinct blocks in a typical WSI classification architecture:

1. A feature extractor module  $f_e$  (typically a CNN architecture) that encodes each tile

$x_{i,j}$  into a descriptor  $d_{i,j} \in \mathbb{R}^N$  with  $N$  the descriptor size (depending on the feature extractor):  $d_{i,j} = f_e(x_{i,j})$ ;

Note that this block is part of pre-processing steps enabling to encode the slide into a bag of tile descriptors.

2. A tile scoring module  $f_s$  that, based on each tile descriptor  $d_{i,j}$ , assigns a single score per tile  $s_{i,j} \in \mathbb{R}$ :  $s_{i,j} = f_s(d_{i,j})$ ;
3. An aggregation module  $f_a$  that, based on all tile scores  $s_{i,j}$ , and sometimes their tile descriptors  $d_{i,j}$ , computes a slide descriptor  $D_i \in \mathbb{R}^M$  with  $M$  the slide descriptor size (depending on the aggregation module):  $D_i = f_a(s_{i,j}, d_{i,j})$ ;
4. A decision module  $f_{cls}$  that, based on the slide descriptor  $D_i$ , makes a class prediction  $P_i \in \mathbb{R}^C$  with  $C$  the number of classes:  $P_i = f_{cls}(D_i)$ .

Heat-maps based on tile scores have been proven to be really efficient to the point of being able to spot cancerous lesions that had been missed by experts (in [Campanella et al. 2019]).

Figure 5.3 illustrates this design and introduces our contributions for improving interpretability.

### 5.2.2 Proposed method: Using gradient-based explanations

Our approach (illustrated in 5.3) consists in rewinding explanations from the decision module to tile information by applying interpretability methods and by answering successively the following three questions:

1. Which features of slide descriptors are relevant for a class prediction?
2. With regard to the aggregation module, which features of tile descriptors are responsible for previously identified relevant slide descriptor features?
3. Are these features of tile descriptors relevant medically and representative of histopathological information?

The first question is answered using attribution vector  $A_c \in \mathbb{R}^M$  (one for each class  $c$ ) computed as the gradient of the component of index  $c$  of  $P_i$  (noted  $P_{i,c}$ ) with respect to  $D_i$ . It enables us to identify a set of relevant positions (corresponding to features extracted)  $K_c = \{K_{c,1}, \dots, K_{c,L}\}$  in slide descriptors, i.e. the  $L$  (empirically determined) positions in  $A_c$  with highest attributions over the slide predicted in class  $c$ . Each attribution  $A_{c,m}$  at position  $m \in [0; M]$  of vector  $A_c$  is computed as:

$$A_{c,m} = \sum_{i \in I_c} \left| \frac{\partial P_{i,c}}{\partial D_{i,m}} \right| = \sum_{i \in I_c} \left| \frac{\partial f_{cls}(D_i)_c}{\partial D_{i,m}} \right|,$$

with  $I_c$  the set of slides predicted to be in class  $c$  and  $|\cdot|$  the absolute value.

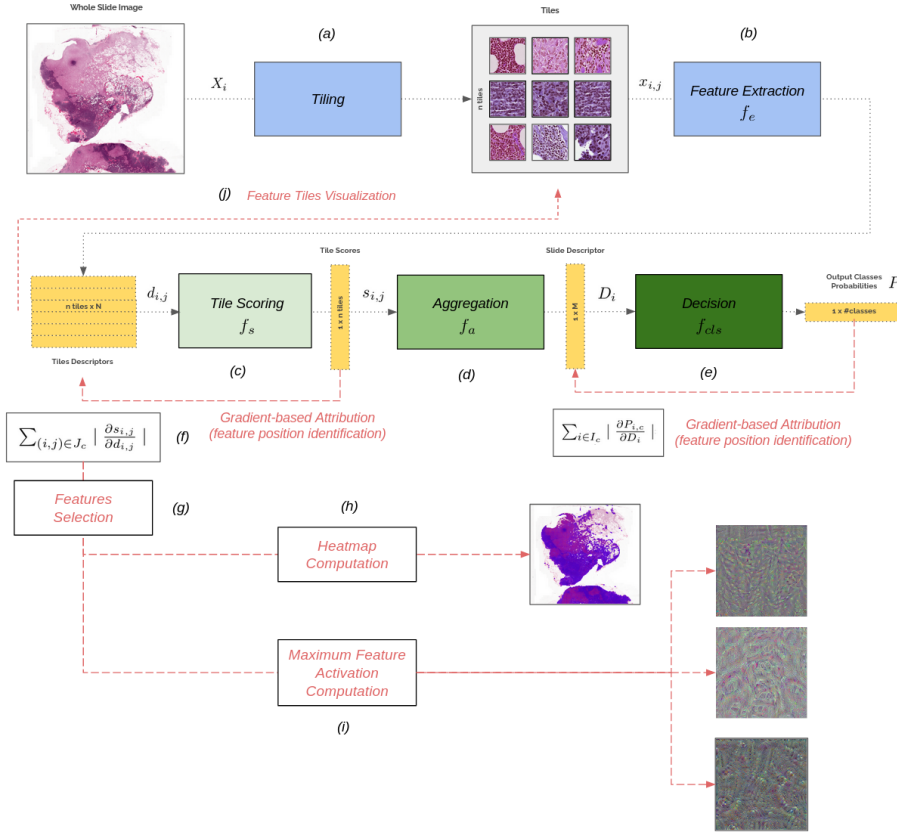


Figure 5.3: Overview of the proposed method. WSI Classification: (a) Tiling from the slide; (b) Features extraction from the tiles; (c) Tile scoring from tile descriptors; (d) Aggregation; (e) Decision from the slide descriptor; Interpretability (in red): (f) Feature identification from gradient-based attributions; (g) Feature selection from feature colocalization; (h) Heat-map computation from activation of selected features; (i) Individual feature visualization through gradient ascent; (j) Individual feature visualization through tile feature activation.

Then, the second question is also answered using an attribution vector  $a_c \in \mathbb{R}^N$  computed as the gradient of tile score  $s_{i,j}$  with respect to tile descriptor  $d_{i,j}$ . This enables to identify feature positions  $k_c = \{k_{c,1}, \dots, k_{c,l}\}$  in tile descriptors, i.e. the  $l$  (empirically determined) tile descriptors that are responsible for high activation at previously identified  $K_c$  positions in slide descriptor. Each attribution  $a_{c,n}$  at position  $n \in [0; N]$  of vector  $a_c$  is computed as:

$$a_{c,n} = \sum_{(i,j) \in J_c} \left| \frac{\partial s_{i,j}}{\partial d_{i,j,n}} \right| = \sum_{(i,j) \in J_c} \left| \frac{\partial f_s(d_{i,j,n})}{\partial d_{i,j,n}} \right|$$

with  $J_c$  the set of tile positions  $(i, j)$  that most activate  $K_c$  positions in slide descriptors (threshold empirically determined, explained in the next subsection).

### 5.2.3 Feature identification on trained CHOWDER and attention-based models

**Dataset and Preprocessing.** We validate our approach using Camelyon-16 dataset that contains 345 WSI divided into 209 “normal” cases and 136 “tumor” cases. This dataset contains slides digitized at 40X magnification from which we perform sample detection using Otsu thresholding [Otsu 1979] on a thumbnail of the slide downscaled by a factor 32 and keeping tiles that contain at least 50% of foreground pixels w.r.t. Otsu segmentation. Then, we extract, with regard to a non-overlapping grid,  $224 \times 224$  pixels tiles at 20X magnification without stain normalization. Then, we pre-compute, for each tile, 2048-tile descriptors using a ResNet-50 model trained on ImageNet as it is done in [Courtiol et al. 2018; Campanella et al. 2019; Naylor et al. 2019; Lu et al. 2020; Campanella, Silva, and Fuchs 2018; Li, Li, and Eliceiri 2020]. 216 slides are used to train our models while 129 slides form the test set to evaluate performances of the different trained models.

After training, both models perform similarly at slide-level classification with an AUC of 0.82 for CHOWDER and 0.83 for attention-based.

**Results on CHOWDER model.** Let us now illustrate and detail the results of our approach on the CHOWDER model guided by the three questions raised in the previous subsection.

The first question is “Which slide descriptors features are relevant for a class prediction?” i.e. for CHOWDER, given the  $M=10$  ( $R=5$ ) tile scores given as slide descriptor (the 5 minimum tile scores and the 5 maximum tile scores), what is the contribution of each of these values to the prediction?

Figure 5.4 shows, as histograms, the distribution of the (5-)min and (5-)max scores w.r.t. predictions over the whole 129 test slides, and highlights that min scores are the ones that contribute to discriminate between the two classes (i.e. the lower min scores, the more the slide is predicted as being “tumor”). A Mann-Whitney U-Test between scores (min and max independently) distributions reveals that min scores distributions per predicted class are statistically different ( $p < 10^{-3}$ ) while max scores are not ( $p = 0.23$ ). The attribution of min and max scores distributions validates this assertion by showing a statistically higher attribution on min tile scores than on max tile scores.

After finding that min scores are the ones describing tumorous regions and thus that max scores are used for the “normal” class, we are interested in identifying which features of tile descriptors are mostly responsible for minimum and maximum scores, i.e. to describe each class. To address this second question, we use the same gradient-based explanation method on tile scoring module.

Most minimal tile scores are under -5 and most maximal tile scores are above 11. For each of these groups of tiles, we compute the average attribution of each of the  $N=2048$  features in tile descriptors (extracted by a ResNet-50 trained on ImageNet). Figure 5.5 shows the distribution of features hence activated and allows us to identify which features are mostly

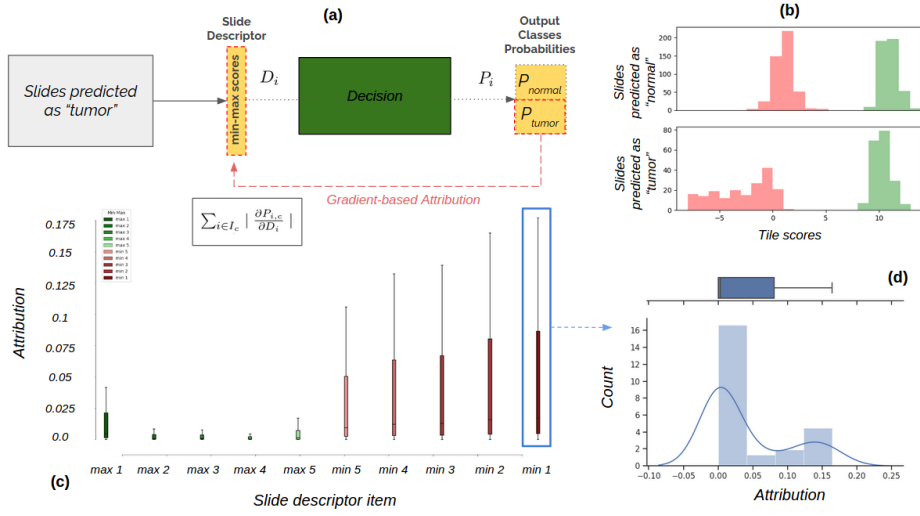


Figure 5.4: Slide descriptor attribution for the “tumor” class; (a) Illustration of gradient on the learned “Decision” block; (b) Distributions of min and max scores w.r.t. the predicted class; (c) Distributions of attribution on each slide descriptor items (5-min scores and 5 max-scores); (d) Detail of the distribution of attribution for min 1 (lowest) tile score item and bimodal Gaussian approximation.

responsible for min and max tile scores, i.e. highest attribution for min and max scored tiles.

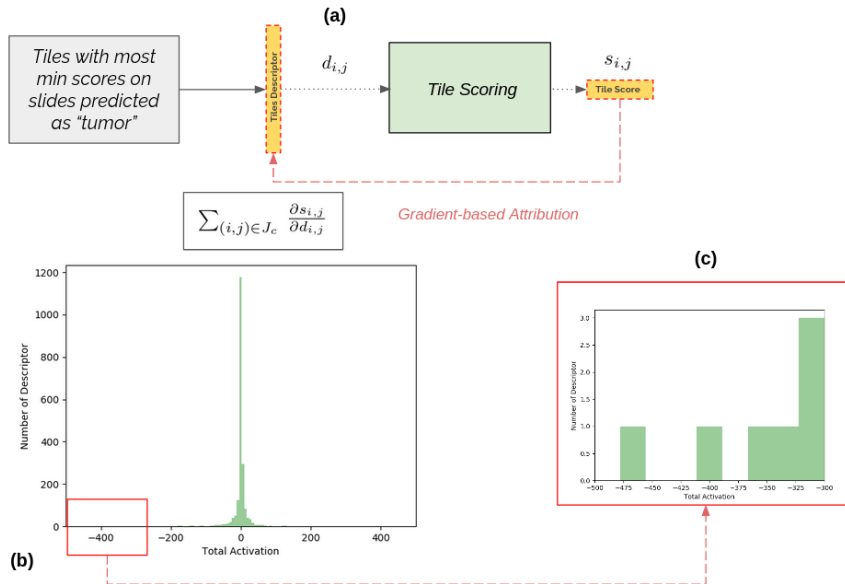


Figure 5.5: Attribution on tile descriptor items for the “tumor” class; (a) Illustration of gradient on the learned “Tile Scoring” block; (b) Distributions of attributions for tile descriptors with lowest score; (c) Zoom on the range of interest.

Thus we are able to claim that features (defined by their position in the descriptor) that

are mostly useful for the trained model for the “tumor” class are 242, 420, 602, 1154, 1644, 1652 and 1866. Following the same process, we identified 565, 628, 647, 1158 and 1247 as being the most contributing features for the “normal” class according to CHOWDER model.

**Results on Attention-based model.** Here, tile scores are used to weight how much each tile is contributing to describe the slide w.r.t. the medical task the model has been trained on. As we understand that high tile scores should put forward tile descriptors that activate relevant features for the diagnosis, we also understand that, if the attention module makes its job well, relevant features should be used by both attention module and decision module. Thus, we propose to select features that have a high attribution in both tile descriptors and slide descriptors..

Using gradient-based attribution, we compute the histogram of average attribution over the 2048 features of both slide descriptors and high scored tile descriptors per class (respectively w.r.t. the class prediction made and the tile score predicted). Figure 5.6 shows the selection of features for “tumor” class i.e. attribution of slide and high (above 0.1) tile descriptors for slides predicted as “tumor”.

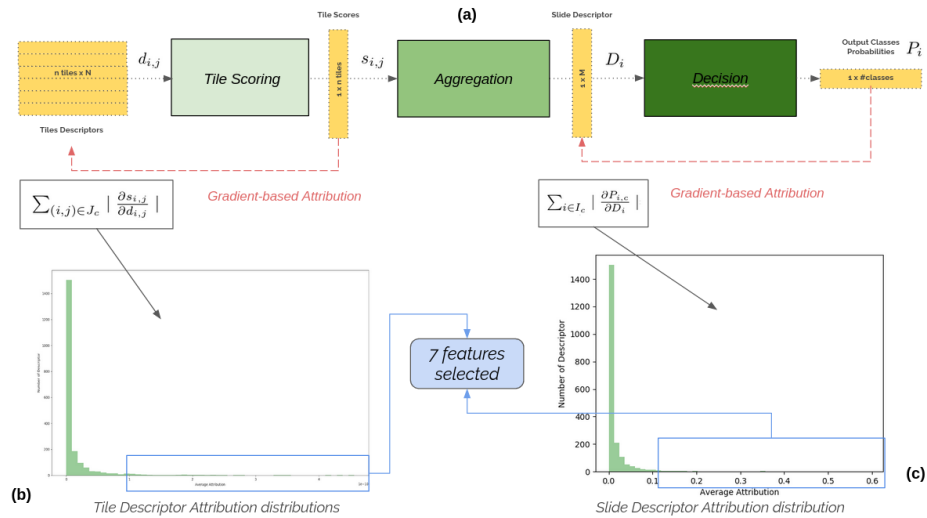


Figure 5.6: Feature selection for the “tumor” class using the attention-based model. (a) Illustration of the model and gradient-based explanation computation; (b) Distributions of attributions for tile descriptors of high scored tiles; (c) Distributions of attributions for slide descriptors of slides predicted as “tumor”.

This process once again enables us to select the 7 features identified as being the most useful for “tumor” class prediction (position 242, 529, 602, 647, 762, 873 and 1543) and 5 features for the “normal” class (position 672, 762, 1151, 1644 and 1676).



### 5.2.4 Tile-level explanations

As exposed in the previous paragraph, based on explanations on decision blocks, we have been able to identify 7 and 5 features that are mostly used by the trained CHOWDER model to make decisions (and we did the same for the attention-based model). Now, we are interested in interpretable information to return to pathologists so that they can use their expertise to understand what these features put forward histopathologically speaking. We benefited from discussions with two experienced pathologists and report their overall feedback on the interpretable visualization we proposed.

To answer the third question put forward in Section 5.2.2, we rely on feature activation to highlight features identified as being discriminative to the task by selecting tiles  $x_{i,j}$  that have the highest activation per feature in  $k_c$  identified over the whole test set. Along with these tiles, we display, for each position in tile descriptors  $k \in k_c$ , a maximum activation  $\mathcal{X}^k$  image obtained by iteratively tuning pixels values to activate the feature by gradient ascent as follows:  $\mathcal{X}^k$  is initialized as a uniformly distributed noise image  $\mathcal{X}_0^k$ ; then while  $f_e(\mathcal{X}_{n-1}^k)_k$  (activation at position  $k$ ) increases, iterate over  $n > 0$ :

$$\mathcal{X}_n^k = \mathcal{X}_{n-1}^k + \frac{\partial f_e(\mathcal{X}_{n-1}^k)_k}{\partial \mathcal{X}_{n-1}^k}.$$

Figure 5.7 shows the 7 tiles that activate the most (over all tiles) each feature and the max activation image, that we expect to reveal what the feature means with regards to the histopathological problem it has been trained on.

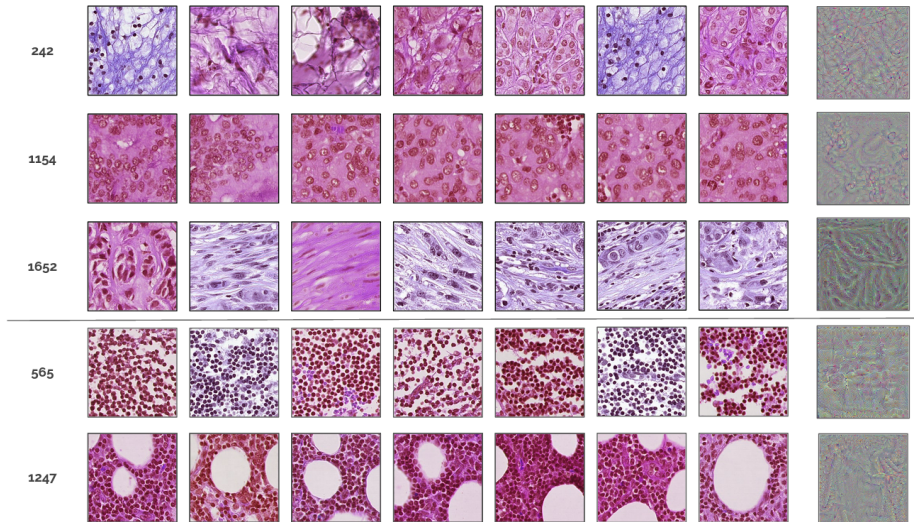


Figure 5.7: Patch-based visualizations obtained for features 242, 1154 and 1652 (for min-scores features); 565 and 1247 (for max-scores features); tiles and max activation images (right).

Pathologists agreed that patch-based tiles visualizations are highly interpretable and exhibit features that are indeed related to each class [Hoon Tan et al. 2019]. For example, feature 1652 tends to trigger spindle-shaped cells that indeed can be a metastatic tissue or-

ganization. For “normal” tissue features, feature 565 describes mainly clustered lymphocytes that are preponderant in normal tissues.

### 5.2.5 Feature-based heat-maps

Furthermore, we also propose a new way to compute heat-maps for each slide  $i$ . We note  $H_{c,i}$  the map that highlights regions on slide  $i$  that explain what has been learned to describe class  $c$  based on the identified features. For each slide  $i$  and tile  $j$ , the heat-map value  $H_{c,i,j}$  is computed as the average of activations  $d_{i,j,k}$  (normalized per feature over all tiles of all slides) over identified features  $k$  in  $k_c$  for class  $c$ :

$$H_{c,i,j} = \frac{1}{|k_c|} \cdot \sum_{k \in k_c} \frac{d_{i,j,k} - \min_k}{\max_k - \min_k}$$

with  $\max_k = \max_{i,j}(d_{i,j,k})$  and  $\min_k = \min_{i,j}(d_{i,j,k})$ .

This heat-map values (between 0 and 1) can be considered as a prediction scoring system, and thus we propose to compute the Area Under the ROC (Receiver Operating Characteristic) Curve to measure how relevant is the interpretability brought by our automatic feature extraction approach using ground truth lesion annotations when given. This localization AUC measures the separability between the class of interest (e.g. “tumor”) and other classes using heat-maps. Indeed for a good heat-map we expect all tiles that are representative of the class of interest to have a high score and all other tiles to have a low score.

### 5.2.6 Measure of interpretability through heat-maps relevance

We also validate results obtained with localization AUC by performing a ROAR analysis adapted to MIL context. Indeed, good heat-maps put forward discriminative tiles, thus removing these tiles from bags should prevent the model to learn. In this context, we propose to gradually (by thresholding the tile scores) remove tiles with a high score and to train a model with these new reduced bags. If heat-maps are relevant (i.e. if highlighted tiles represent the class of interest) and complete (i.e. if tiles representing the class of interest all have high tile scores) then slide classification performances should drop, while if heat-maps are not relevant or not complete the performances should remain stable through training.

Further and deeper analysis on the impact of the number of features selected on the quality of generated heat-maps presented in the next section enabled us to propose an additional *feature selection* block (in Figure 5.3) to filter out the selected outliers. We will present this method after motivating it by our results.

The coherence between patches extracted for a better interpretability led us to think about another way to present features to pathologists. Indeed, since tissues have a coherent and somehow organized structure, a relevant feature for histological problems would be activated in a coherent and somehow organized way over slides. Thus, along with patch-based visualization, we propose to access feature activation heat-maps  $H_{c,i}$  over slides, as presented in

Section 5.2.5.

Figure 5.8 illustrates qualitative results, and highlights how our feature-based heat-maps enable to extensively put forward “tumor” regions and ignore “normal” tissue. Quantitatively, we report a tile-level localization AUC of 0.884 for CHOWDER model and 0.739 for Attention-based model, using feature-based heat-map values (that are the average normalized feature activation over all features identified for the “tumor” class, see  $H_{c,i,j}$  computation in Section 5.2.2) as a “tumor” prediction score and using lesion annotation provided by Camelyon-16 dataset to get the ground-truth label per tile. Both AUCs are significantly high, which validates our approach of identifying features that are relevant and of computing heat-maps for interpretation and explanation. Note that the AUC computed using tile scores is 0.684 for CHOWDER model and 0.421 for Attention-based model (see Table 5.1). We can also note that there is a gap in interpretability between CHOWDER model and Attention-based model while classification performances are similar (AUC of 0.82 for the CHOWDER model and 0.83 for the Attention-based model). The gap can be explained by the fact that, in the context of Camelyon-16, identifying one tumorous tile is enough to label a slide as “tumor”, so implicitly the tile classification does not need to be exhaustive to provide meaningful information to the slide level decision module.

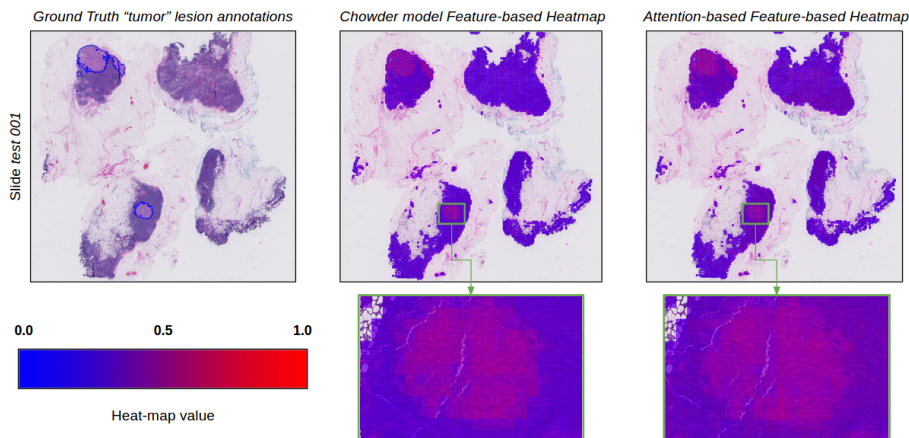


Figure 5.8: Slide-based visualizations: Heat-maps explaining the “tumor” class obtained by computing average normalized activation over identified features; ground-truth annotations for “tumor” tissue (left); CHOWDER model feature-based heat-maps (middle); attention-based model feature-based heat-maps (right).

Table 5.1: Results: classification and localization AUC using both methods (improvement of localization AUC by 0.200 for CHOWDER and 0.318 for Attention-based model).

Model	Classification AUC	Heat-map method	Localization AUC
CHOWDER	0.82	Tile scores	0.684
		Feature-based (ours)	<b>0.884</b>
Attention-based	<b>0.83</b>	Tile scores	0.421
		Feature-based (ours)	0.739

We also validate the better explanation given by our feature-based heat-maps with a ROAR approach adapted to the MIL context. In the context of explanation heat-maps, we expect hot colors regions (i.e. with high scored tiles) to be informative and cold colors regions to be non-informative. So we propose to remove tiles with an increasing threshold and to retrain from scratch (still pre-training from ImageNet) a model that we evaluate. Thus, for a complete and relevant heat-map method the performances should dramatically drop as high scored tiles, which would be the informative tiles, are removed. By contrast, for irrelevant or incomplete heat-maps, performances should remain unchanged since informative tiles are still available for learning (for that we included a control experiment consisting of randomly distributed tile scores).

Figure 5.9 shows the performances of models retrained after the removal of tiles with different thresholds on the heat-maps obtained from the trained model on the full bags. We can observe that our feature-based heat-maps are the ones impacting the most the performances, which confirms the results in Table 5.1. Also it confirms that CHOWDER tile score heat-maps are complete and relevant while attention-based tile scores heat-maps are equivalent to random heat-maps due to the important number of positive tiles being scored with a low score by the attention module.

Note that attention-based tile scores are not irrelevant but not complete. Indeed, these scores are learned and optimized for slide classification with a competitive approach which makes them not complete, and generally pushed most tile scores to a zero value, and one or two (still relevant) tiles with high scores.

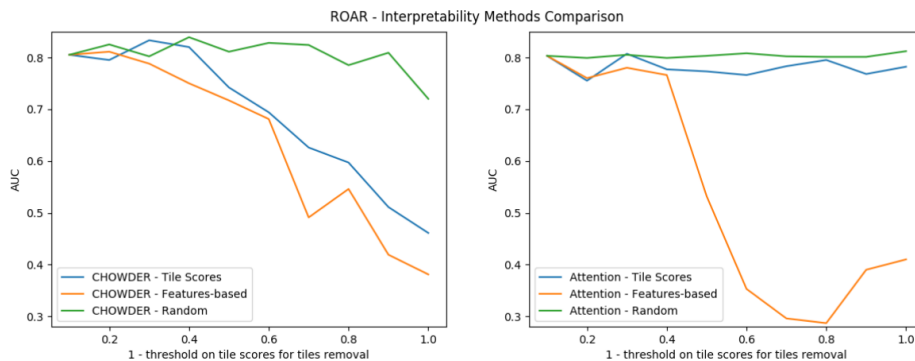


Figure 5.9: RemOve and Retrain experiment results: Impact of tile removal using heat-maps from CHOWDER (left) and Attention-based model (right) on slide classification performances.

### 5.2.7 Analysis of the number of features

Up to now the number of selected features is fixed by hand. We propose to thoroughly study the impact of the number of selected features on the quality of our feature-based heat-maps.

First, we measured this impact with a small number of features from using only the one

most contributing feature up to the first 7 most contributing identified features. We can observe, in Figure 5.10, that there is an important variability of localization AUC performances depending on the number of features with a variation between 0.903 and 0.82 (which are still great performances). We can also interpret that there are features of interest that make the localization AUC increase (such as feature 602 or feature 1866) and adversarial features that make the localization AUC decrease (such as feature 1644 or feature 420).

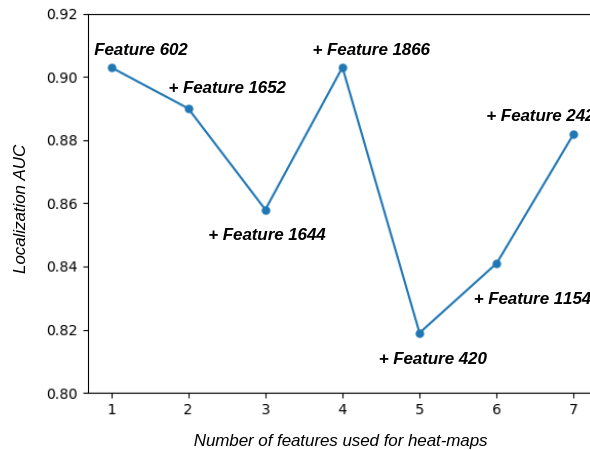


Figure 5.10: Impact of the number of selected features on localization AUC (between 0.80 and 0.92) for small numbers of features.

Thus, it seems critical to study more deeply this problem. So, by thresholding contribution scores at different values, we select from 1 to (all) 2048 features (according to the distribution in Figure 5.5) and show the evolution of localization AUC as we use more features to compute heat-maps (see Figure 5.11).

Three behaviors can be identified depending on the number of selected features:

1. If the number of selected features is really low (here between 1 and 12 features), the localization performances are unstable;
2. If the number of selected features is between 1% and 5% of features, we have a pretty constant regime of performances;
3. If the number of selected features is too high, localization AUC performances drop.

This leads to the conclusion that our method enables to select statistically a majority of features of interest among top-features identified. Thus when the number of selected features is low the performances are really impacted by the few adversarial features. So we could propose to select a fair amount of features that ensure good heat-maps. However, being able to study individually a small number of features (that lead to about 10 minutes of discussion per feature) really convinced pathologists, while it is not conceivable to ask a medical expert to analyze deeply a lot of features individually.

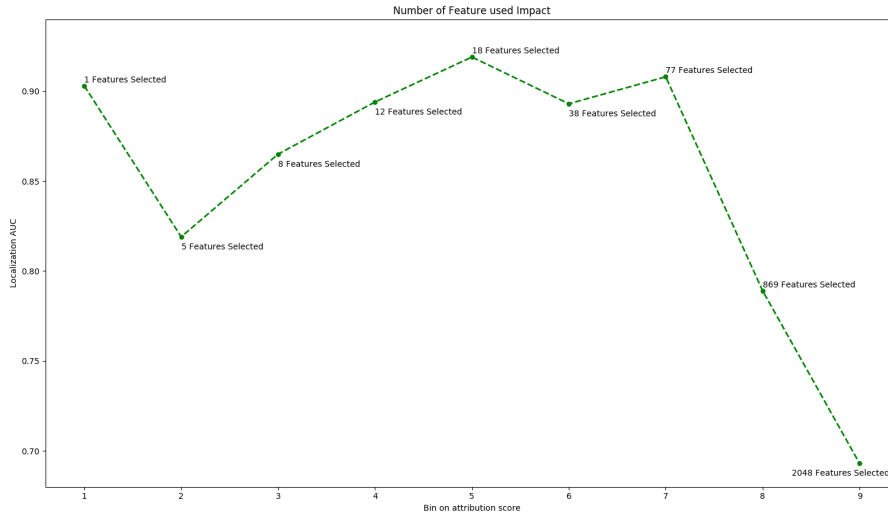


Figure 5.11: Impact of the number of selected features on localization AUC (between 0.7 and 0.9) for high numbers of features.

### 5.2.8 Colocalization filtering

This study about the impact of the number of selected features on heat-maps quality shows that three kinds of features stand out for Camelyon-16 dataset among ImageNet features: features of interest that activate homogeneously mostly in tumorous regions, adversarial features that activate homogeneously not only in tumorous regions, and unrelated features that either activate non homogeneously or almost do not activate over slides. Figure 5.12 illustrates the difference between features of interest and adversarial features. We can observe that features of interest indeed activate homogeneously and densely mostly in “tumor” regions, and that adversarial features either activate homogeneously outside of “tumor” regions or non homogeneously. It also gives another way to think about the third question we put forward (“Are these features of tile descriptors relevant medically and representative of histopathological information?”) by introducing a manner to measure the potential transfer of each individual feature to a given histopathological problem.

Under the hypothesis that our feature-based method enables to select statistically a majority of features of interest, we should be able to filter out adversarial features that do not colocalize with the feature-based heat-maps (computed as the normalized average of selected features).

To do so we propose to measure, for each selected feature individually, the Mean Absolute Error (MAE) between the feature  $k$  activation (normalized) heat-map  $H_{k,i}$  and the feature-based heat-maps  $H_{c,i}$  over whole slide  $i$  (given  $N_i$  the number of tiles on the slide  $i$  and  $N_s$  the number of slides) as following:

$$MAE^k = \frac{1}{N_i \times N_s} \times \sum_{i=1}^{N_s} \sum_{j=1}^{N_i} |H_{k,i,j} - H_{c,i,j}|.$$



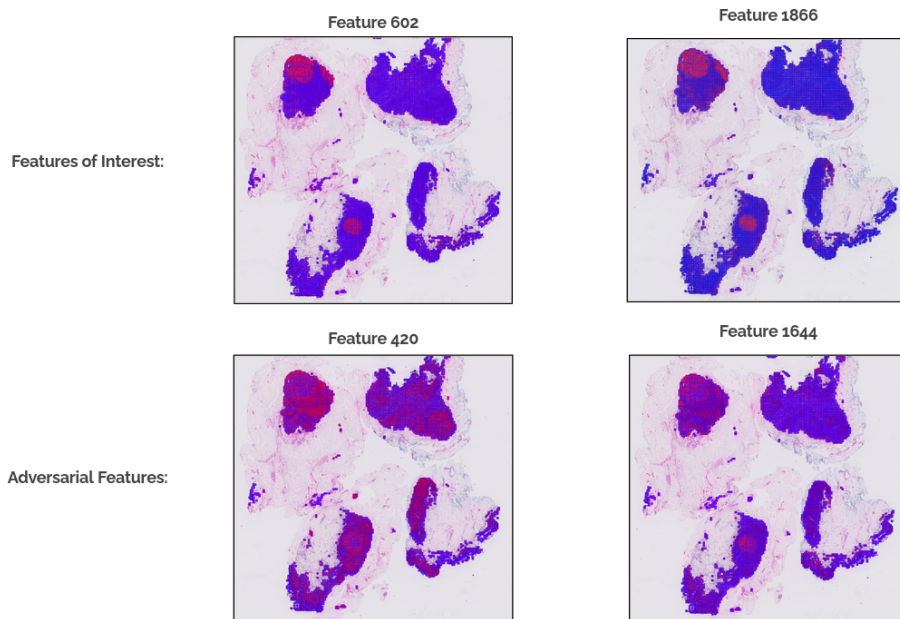


Figure 5.12: Contributing feature activation over slides.

Thus we obtain a distribution of MAE through selected features and we propose to keep only the ones that are on the lower half of the distribution. Figure 5.13 illustrates the method in the case of the 7 selected features for the CHOWDER model. We can observe the normalized activation of each single feature selected from which MAEs are computed. We can see that one feature (the one at position 420) has a high MAE (compared to the others), thus this feature is identified as adversarial and removed from the selection. Moreover, it can be noted that feature 420 is the one associated with the most important drop in localization AUC in Figure 5.10.

### 5.2.9 Application to the SFP Challenge

**Datasets & Evaluation** To validate the usefulness of our solution and come closer to our use case of cervical cancer screening, we ran a qualitative and quantitative evaluation on the SFP Challenge dataset [Pathologie (SFP) 2020]. This challenge consists in classifying histology WSI that come from biopsies from cervix between four (ordered w.r.t. cancer severity) classes: “normal”, “low grade lesion”, “high grade lesion” and “carcinoma”.

The dataset contains 1015 WSIs, uniformly distributed among classes. We use 80% slides (810) to train and validate our model and 20% (205 slides) to evaluate its performances.

The evaluation metric is a custom one that takes into account distance between classes and is computed as one minus the average error w.r.t. an error table that is detailed in Figure 5.14 (aside with tiles that are representative of classes of interest):

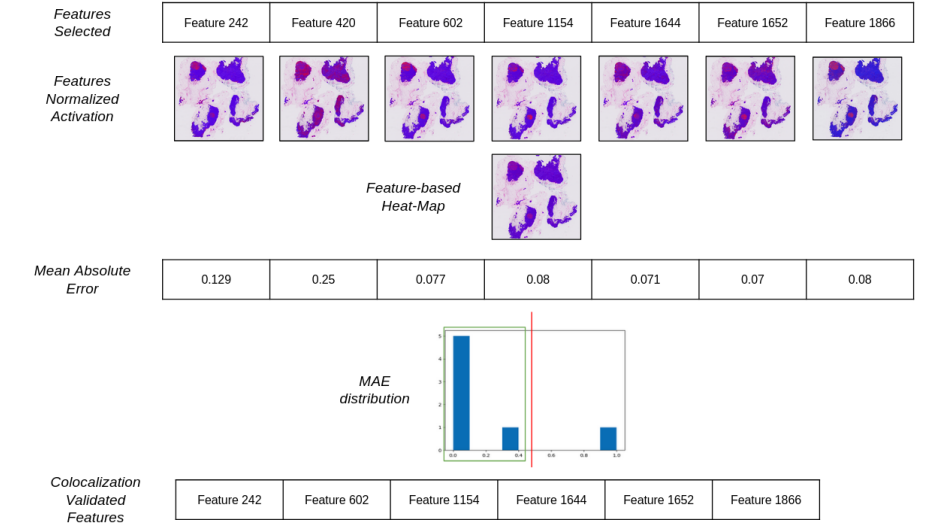


Figure 5.13: Illustration of the colocalization filtering method.

Given  $M_{sfp}$  the error table,  $M_{model}$  the confusion matrix obtained by a predictive model and  $\odot$  the element-wise multiplication, the SFP metric  $s$  is computed as:

$$s = 1 - \left( \frac{1}{\sum_{i=1}^4 \sum_{i=1}^4 M_{model}} \times \sum_{i=1}^4 \sum_{i=1}^4 M_{sfp} \odot M_{model} \right)$$

It can be observed that most of classes are diagnosed through a visual inspection of the epithelial surface. “Normal” slides will only have normal cells, stacked and well organized, while “low grade lesions” and “high grade lesions” will reveal dysplastic cells (if the thickness of the dysplastic layer fills more than two thirds of the epithelial surface then it is classified as “high grade lesion”). The “carcinoma” class is triggered when the dysplasia invades the tissue outside of the epithelial surface (thus a “carcinoma” diagnosis requires a screening on the whole tissue).

**Model training** We trained an attention-based model using a linear regression constraint (since classes are ordered) and tiling between 5X and 10X (standardizing on individual tiles provided by the SFP challenge). We report an overall accuracy of 61.8%, a mean AUC of 0.845 and a SFP metric of 0.9 (see Figure 5.15) on the test set which reveals a good training.

**Heat-maps and Interaction with Pathologists** We had the chance to work with an additional Data Scientist (Melanie Lubrano) and three pathologists (Yaelle Harrar, Raphael Bourgade and Delphine Loussouarn) which enabled us to interpret explanation heat-maps. We proposed to use both tile scores and feature-based heat-maps that revealed to be somehow complementary. Indeed, feature-based heat-maps enabled to identify overall tissue and cell organization that are responsible for the predicted labels, while tile score heat-maps, even if less interpretable, highlight the reason of the misclassification in case of error in prediction.



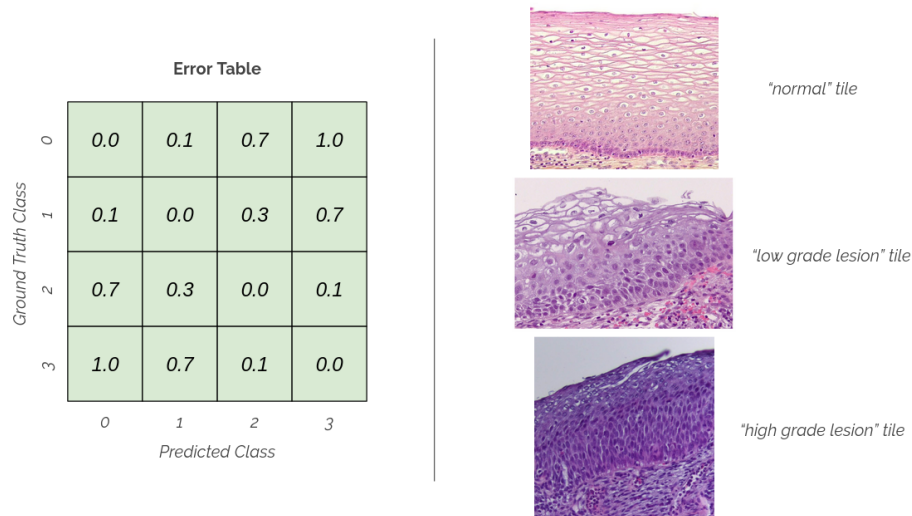


Figure 5.14: Error table for SFP metric computation and classes illustrations.

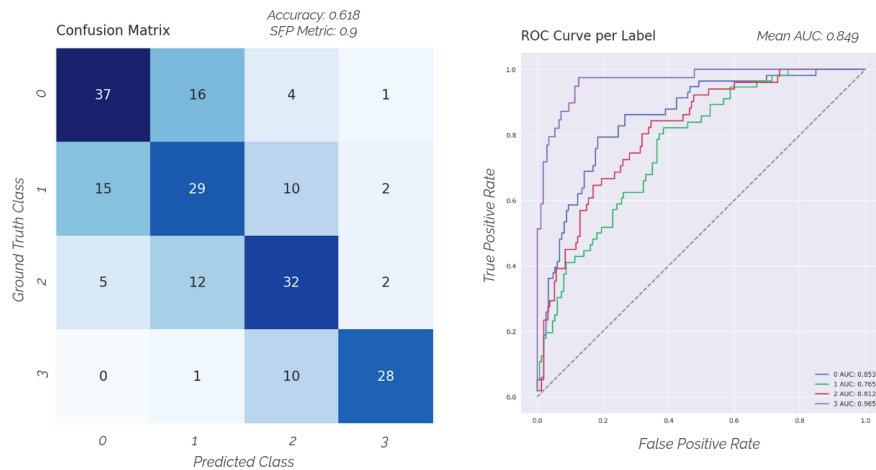


Figure 5.15: Attention-based model performances on SFP Challenge; the confusion matrix (left); the ROC curve (right)

Figure 5.16 shows tile score heat-maps for a “normal” slide and a “high-grade lesion”, both well classified. It can be observed and appreciated that these decisions are made in the epithelial surface region and that most contributing tiles (w.r.t. tile scores) are coherent with experts decisions.

Figure 5.17 shows a “high grade lesion” slide wrongly predicted as “carcinoma”. We can observe, using tile scores heat-maps, that, even if the most contributing tile is not relevant for the expected diagnosis, features associated with the “carcinoma” class (identified using our method) highlight in the feature-based heat-map the relevant region for the expected diagnosis. This confirms the relevance of features used by the model and validates the useful-

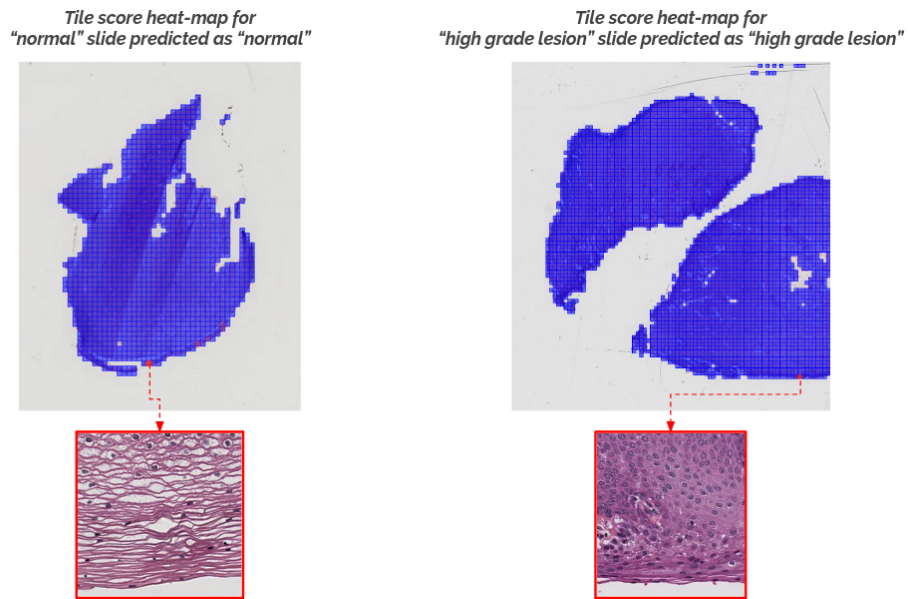


Figure 5.16: Tile score heat-maps for well classified “normal” and “high-grade lesion” slides.

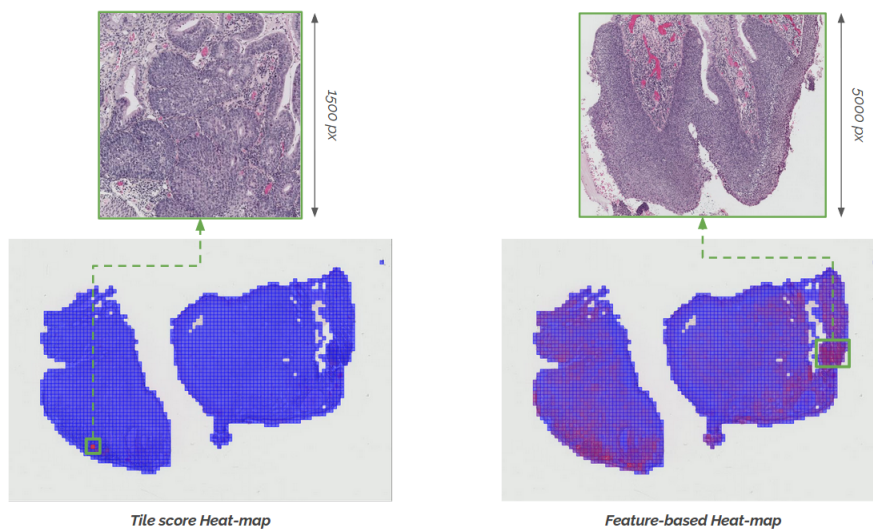


Figure 5.17: Tile score and feature-based heat-maps for a “high grade lesion” slide wrongly predicted as “carcinoma in situ”.

ness of our interpretability study and proposed method for slide reviews. Thus pathologists were able to identify histomorphological features that are responsible for errors made by the model. For example, in Figure 5.17, the tile that has a high tile score reveals a colonization of endocervical glands by the dysplasia, so it can be interpreted that the model gives a high contribution to the “infiltration” but does not take into account the cellular aspect, for this case, preventing it from differentiating between glandular infiltration and epithelial carcinoma in situ.

Perspectives on the potential use of this information to improve performances are considered and developed in Chapter 6.

We showed, using Camelyon-16 and the SFP Challenge histopathological datasets, that we were able to improve interpretability from trained WSI classification models but we also showed that features transferred from ImageNet were mostly textures that enable to capture features such as infiltrations but no cellular level features. This observation may limit the direct application of these methods to the LBC datasets, since, as explained in Chapter 2, the decision can be made on single cells visual analysis.

## 5.3 LBC slides classification

In this section, we apply and adapt MIL-based WSI classification approaches to cytology images. In particular, we use a dataset of about 400 LBC abnormal cervical smear slides provided by a partner (Medipath). Then, we experiment the automatic classification of these slides using the attention-based architecture and confirm the limitation identified before. We overcome this limitation using a weak “abnormality” detector that enables us to relax the difficult MIL context. In the end, we are able to train an efficient slide classifier and to detect individual cells (at 40X) that are responsible for the prediction at slide-level and that could be used to guide cyto-pathologists for slide reviews in routine.

### 5.3.1 Medipath dataset

Medipath is a group of anatomic-pathology laboratories in France that mainly performs cancer diagnosis (breast, thyroid, cervix ...).

This group provided 393 slides distributed into 4 classes that are the malpighian abnormalities: Atypical Squamous Cells of Undetermined Significance (ASCUS), Low-grade Squamous Intraepithelial Lesion (LSIL), Atypical Squamous Cells that cannot exclude High-grade lesion (ASC-H) and High-grade Squamous Intraepithelial Lesion (HSIL) (listed in order of severity). Definitions, descriptions and examples can be found in Chapter 2. Figure 5.18 shows a LSIL slide and a HSIL slide, and illustrates the difficulty of the task of classifying these slides that is equivalent to searching for a needle in a haystack (given, once again, that most of the time there is no needle in the haystack).

### 5.3.2 MIL classical approach

First, we naively train an attention-based model on this dataset. We know from practitioners that the decision is made at 40X magnification, thus we perform tiling at this level. It gives an average number of 31945 tile per slide (with a maximum of 39058) which is a difficult MIL context knowing that discriminative information is really sparse. At 20X, differentiation

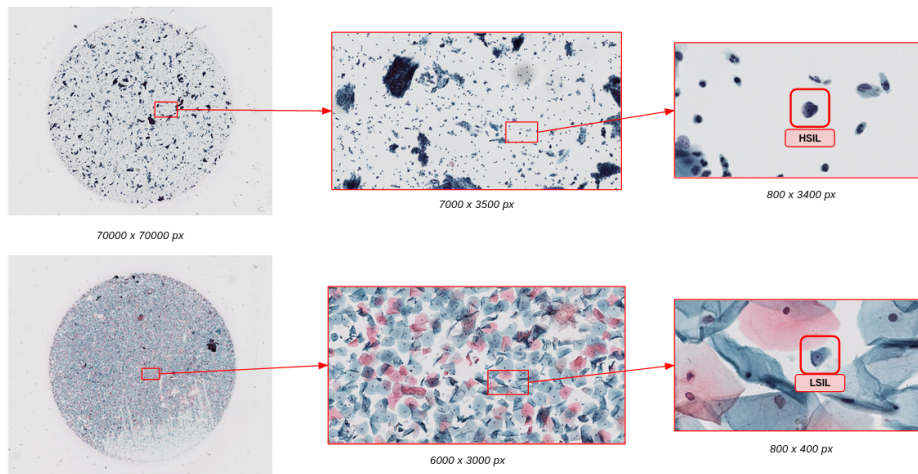


Figure 5.18: “LSIL” and “HSIL” slide examples (whole slide, about 10X view and about 40X view).

between classes is still possible but more difficult visually but the MIL context is more relaxed with “only” an average of 7673 tiles per slide (maximum of 9810). For 10X tiling, important cells are really hard to find and the MIL context is much relaxed with 1998 tiles per slide on average (maximum of 2478 tiles).

Figure 5.19 shows the visual content associated with a relevant tile for a “HSIL” decision at the three magnifications we consider.

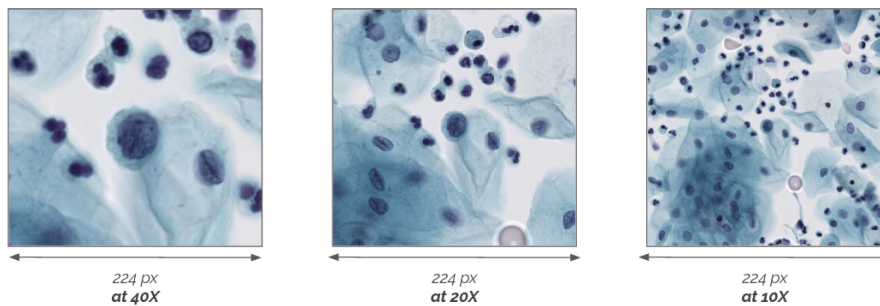


Figure 5.19: Example of content of single tiles at 10X, 20X and 40X.

10X tiling gave poor results with a 29.5% accuracy and a 0.657 average ROC-AUC that are barely better than a random predictor.

20X tiling gave better results with an accuracy of 34.6% and an average ROC-AUC of 0.68. Looking at the confusion matrix, on the left of Figure 5.20, it can be appreciated that the submatrix that concerns only the three first classes is nicely diagonal. Only “HSIL” slides have not been well learned, which makes the KAPPA measure drop to -0.052. Therefore, we applied a linear regression constraint to increase the importance of the distance between classes in the training. This improved accuracy a bit with 35.9% and drastically KAPPA

measure with 0.14 (see the right confusion matrix in Figure 5.20).

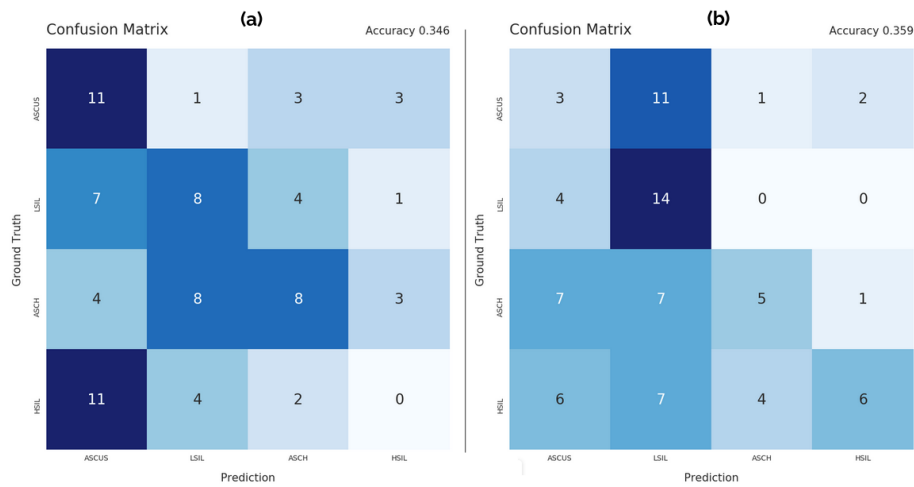


Figure 5.20: Confusion matrix obtained using a tiling at 20X; (a) Using softmax cross-entropy approach; (b) using classifier under regression constraint.

A similar approach applied at 40X gave promising results with an accuracy of 42.2% and a KAPPA of 0.320, which shows a better agreement between our predicting system and the expertise of cytopathologists. The confusion matrix can be observed in Figure 5.21 aside with explicative attention-based tile score heat-maps. Although classification metrics hint interesting results, these heat-maps do not show relevant information when the label is well predicted (see Figure 5.21-b). Sometimes they show relevant cells but a wrong prediction is made (see Figure 5.21-c). Good classification performances for bad reasons can be explained by the large number of tiles and the low number of slides, aside with the complexity and large variability that increase the possibility of overfit or irrelevant learning “short-cuts”.

### 5.3.3 “Abnormality” detector sampling approach

We showed that there is a limitation in the direct application of MIL approach for cytology use case, either because of the too important number of tiles or because we have to work at a magnification level that is high. Also after studying the transfer of ImageNet features to histopathology problems, we know that most features seem to be texture features. Therefore, we understand they can be transferred to describe histology slides but become less useful for LBC problem where tissue structure and organization are lost and cells are analyzed individually.

Thus, guided by the motivation to work at 40X to benefit as much as possible of texture features, we chose to relax the MIL context by reducing the number of tiles using weakly supervised localization.



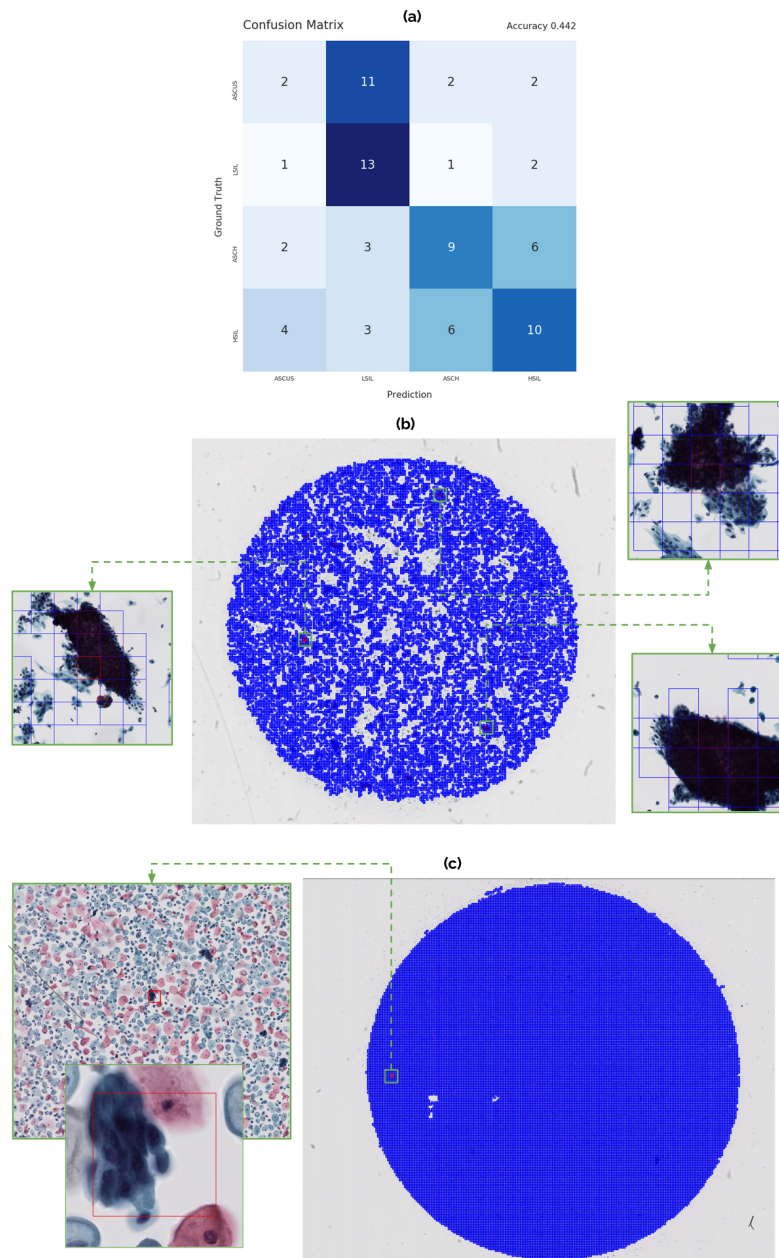


Figure 5.21: Confusion matrix obtained using a tiling at 40X; (a) using classifier under regression constraint; (b) Explanation heat-map of a “ASC-H” slide predicted as “ASC-H”; (b) Explanation heat-map of a “ASC-H” slide predicted as “ASC-US”.

### 5.3.3.1 Faster-RCNN training vs tile classifier

In Section 4.3, we presented two approaches for detecting potentially abnormal cells on LBC slides: an object detection-based method using a Faster-RCNN and a classification-based

method using a tile classifier and gradient-based explanations.

We showed that Faster-RCNN is not adapted for direct computer-aided diagnosis tools by being over-sensitive and proposing way too much cells to review. Indeed, all abnormal cells were detected but there were a lot of false positive detections which would not enable to reduce the workload of cytopathologists and guide them efficiently.

On the other hand, we showed that classification and explanation-based method were adapted for reducing the burden of analyzing all cells but were missing most abnormalities (32.8% on localization accuracy).

To reduce the MIL context, we want to be sure to respect the fact that positive bags contain at least one positive sample, thus Faster-RCNN approach is more adapted for this purpose.

### 5.3.3.2 Count of cells

First, we analyze the count of “abnormal” cells per slide w.r.t. the slide label. Figure 5.22 (a) shows the distribution of the number of detections per slide. We can observe, first, that there is no correlation between the number of detections and the severity of the label associated with the slide. Secondly, we can notice that there are few slides with a really low number of detections (under 100 detections). Figure 5.22 (b) shows one of these slides and highlights that it is actually a fully blurry slide due to a digitalization bug, thus these slides were removed from the datasets. Note that this study was actually made before the other studies, so blurry slides were also not used for the training of the other models presented previously in this section.

### 5.3.3.3 WSI classification results based on weak “abnormal” cell detection

Figure 5.23 shows the confusion matrix obtained using the model trained on bags composed only of bottlenecks computed from detected cells. We get an accuracy of 48.1% along with a KAPPA of 0.407, which is above the “moderate accordance” threshold. These results give us good hopes for a model that identified medically relevant features.

### 5.3.3.4 Attention-based tile score heat-maps

Moreover, tile score heat-map analysis reveals interesting results. Indeed, two tile scores heat-maps can be observed in Figure 5.24. The first one (a) shows a “ASC-H” slide predicted as “ASC-H” and the second one is a “ASC-H” slide predicted as “ASC-US”. We can appreciate the consistency in tiles with high scores that all show groups of cells with dark blue nucleus.

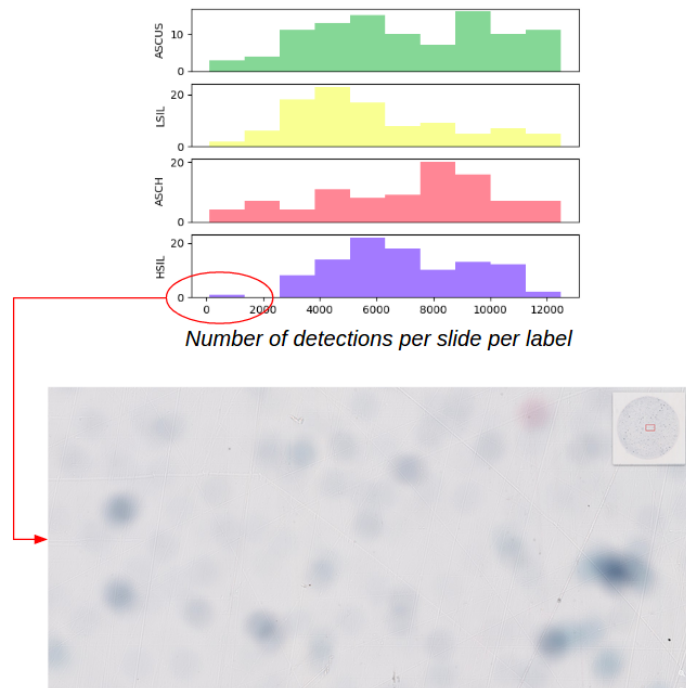


Figure 5.22: Study of the number of detections by the “abnormality” detector; (a) Distribution of the number of detections per slide w.r.t. slide label; (b) Outlier slide visualization.

Confusion Matrix Accuracy 0.481

Ground Truth	ASCUS	9	4	1	4
	LSIL	9	10	2	2
	ASCH	0	2	7	2
	HSIL	5	1	8	11
		ASCUS	LSIL	ASCH	HSIL

Prediction

Figure 5.23: Confusion matrix obtained using a tiling at 40X based on “abnormality” detection

### 5.3.3.5 Features and feature-based heat-maps

Now that we validated the relevance of the trained model, we apply our feature-based method to extract more information about what has been learned by this model.



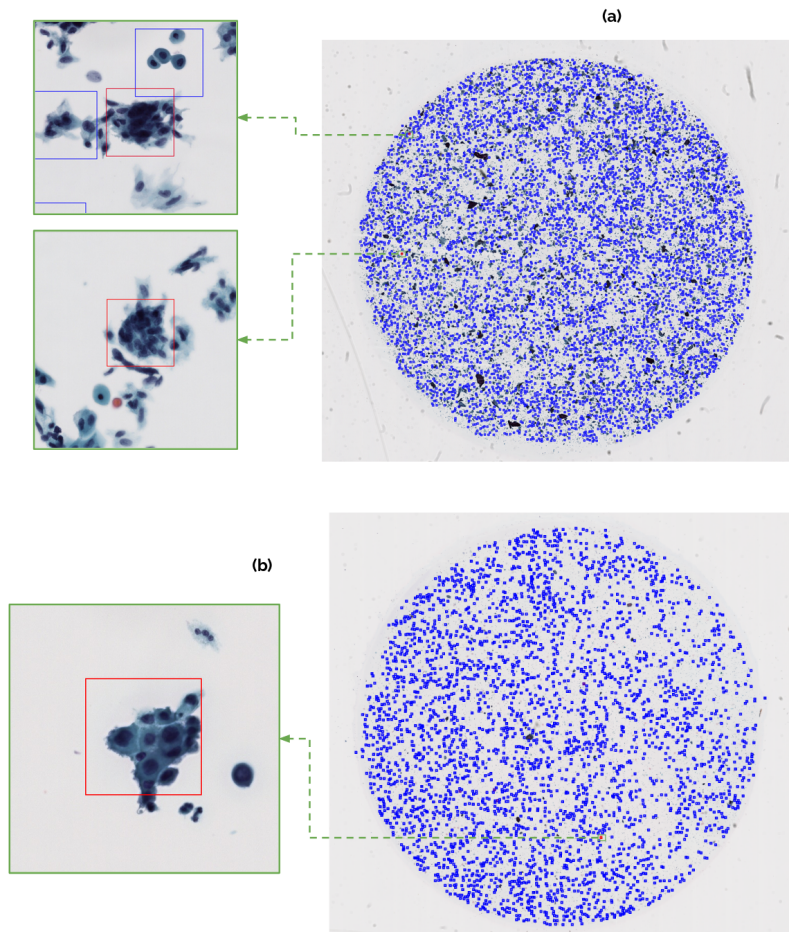


Figure 5.24: Tile score explanation heat-maps. (a) Explanation heat-map of a “ASC-H” slide predicted as “ASC-H”; (b) Explanation heat-map of a “HSIL” slide predicted as “ASC-US”.

Following the same process presented previously, we identify four features that mostly contribute to describe cells used for the “HSIL” class (the most malignant one). Figure 5.25 shows feature no 420, 434, 860 and 1652 with tiles that activate them mostly and a maximum activation generated tile for interpretation. Interestingly, two of these features were part of features identified for the “tumor” class in Camelyon-16 dataset study.

Using these four features, we compute our feature-based heat-maps that show meaningful hot regions with groups of abnormal cells.

Note: An observation that we made is that often groups of cells are identified and not often single cells.

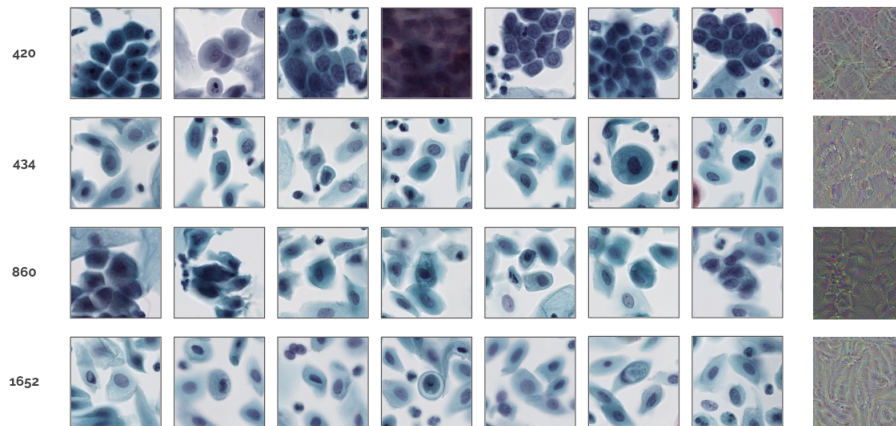


Figure 5.25: Identified features (420, 434, 860 and 1652); The dataset tiles that activate them mostly and the maximum activation generated tile.

## 5.4 Conclusion

In this chapter, we presented our interpretability approach and researches that apply to WSI classification architectures. We proposed a unified design that gather a vast majority of WSI classification methods relying on MIL learning. We motivated and applied a gradient-based attribution method to identify features that have been learned to be relevant in intermediate (tile and slide) descriptors. Then, using on Camelyon-16 dataset, we showed the relevance of these features by visualization (with dataset patches and max activation) and validation by pathologists. These discussions made us consider measuring interpretability by computing explainability heat-maps over whole slides taking into account only identified features. Allaying patch-based and slide-based visualization took interpretability to a next level for pathologists to understand histological meaning of features used by trained models. They confirmed that our approach gives explanations that are highly meaningful and interpretable, and convinced them that characteristics used by the model are aligned with the experience of pathologists. Our per-block approach can be used for all WSI classification pipelines trained on histopathological problems (and probably more, such as biomarker discoveries or treatment response) that follow the general design defined in this work, and shed a light on how WSI pipeline learn. Validating our approach on two distinct architectures enabled us to claim its generalizability. Quantifying the improvement of heat-maps generated through two interpretability measures strengthen this point. Finally, our individual analysis of each feature selected at slide-level enabled us to filter out outlier features, to stabilize interpretability performances, and to automatically select the right number of features needed for good heat-maps. This work digs deeper in the interpretability of WSI classification trained models. Then we validated this work on a new dataset and with pathologists. This dataset deals with cervical cancer screening in a context of histopathology. In the end, we entered the LBC context and highlighted the limitations for a direct transfer of previously presented methods for this application. We proposed to use a weak "abnormality" detector to relax the MIL context. This enabled us to reach acceptable performance and to prove, through our feature-based in-

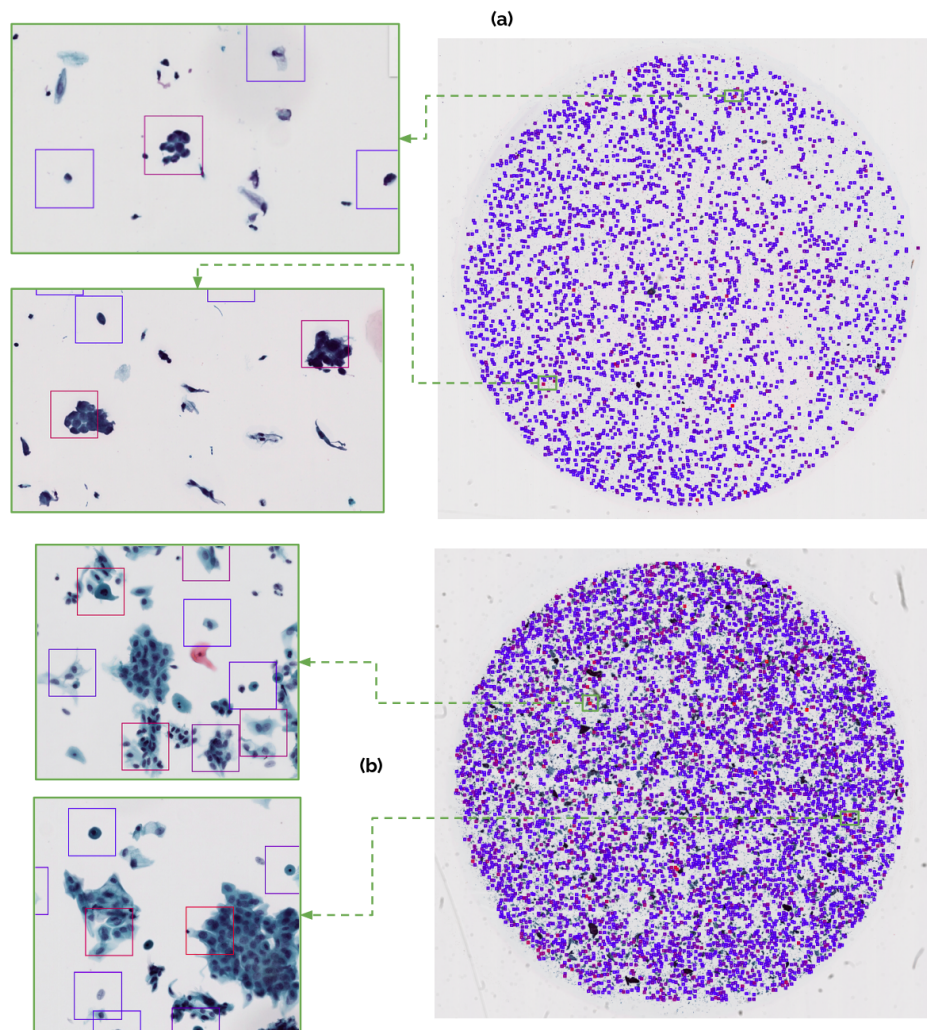


Figure 5.26: Feature-based explanation heat-maps. (a) Feature-based explanation heat-map of a “ASC-H” slide predicted as “ASC-H”; (b) Feature-based explanation heat-map of a “HSIL” slide predicted as “ASC-US”.

interpretability approach, the relevance of what has been learned, and to highlight the potential limitations.



# Conclusion and Perspectives

---

In this manuscript we presented our work that aims at proposing new methods to build an efficient Computer-Aided Diagnosis (CAD) tool for cervical cancer screening. Cervical cancer screening, that tackles one of the most devastating cancer worldwide for women, is performed by medical experts called cytopathologists that try to detect pre-cancerous changes on microscopy slides that may contain more than 100,000 cells. While about 90% of the time there is “nothing” to find, the 10% left are medically potentially critical cases that should not be missed. The recent efficient process of Whole Slide Imaging enables to digitize microscopy slides at a high resolution. Thus, it represents an amazing opportunity for medical experts to benefit from most advanced Artificial Intelligence (AI). Indeed, latest Machine Learning ML methods promise efficient and personalized Computer-Aided Diagnosis (CAD) tools.

In that sense, Chapter 2 and Chapter 3 respectively offer a review of the medical context and a state-of-the art study of deep learning neural networks for image classification, object detection, WSI classification and interpretability (that we identify as a crucial property for our proposed methods).

Through these first two chapters, we can clearly understand that there are four desired properties for the direct use of the most popular WSI classification methods to Liquid-Based Cytology (LBC) applications, and there is accordingly a need for:

1. An efficient single cell classifier since the medical decision may rely on few individual cells;
2. A method that can deal with WSI containing discriminative information at 40X;
3. A method that can make the decision relatively quick (experts generally take less than 5 minutes per case);
4. A method that is as interpretable as possible in order to efficiently guide medical doctors toward relevant regions/cells.

Our first contribution, in Chapter 4, tackles the question of having an efficient cell classifier that can be integrated in a CAD tool. We rely heavily on a public dataset called Herlev dataset, that contains about a thousand images of single cells extracted from Pap smear slides and divided into 7 classes. We turned this problem into a severity problem and experimented a classification and regression approach before proposing a solution that unifies these two

approaches and outperforms them: a regularization term that is implemented as a regression constraint on top of a classifier, which enables to train a classifier while taking into account the notion of “distance” between classes. Then, we exported this method closer to a WSI system by working on simulated regions containing several cells. We validated the relevance of our regression constraint and demonstrated that it qualitatively improved interpretability by performing cell localization and abnormality detection in a weakly-supervised context. Finally, we embedded this work into a single WSI classification system that can be used to guide slide reviews for cytopathologists.

Thus, in this chapter, we tackle the two first properties. The point 1 is completed with great performances. The point 2 could be improved with a better supervised object detector (e.g. a multi-class detector). The point 3 regarding timing is still limiting with a prediction time around 20 minutes per slide.

Our second contribution, in Chapter 5, aims at using WSI classification architectures based on Multiple Instance Learning (MIL) for LBC cervical cancer screening. We started by questioning the relevance of interpretability as usually defined in this context and brought a new insight on how it could be improved. We relied on Camelyon-16 dataset to show how we improved interpretability by extracting tile-level information, that we proved to be highly relevant for medical experts, and slide level explanation heat-maps. We validated the method by applying the exact same method to a cervical cancer histopathology dataset. In a second step, we used a 400 WSI dataset of “abnormal” LBC slides for cervical cancer, and established the limitations of current methods to perform on this dataset. Noticing that the MIL context is too complicated, with over 30,000 tiles per slide at 40X, we overcame this by using a weakly trained “abnormality” object detector for sampling. This enabled to reach acceptable performances with an inference time around 3 minutes (point 3). Finally, our interpretability methods enabled us to highlight the most contributing cells and to show the relevance of the learning (point 4).

In the end, we answered all points listed above with an efficient and interpretable cell classifier and a good WSI classification system with improved interpretability. However, all points could also be taken further.

There are still several validation steps to complete to create a tool that is fully suited to enter the routine of cytopathologists. Obviously, including “normal” slides and other abnormalities, notably glandular atypia such as Atypical Glandular cells of Undetermined Significance (AGUS) and Adenocarcinoma In Situ (AIS), would be the next step to create a model that covers all cases. We would suggest using a hierarchical classification approach (see Figure 6.1) with regression constraint on each “atypia” branch. We could imagine training a model per node in that tree.

Training the feature extractor with a self-supervision method such as simCLR [Dehaene et al. 2020; Chen et al. 2020b] would be a necessary step since it would improve the results without any additional supervision. It could be applied to both cell-level classification and WSI classification. Indeed, what we are looking for is not to describe the whole content of the image of cells but to find features that cells from a certain class have in common.



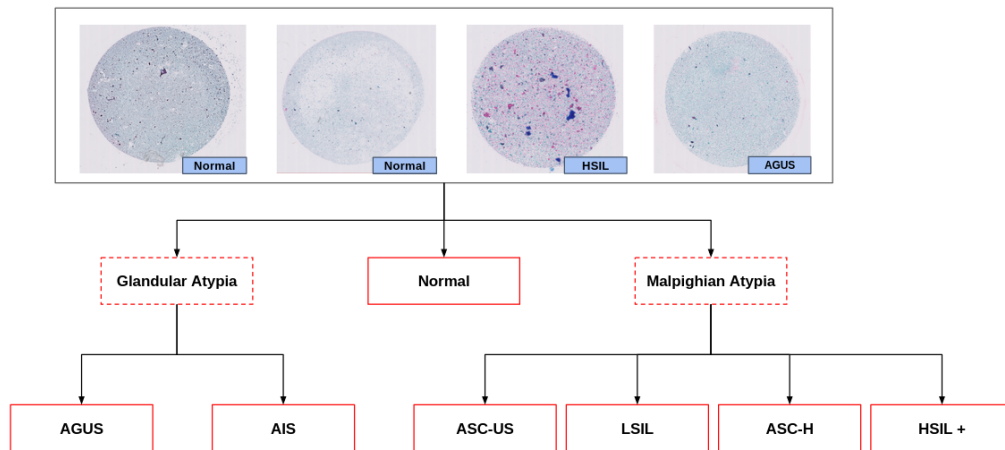


Figure 6.1: Hierarchical design for a complete CAD tool for cervical cancer screening.

Another approach that was considered and seemed promising to us in order to improve interpretability and performances was to add even more prior knowledge (such as the fact that discriminative features are localized) in the WSI classification architecture by including gradient-based localization.

The motivation is to turn the workflow of cytopathologists into a deep learning pipeline to bring explainability and medical feedbacks to practitioners when they are using the CAD tool built on a trained model. Indeed, when a cytopathologist screens a slide, he/she generally screens the full slide at (around) 10X magnification looking for potentially abnormal cells. When he/she localizes a potentially abnormal cell or group of cells, it zooms on this region and inspects the details of the nucleus at around 40X magnification. Finally, after repeating this process over the whole slide, based on the presence or not of abnormal cells and their malignancy (dysplasia), he/she makes a diagnosis. Thus, the proposed approach would be divided into four stages:

1. A tile classifier that scores each tile at 10X magnification;
2. A 40X region/cell sampler that selects potentially abnormal cells based on 10X tiles scores;
3. A cell level classifier that outputs a descriptor for each of top- $N$  regions/cells previously selected;
4. A MIL diagnosis classifier based on previous descriptors that aggregates them into a slide descriptor and classifies the slide.

We now suggest a potential pipeline (see Figure 6.2):

The pipeline takes as input a bag of tiles of size 896x896 pixels (and the associated diagnosis at training time) which are resized at 224x224 size to be 10X tiles. These tiles

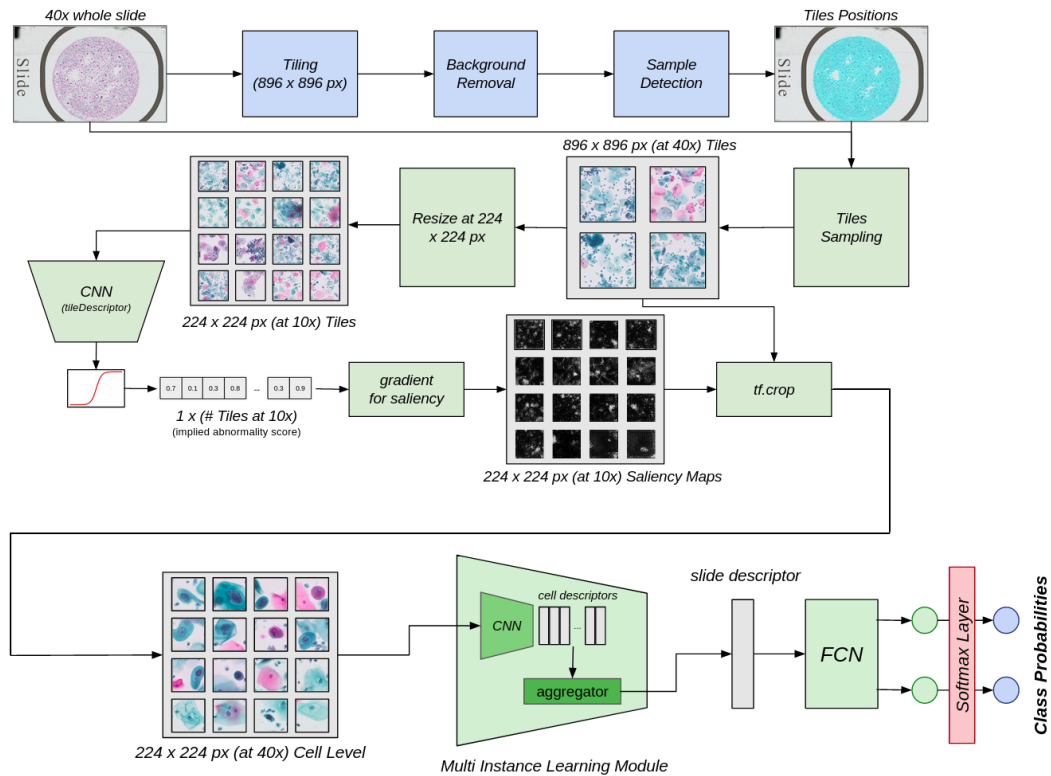


Figure 6.2: Suggested WSI classification architecture guided by localization and interpretability.

are fed to the tile classifier which outputs a descriptor for each tile. These descriptors are turned into a single score (implicitly abnormality score). Based on these “abnormality” scores, the top- $N$  tiles are selected, and the gradient of the score with respect to the tile pixels at 10X is computed. The value of  $N$  could be set to 22 because a method called ThinPrep Images System [Heard et al. 2012], approved by FDA (Food and Drug Administration), is currently used in the routine of cytopathologists, and proposes 22 regions for the expert to make the decision. The absolute values of the resulting attribution maps are multiplied by the negative of the 10X selected tile (since white background is known to be uninformative). Thus we obtain attribution maps which are grayscale maps where the whiter a pixel the more it contributes to the abnormality score. We propose to extract, per selected tile, the  $56 \times 56$  region that has the more attribution (corresponding to a  $224 \times 224$  region at 40X). A crop is then made on input ( $896 \times 896$  pixels) images for each selected region (at 40X). Thus we obtain  $N$  tiles at 40X of size  $224 \times 224$  (hopefully most abnormal cells) that we call cell level tiles. These images are then fed to the second “diagnosis” classifier that also outputs a descriptor for each image. These descriptors are given to an attention module (described in the next paragraph) that aggregates these cell-level descriptors into a single slide descriptor that is finally given to a two layers multi-layers perceptron that outputs the logits pre-softmax. The softmaxed class scores are the proposed diagnosis.



In the end, our interpretability method could be used to improve slide-level classification performances through active learning. Indeed in Subsection 5.2.9, we explained how pathologists used both tile score and feature-based heat-maps to understand errors and validate good decisions made by the network. Thus, we considered using these heat-maps to create editable annotations that could be adapted by experts reviewing slides to correct the model and induce discriminative features that have not been learned. The idea here is to benefit from information that can be learned directly from slide-level labels. Then, using interpretability from trained models, to rely on the expertise of pathologists to build new MIL bags that come to correct errors. In current WSI classification architecture relying on MIL context, there are two possible sources of errors: localization errors and decision errors. The first ones are when the heat-map shows that the model based its decision on a non-informative tile. The second ones are when the heat-map shows that the model based its decision on a relevant tile but still cannot predict the right label. It is also important not to be biased by ground-truth labels since a model can make a good prediction but for bad reasons, which is generally the case if there is an overfitting due to the small size of validation set. The success of such an approach would be highly beneficial since it would mean that we would be able to correct models as pathologists are trained, and in general would mean a closer collaboration between medical experts and computer scientists which is crucial for efficient CAD tools. A promising approach to induce identified missing features in models is through multi-task and adversarial training as it is done in [Graziani et al. 2020] which would require finding ways to quantify it. For example, we could find a way to quantify the concept of epithelial cells vs glandular cells that seems to have been missed by our trained model for SFP Challenge.

Thus, our work answered critical points for an efficient CAD tool for cervical cancer. Still there is room for improvement through the new methods being proposed every day. It can also be noted that cytology exams are widely used for other indications such as bladder cancer or thyroid cancer screening which could be an interesting validation step for our methods.



# From Machine Learning to Deep Learning

---

In the wide field of Artificial Intelligence (AI), Machine Learning (ML) gathers methods that are able to learn through a set of examples also called dataset. Deep Learning (DL) is a specific group of ML methods relying on deep neural network architectures, thus allowing the learning of specific features to solve complex tasks.

DL recently gained a lot of interest especially for image processing with architectures called CNN and this is due to an event that happened in 2012.

## A.1 Machine Learning and image classification before 2012

Every year since 2010, the Large Scale Visual Recognition Challenge (LSVRC) [Russakovsky et al. 2015] enables the best computer vision teams worldwide to compare their image classification methods on ImageNet dataset [Deng et al. 2009] (see Fig. A.1). This dataset contains more than 1,400,000 images divided into 1000 categories, also called classes, inspired from WordNet [Fellbaum 1998] hierarchy “leaf” classes such as dog breeds (“dalmatian”, “golden retriever”, “border collie”, ...), car kinds (“race car”, “minivan”, “cab”, ...) or landscape types (“cliff”, “valley”, “volcano”, ...). The challenge consists in automatically classifying these natural images in the ground truth class. It is mainly evaluated using top-5 accuracy i.e. percentage of images well classified among 5 allowed predicted classes, and the hierarchical cost i.e. the average distance to the closest common ancestor between predicted class and ground-truth class according to WordNet semantic hierarchy.

### A.1.1 Hand-crafted feature extraction

In 2010, the winning method, proposed by a team from University of Illinois (USA) and NEC (Japan), performed with a top-5 accuracy of 71.8%, an error rate of 47.1% and a hierarchical cost of 2.1144. This method [Lin et al. 2010] consists in extracting features from the image in a dense grid descriptor using Histogram of Oriented Gradients (HOG) [Freeman and Roth 1995] and Local Binary Pattern (Local Binary Pattern) [Harwood et al. 1995], then encoding this descriptor in a high-level descriptor using local coordinate coding [Yu, Zhang, and Gong

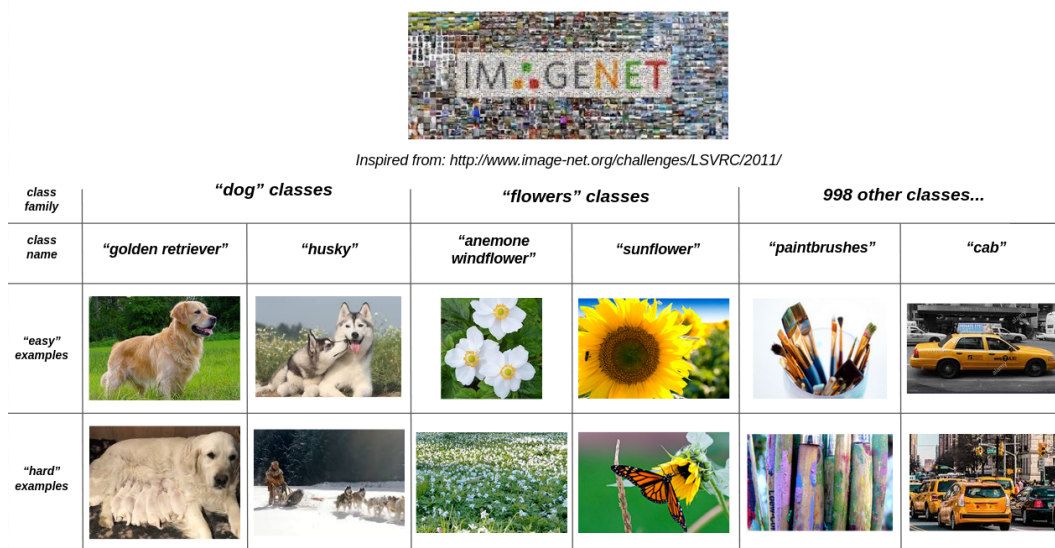


Figure A.1: LSVRC ImageNet dataset images/classes examples

2009] and super-vector coding [Zhou et al. 2010], reducing computational needs by reducing this descriptor size using spatial pyramid pooling [Lazebnik, Schmid, and Ponce 2006], and predicting a class based on this pooled descriptor training efficiently a linear Support Vector Machine (SVM) [Cortes and Vapnik 1995] with averaged stochastic gradient descent [Polyak and Juditsky 1992].

The same year, the second best method, by Xerox Research Centre Europe team, proposed to rely on Fisher vectors [Perronnin, Sánchez, and Mensink 2010]. First, features are computed using Scale-Invariant Feature Transform (SIFT) [Lowe 2004] and color features, and then reduced using Principal Component Analysis (PCA). A Gaussian Mixture Model (GMM) is trained to compute Fisher vectors. Spatial pyramid pooling is also used and finally a linear classifier is trained over these pooled Fisher vectors to make the decision using Stochastic Gradient Descent (SGD) with two regularizations: L2 normalization, and power normalization to remove image-specific information and to avoid Fisher vector sparsity. Thus these methods reached a top-5 accuracy of 66.4% and a hierarchical cost of 2.5553.

The year after, this same team won by improving their classification method with a step of compressing Fisher vectors using product quantization [Sánchez and Perronnin 2011], which consists in splitting vectors in small sub-vectors, clustering these using k-means [Steinhaus 1956], and representing each sub-vector by its centroid (encoded by an index). This reduces the size of the dataset, and thus enables to take the most out of computational and storage power. The performance reported are a top-5 accuracy of 74.3% and a hierarchical cost of 0.10980.

The 2011 edition second best method was proposed by a team from University of Amsterdam. Their localization method [Uijlings, Sande, and Gevers 2013] (interestingly this method won 2011 localization LSVRC competition) combines graph-based segmentation [Felzenszwalb and Huttenlocher 2004] and grouping methods (color-based [Sande, Gevers, and Snoek 2010],

texture-based [Lowe 2004] and size-based) for hierarchical boxes proposals. Their classification pipeline [Sande, Gevers, and Snoek 2010] uses key-point sampling [Mikolajczyk and Schmid 2004] then a set of SIFT descriptors with spatial pyramid and vector quantization. These descriptors are used to train an intersection kernel support vector classifier [Maji, Berg, and Malik 2008]. The performances are a top-5 accuracy of 69% and a hierarchical cost of 0.13270.

As we can note, all of these methods rely on two stages:

1. Extracting hand-crafted features to describe the image;
2. Training a ML architecture to make the decision based on the image descriptor.

Indeed, due to the lack of computational and memory limitations, the datasets must be down-sized drastically through embeddings, encodings, quantizations and poolings with hand-crafted features, while keeping task-related relevant information to train a classifier (and it works fairly well [Bristow and Lucey 2014]).

And more generally, beyond this LSVRC competition, this type of ML approach has been extensively used for object recognition: In [Dalal and Triggs 2005], HOG [Freeman and Roth 1995] are used for face detection. In [Mu et al. 2008], Local Binary Pattern (LBP) [Harwood et al. 1995] are used for human detection. [Jegou et al. 2010] propose Vector of Locally Aggregated Descriptors (VLAD) as encoding method and use it for scene classification.

### A.1.2 ML classification methods

At this time, SVMs [Cortes and Vapnik 1995] seemed to be the most popular ML classification method. It is an extension of linear classifiers that enables to deal with non-linearly separable problems. To do so, it projects samples into a higher dimension space using a kernel function which makes the problem linearly separable and then tries to find the optimal separation hyperplane i.e. the one that implies the largest margin (smallest distance between hyperplane and a class sample) using other samples as support vectors (see Fig. A.2).

There are other popular and efficient classification methods [Wu et al. 2007, Dreiseitl and Ohno-Machado 2002]. For instance, decision trees [Breiman et al. 1984] and random forest [Brieman 2001] (ensemble of decision trees), that basically aim at dividing the decision into small if/else condition problems on features values, were also very popular (e.g. [Rajendran and Madheswaran 2001]).

Logistic regression [Cabrera 1994] and Perceptron [Rosenblatt 1958] are methods that try to predict the conditional probability of a class with regard to the given input by tuning parameters with maximum-likelihood estimation. Perceptron method is also called an Artificial Neural Network (ANN) that consists in multiplying each input feature  $x_i$  by a weight  $w_i$  that is used to compute the weighted sum of the input features that is given to an activation function, for example a sigmoid:

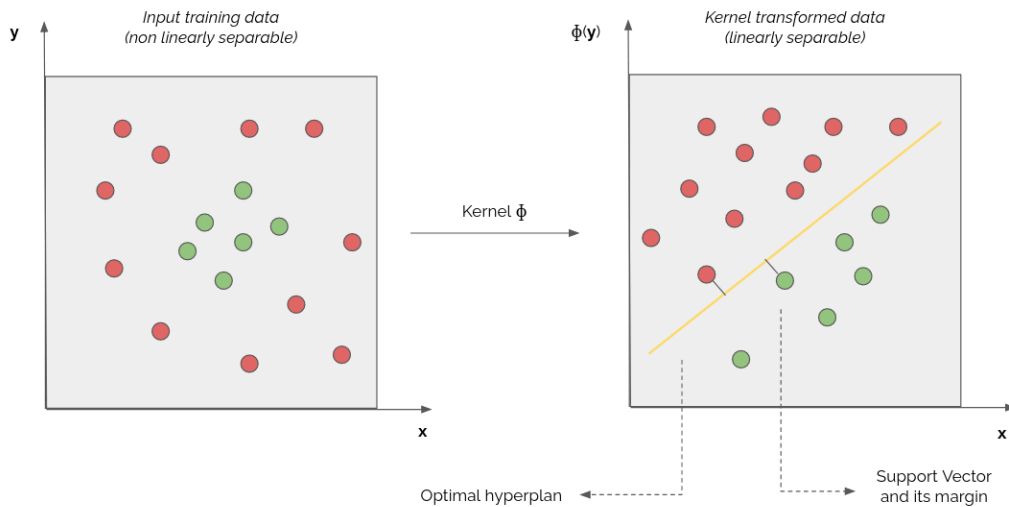


Figure A.2: SVM principle.

$$P(y|x) = 1/(1 - e^{-\sum_i x_i \cdot w_i})$$

We will go into detail about how these weights are computed in the next Appendix B.

Adaboost [Freund and Schapire 1997] proposes to use an ensemble learning strategy by training a first classifier and then iteratively training new classifiers giving more weight to training samples that are poorly classified by previous classifiers.

The k-Nearest Neighbors [Kowalski and Bender 1972] method defines a distance measure between samples and uses it to map the feature space to a class based on the k closest neighbors.

### A.1.3 2012 LSVRC Edition

The specificity of LSVRC is that with its million of images and 1,000 classes, having hand-crafted features to describe images is limiting. Indeed features cannot be class-specific but need to be generalizable to describe a large range of objects and of types of images.

In 2012, while the second best proposed method [Harada and Kuniyoshi 2012] still performs around 74% top-5 accuracy using the same ML approach, a DL approach, proposed by SuperVision team, succeeded with a top-5 accuracy of 84.7% and even a 62.5% top-1 accuracy. And the year after, in 2013, all top methods were using DL methods and performed at this level (see Fig. A.3).

Thus, the 2012 edition of LSVRC shook the field and turned the spotlight on another way of working with promises of greater results, namely using CNNs architectures.

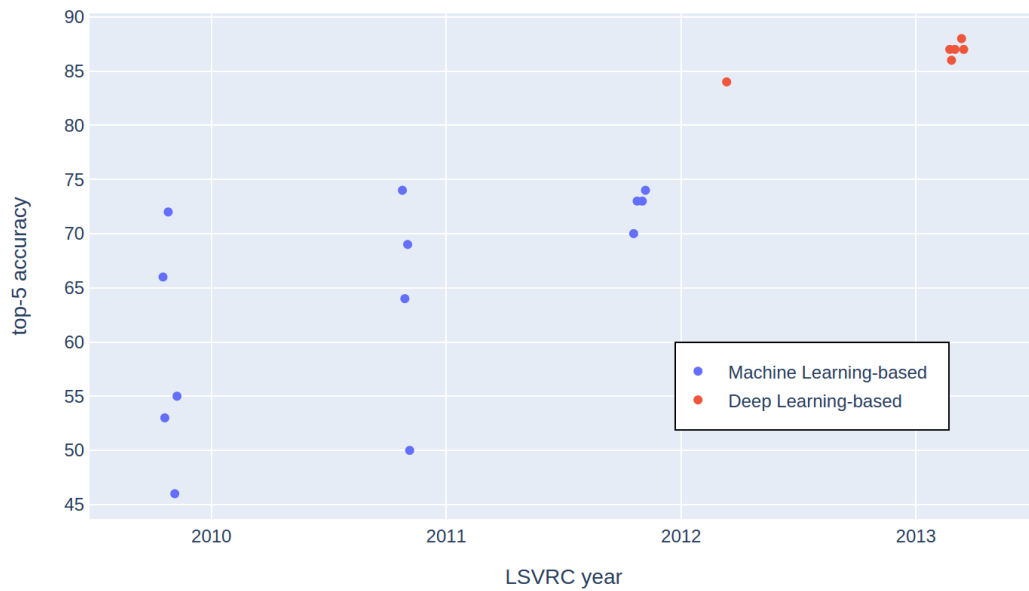


Figure A.3: LSVRC Results over the years; Inspired from: <http://www.image-net.org/challenges/LSVRC/2011/>





# Introduction to Convolutional Neural Networks and their training strategies

---

In this appendix, we detail the architecture called AlexNet which will enable us to introduce important concepts to understand Convolutional Neural Networks (CNN) architectures and their training strategies.

## B.1 Convolution layers

The architecture named AlexNet [Krizhevsky, Sutskever, and Hinton 2012] is mainly known to be the first convolutional neural network architecture to have been successfully trained on a large scale dataset. Such architectures were introduced 30 years before [LeCun et al. 1989] with convolutional layers that were used for zip code recognition.

As a reminder, given  $I$  a grayscale image of size  $(X_I, Y_I)$  and  $F$  a linear filter (also called kernel) of size  $(X_F, Y_F)$ , the response  $A_{(I,F)}$  of the image to this kernel also called activation map is computed using a convolution product as follows:

$$A_{(I,F)}(x, y) = \sum_{i=1}^{X_F} \sum_{j=1}^{Y_F} I(x - \frac{X_F-1}{2} + i, y - \frac{Y_F-1}{2} + j) \times F(i, j)$$

Intuitively, it can be seen as screening the image for a low-level pattern represented by the kernel. For example, Sobel filters are known to be good vertical and horizontal edge detectors (see the top of Figure B.1).

CNNs are simply a stack of trainable layers among which there are convolutional layers which contain trainable kernels that can be thus adapted to the task they are trained on (see the bottom of Figure B.1). The intuition behind CNN relies on the fact that the mammal visual cortex works approximately this way [Hubel and Wiesel 1962, Fukushima 1980, Lindsay 2020] with visual cortexes from the first, detecting orientation and edges, to the sixth that mixes information from previous cortexes to build a global representation of visual information. The capacity that brain synapses have to strengthen their link when being often stimulated is called the synaptic plasticity and can be seen as similar to the weights that are learned to mix information from different features learned.

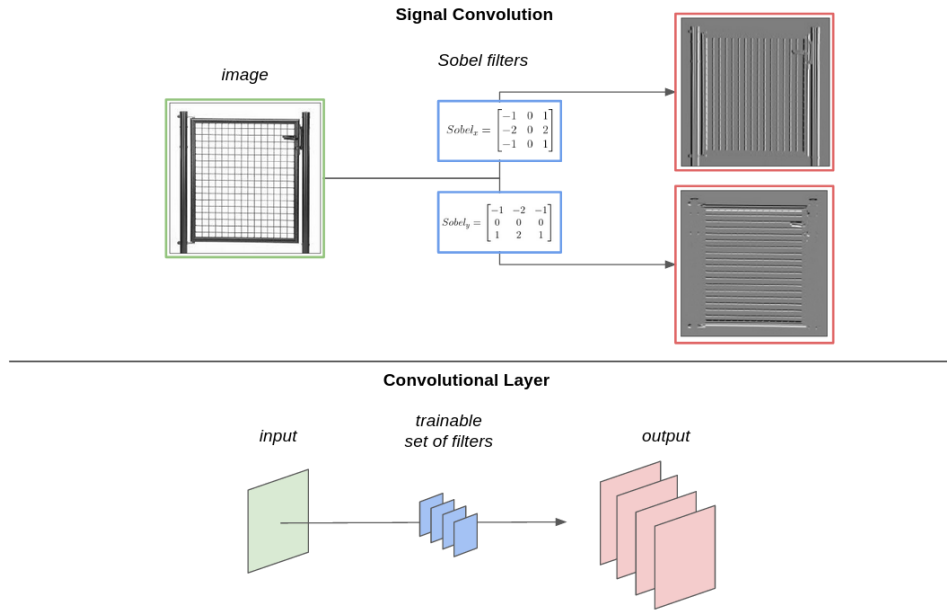


Figure B.1: Illustration of convolutional layer. Edge detector using Sobel filters (top); Convolutional layer and its trainable filters (bottom)

Convolutional Neural Networks layers are mainly composed of four type of layers:

1. Convolutional layers which are composed of sets of trainable kernels (kernel size, number of kernels, stride size and padding method are hyper-parameters). A convolutional layer is defined by its number of filters (also called depth), its stride, i.e. convolution “steps” size or the number of pixels the kernel is translated at each step, and its padding, i.e. the number of pixels that are added outside the image to compute convolution for pixels close the border or the image (generally zero-padding or same-padding are used). For  $c$  the number of channels of the input,  $k$  the kernel size and  $d$  the depth of the layer, the number of trainable parameters in a convolutional layer is  $(k \times k \times c + 1) \times d$ . The +1 term represents biases that are added to the pre-activation thus enabling a shift of the output. These layers output features responses in the image space that are called feature maps.
2. Pooling layers, that, given a rule (mean or max), downscale feature maps. Their role is to reduce computational complexity, add some translation invariance and increase the receptive field of deeper neurons (the type of pooling, which is generally mean or max pooling, and the downscaling ratio are hyperparameters);
3. Activation layers that introduce non linearity in the network. The most popular activation layer is Rectified Linear Unit (ReLU) [Nair and Hinton 2010]:  $ReLU(x) = \max(0, x)$ ;
4. Fully connected layers that mix learned high level features activations. A fully connected layer (see Figure B.2) is defined by the number of its inputs and its number of neurons.

It outputs a vector of size equal to its number of neurons, and it can be seen as a matrix multiplication. For an input vector  $X = (x_1, \dots, x_N) \in \mathbb{R}^N$ , output  $Y = (y_1, \dots, y_M) \in \mathbb{R}^M$  using learned weights  $W \in \mathbb{R}^{N \times M}$ , for  $j \in [1, \dots, M]$ :

$$y_j = \sum_{i=1}^N W_{i,j} \times x_i.$$

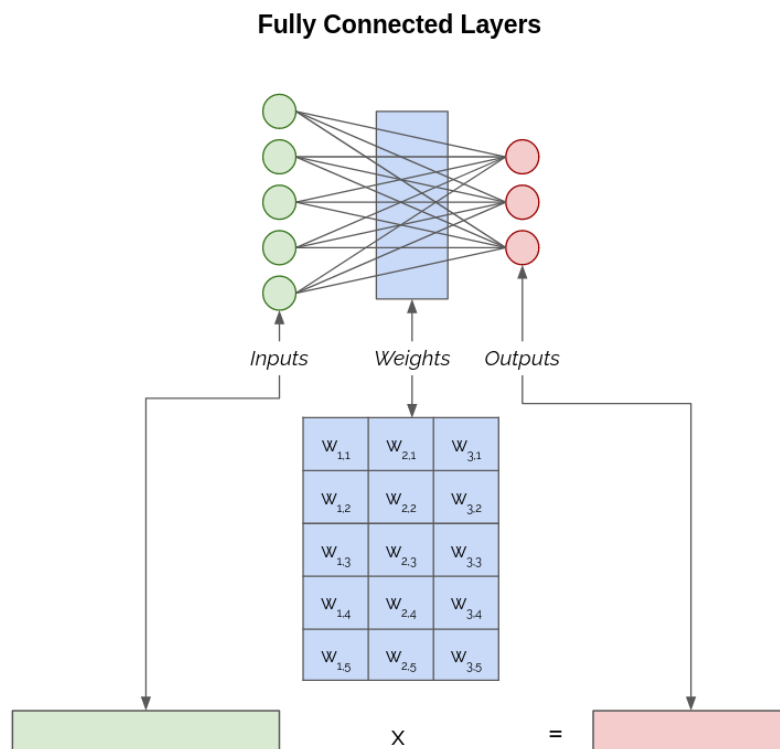


Figure B.2: Fully connected layer.

Mostly due to computational power limitations that have been overcome with the use of Graphical Processors Units (GPUs) (e.g. two Nvidia Geforce GTX) for training (method introduced in [Raina, Madhavan, and Ng 2009]), these architectures had to wait for 2012 and AlexNet [Krizhevsky, Sutskever, and Hinton 2012] to be proved efficient at a large scale.

AlexNet architecture (see Figure B.3), a deeper (more kernels per layer) version on LeNet from [LeCun et al. 1989], consists of a series of a  $11 \times 11$  convolution layers with 96 filters, a  $3 \times 3$  max pooling layer, a  $5 \times 5$  convolution layer with 256 filters, a  $3 \times 3$  max pooling layer, three  $3 \times 3$  convolutional layers (twice 384 and 256 filters), a  $3 \times 3$  max pooling layer. The feature map thus obtained is flattened and is iteratively given to two 4096-fully connected layers. Finally a 1000-fully connected layer outputs pre-output logits that are turned into class probabilities through a softmax layer (defined in Section B.2).

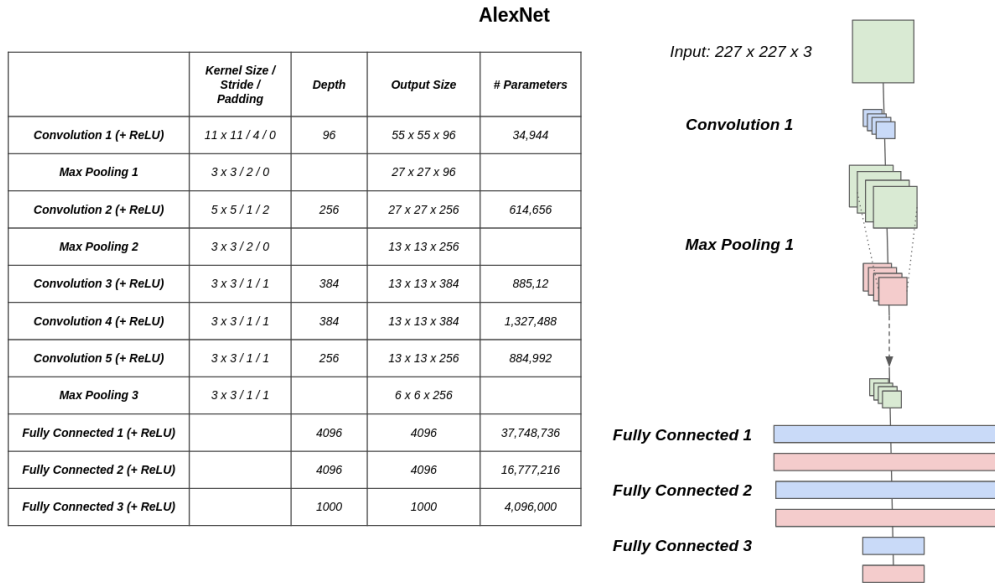


Figure B.3: AlexNet architecture.

## B.2 Supervised learning

Supervised training (see Figure B.4) consists in iteratively passing examples (e.g. images) through the network to get the predicted output, compute the error (with respect to the ground truth label) through a loss function, and then adapt the weights to reduce the error. This is done through backpropagation and gradient descent [Rumelhart, Hinton, and Williams 1986].

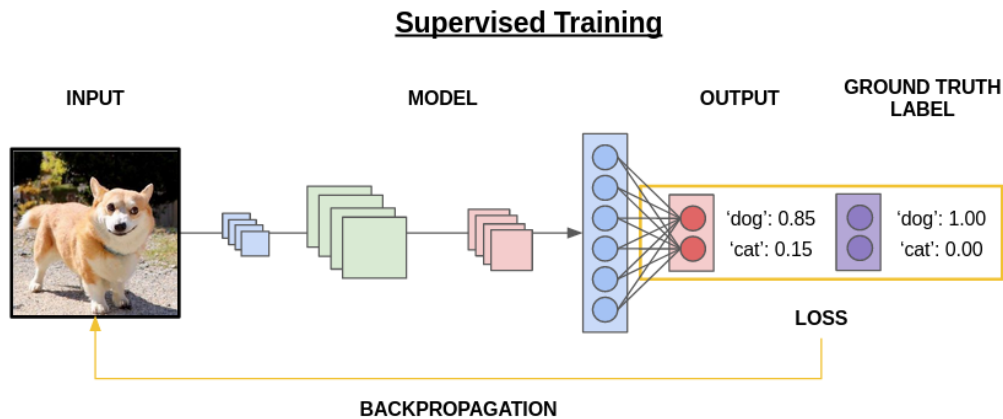


Figure B.4: Supervised learning.

Given an input  $X$  and its associated label  $Y^*$ , the model  $f_{model}$  outputs a predicted label  $Y = f(X)$ . The error  $L$  is computed using a loss function  $f_{loss}$ :  $L = f_{loss}(Y, Y^*)$ . Using a gradient descent-based optimizer, the inner-parameters  $\theta$  are updated. For example,

a standard method is Stochastic Gradient Descent (SGD) which consists in computing the partial derivatives of the loss function  $f_{loss}$  with respect to the parameters of the model  $\theta$  at the point of interest  $X$  and updating each parameter with a “step” in the opposite direction of this gradient. Thus  $\theta$  becomes  $\theta - \eta \Delta_{\theta} f_{loss}(Y, Y^*)$ . This step is repeated until a stopping criterion is reached. Parameter  $\eta$  is called learning rate, and defines the “size of the step” that is taken at each iteration (it will be further explained later in Section B.3). This is called backpropagation, defined by the optimization method (optimizer) and the loss function.

The loss function depends on the task that the model is trained to perform. For classification, the most standard loss is categorical cross-entropy which consists of a softmax layer and a cross-entropy loss computation. The softmax layer turns values from a vector  $p_i$  of size  $N$  into a value  $P_i$  as follows:  $P_i = \frac{e^{p_i}}{\sum_{j=1}^N e^{p_j}}$ . A great interest of softmax layer is that it outputs normalized values ( $\sum_i P_i = 1$ ) and can be interpreted as class probabilities. The cross-entropy, inspired from entropy measure in information theory, quantifies the difference between two probability vectors. Given a ground-truth vector  $G$  and a probability vector  $P$ , the cross-entropy is defined as:  $\mathcal{CE}(P, G) = -\sum_{i=1}^N G_i \times \log(P_i)$ . Thus, the derivative of the categorical softmax cross-entropy w.r.t. the softmax input  $p$  is:

$$\frac{\partial \mathcal{CE}(P, G)}{\partial p} = \begin{cases} \frac{e^{pg}}{\sum_{j=1}^N e_j^p} - 1; & \text{at position } g \text{ corresponding of the ground-truth label index} \\ \frac{e^{pn}}{\sum_{j=1}^N e_j^p}; & \text{at position } n \text{ for } n \neq g \end{cases}$$

The same kind of calculus can be made for layers defined above to compute errors and gradients for every parameter of the architectures.

An optimizer is used to update weights with regards to the computed gradients. As an example above, we presented SGD that is the most standard optimization method used in deep learning. It simply consists in iteratively updating weights with gradients computed on a single training sample. On the other hand, Batch Gradient Descent (or Vanilla Gradient Descent) computes and accumulates the gradients over all training samples before updating the weights. Batch Gradient Descent ensures a convergence to a local minimum but its computational cost makes it not suitable for large-scale deep learning applications. By contrast, SGD is way lighter in terms of computational cost, but is more noisy in its convergence, although it has been shown to be pretty efficient for deep learning applications. Mini-batch Gradient Descent proposes a compromise between these two approaches by updating parameters with gradients accumulated on a small subsample of training samples (a mini-batch). Thus it converges more efficiently and remains light in its computation.

SGD is the basis of a lot of works that improved it. For example, a standard extension of SGD is Momentum [Qian 1999]. It consists, as its name hints, in adding to the gradient direction currently computed on the current training sample a fraction of the previously computed gradient, thus enabling to avoid oscillating and making the model converge faster.

The main limitation using these methods is that the results highly depend on the learning rate value. Indeed, a too high learning rate would imply a divergence of the loss, while a too low learning rate would produce a slow convergence and would fall into a local minimum (see

Figure B.5).

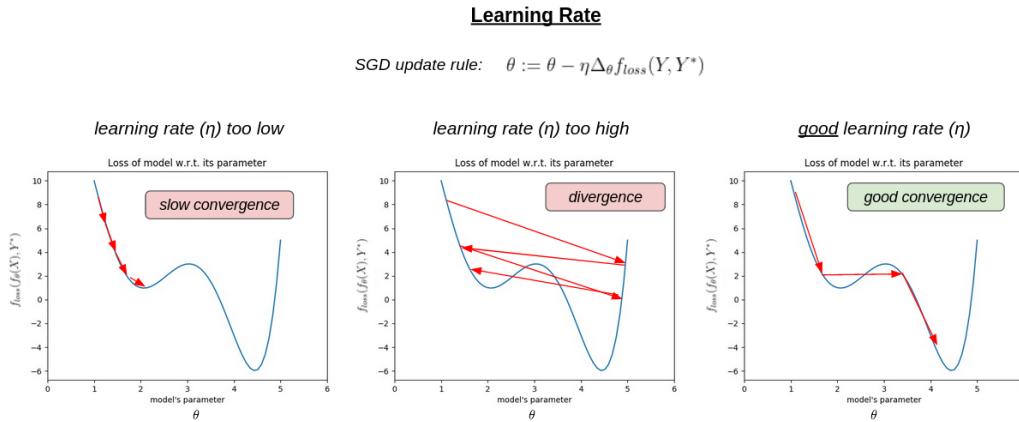


Figure B.5: Learning rate impact on model convergence.

Thus, some methods propose adaptive learning rate approaches. For example AdaGrad [Duchi, Hazan, and Singer 2011] proposes, in addition to apply a different learning rate to each parameter, to automatically adapt the learning rate with respect to the past gradient. For each parameter  $\theta_i$ , the learning rate at an iteration  $T$  is:  $\frac{\eta}{\sqrt{\sum_{t=0}^T (G_{i,t})^2 + \epsilon}}$  with  $G_{i,t}$  the gradient w.r.t. parameter  $\theta_i$ ,  $\eta$  the initial global learning rate and  $\epsilon$  a positive term to avoid dividing by zero. Adam [Kingma and Ba 2015] extends Adagrad by computing an average exponentially decay of past squared gradients (AdaDelta [Zeiler 2012]) and an average exponentially decay of past gradients for momentum computation. It is considered as the gold standard of optimizers.

[Ruder 2016] gives a good overall study of most popular gradient-based optimization methods.

### B.3 Training strategies to reduce overfitting: regularization and transfer learning

The dataset, e.g. a set of {image, label} couples also called samples, is generally divided into three sub-sets:

1. The training set: samples used for the actual training of the model, i.e. to compute gradients and perform backpropagation. The model is trained to perform on these samples. This set represents the majority of the dataset (generally around 70%);
2. The validation set: samples used to test the model regularly during the training process to ensure its good learning capabilities, and detect problems that can happen during training. It is particularly used to ensure the generalizability of features learned by

the model. Indeed, because of the high number of parameters, models may learn non-generalizable training set biases (knowledge that is not transferable to other sets), and thus learn to perform really well on training samples (since it is optimized on them) while performing poorly on other sets: this is called overfitting (see Figure B.6). The validation set enables to check regularly during training the generalizability of what is learned by the model on the training set. It is the smallest sub-set of the dataset (generally around 10%);

3. The testing set: it is kept completely independent from the training process and is used to evaluate the model, i.e. to compute measures of performances of the model for the task it has been trained on. It gathers generally around 20% of the dataset.

Figure B.6 illustrates the overfitting phenomenon due to over parametrization. Generally in deep learning the dataset is not large enough in comparison to the millions of parameters, so overfitting is often a good statistical solution for a model during training. To ensure the generalizability of what is learned, we generally rely on the validation set and more generally on what is called regularizations [Caruana, Lawrence, and Giles 2000] that are methods to avoid or at least fight the overfit.

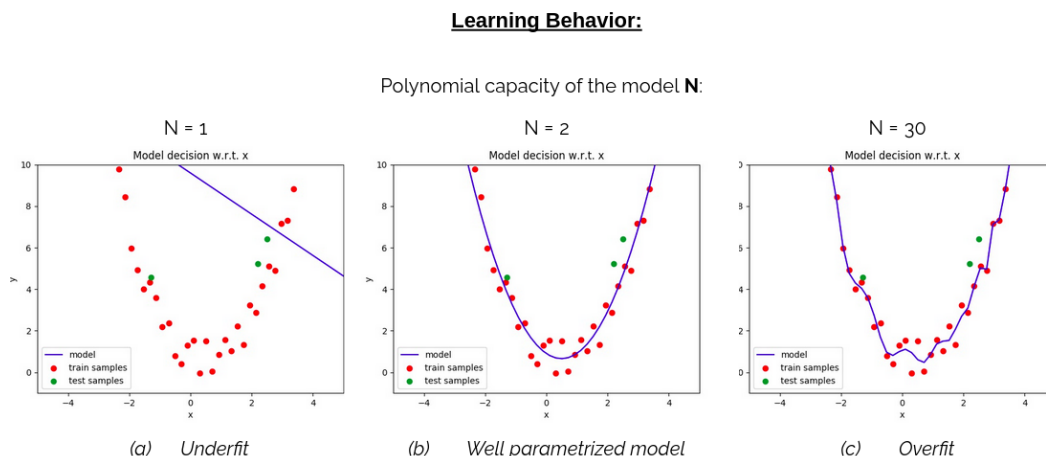


Figure B.6: Illustration of overfitting on a second degree polynomial problem; (a) underfit: the model has not enough learning capacity; (b) the model is well designed for the problem; (c) overfit: the model overfits the problem and matches training samples.

Indeed, the validation set is made to measure the generalizability during training. Thus monitoring the average loss over the validation set during training allows detecting the overfit. Figure B.7 shows the evolution of the average loss on the training set and on the validation set computed at regular intervals. Overfit can be “diagnosed” when these two measures diverge. During the optimization, the average loss on the training set decreases, and at some point the average loss on the validation set starts to increase, which means that what is learned does not apply to this set anymore. As illustrated, these measures can be used to perform

early stopping, which consists in stopping the training right before these two losses start to diverge. It is implemented using a patience threshold, i.e. a number of intervals during which the minimum validation loss obtained should decrease at least once.

**Early Stopping:**

Evolution of average loss on training and validation set

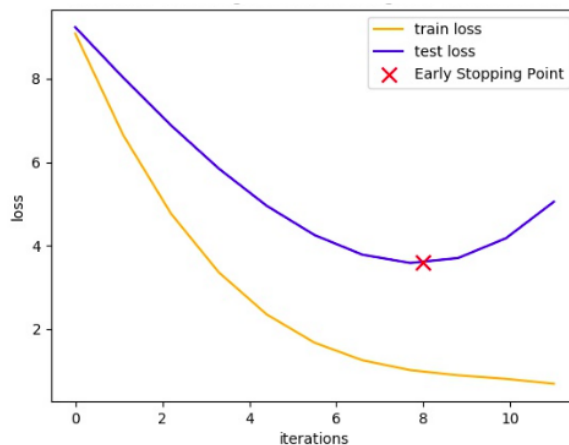


Figure B.7: Detecting overfit with average loss computed on the validation set.

There are other ways to regularize the training of a model through the loss, the architecture design or by transfer learning in order to inducing knowledge or expected behavior. Note that regularization techniques are used during training and fixed at inference time.

For instance, a popular regularization consists in adding to the training loss a  $L^2$  norm term computed on the squared weights of the model. The total loss thus becomes  $L + \lambda \cdot \|W\|^2$  with  $L$  the task related loss,  $\lambda$  the importance given to the regularization and  $W$  the model weight matrix. Training a model with this regularization will guide the model to have weights of the models to be small (close to zero), which disables the capacity of the model to give too much importance to specific features.

Along the same idea of having all weights contributing to the prediction (and not giving too much importance to a subset of weights) with a more design-driven approach, dropout [Srivastava et al. 2014] regularization proposes to cut some connections, i.e. to set to zero a given percentage of weights during training (see Figure B.8). It forces the network to learn how to rely on all connections to predict and avoid “shortcuts” i.e. path of strong connections that fit the training data.

Batch normalization [Ioffe and Szegedy 2015] (and later group normalization [Wu and He 2018]) is a method that proposes to standardize feature maps so that the network does not



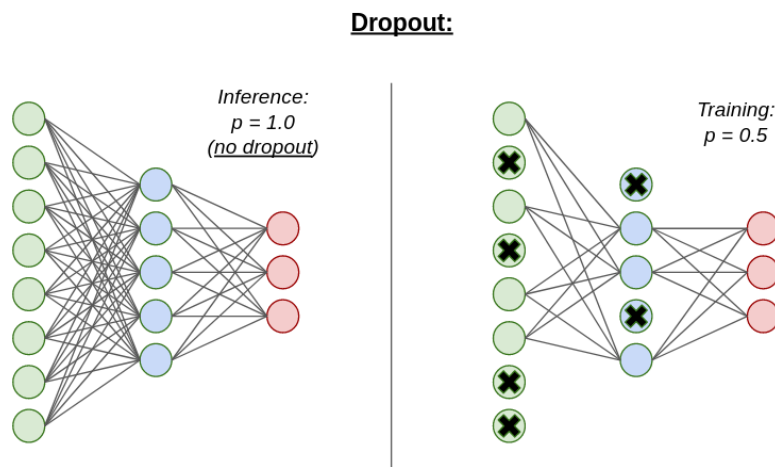


Figure B.8: Dropout illustration; Fully connected layers without dropout (left); Fully connected layers with 50% dropout (right).

have to adapt to distribution shifts (also called covariate shift), which makes deeper neurons more robust to early neuron changes, and in practice speed up and stabilize trainings. This method consists in learning standardization parameters  $\gamma$  for scaling and  $\beta$  for shifting that are applied as follows for a sample input  $x$  of a batch  $B$ :  $BatchNorm_{\gamma,\beta}(x) = \gamma \cdot \frac{x - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}} + \beta$  with  $\mu_B$  the mean over the batch and  $\sigma_B^2$  the standard deviation over the batch. For inference, precomputed values of  $\mu$  and  $\sigma^2$  are used. Note that a simple batch normalization trick enabled to improve performances on ImageNet of 0.7%, leading to a new state-of-the-art method reaching a top-1 accuracy of 85.5% [Xie et al. 2020b].

Another regularization method that is used in most works is called transfer learning (see Figure B.9). It consists in using a model trained on a (source) task to help training a second model on another (destination) task. A great percentage of medical applications of deep learning are performing transfer learning from a model pre-trained on ImageNet. It might seem surprising to transfer knowledge and feature learned on natural images to medical specific applications, but due to the very high number of classes and images per class in ImageNet, dataset features learned on early layers by an efficient model can not afford to be class-specific and need to be applicable for a wide range of cases. In practice this approach works really well and enables to avoid quick overfit of the model due to a generally small dataset.

There are two main methods to perform transfer learning. The first method reminds of early Machine Learning (ML) methods and consist in using the pre-trained model to encode descriptors that are then used to train a classifier (generally some fully connected layers). The second method, called “fine-tuning”, simply uses the pre-trained model as the starting point for the training of the second model, and thus the learned filters will be adapted to the new task. This requires source task and destination task to have the same number of classes. Both methods can be used in the same training by fine-tuning the model while training a

classifier.

Interestingly, transfusion work [Raghu et al. 2019] exposed that pre-training on ImageNet was purely regularization since the same performances on several medical datasets were obtained with a deep architecture pre-trained on ImageNet and with a smallest architecture initialized with random weights. Using interpretability methods (see more details in Section 3.4) the authors showed that features learned and used by both models in early low-level layers were the same. More recently, transfer learning from a pre-trained network has been questioned and outperformed using self-supervision in [Zoph et al. 2019].

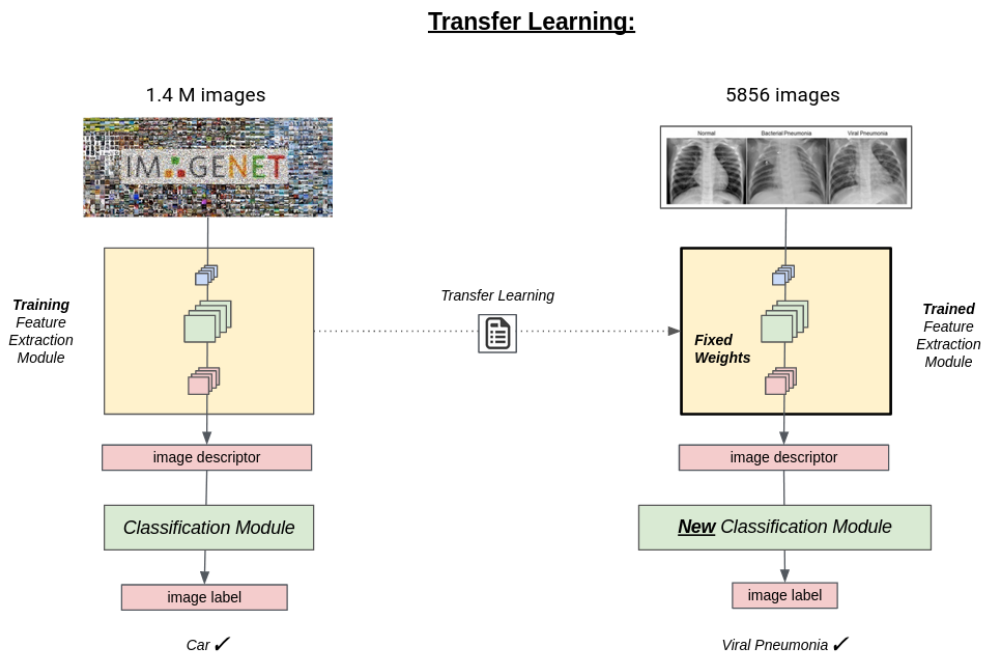


Figure B.9: Transfer learning: transfer from ImageNet to Chest X-Ray Pneumonia dataset (<https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia>).

Finally, a popular practice in computer vision challenges is called ensembling and consists in training several models and combining their predictions to obtain more generalizable results (e.g. [Breiman 1996]).

## B.4 Other kinds of learning

There are different levels of supervision that can be used for training:

1. Supervised training (presented extensively above) refers to training processes where every input is paired with its expected output or ground truth;
2. Weakly supervised training is a training where the expected output is not given but a

more global information is used (e.g. only image label is given to perform localization);

3. Semi-supervised training defines a training where some inputs have their ground truth labels and some others do not;
4. Unsupervised learning is a training where no information except the inputs is given.



# Bibliography

- Adebayo, J. et al. (2018). “Sanity Checks for Saliency Maps”. In: NIPS’18, pp. 9525–9536.
- Ahmady Phoulady, H. and P. R. Mouton (2018). “A New Cervical Cytology Dataset for Nucleus Detection and Image Classification (Cervix93) and Methods for Cervical Nucleus Detection”. In: abs/1811.09651. URL: <http://arxiv.org/abs/1811.09651>.
- Ancona, M. et al. (2017). “Towards better understanding of gradient-based attribution methods for Deep Neural Networks”. In: arXiv: 1711.06104 [cs.LG].
- Babes, A. (1928). “Diagnostic du cancer du col de uterin par les frottis”. In: *Presse Medicale* 36, pp. 451–454.
- Barker, J. et al. (2016). “Automated classification of brain tumor type in whole-slide digital pathology images using local representative tiles”. In: *Medical Images Analysis* 20, pp. 60–71. DOI: 10.1016/j.media.2015.12.002.
- Bel, T. de et al. (2019). “Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology”. In: *Medical Imaging with Deep Learning Conference*.
- Bera, K., K. A. Schalper, and D. L. et al. Rimm (2019). “Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology”. In: *Nature Reviews Clinical Oncology* 16, pp. 703–715. DOI: 10.1038/s41571-019-0252-y.
- Bora, K. et al. (2016). “Pap Smear Image Classification Using Convolutional Neural Network”. In: *10th Indian Conference on Computer Vision, Graphics and Image Processing* 55, pp. 1–8.
- Breiman, L. (1996). “Bagging predictors”. In: *Mach. Learn.* 24, pp. 123–140. DOI: 10.1007/BF00058655.
- Breiman, L. et al. (1984). “Classification and regression trees”. In: *Wadsworth and Brooks/Cole Advanced Books and Software*.
- Brennan, R. L. and D. J. Prediger (1981). “Coefficient Kappa: Some Uses, Misuses, and Alternatives”. In: *Educational and Psychological Measurement* 41(3), pp. 687–699.
- Brieman, L. (2001). “Random Forests”. In: *Machine Learning* 45, pp. 5–32. DOI: 10.1023/A:1010933404324.
- Bristow, H. and S. Lucey (2014). “Why do linear SVMs trained on HOG features perform so well?” In: arXiv: 1406.2419 [cs.CV].
- Cabrera, A. F. (1994). “Logistic regression analysis in higher education: an applied perspective”. In: *Higher Education: Handbook of Theory and Research* 10, pp. 225–256.
- Campanella, G., V. W. K. Silva, and T. J. Fuchs (2018). “Terabyte-scale Deep Multiple Instance Learning for Classification and Localization in Pathology”. In: *Computing Research Repository (CoRR), Arxiv*. eprint: 1805.06983.
- Campanella, G. et al. (2019). “Clinical-grade computational pathology using weakly supervised deep learning on whole slide images”. In: *Nature Medicine* 25, p. 1.
- Caruana, R., S. Lawrence, and L. Giles (2000). “Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping”. In: *Proceedings of the 13th International Conference on Neural Information Processing Systems*, pp. 381–387.

- Chen, C. L., C. C. Chen, and W.H. et al Yu (2021). “An annotation-free whole-slide training approach to pathological classification of lung cancer types using deep learning”. In: *Nature Communications* 12. DOI: 10.1038/s41467-021-21467-y.
- Chen, T. et al. (2020a). “A Simple Framework for Contrastive Learning of Visual Representations”. In: *Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research* 119, pp. 1597–1607.
- Chen, X. et al. (2020b). “Improved Baselines with Momentum Contrastive Learning”. In: *ArXiv* abs/2003.04297.
- Cheng, Hsien-Tzu et al. (2020). “Self-similarity Student for Partial Label Histopathology Image Segmentation”. In: arXiv: 2007.09610 [eess.IV].
- Cheng, J., Z. Wang, and G. Pollastri (2008). “A neural network approach to ordinal regression”. In: *IEEE International Joint Conference on Neural Networks*, pp. 1279–1284.
- Chicco, D. and G. Jurman (2020). “The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation”. In: *BMC Genomics* 21(6). DOI: 10.1186/s12864-019-6413-7.
- Chivukula, M. and V. Shidham (2006). “ASC-H in Pap test—definite categorization of cytomorphological spectrum”. In: *CytoJournal*. DOI: 10.1186/1742-6413-3-14.
- Ciampi, F. et al. (2017). “The importance of stain normalization in colorectal tissue classification with convolutional networks”. In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. DOI: 10.1109/isbi.2017.7950492. URL: <http://dx.doi.org/10.1109/ISBI.2017.7950492>.
- Combalia, M. and V. Vilaplana (2018). “Monte-Carlo Sampling applied to Multiple Instance Learning for Histological Image Classification”. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*.
- Cortes, C. and V. Vapnik (1995). “Support-vector networks”. In: *Machine Learning* 20(3), pp. 273–297. DOI: 10.1007/bf00994018.
- Coudray, N., P.S. Ocampo, and T. et al Sakellaropoulos (2018). “Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning”. In: *Nature Medicine* 24, pp. 1559–1567. DOI: 10.1038/s41591-018-0177-5.
- Courtiol, P. et al. (2018). “Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach”. In: *Computing Research Repository (CoRR), Arxiv*.
- Couteaux, V. et al. (2019). “Towards Interpretability of Segmentation Networks by analyzing DeepDreams”. In: *iMIMIC Workshop at MICCAI 2019: Interpretability of Machine Intelligence in Medical Image Computing*, pp. 56–63.
- Dalal, N. and B. Triggs (2005). “Histograms of oriented gradients for human detection”. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 1, pp. 886–893. DOI: 10.1109/CVPR.2005.177.
- Dehaene, O. et al. (2020). “Self-Supervision Closes the Gap Between Weak and Strong Supervision in Histology”. In: *ArXiv* abs/2012.03583.
- Deng, J. et al. (2009). “ImageNet: A large-scale hierarchical image database”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255.

- Diamantis, A., E. Magiorkinis, and G. Androutsos (2010). “What’s in a name? Evidence that Papanicolaou, not Babes, deserves credit for the Pap test”. In: *Diagnosis Cytopathology* 38(7), p. 473. DOI: 10.1002/dc.21226.
- Diaz, R. and A. Marathe (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. et al. (2021). “Weakly supervised instance learning for thyroid malignancy prediction from whole slide cytopathology images”. In: *Medical Image Analysis* 67, pp. 772–780. DOI: 10.1016/j.media.2020.101814.
- Dreiseitl, S. and L. Ohno-Machado (2002). “Logistic regression and artificial neural network classification models: a methodology review”. In: *Journal of Biomedical Informatics* 35(5), pp. 352–359. DOI: 10.1016/S1532-0464(03)00034-0.
- Duchi, J., E. Hazan, and Y. Singer (2011). “Adaptive Subgradient Methods for Online Learning and Stochastic Optimization”. In: *Journal of Machine Learning Research* 12, pp. 2121–2159.
- Durand, T., N. Thome, and N. Cord (2016). “WELDON: Weakly Supervised Learning of Deep Convolutional Neural Network”. In: *29th IEEE Conference on Computer Vision and Pattern Recognition*.
- Eggert, C. et al. (2017). “A closer look: Small object detection in faster R-CNN”. In: *IEEE International Conference on Multimedia and Expo*. DOI: 10.1109/icme.2017.8019550.
- Ehteshami Bejnordi, B. et al. (2016). “Stain Specific Standardization of Whole-Slide Histopathological Images”. In: *IEEE Transactions on Medical Imaging* 35(2), pp. 404–415. DOI: 10.1109/tmi.2015.2476509.
- Ehteshami Bejnordi, B. et al. (2017). “Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer”. In: *Journal of the American Medical Association* 312(22), pp. 2199–2210.
- Everingham, M., L. Van Gool, and C. K. I. et al. Williams (2010). “The Pascal Visual Object Classes (VOC) Challenge”. In: *International Journal of Computer Vision* 88, pp. 303–338.
- Farahani, N., A. Parwani, and L. Pantanowitz (2015). “Whole slide imaging in pathology: advantages, limitations, and emerging perspectives”. In: *Pathology and Laboratory Medicine International* 7, pp. 23–33. DOI: 10.2147/PLMI.S59826.
- Fellbaum, Christiane (1998). “WordNet: An Electronic Lexical Database”. In: *Language, Speech, and Communication*.
- Felzenszwalb, P. and D. Huttenlocher (2004). “Efficient Graph-Based Image Segmentation”. In: *International Journal of Computer Vision* 59(2), pp. 167–181.
- Fong, R. C. and A. Vedaldi (2017). “Interpretable Explanations of Black Boxes by Meaningful Perturbation”. In: *IEEE International Conference on Computer Vision*, pp. 3449–3457.
- Forslid, G. et al. (2017). “Deep Convolutional Neural Networks for Detecting Cellular Changes Due to Malignancy”. In: *IEEE International Conference on Computer Vision Workshops*, pp. 82–89.
- Freeman, W. T. and M. Roth (1995). “Orientation Histograms for Hand Gesture Recognition”. In: *International Workshop on Automatic Face and Gesture Recognition*.

- Freund, Y. and R. E. Schapire (1997). “A decision-theoretic generalization of on-line learning and an application to boosting”. In: *Journal of Computer and System Sciences* 55(1), pp. 119–139.
- Fu, Y. et al. (2020). “Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis”. In: *Nature Cancer*. DOI: 10.1038/s43018-020-0085-8.
- Fukushima, K. (1980). “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position”. In: *Biological cybernetics* 36(4), pp. 193–202.
- Gal, Y. and Z. Ghahramani (2016). “Dropout as a Bayesian approximation: Representing model uncertainty in deep learning”. In: *International Conference on Machine Learning*, pp. 1050–1059.
- Gay, J. D., L. D. Donaldson, and J. R. Goellner (1985). “False-negative results in cervical cytologic studies”. In: *Acta Cytologica* 29, pp. 1043–1046.
- Girshick, R. (2015). “Fast R-CNN”. In: *2015 IEEE International Conference on Computer Vision (ICCV)*. DOI: 10.1109/iccv.2015.169. URL: <http://dx.doi.org/10.1109/ICCV.2015.169>.
- Girshick, R. et al. (2014). “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation”. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. DOI: 10.1109/cvpr.2014.81. URL: <http://dx.doi.org/10.1109/CVPR.2014.81>.
- Goh, G. S. W. et al. (2020). “Understanding Integrated Gradients with SmoothTaylor for Deep Neural Network Attribution”. In: *Computing Research Repository (CoRR), Arxiv*. URL: <https://arxiv.org/abs/2004.10484>.
- Goodfellow, I. et al. (2014). “Generative Adversarial Networks”. In: *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 2672–2680.
- Graziani, M. et al. (2020). “Guiding CNNs towards Relevant Concepts by Multi-task and Adversarial Learning”. In: arXiv: 2008.01478 [cs.CV].
- Harada, T. and Y. Kuniyoshi (2012). “Graphical Gaussian Vector for Image Categorization”. In: *Advances in Neural Information Processing Systems 25*, pp. 1547–1555.
- Harinarayanan, K. K. and J. Nirmal (2018). “Classification driven Assisted Screening for cervical cancer using deep neural network”. In:
- Harwood, D. et al. (1995). “Texture classification by center-symmetric auto-correlation, using Kullback discrimination of distributions”. In: *Pattern Recognition Letters* 16(1), pp. 1–10. DOI: 10.1016/0167-8655(94)00061-7.
- Hata, H. et al. (2019). “A Comparison of Cytomorphological Features of ASC-H Cells Based on Histopathological Results Obtained from a Colposcopic Target Biopsy Immediately after Pap smear Sampling.” In: *Asian Pacific Journal of Cancer Prevention* 20(7), pp. 2139–2143. DOI: 10.31557/APJCP.2019.20.7.2139.
- He, K. et al. (2016). “Deep Residual Learning for Image Recognition”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.
- Heard, T. et al. (2012). “Use of the ThinPrep Imaging System for internal quality control of cervical cytology”. In: *Cytopathology : official journal of the British Society for Clinical Cytology* 24. DOI: 10.1111/*\emph{cyt}*.12010.



- Hinterreiter, A., M. Streit, and B. Kainz (2020). “Projective Latent Interventions for Understanding and Fine-tuning Classifiers”. In: *Workshop on Interpretability of Machine Intelligence in Medical Image Computing at MICCAI 2020*.
- Hoffman, R.A. et al. (2014). “A High-Resolution Tile-Based Approach for Classifying Biological Regions in Whole-Slide Histopathological Images”. In: *International Federation for Medical and Biological Engineering proceedings* 42, pp. 280–283. DOI: 10.1007/978-3-319-03005-0\_71.
- Hooker, S. et al. (2018). “Evaluating Feature Importance Estimates”. In: *arXiv*. URL: <https://arxiv.org/pdf/1806.10758.pdf>.
- Hoon Tan, P. et al. (2019). “Breast Tumours, WHO Classification of tumours”. In: *International Collaboration on Cancer Reporting (5th edition)*.
- Howard, A. G. et al. (2017). “Mobilenets: Efficient convolutional neural networks for mobile vision applications”. In: arXiv: 1704.04861.
- Hu, L. et al. (2019). “An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening”. In: *Journal of the National Cancer Institute* 111,9, pp. 923–932.
- Huang, G. et al. (2017a). “Densely Connected Convolutional Networks”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.1109/cvpr.2017.243. URL: <http://dx.doi.org/10.1109/CVPR.2017.243>.
- Huang, J. et al. (2017b). “Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.1109/cvpr.2017.351. URL: <http://dx.doi.org/10.1109/CVPR.2017.351>.
- Hubel, D. H. and T. N. Wiesel (1962). “Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex”. In: *The Journal of physiology* 160(1), pp. 106–154.
- Hyeon, J. et al. (2017). “Diagnosing cervical cell images using pre-trained convolutional neural network as feature extractor”. In: *IEEE International Conference on Big Data and Smart Computing*, pp. 390–393.
- Ianni, J. D. et al. (2020). “Tailored for Real-World: A Whole Slide Image Classification System Validated on Uncurated Multi-Site Data Emulating the Prospective Pathology Workload”. In: *Nature Scientific Report* 10.
- Idlahcen, F., M. M. Himmi, and A. Mahmoudi (2020). “CNN-based approach for cervical cancer classification in whole-slide histopathology images”. In: *ICLR Workshop on AI for Overcoming Global Disparities in Cancer Care*.
- Iizuka, O., F. Kanavati, and K. et al Kato (2020). “Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours”. In: *Scientific Reports* 10. DOI: 10.1038/s41598-020-58467-9.
- Ilse, M., J. M. Tomczak, and M. Welling (2018). “Attention-based deep multiple instance learning”. In: *Proceedings of the International Conference on Machine Learning*.
- Ioffe, S. and C. Szegedy (2015). “Batch normalization: accelerating deep network training by reducing internal covariate shift”. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning* 37, pp. 448–456.
- ISBI (2015). “The Second Overlapping Cervical Cytology Image Segmentation Challenge”. In: URL: [https://cs.adelaide.edu.au/~zhi/isbi15\\_challenge/](https://cs.adelaide.edu.au/~zhi/isbi15_challenge/).

- Janerich, D. T., O. Hadjimichael, and P. E. Schwarz (1995). "The screening histories of women with invasive cervical cancer". In: *American Journal of Public Health* 85, pp. 791–794.
- Jantzen, J. et al. (2005). "Pap-smear Benchmark Data For Pattern Classification". In: *Nature inspired Smart Information Systems*, pp. 1–9.
- Jegou, H. et al. (2010). "Aggregating local descriptors into a compact image representation". In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8.
- Joste, N. E., C. P. Crum, and E. S. Cibas (1995). "Cytologic/histologic correlation for quality control in cervicovaginal cytology: Experience with 1582 paired cases". In: *American Journal of Clinical Pathology* 103, pp. 32–34.
- Karimi-Zarchi, M. et al. (2013). "A Comparison of 3 Ways of Conventional Pap Smear, Liquid-Based Cytology and Colposcopy vs Cervical Biopsy for Early Diagnosis of Premalignant Lesions or Cervical Cancer in Women with Abnormal Conventional Pap Test". In: *International Journal of Biomedical Science* 9(4), pp. 205–215.
- Karras, T., S. Laine, and T. Aila (2019). "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4401–4410.
- Kather, J. N. et al. (2020). "Pan-cancer image-based detection of clinically actionable genetic alterations". In: *Nature Cancer*. DOI: 10.1038/s43018-020-0087-6.
- Khan, A. et al. (2014). "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution". In: *IEEE Transactions on Biomedical Engineering* 61(6), pp. 1729–1738.
- Kindermans, P. et al. (2019). "The (Un)reliability of Saliency methods". In: pp. 267–280.
- Kingma, D. P. and J. L. Ba (2015). "Adam: a Method for Stochastic Optimization". In: *International Conference on Learning Representations*, pp. 1–13.
- Kowalski, B. R. and C. F. Bender (1972). "K-Nearest Neighbor Classification Rule (pattern recognition) applied to nuclear magnetic resonance spectral interpretation". In: *Analytical Chemistry*, pp. 1405–1411. DOI: 10.1021/ac60316a008.
- Kristensen, G.B. et al. (1991). "Analysis of cervical smears obtained within three years of the diagnosis of invasive cervical cancer". In: *Acta Cytologica* 35, pp. 47–50.
- Krizhevsky, A., I. Sutskever, and G. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems* 25, pp. 1097–1105.
- Kumar, N., R. Gupta, and S. Gupta (2020). "Whole Slide Imaging (WSI) in Pathology: Current Perspectives and Future Directions". In: *Journal of Digital Imaging* 33(4), pp. 1034–1040. DOI: 10.1007/s10278-020-00351-z.
- Kwon, M. et al. (2018). "Multi-label Classification of Single and Clustered Cervical Cells Using Deep Convolutional Networks". In: *14th International Conference on Data Science*.
- Lazebnik, S., C. Schmid, and J. Ponce (2006). "Image Classification Using Super-Vector Coding of Local Image Descriptors". In: pp. 2169–2178. DOI: 10.1109/CVPR.2006.68..
- LeCun, Y. et al. (1989). "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1(4), pp. 541–551.
- Lee, B. and K. Paeng (2018). "A Robust and Effective Approach Towards Accurate Metastasis Detection and pN-stage Classification in Breast Cancer". In: *CoRR* abs/1805.12067.

- Li, B., Y. Li, and K. W. Eliceiri (2020). “Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning”. In: arXiv: 2011.08939 [cs.CV].
- Li, J. et al. (2019). “An attention-based multi-resolution model for prostate whole slide image classification and localization”. In: arXiv: 1905.13208 [cs.CV].
- Li, Y., J. Wu, and Q. Wu (2019). “Classification of Breast Cancer Histology Images Using Multi-Size and Discriminative Patches Based on Deep Learning”. In: *IEEE Access* 7, pp. 21400–21408.
- Lin, H. et al. (2019). “Fine-Grained Classification of Cervical Cells Using Morphological and Appearance Based Convolutional Neural Networks”. In: *IEEE Access*.
- Lin, T. et al. (2014). “Microsoft COCO: Common Objects in Context”. In: *Computer Vision - ECCV 2014*, pp. 740–755.
- Lin, T. et al. (2017). “Feature Pyramid Networks for Object Detection”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.1109/cvpr.2017.106. URL: <http://dx.doi.org/10.1109/CVPR.2017.106>.
- Lin, T. et al. (2020). “Focal Loss for Dense Object Detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42,2, pp. 318–327.
- Lin, Y. et al. (2010). “Imagenet classification: fast descriptor coding and large-scale svm training”. In: *Large scale visual recognition challenge*.
- Lindsay, G. (2020). “Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future”. In: *Journal of Cognitive Neuroscience*, pp. 1–15. DOI: 10.1162/jocn\_a\_01544.
- Liu, W. et al. (2016). “SSD: Single Shot MultiBox Detector”. In: *Lecture Notes in Computer Science*, pp. 21–37. ISSN: 1611-3349. DOI: 10.1007/978-3-319-46448-0\_2. URL: [http://dx.doi.org/10.1007/978-3-319-46448-0\\_2](http://dx.doi.org/10.1007/978-3-319-46448-0_2).
- Liu, Y. et al. (2017). “Detecting Cancer Metastases on Gigapixel Pathology Images”. In: *International Journal of Computer Vision* 60(2), pp. 91–110.
- Lowe, D. G. (2004). “Distinctive Image Features from Scale-Invariant Keypoints”. In: *International Journal of Computer Vision* 60(2), pp. 91–110.
- Lu, Ming Y. et al. (2020). “Data Efficient and Weakly Supervised Computational Pathology on Whole Slide Images”. In: arXiv: 2004.09666 [eess.IV].
- Maaten, L. van der and G. Hinton (2008). “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9, pp. 2579–2605.
- Macenko, M. et al. (2009). “A method for normalizing histology slides for quantitative analysis”. In: *IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 23, pp. 1977–1986.
- Maji, S., A. C. Berg, and J. Malik (2008). “Classification using intersection kernel support vector machines is efficient”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. DOI: 10.1109/CVPR.2008.4587630.
- Maraqa, B., I. Lataifeh, and L. et al Otay (2017). “Prevalence of Abnormal Pap Smears: A Descriptive Study from a Cancer Center in a Low-Prevalence Community”. In: *Asian Pacific Journal of Cancer Prevention* 18(11), pp. 3117–3121. DOI: 10.22034/APJCP.2017.18.11.3117.
- Maron, O. and T. Lozano-Pérez (1997). “A Framework for Multiple-Instance Learning”. In: *NIPS’97*, pp. 570–576.

- Marshall, P. N. (1983). “Papanicolaou staining—a review”. In: *Microscopica Acta* 87(3), pp. 233–243.
- McInnes, L. and J. Healy (2018). “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”. In: *ArXiv*. arXiv: 1802.03426.
- Meiquan, X. et al. (2018). “Cervical cytology intelligent diagnosis based on object detection technology”. In: *Medical Imaging with Deep Learning Conference*.
- Mikolajczyk, K. and C. Schmid (2004). “Scale and affine invariant interest point detectors”. In: *International Journal of Computer Vision* 60(1), pp. 63–86.
- Morell, N. D. et al. (1982). “False-negative cytology rates in patients in whom invasive cervical cancer subsequently developed”. In: *Obstetrics and Gynecology* 60, pp. 41–45.
- Mu, Y. et al. (2008). “Discriminative local binary patterns for human detection in personal album”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8.
- Nachar, N. (2008). “The Mann Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution”. In: *Tutorials in Quantitative Methods for Psychology* 4(1), pp. 13–20.
- Naik, N., A. Madani, and A. et al. Esteva (2020). “Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains”. In: *Nature Communications* 11. DOI: 10.1038/s41467-020-19334-3.
- Nair, V. and G. E. Hinton (2010). “Rectified Linear Units Improve Restricted Boltzmann Machines”. In: *Proceedings of the 27th International Conference on Machine Learning*, pp. 21–24.
- Nanda, K., D. C. McCrory, and E. R. et al Myers (2000). “Accuracy of the Papanicolaou test in screening for and follow-up of cervical cytologic abnormalities: a systematic review”. In: *Annals of Internal Medicine* 132, pp. 810–819.
- Nayar, R. and C. D. Wilbur (2015). “The Pap Test and Bethesda 2014”. In: *Acta Cytologica* 59, pp. 121–132.
- Naylor, P. et al. (2019). “Predicting Residual Cancer Burden in a triple negative breast cancer cohort”. In: *IEEE 16th International Symposium on Biomedical Imaging*, pp. 933–937.
- Nie, W., Y. Zhang, and A. Patel (2018). “A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations”. In: *Proceedings of Machine Learning Research* 80. Ed. by Jennifer Dy and Andreas Krause, pp. 3809–3818. URL: <http://proceedings.mlr.press/v80/nie18a.html>.
- Nishat, R. et al. (2017). “Digital cytopathology”. In: *Journal of Oral and Maxillofacial Pathology* 21(1), pp. 99–106. DOI: 10.4103/0973-029X.203767.
- Olah, C., A. Mordvintsev, and L. Schubert (2017). “Feature Visualization”. In: *Distill*. <https://distill.pub/2017/visualization>. DOI: 10.23915/distill.00007.
- Olah, C. et al. (2020). “Zoom In: An Introduction to Circuits”. In: *Distill*. <https://distill.pub/2020/circuits/zoom-in>. DOI: 10.23915/distill.00024.001.
- Otsu, N. (1979). “A threshold selection method from gray-level histograms”. In: *IEEE Transactions on Systems, Man, and Cybernetics* 9, pp. 62–66. DOI: 10.1109/TSMC.1979.4310076.
- Papanicolaou, G. N. and H. F. Traut (1943). “Diagnosis of Uterine Cancer by the Vaginal Smear”. In: *The Yale Journal of Biology and Medicine* 15(6), p. 127.

- Pathologie (SFP), Société Française de (2020). “TissueNet: Detect Lesions in Cervical Biopsies”. In: URL: <https://www.drivendata.org/competitions/67/competition-cervical-biopsy/>.
- Peng, H., F. Long, and C. Ding (2005). “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, pp. 1226–1238.
- Perronnin, F., J. Sánchez, and T. Mensink (2010). “Improving the Fisher Kernel for Large-Scale Image Classification”. In: *11th European Conference on Computer Vision*. DOI: 10.1007/978-3-642-15561-1\_11.
- Pinckaers, H., B. van Ginneken, and G. J. S. Litjens (2019). “Streaming convolutional neural networks for end-to-end learning with multi-megapixel images”. In: *ArXiv abs/1911.04432*.
- Polyak, B. T. and A. Juditsky (1992). “Acceleration of Stochastic Approximation by Averaging”. In: 30(4), pp. 838–855. DOI: 10.1137/0330046.
- Qian, N. (1999). “On the momentum term in gradient descent learning algorithms”. In: *Neural Networks : The Official Journal of the International Neural Network Society* 12(1), pp. 145–151. DOI: 10.1016/S0893-6080(98)00116-6.
- Qureshi, Sabuhi et al. (2017). “Liquid-based Cytology vs Conventional Cytology as a Screening Tool for Cervical Cancer in Postmenopausal Women”. In:
- Raffel, C. and D. P. W. Ellis (2015). “Feed-Forward Networks with Attention Can Solve Some Long-Term Memory Problems”. In: arXiv: 1512.08756 [cs.LG].
- Raghu, M. et al. (2019). “Transfusion: Understanding transfer learning for medical imaging”. In: *Proceedings of the Neural Information Processing Systems*, pp. 3342–3352.
- Rahaman, M. M. et al. (2020). “A Survey for Cervical Cytopathology Image Analysis Using Deep Learning”. In: *IEEE Access* 8, pp. 61687–61710.
- Raina, R., A. Madhavan, and A. Y. Ng (2009). “Large-Scale Deep Unsupervised Learning Using Graphics Processors”. In: *Proceedings of the 27th International Conference on Machine Learning*, pp. 873–880. DOI: 10.1145/1553374.1553486.
- Rajendran, P. and M. Madheswaran (2001). “Hybrid Medical Image Classification Using Association Rule Mining with Decision Tree Algorithm”. In: *Machine Learning* 45, pp. 5–32. DOI: 10.1023/A:1010933404324.
- Redmon, J. and A. Farhadi (2016). “YOLO9000: Better, Faster, Stronger”. In: *Computing Research Repository (CoRR), Arxiv*. URL: <http://arxiv.org/abs/1612.08242>.
- (2018). “YOLOv3: An Incremental Improvement”. In: *Computing Research Repository (CoRR), Arxiv*. URL: <http://arxiv.org/abs/1804.02767>.
- Redmon, J. et al. (2016). “You Only Look Once: Unified, Real-Time Object Detection”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788. DOI: 10.1109/CVPR.2016.91.
- Reinhard, E. et al. (2001). “Color transfer between images”. In: *IEEE Computer Graphics and Applications* 21(5), pp. 34–41.
- Ren, S. et al. (2015). “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, pp. 1440–1448.
- Riotton, G. L. J. et al. (1973). “Cytology of the female genital tract”. In: *International histological classification of tumours* 8, p. 41.

- Ronneberger, O., P. Fischer, and T. Brox (2015). “U-Net: Convolutional Networks for Biomedical Image Segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241.
- Rosenblatt, F. (1958). “The perceptron: a probabilistic model for information storage and organization in the brain”. In: *Psychological review* 65(6), pp. 386–408.
- Ruder, S. (2016). “An overview of gradient descent optimization algorithms”. In:
- Rumelhart, D., G. Hinton, and R. Williams (1986). “Learning representations by back-propagating errors”. In: *Nature* 323, pp. 533–536. DOI: 10.1038/323533a0.
- Russakovsky, Olga et al. (2015). “ImageNet Large Scale Visual Recognition Challenge”. In: *International Journal of Computer Vision (IJCV)* 115.3, pp. 211–252. DOI: 10.1007/s11263-015-0816-y.
- Sande, K. E. A. van de, T. Gevers, and C. G. M. Snoek (2010). “Evaluating color descriptors for object and scene recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, pp. 1582–1596.
- Schutte, K. et al. (2020). “Using StyleGAN for Visual Interpretability of DeepLearning Models on Medical Images”. In: *Neural Information Processing Systems Conference*.
- Schwartz, D. (2002). “Dépistage cytologique du cancer du col de l’utérus par prélèvement en milieu liquide”. In: *Thèse de doctorat : Univ. Genève, 2002 - Méd. 10250*. URL: [http://www.unige.ch/cyberdocuments/theses2002/SchwartzD/these\\_body.html](http://www.unige.ch/cyberdocuments/theses2002/SchwartzD/these_body.html).
- Selvaraju, R. R. et al. (2017). “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization”. In: *IEEE International Conference on Computer Vision*, pp. 618–626.
- Sermanet, P. et al. (2014). “OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks”. In: *2nd International Conference on Learning Representations*.
- Sherman, M. et al. (2007). “The Bethesda interobserver reproducibility study (BIRST)”. In: *Cancer Cytopathology* 111(1), pp. 15–25.
- Shi, X. et al. (2020). “Graph temporal ensembling based semi-supervised convolutional neural network with noisy labels for histopathology image analysis”. In: *Medical Images Analysis* 60.
- Simonyan, K., A. Vedaldi, and A. Zisserman (2014). “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”. In: URL: <http://arxiv.org/abs/1312.6034>.
- Simonyan, K. and A. Zisserman (2015). “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *International Conference on Learning Representations*.
- Sing, T. et al. (2019). “A deep learning-based model of normal histology”. In: *bioRxiv*. DOI: 10.1101/838417.
- Singh, R. V. (2016). “ImageNet Winning CNN Architectures - A Review”. In: URL: [https://rajatvikramsingh.github.io/media/DeepLearning\\_ImageNetWinners.pdf](https://rajatvikramsingh.github.io/media/DeepLearning_ImageNetWinners.pdf).
- Singh, U., Anjum, and S. et al Qureshi (2018). “Comparative study between liquid-based cytology and conventional Pap smear for cytological follow up of treated patients of cancer cervix”. In: *Indian Journal on Medical Research* 147(3), pp. 263–267. DOI: 10.4103/ijmr.IJMR\_854\_16.

- Smilkov, D. et al. (2017). “Smoothgrad: removing noise by adding noise”. In: *Workshop on Visualization for Deep Learning at ICML*.
- Solomon, D. et al. (2002). “The 2001 Bethesda System Terminology for Reporting Results of Cervical Cytology”. In: *Journal of the American Medical Association* 287(16), pp. 2114–2119.
- Sornapudi, S. et al. (2019). “Comparing Deep Learning Models for Multi-cell Classification in Liquid-based Cervical Cytology Images”. In: arXiv: 1910.00722 [eess.IV].
- Srinidhi, C. L., C. Oxan, and Martel A. L. (2021). “Deep neural network models for computational histopathology: A survey”. In: *Medical Images Analysis* 67. DOI: 10.1016/j.media.2020.101813.
- Srinivas, S. and F. Fleuret (2019). “Full-Gradient Representation for Neural Network Visualization”. In:
- Srivastava, N. et al. (2014). “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *Journal of Machine Learning Research* 15(1), pp. 1929–1958. ISSN: 1532-4435.
- Steinhaus, H. (1956). “Sur la division des corps matériels en parties”. In: *Bulletin de l’Académie Polonaise des Sciences, Classe III* 4(12), pp. 801–804.
- Stoler, M. H. and M Schiffman (2001). “Interobserver Reproducibility of Cervical Cytologic and Histologic Interpretations Realistic Estimates From the ASCUS-LSIL Triage Study”. In: *Journal of the American Medical Association* 285(11), pp. 1500–1505.
- Sundararajan, M., A. Taly, and Q. Yan (2017). “Axiomatic Attribution for Deep Networks”. In: *34th International Conference on Machine Learning* 70, pp. 3319–3328.
- Szegedy, C. et al. (2015). “Going Deeper With Convolutions”. In: *IEEE Conference on Computer Vision and Pattern Recognition*.
- Szegedy, C. et al. (2016). “Rethinking the Inception Architecture for Computer Vision”. In: *IEEE Conference on Computer Vision and Pattern Recognition*.
- Sánchez, J. and F. Perronnin (2011). “High-dimensional signature compression for large-scale image classification”. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1665–1672.
- Taha, B. et al. (2017). “Classification of Cervical-Cancer Using Pap-Smear Images: A Convolutional Neural Network Approach”. In: *Medical Image Understanding and Analysis*, pp. 261–272.
- Tan, M. et al. (2019). “MnasNet: Platform-Aware Neural Architecture Search for Mobile”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. DOI: 10.1109/CVPR.2019.00293.
- Tellez, D. et al. (2019). “Neural image compression for gigapixel histopathology image analysis”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1.
- Tellez, D. et al. (2020). “Extending Unsupervised Neural Image Compression With Supervised Multitask Learning”. In: *Medical Imaging with Deep Learning Conference*.
- Tomczak, K., P. Czerwińska, and M. Wiznerowicz (2015). “The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge”. In: *Contemporary Oncology (Pozn)* 19(1A), pp. 68–77. DOI: 10.5114/wo.2014.47136.

- Uijlings, J. R. R., K. E. A. van de Sande, and T. et al. Gevers (2013). “Selective Search for Object Recognition”. In: *International Journal of Computer Vision* 104, pp. 154–171. DOI: 10.1007/s11263-013-0620-5.
- Veta, M. et al. (2019). “Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge”. In: *Medical Images Analysis* 54, pp. 111–121. DOI: 10.1016/j.media.2019.02.012.
- WHO, World Health Organization (2014). “Comprehensive cervical cancer control: a guide to essential practice, 2nd edition”. In:
- Wilbanks, G. D. et al. (1968). “An evaluation of a one-slide cervical cytology method for the detection of cervical intraepithelial neoplasia”. In: *Acta Cytologica* 12(2), pp. 157–165.
- Wright T. C., Jr. et al. (2002). “2001 Consensus Guidelines for the Management of Women With Cervical Cytological Abnormalities”. In: *Journal of the American Medical Association* 287(16), pp. 2120–2129.
- Wu, X. et al. (2007). “Top 10 algorithms in data mining”. In: *Knowledge and Information Systems* 14(1), pp. 1–37. DOI: 10.1007/s10115-007-0114-2.
- Wu, Y. and K. He (2018). “Group Normalization”. In: *Lecture Notes in Computer Science*, pp. 3–19. DOI: 10.1007/978-3-030-01261-8\_1.
- Xie, C. et al. (2019). “VOCA: Cell Nuclei Detection In Histopathology Images By Vector Oriented Confidence Accumulation”. In: *Proceedings of Machine Learning Research* 102. Ed. by M. Jorge Cardoso et al., pp. 527–539. URL: <http://proceedings.mlr.press/v102/xie19a.html>.
- Xie, C. et al. (2020a). “Beyond Classification: Whole Slide Tissue Histopathology Analysis By End-To-End Part Learning”. In: *Medical Imaging with Deep Learning Conference*.
- Xie, Ci. et al. (2020b). “Adversarial Examples Improve Image Recognition”. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.1109/cvpr42600.2020.00090. URL: <http://dx.doi.org/10.1109/cvpr42600.2020.00090>.
- Yu, K., T. Zhang, and Y. Gong (2009). “Nonlinear Learning using Local Coordinate Coding”. In: ed. by Y. Bengio et al., pp. 2223–2231. URL: <http://papers.nips.cc/paper/3875-nonlinear-learning-using-local-coordinate-coding.pdf>.
- Zarella, M. D. et al. (2019). “A Practical Guide to Whole Slide Imaging: A White Paper From the Digital Pathology”. In: *Archives of Pathology and Laboratory Medicine* 143(2), pp. 222–234. DOI: 10.5858/arpa.2018-0343-RA.
- Zeiler, M. D. (2012). “ADADELTA: An Adaptive Learning Rate Method”. In: *CoRR*. arXiv: 1212.5701. URL: <http://arxiv.org/abs/1212.5701>.
- Zeiler, M. D. and R. Fergus (2014). “Visualizing and Understanding Convolutional Networks”. In: *European Conference on Computer Vision* 8689, pp. 818–833.
- Zhang, L. et al. (2014). “Automation-Assisted Cervical Cancer Screening in Manual Liquid-Based Cytology With Hematoxylin and Eosin Staining”. In: *Cytometry. Part A : the journal of the International Society for Analytical Cytology* 85.
- Zhang, L. et al. (2017). “DeepPap: Deep Convolutional Networks for Cervical Cell Classification”. In: *IEEE Journal of Biomedical and Health Informatics* 21, pp. 1633–1643.
- Zhou, B. et al. (2016). “Learning Deep Features for Discriminative Localization”. In:



- Zhou, R., E. H. Hammond, and D. L. Parker (1996). “A multiple wavelength algorithm in color image analysis and its applications in stain decomposition in microscopy images”. In: *Medical Physics* 23, pp. 1977–1986.
- Zhou, X. et al. (2010). “Image Classification Using Super-Vector Coding of Local Image Descriptors”. In: ed. by K. Daniilidis, P. Maragos, and N. Paragios, pp. 141–154.
- Zhu, J. et al. (2017). “Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks”. In: pp. 2223–2232.
- Zitnick, C. L. and P. Dollar (2014). “Edge Boxes: Locating Object Proposals from Edges”. In: *Computer Vision - ECCV 2014*, pp. 391–405.
- Zoph, B. et al. (2019). “Rethinking Pre-training and Self-training”. In: *Proceedings of the Neural Information Processing Systems*.



**Titre :** Méthodes de diagnostic assisté par ordinateur pour le dépistage du cancer du col de l'utérus sur lames de frottis vaginal en milieu liquide basées sur les réseaux de neurones à convolutions : conception, optimisation et interprétabilité.

**Mots clés :** Apprentissage profond ; Réseaux de neurones à convolutions ; Cytologie ; Classification de lames entières ; Interprétabilité.

**Résumé :** Le cancer du col de l'utérus est le deuxième cancer le plus important pour les femmes après le cancer du sein. En 2012, le nombre de cas recensés dépasse 500,000 à travers le monde, dont la moitié se sont révélés mortels.

Jusqu'à maintenant, le dépistage primaire du cancer du col de l'utérus est réalisé par l'inspection visuelle de cellules, prélevées par frottis vaginal, par des cytopathologistes utilisant la microscopie en fond clair dans des laboratoires de pathologie. En France, environ 5 millions de dépistage sont réalisés chaque année et environ 90% mènent à un diagnostic négatifs (i.e. pas de changements précancéreux détectés).

Pourtant, ces analyses au microscope sont extrêmement fastidieuses et couteuses en temps pour les cyto-techniciens et peut nécessiter l'avis conjoint de plusieurs experts. Ce processus impacte la capacité à traiter cette immense quantité de cas et à éviter les faux négatifs qui sont la cause principale des retards de traitements médicaux. Le manque d'automatisation et de traçabilité des dépistage deviennent ainsi de plus en plus critique à mesure que le nombre d'experts diminue.

En ce sens, l'intégration d'outils numériques dans les laboratoires de pathologie devient une réelle problématique de santé publique et la voie privilégiée pour l'amélioration de ces laboratoires.

Depuis 2012, l'apprentissage profond a révolutionné le domaine de la vision par ordinateur, en particulier grâce aux réseaux de neurones à convolutions qui se sont montrés fructueux sur un large panel d'applications parmi lesquelles plusieurs en imagerie biomédicale. Parallèlement, le processus de digitalisation de lames entières a ouvert l'opportunité pour de nouveaux outils et de nouvelles méthodes de diag-

nostic assisté par ordinateur.

Dans cette thèse, après avoir motivé le besoin médical et introduit l'état de l'art en terme de méthodes d'apprentissage profond pour le traitement de l'image, nous présentons nos contribution au domaine de la vision par ordinateur traitant le dépistage du cancer du col de l'utérus dans un contexte de cytologie en milieu liquide.

Notre première contribution consiste à proposer une méthode simple de régularisation pour l'entraînement de modèles dans le contexte d'une classification ordinaire (i.e. classes suivant un ordre). Nous démontrons l'avantage de notre méthode pour la classification de cellules utérines en utilisant sur le jeu de données Herlev. De plus, nous proposons de nous appuyer sur des explications basées sur le gradient pour réaliser une localisation faiblement supervisée et plus généralement une détection d'anormalité. Finalement, nous montrons comment nous intégrons ces méthodes pour créer un outil assisté par ordinateur qui pourrait être utilisé afin de réduire la charge de travail des cytopathologistes.

La seconde contribution se concentre sur la classification de lames entières et l'interprétabilité de ces approches. Nous présentons en détails les méthodes de classification de lames entières s'appuyant sur l'apprentissage multi-instances, et améliorons l'interprétabilité dans un contexte d'apprentissage faiblement supervisé via des visualisations de caractéristiques au niveau de la tuile et une nouvelle manière de calculer des cartes de chaleur explicatives. Finalement, nous appliquons ces méthodes pour le dépistage du cancer du col de l'utérus en utilisant un détecteur d' "anormalité" qui guide l'entraînement pour l'échantillonnages de régions d'intérêt.

**Title :** Computer-aided diagnosis methods for cervical cancer screening on liquid-based Pap smears using Convolutional Neural Networks : design, optimization and interpretability.

**Keywords :** Deep Learning ; Convolutional neural networks ; Cytology ; Whole Slide Images Classification ; Interpretability.

**Abstract :** Cervical cancer is the second most important cancer for women after breast cancer. In 2012, the number of cases exceeded 500,000 worldwide, among which half turned to be deadly.

Until today, primary cervical cancer screening is performed by a regular visual analysis of cells, sampled by pap-smear by cytopathologists under bright-field microscopy in pathology laboratories. In France, about 5 millions of cervical screening are performed each year and about 90% lead to a negative diagnosis (i.e. no pre-cancerous changes detected).

Yet, these analyses under microscope are extremely tedious and time-consuming for cytotechnicians and can require the joint opinion of several experts. This process has an impact on the capacity to tackle this huge amount of cases and to avoid false negatives that are the main cause of treatment delay. The lack of automation and traceability of screening is thus becoming more critical as the number of cyto-pathologists decreases.

In that respect, the integration of digital tools in pathology laboratories is becoming a real public health stake for patients and the privileged path for the improvement of these laboratories.

Since 2012, deep learning methods have revolutionized the computer vision field, in particular thanks to convolutional neural networks that have been applied successfully to a wide range of applications among which biomedical imaging. Along with it, the whole slide imaging digitization process has opened the op-

portunity for new efficient computer-aided diagnosis methods and tools.

In this thesis, after motivating the medical needs and introducing the state-of-the-art deep learning methods for image processing and understanding, we present our contribution to the field of computer vision tackling cervical cancer screening in the context of liquid-based cytology.

Our first contribution consists in proposing a simple regularization constraint for classification model training in the context of ordinal regression tasks (i.e. ordered classes). We prove the advantage of our method on cervical cells classification using Herlev dataset. Furthermore, we propose to rely on explanations from gradient-based explanations to perform weakly-supervised localization and detection of abnormality. Finally, we show how we integrate these methods as a computer-aided tool that could be used to reduce the workload of cytopathologists.

The second contribution focuses on whole slide classification and the interpretability of these pipelines. We present in detail the most popular approaches for whole slide classification relying on multiple instance learning, and improve the interpretability in a context of weakly-supervised learning through tile-level feature visualizations and a novel manner of computing explanations of heat-maps. Finally, we apply these methods for cervical cancer screening by using a weakly trained “abnormality” detector for region of interest sampling that guides the training.