AI-based selection of imaging and biological markers predictive of therapy response in lung cancer

Paul Tourniaire

UNIVERSITÉ CÔTE D'AZUR

ÉCOLE DOCTORALE
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

Thèse de Doctorat

Sélection de biomarqueurs basée sur l'IA pour prédire la réponse au traitement du cancer du poumon

Paul Tourniaire

Inria, Équipe Epione

Thèse dirigée par Hervé Delingette et co-dirigée par Nicholas Ayache et Paul Hofman

Soutenue le 12 juin 2023

Présentée en vue de l'obtention du grade de Docteur en Automatique, Traitement du Signal et des Images de l'Université Côte d'Azur.

Devant le jury composé de :

Elsa Angelini	Télécom Paris	Rapporteuse
Hugues Talbot	CentraleSupélec	Rapporteur
Nasir Rajpoot	University of Warwick	Examinateur
Maria Vakalopoulou	CentraleSupélec	Examinatrice
Laure Blanc-Féraud	I3S Laboratory	Présidente
Paul Hofman	Centre Hospitalier Universitaire de Nice	Co-encadrant
Marius Ilié	Centre Hospitalier Universitaire de Nice	Co-encadrant
Nicholas Ayache	Centre Inria d'Université Côte d'Azur	Co-directeur de thèse
Hervé Delingette	Centre Inria d'Université Côte d'Azur	Directeur de thèse

Inria

3iA Côte d'Azur
Institut interdisciplinaire d'intelligence artificielle

chu de Nice

Sélection de biomarqueurs basée sur l'IA pour prédire la réponse au traitement du cancer du poumon

AI-based Selection of Imaging and Biological Markers Predictive of Therapy Response in Lung Cancer

Jury

Présidente

Laure Blanc-Féraud	Directrice de Recherche	I3S Laboratory

Rapporteurs

Elsa Angelini	Professeur	Télécom Paris
Hugues Talbot	Professeur	CentraleSupélec

Examinateurs

Nasir Rajpoot	Professeur	University of Warwick
Maria Vakalopoulou	Assistant Professor	CentraleSupélec
Laure Blanc-Féraud	Directrice de recherche	I3S Laboratory
Paul Hofman	Professeur	Centre Hospitalier Universitaire de Nice
Nicholas Ayache	Directeur de recherche	Centre Inria d'Université Côte d'Azur
Hervé Delingette	Directeur de recherche	Centre Inria d'Université Côte d'Azur

Invité

Marius Ilié	Professeur	Centre Hospitalier Universitaire de Nice

Résumé

L'objectif de cette thèse est de développer des modèles d'apprentissage automatique capables d'exploiter des lames histologiques et des données cliniques pour prédire le résultat des traitements par immunothérapie contre le cancer du poumon. À cette fin, plusieurs défis doivent être relevés, tels que la classification et la localisation simultanées d'informations dans des images de lames entières de grande taille, ou l'interprétation des prédictions faites par les modèles. Dans ce qui suit, nous proposerons plusieurs contributions pour relever ces défis.

Le chapitre 2 introduit le concept de supervision mélangée en histopathologie. L'objectif de cette méthode est de tirer parti de plusieurs niveaux de supervision (c'est-à-dire la supervision globale et locale) pour rendre le modèle plus efficace à la fois en classification et en localisation. Sur la base d'un modèle d'apprentissage profond basé sur l'attention et adapté à la classification globale et locale de tissu dans des coupes histologiques, nous montrons qu'il est possible d'améliorer non seulement les performances du modèle en matière de classification des lames, mais aussi et surtout sa capacité à localiser avec précision les régions d'intérêt dans le tissu disponible, lorsque seules quelques annotations disponibles.

Le chapitre 3 étend le travail présenté dans le chapitre 2, en consolidant les branches de classification et de localisation simultanées du modèle avec des fonctions de coût adaptées qui contraignent la distribution de l'attention à suivre la distribution réelle des labels d'après les annotations disponibles. Une stratégie d'échantillonnage des images est également proposée pour renforcer les performances de localisation et simplifier la procédure d'apprentissage afin qu'elle s'inscrive dans un processus unique.

Dans le chapitre 4, nous présentons un ensemble de données multicentriques sur le cancer du poumon dédié à la prédiction de la réponse à l'immunothérapie. Nous documentons les différentes étapes suivies pour éliminer les échantillons de faible qualité, ainsi que les cas indéterminés, et nous discutons de la définition de ce qu'est une réponse positive ou négative par rapport aux évaluations cliniques actuelles de référence. Enfin, nous évaluons plusieurs modèles permettant de prédire directement la réponse au traitement à partir de lames histologiques et discutons des écueils des approches envisagées.

Dans le chapitre 5, nous passons de la prédiction binaire de la réponse au traitement à la prédiction de survie, et nous utilisons l'apprentissage par contraste ainsi que le regroupement non paramétrique profond pour générer un ensemble de caractéristiques pronostiques de manière non supervisée. Nous montrons que l'ensemble des caractéristiques obtenu est un puissant indicateur de survie et qu'il conserve un bon niveau de performance lorsque l'on choisit un seul centre comme ensemble de test. Nous discutons également de l'interprétation histologique faite du résultat de l'algorithme de regroupement, en particulier pour les groupes les plus corrélés à la survie.

Pour conclure, nous abordons les questions et les défis en suspens, et nous discutons des orientations futures qui pourraient être prises afin de répondre aux questions restées sans réponse.

Mots-clés: histopathologie numérique, apprentissage multi-instance, supervision mélangée, apprentissage profond, analyse de survie, cancer du poumon, immunothérapie.

Abstract

The purpose of this thesis is to develop machine learning models that can leverage histology slides and clinical data to predict the outcome of immunotherapies against lung cancer. To this end, there are several challenges to overcome, such as the concurrent classification and localization of information within whole-slide images of large size, or the interpretability of the predictions made by the models. The thesis proposes several contributions to address such challenges.

Chapter 2 introduces the concept of mixed supervision in histopathology. The purpose of this framework is to leverage several levels of supervision (i.e., global and local supervision) to make the model more efficient in both classification and localization tasks. Set on an attention-based deep learning model fit for global and local classification of tissue in whole-slide images, we show that it is possible to improve not only the model slide-level classification performance, but also and most importantly its ability to accurately locate regions of interest in the tissue, with only a few available local annotations.

Chapter 3 extends the work presented in chapter 2, by consolidating the simultaneous classification and localization branches of the model with tailored loss functions that enforce attention distribution to follow the actual label distribution with respect to the available annotations. A slide sampling strategy is also proposed to strengthen the localization performance, and simplify the training procedure to have it fit in a single process.

In Chapter 4, we present a multicentric lung cancer dataset dedicated to the prediction of immunotherapy response. We document the various steps followed to filter out low-quality samples, as well as undetermined outcomes, and we discuss the interpretation of what is a positive or a negative response with respect to the current gold standard clinical evaluations. Finally, we evaluate several models to directly predict the treatment response out of pathology slides, and discuss the caveats of the tested approaches.

In Chapter 5, we switch from binary treatment response prediction to survival analysis, and use contrastive learning along with deep nonparametric clustering to generate a set of prognostic features in an unsupervised manner. We show that the obtained set of characteristics is a powerful indicator of survival, and that it maintains a good level

of performance when picking one acquisition center as the test set. We also discuss the histological interpretation of the most prominent discovered clusters.

To conclude, we address the remaining issues and challenges, and debate what future directions could be taken in order to tackle unanswered questions.

Keywords: digital pathology, multiple instance learning, mixed supervision, deep learning, survival analysis, lung cancer, immunotherapy.

Remerciements

Après trois ans et demi passés sur cette thèse, il me faudrait remercier tant de monde qu'un manuscrit annexe serait nécessaire. Je ne m'offrirai pas ce luxe, et me limiterai à quelques dédicaces seulement. Que ceux que je ne citerai pas me pardonnent, mais sous l'averse chaleureuse des rencontres que j'ai faites, il m'est impossible d'évoquer toutes les gouttes.

Avant toute chose, je tiens à remercier les deux personnes qui m'ont accueilli et suivi tout au long de cette thèse. Hervé, Nicholas, malgré un début de collaboration sous le signe du confinement, de la distance, et de premiers résultats guère encourageants, vous avez su préserver votre confiance et votre optimisme. Au fil des avancées et des échecs, j'ai toujours senti votre soutien et apprécié votre aide.

Quant à Marius et à Paul, je vous remercie chaleureusement pour votre accueil, votre expertise, ainsi que pour m'avoir permis de découvrir ce domaine si fascinant qu'est l'anatomo-pathologie. Peut-être n'ai-je fait qu'en effleurer la surface, mais je suis heureux d'avoir pu le faire en votre compagnie.

Je souhaite remercier chacun des membres du jury d'avoir accepté d'évaluer mon travail, et en particulier les deux rapporteurs, Elsa Angelini, et Hugues Talbot, pour leur relecture attentive.

A Maria, et Laure, je vous suis infiniment reconnaissant d'être apparues au moment opportun pour me tirer de mon découragement. Je dirai même, de mon marasme. Il n'aura suffit que d'un court entretien, ô combien précieux ! Si j'ai pu achever ce travail, c'est aussi grâce à vous.

C'est également mon équipe, Epione, que je souhaite faire figurer ici. Je crois n'avoir jamais connu pareil cadre de travail, et j'ai bien peur de ne jamais en connaître de meilleur. Vous avez contribué de bien des manières au plaisir que j'ai ressenti au cours de mon séjour ici, et j'espère pouvoir, à l'issue de ma thèse, vous dire au revoir, plutôt qu'adieu.

A tous mes amis qui m'ont soutenu pendant ces trois années, à Jahed et Manel, qui m'ont rendu de nombreuses visites, à Thomas, qui a tout fait pour m'arracher à la thèse pour la

pratique de la guitare, et à Laure, qui aura rendu mes derniers mois de travail le plus agréable possible. Je vous réserve ma plus profonde affection.

Je n'oublie pas bien sûr les Avignonnais, là dès bien avant la thèse, et bien après j'espère. Je n'égrènerai pas chaque nom : ils se reconnaîtront.

A toute ma famille, à mes côtés depuis toujours.

Enfin, c'est inévitablement à Hind que reviennent les ultimes remerciements. Je ne saurais rendre compte en quelques lignes de l'importance que tu as eue au cours de ces années. J'espère seulement que tu en as conscience.

Acknowledgement

We would like to thank Hamila Marame, Julien Fayada, Marine Pedro, Cyrielle Falduzza, Olivier Carruggi, and Pascal Grier for their contribution to the preparation and digitization of the whole-slide images at the Nice hospital.

We also thank the members of the medical centers who contributed to the dataset presented in this work: Julien Mazières, MD., PhD., Anna Vigier, François Ghiringhelli, Nicolas Piton, Jean-Christophe Sabourin, Frédéric Bibeau

Financial Support

This work has been supported by the French government, through the 3IA Cote d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

Contents

Abbreviations:

Medical

CR	Complete Response
ICI	Immune-Checkpoint Inhibitor
IHC	Immuno-Histo Chemistry
NGS	Next Generation Sequencing
NSCLC	Non-small Cell Lung Cancer
PD	Progressive Disease
PD-1	Programmed cell death protein 1
PD-L1	Programmed death-ligand 1
PFS	Progression-free Survival
PR	Partial Response
OS	Overall Survival
RECIST	Response Evaluation Criteria in Solid Tumors
SD	Stable Disease
TMB	Tumor Mutational Burden
TPS	Tumor Proportion Score
WES	Whole-exome Sequencing
WSI	Whole-slide Image

Machine Learning

CLAM	Clustering-constrained-attention Multiple Instance Learning
CNN	Convolutional Neural Network
DPM	Dirichlet Process Mixture
MIL	Multiple Instance Learning

Metrics

AUC	area under the receiver operating characteristic curve
C-index	Concordance index

1 Introduction

Contents

This thesis explores how AI-based data analysis can help develop new biomarkers for the prediction of lung cancer response to immunotherapy.

1.1 Clinical Context

1.1.1 An overview of lung cancer

According to the GLOBOCAN 2020 estimates [Sung, 2021], there were 19.3 million new cases of cancer that year, for 9.9 million cancer deaths. The number of new cases per year is also expected to rise by 47% in 2040. Although lung cancer is only the second most prevalent form of cancer for both sexes behind breast, accounting for 2.2 million new cases a year, it is nonetheless the most lethal one, with nearly 1.8 million cancer deaths a year. While two thirds of lung cancer deaths are attributable to smoking, other environmental hazards are risk factors, such as air pollution. The 5-year survival rate of lung cancer patients stands between 10 and 20%, although some countries sport a slightly more optimistic trend. Lung cancer is generally separated between two kinds: Non-small Cell (NSCLC, 80-85%), and Small Cell (10-15%)[1]. Depending on the stage of the disease, a certain number of treatment options are available. For early-stage lung cancers, surgery is often recommended, generally accompanied by an adjuvant treatment, such as chemotherapy, while for stages IIIA and above, surgery is generally performed *after* drug administration whenever possible, the latter being "neo-aduvant"[2].

[1] https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html. Accessed in March 2023
[2] https://www.cancer.gov/types/lung/patient/non-small-cell-lung-treatment-pdq. Accessed in March 2023

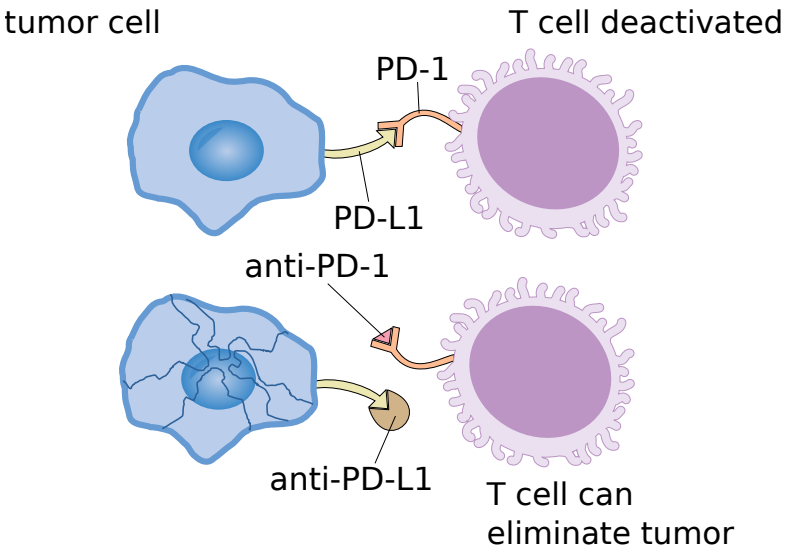


Fig. 1.1.: A representation of the immune-checkpoint inhibitor principle. Under the action of immune-checkpoint blockade on either PD-1 or PD-L1 receptors, the T cell can accurately identify and neutralize the tumor cell. Tumor cell and lymphocyte diagrams come from the Wikimedia Commons repository[3]

Other than that, targeted therapies, radiation, laser, and, more recently, immunotherapy, are all treatment options which can improve a patient's prognosis. Treatments are scheduled in *lines*, given the outcomes of the previous one, and the current state of the patient. They can also be used together, to maximize the chance of getting an inhibiting effect, for instance administering immunotherapy with chemotherapy. The choice of an appropriate cure is of paramount importance: efficiency, side effects, but also cost depend on the correct option. No treatment works perfectly for every case: to help caregivers select the most appropriate cure, a broad range of biomarkers have been developed to identify which drug or surgery may work best. From radiological examination to histological or biological analysis, each biomarker requires different kinds of modalities to be identified.

1.1.2 Immunotherapy

Among all the treatment options available, immunotherapy is one of the most impactful solution that was proposed recently. For the scope of this thesis, we will focus on Immune-Checkpoint Inhibitors (ICIs). The mechanism behind ICI treatment can be seen in Figure 1.1. The objective of the treatment is to help T lymphocytes correctly identify cancer cells and eliminate them. When lymphocytes approach cancer cells, a certain number of proteins bind between the two: the checkpoints. Among them, the connection between the Programmed cell death protein 1 (PD-1, lymphocyte) and the Programmed death-ligand 1 (PD-L1, tumor cell) acts as a deterrent of the immune

[3] Cancer Research UK, CC BY-SA 4.0 https://creativecommons.org/licenses/by-sa/4.0, via Wikimedia Commons.

response. To prevent this from happening, ICIs block one of the two receptors, such that the lymphocyte can correctly identify and eliminate the tumor. Apart from the PD-1/PD-L1 connection, other checkpoints can be blocked, such as the one involving the CTLA-4 protein. Since the introduction of ICIs in the treatment pipeline, lung cancer prognosis has considerably improved. Numerous large-scale studies have been conducted to report the efficacy of immunotherapies at different lines and in combination with other treatments [Malhotra, 2017; Horn, 2017; Forde, 2022; Paz-Ares, 2021]. Since then, ICIs have been approved by several public health services as early as in the first line of treatment[4,5]. And yet, several challenges remain to be taken up with respect to the treatment effect. Indeed, the response rate of patients to ICIs for lung cancer is rather low: only around 18% of them manifest the signs of a positive impact [Mazieres, 2019; Berghmans, 2020]. The treatment response, assessed radiologically via standardized Response Evaluation Criteria in Solid Tumors (RECIST) [Eisenhauer, 2009], belongs to one out of four categories – complete response (CR), partial response (PR), stable disease (SD) and progressive disease (PD) – depending on the growth or the shrinkage of the lesions. Given the burden that are cancer treatments for patients, it is crucial to avoid as much as possible the administration of drugs with little to no impact. As for any other medication, immunotherapies come with their lot of side effects which, although usually benign (e.g., fatigue, skin inflammation), can sometimes be particularly aggressive, until they become the cause of death (e.g., hepatitis, myocarditis) [Sumi, 2022; Martins, 2019]. Strong side-effects can also cause the interruption of treatment following the patient's decision, which in turn diminishes its effect. Once again, it is essential to choose wisely which patient should partake in such an option.

1.1.3 Current biomarkers

To help clinicians decide which treatment to give to patients, specific biomarkers have been proposed. For immunotherapy in particular, the current gold standard is the Immuno-Histo-Chemistry-based (IHC) assessment of the PD-L1 expression of tumor cells (also denoted as the Tumor Proportion Score, or TPS), for PD-1/PD-L1 targeting drugs (e.g., nivolumab, pembrolizumab) [Mok, 2019]. The expression, ranging from 0 to 100%, is positively correlated to the response rate. However, two thresholds in particular – 1% and 50% – are usually pointed at for yielding groups with statistically significant differences in survival among patients [Grigg, 2016; Garon, 2015]. Indeed, patients above these thresholds show higher response rates than others (27% and 39% respectively) [Reck, 2019; Mok, 2019]. To this day, the TPS is the only clinically used biomarker to select patients for anti-PD-1/PD-L1 therapy. Its efficiency nonetheless calls for more powerful markers, so as to pick even more accurately the patients fit for immunotherapy. More recently, sequencing techniques such as RNA, Whole Exome, or

[4] FDA website. Accessed in March 2023
[5] EMA website. Accessed in March 2023

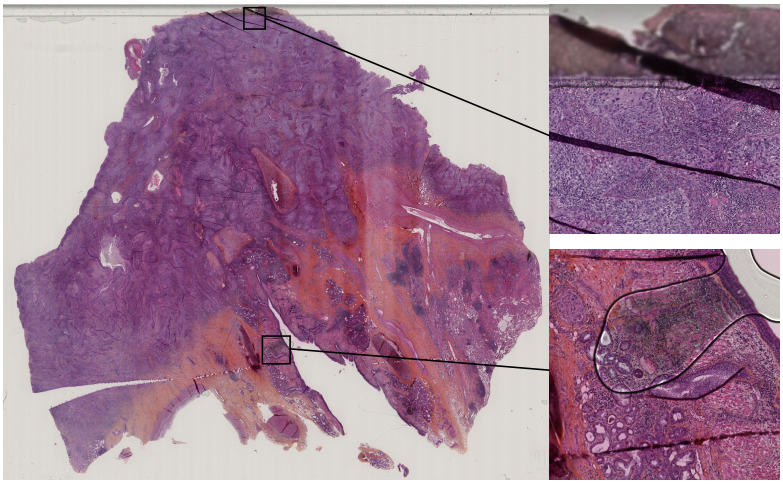


Fig. 1.2.: An example of a WSI next to some zoomed-in regions containing artifacts: glass crack and tissue fold (top), air bubble (bottom).

Next Generation Sequencing (RNA-seq, WES, NGS) have allowed for the development of a new biomarker called Tumor Mutational Burden (TMB) [Hellmann, 2018; Marabelle, 2020; Klein, 2021]. Computed as the number of somatic mutations per megabase in the tumor DNA, the TMB informs the clinician about the mutational status of the tumor. Patients with $\geq$10 mutations per megabase have significantly higher response rates and progression-free as well as overall survival (PFS, OS). The American Food and Drug Administration recently approved the FoundationOneCDx assay for TMB evaluation[6], but this one only. The lack of a real standardization of assays and measures is still an impediment to a broad application in clinical practice [Stenzinger, 2019; Heeke, 2020]. Moreover, the sequencing techniques are not as common or affordable as the other diagnostic tools, which is another obstacle to the adoption of such a biomarker. TPS evaluation, on the other hand, remains more accessible, but is not immune from blame: its robustness and exposure to inter-rater variability call for cautious evaluation [Ilie, 2017; Cooper, 2017]. It appears that several reasons – efficiency, reproducibility, cost – motivate the search for new biomarkers.

1.2 Computational Pathology

Digital whole-slide imaging

The development of whole slide scanners has brought histopathology to the field of digital image analysis. Given the nature of whole-slide images (WSIs), several specific challenges can prevent the straightforward application of standard image processing methods. First and foremost, WSIs are typically up to 100,000-pixel wide or high: such dimensions require specific care, since no machine learning – let alone deep learning – framework can process such huge images all at once. Moreover, several artifacts can

[6] See this link. Accessed in March 2023

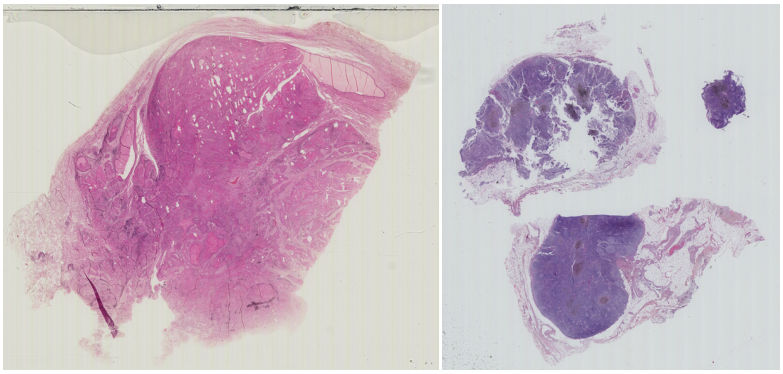


Fig. 1.3.: Two WSIs acquired in the same hospital, by the same operators. Yet, colors differ quite substantially between the two.

hamper the analysis of WSIs, such as tissue fold, tear, air bubbles, glass cracks, etc... Examples of artifacts can be seen in Figure 1.2. Finally, WSIs are stained using two or three dyes: Hematoxylin (purple), Eosin (pink) (H&E), with sometimes Saffron (HES), which ease the interpretation of the pathologist under a microscope. Unfortunately, the staining procedure is not really standardized, which can lead to large variations in the resulting colors after scanning [Macenko, 2009; Vahadane, 2016] (Figure 1.3).

Methods and applications

In spite of all these obstacles, a great number of works have focused on the automated analysis of whole-slide images, and in particular deep learning methods [Srinidhi, 2021; Viswanathan, 2022]. The one task that is tackled the most is unsurprisingly tumor classification and detection [Lu, 2021; Courtiol, 2018; Campanella, 2019; Dehaene, 2020; van Rijthoven, 2021], or associated objectives like histological subtyping [Chen, 2022]. Given the aforementioned problems related to dimensions, several approaches have been considered to adapt existing methods to computational pathology. One of the most widely used is Multiple Instance Learning (MIL) [Dietterich, 1997], in which each slide is seen as a bag, that contains several instances which in the case of histopathology are small regions called *tiles* or *patches* sampled in the image. The challenge then is to perform a suitable aggregation of the instances to recover the bag – or slide – label. As a matter of fact, the tile-to-slide label association is also often a problem the other way around, since most of the pathology tasks are weakly-supervised, i.e., the information is only available at the slide or patient level. This is because obtaining annotations from pathologists can be very time-consuming and subject to variability between annotators. Consequently, this causes sometimes good bag-level classifiers to offer poor tile-level performance, and vice versa. Beyond simple histological pattern recognition, other works have tried to yield non-histological information from WSIs. In particular, the search for genetic mutations has been addressed several times [Coudray, 2018; Schmauch, 2020], as well as the look for specific biomarkers such as microsatellite instability or IFN-γ [Kather, 2020; Saillard, 2021]. Among the most addressed subjects is survival prediction

directly from the tissue slides [Yao, 2020; Li, 2018; Shao, 2021a]. Most of the approaches regarding survival prefer a different path from MIL, and privilege the use of graphs to represent the spatial interactions between either different histomorphological regions, or even cells themselves. This, however, requires to be able to automatically detect or segment entire cells and nuclei, for which several methods have been proposed [Habis, 2022; Le Bescond, 2022].

Of course, treatment outcome prediction and treatment-related biomarkers have been sought already, and in particular with respect to ICIs. [Sha, 2019] have proposed a method to bypass the IHC standard analysis to predict the PD-L1 status, while [Jain, 2020] explore the feasibility of TMB status prediction from WSIs directly, without resorting to sequencing. On the topic of outcome prediction, [Harder, 2019] and [Johannet, 2021] developed deep learning frameworks to output treatment response from WSI analysis in the case of melanoma. Through the combination of recent discoveries on the role of Tumor Infiltrating Lymphocytes (TILs) [Tumeh, 2014], with cells or tissue interactions analysis in slides, recent attempts to stratify patients in survival groups were published [Park, 2022; Wang, 2022].

1.3 Objectives and Organization of the Thesis

Although there have been numerous works published on the automated WSI analysis for either diagnosis or prognosis, there are still several limitations or challenges that need to be addressed:

1. Efficient and coherent classification and localization of tissue in WSIs, i.e., the simultaneous assessment of global and local information by a single model, is not yet achieved. Most of the time, either WSI classification or pixel-level segmentation is performed, but not at the same time. Nevertheless, it is difficult to imagine that a model with accurate slide-level prediction could be used without localization guarantees with respect to the overall prediction.

2. The annotation of histological slides is a time-consuming and noisy task. To limit the effect of rater-dependent labels, consensus can be a potential solution, but it is a trade-off between robustness and manpower. However, annotations greatly help learning tasks, by giving precious information to the models, or reducing the size of the region it has to explore. The development of annotation-efficient models could help reduce the burden of such a process, while keeping domain knowledge at hand to ensure good performance.

3. New biomarker discovery: whole-slide images are acquired routinely in clinical practice for diagnostic purposes. The acquisition of prognostic biomarkers usually relies on other modalities, which add cost and workload to the entire patient follow-up. Leveraging histology to derive new biomarkers predictive of treatment outcomes could help ease and improve the treatment decision process. While current hypotheses on predictive signal within HES slides like TILs have been tested, other interpretable and data-driven approaches exploiting tissue morphology are still sought after.

In this thesis, we intend to address each of these three points through several methodological contributions. The manuscript is organized as follows, in accordance with the aforementioned research objectives:

In Chapter 2, we introduce the concept of *mixed supervision* for digital pathology, i.e., the concurrent use of both slide-level and patch- or tile-level labels. We show that with little annotation effort, we can improve the classification, and, more importantly, the localization results of an attention-based, weakly-supervised, MIL model for digital pathology, CLAM [Lu, 2021]. The latter is chosen for its architecture that allows to train for both tasks at the same time, which is not often the case, as models are usually design to tackle either slide-level or tile-level classification only. It also fits the annotation "efficiency" we are aiming for, since it is said to be already data-efficient, requiring only a few slides to reach good performance. This work was published at the COMPAY workshop of the MICCAI conference in 2021 [Tourniaire, 2021].

In Chapter 3, we extend the work introduced previously with a broader implementation of mixed-supervision in CLAM, which we call MS-CLAM. This time, a global slide-level loss function enforces uniform distribution of the attention weights on all key instances in the slide. A new slide sampling strategy is also implemented to balance the tile-level labels and avoid a collapse of the tile-level classifier. We compare our method with additional state-of-the-art baselines, and show the benefits of the use we make of labels compared to a more classical pretraining approach on two different tumor classification and localization datasets. This work was published in Medical Image Analysis [Tourniaire, 2023a].

In Chapter 4, we introduce and motivate the need for a new multicentric dataset specifically designed for the prediction of immunotherapy response among lung cancer patients. We present the inclusion criteria, and showcase the discrepancies between the centers with respect to patient characteristics and survival. We also discuss the uncertainty surrounding the definition of the treatment response, and evaluate several state-of-the-art whole-slide image classification baselines to predict the binary outcome of the treatment. Finally, we show that such straightforward approaches fail to yield satisfactory results,

and discuss the reasons behind this, which leads us to consider a different take on the subject, by addressing a proxy task to the response prediction, i.e., survival analysis.

In Chapter 5, we move from WSI classification and localization to survival analysis. The MIL framework is also dropped, for the benefit of an unsupervised and fully data-driven approach. In this section, we use tile-level contrastive learning and clustering to reduce the dimensionality and identify common patterns in the slides of our dataset. A feature selection method is also presented to pick the cluster information we need to accurately stratify patients in survival groups, a task seen as a proxy to treatment response prediction. We show that without any histology prior or annotation, we are able to generate a set of features with significant prognostic power regarding patient survival consequent to ICI treatment. The scrutiny of the tissues found in each cluster by an expert pathologist shows coherence with the current literature on the suspected origin of the immune response to treatment against lung cancer. This work was submitted to a journal [Tourniaire, 2023b].

In Chapter 6, we summarize the contributions of this thesis, and discuss its potential outcomes, as well as the new perspectives and the future challenges that are left to explore.

This thesis is conducted in partnership with the Laboratory of Clinical and Experimental Pathology of the Nice university hospital.

1.4 Publications

The described contributions led to the following peer-reviewed publications in both conferences and journals.

Journal Articles

- [Tourniaire, 2023b] **Tourniaire, P.**, Ilie, M., Mazières, J., Vigier, A., Ghiringhelli, F., Piton, N., Sabourin, J.-C., Bibeau, F., Hofman, P., Ayache, N. & Delingette, H. (2023). WhARIO: Whole-slide image-based survival Analysis for patients tReated with ImmunOtherapy. *Submitted to a journal*.

- [Tourniaire, 2023a] **Tourniaire, P.**, Ilie, M., Hofman, P., Ayache, N. & Delingette, H. MS-CLAM: Mixed Supervision for the classification and localization of tumors in Whole Slide Images. *Medical Image Analysis, Volume 85, 102763 (2023)*

- [Ilié, 2022] Ilié, M., Benzaquen, J., **Tourniaire, P.**, Heeke, S., Ayache, N., Delingette, H., Long-Mira, E., Lasalle, S., Hamila, M., Fayada, J., Otto, J., Cohen, C., Gomez-Caro, A., Berthet, J.-C., Marquette, C.-H., Hofman, V., Bontoux, C. & Hofman, P. (2022). Deep learning facilitates distinguishing histologic subtypes of pulmonary neuroendocrine tumors on digital whole-slide images. *Cancers, 14(7), 1740*

Conference Papers

- [Tourniaire, 2021] **Tourniaire, P.**, Ilie, M., Hofman, P., Ayache, N. & Delingette, H. Attention-based Multiple Instance Learning with Mixed Supervision on the Camelyon16 Dataset. *Proceedings of the MICCAI Workshop on Computational Pathology, in Proceedings of Machine Learning Research, 156:216-226*

Abstracts

- [Tourniaire, 2022] **Tourniaire, P.**, Ilie, M., Hofman, P., Ayache, N., & Delingette, H. (2022). Mixed supervision to improve the classification and localization: Coherence of tumors in histological slides. *Cancer Research, 82(12_Supplement), 461-461*

2 Attention-based Multiple Instance Learning with Mixed Supervision on the Camelyon16 Dataset

Contents

Abstract Since the standardization of Whole Slide Images (WSIs) digitization, the use of deep learning methods for the analysis of histological images has shown much potential. However, the sheer size of WSIs is a real challenge, as they are often up to 100,000 pixels wide and high at the highest resolution, and therefore cannot be processed directly by any model. Moreover, as the manual delineation of structures within WSIs is tedious, histological datasets often only contain slide-level labels, or a limited amount of delineated slides. In this context, multiple-instance learning (MIL) approaches have been proposed, considering WSIs as bags of smaller images, designated as tiles or patches. Among these methods, the attention-based MIL aims at learning the importance of each tile for the slide final classification while at the same time performing a clustering of those tiles. In this chapter, we introduce the concept of mixed supervision within this framework, by exploiting tile-level labels in addition to slide-level labels to improve the classification of slides. More precisely, we show on the Camelyon16 dataset that even a small proportion of slides with pixel-wise annotations can improve their classification but also the localization of tumorous regions. This improves the consistency of the results between the tile and slide levels and the interpretability of the algorithm. This chapter was published as a conference paper in [Tourniaire, 2021].

2.1 Introduction

In terms of tumor assessment, histopathology is currently the clinical gold-standard diagnosis technique. Nonetheless, pathologists face multiple challenges when confronted to Whole Slide Images (WSIs), and often require careful and time-intensive efforts. WSIs are usually stained with Hematoxylin and Eosin (H&E) and scrutinized through a microscope by the pathologist in order to detect cancerous tissue. Given the size of tissue samples, and the potential artifacts that may occur, such as tissue folds or tears, the diagnosis is often tedious. Since the digitization of pathological slides, machine and deep learning algorithms have offered automated solutions for the diagnosis of tumors [Bejnordi, 2017; Wang, 2019]. However, pixel-level annotations such as tumor segmentation are usually unavailable, since the annotation process requires both time and medical expertise: datasets often only display slide-level labels. Among deep learning algorithms for weakly-supervised learning, some of the most popular are based on the multiple instance learning assumption (MIL). As such, the WSI is divided into tiles or patches of smaller size (e.g. 512 x 512 pixels^2) and the slide is considered normal if all instances (tiles) are normal, and tumorous if at least one tile contains tumor. The tile-level predictions are then aggregated following various mechanisms to provide the slide-level or patient-level label [Campanella, 2019; Courtiol, 2018].

The attention mechanism was recently proposed as the tile pooling function [Ilse, 2018]. The authors propose to extract tile-level features with a convolutional network [Sirinukunwattana, 2016], and use a two-layered neural network to calculate a weighted average and select key instances for the slide-level prediction. A recent improvement of the attention-based MIL was proposed in the CLAM (Clustering-constrained Attention Multiple instance learning) algorithm [Lu, 2021]: the introduction of instance-level clustering during training, where instances given the highest (resp. lowest) attention scores were considered positive (resp. negative) evidence of the slide class. A smooth SVM loss [Berrada, 2018] calculated on the instance-level prediction using an instance classifier is then added to the overall slide-level cross-entropy. For the tile-level feature extraction, a frozen ImageNet pre-trained ResNet50 [He, 2016] is used, so as to accelerate the training procedure. This, however, was identified as a major caveat by Dehaene et al. [Dehaene, 2020], who decided to replace the backbone with a self-supervised deep neural network, trained on histological images using contrastive loss [He, 2020]. On the Camelyon16 dataset [Bejnordi, 2017], they showed performance close to the best performing fully-supervised method [Wang, 2016], but at the cost of heavy and long GPU training.

When pixel-level labels are available, the combined use of pixel-level and image-level labels is coined *mixed supervision*. It is especially useful when the amount of pixel-level annotated images is low, in regards to the total amount of available images. Mlynarski et al. [Mlynarski, 2019] showed that using weakly and fully annotated data to train a deep learning model for brain tumor segmentation improved the segmentation performance compared to the baseline trained only on fully annotated data. Ciga et al. [Ciga, 2021] simultaneously trained a ResNet18 model on classification and segmentation tasks, leveraging only a part of available fully annotated data, to obtain the same performance as models trained on the entire fully labeled dataset. However, both classification and segmentation tasks relied on tile-level information, the authors having devised a strategy to select tiles suited for either task.

In this work, starting from the CLAM framework, we propose the first mixed supervision approach within the attention-based MIL framework. Our main goal is not only to improve the slide-level classification, but also the localization performance of the model. Our contributions are listed as follows:

- First, we introduce mixed supervision for the MIL framework inside the CLAM algorithm.

- Second, we propose an improvement of the instance-level clustering method in the case of non-pathological slides.

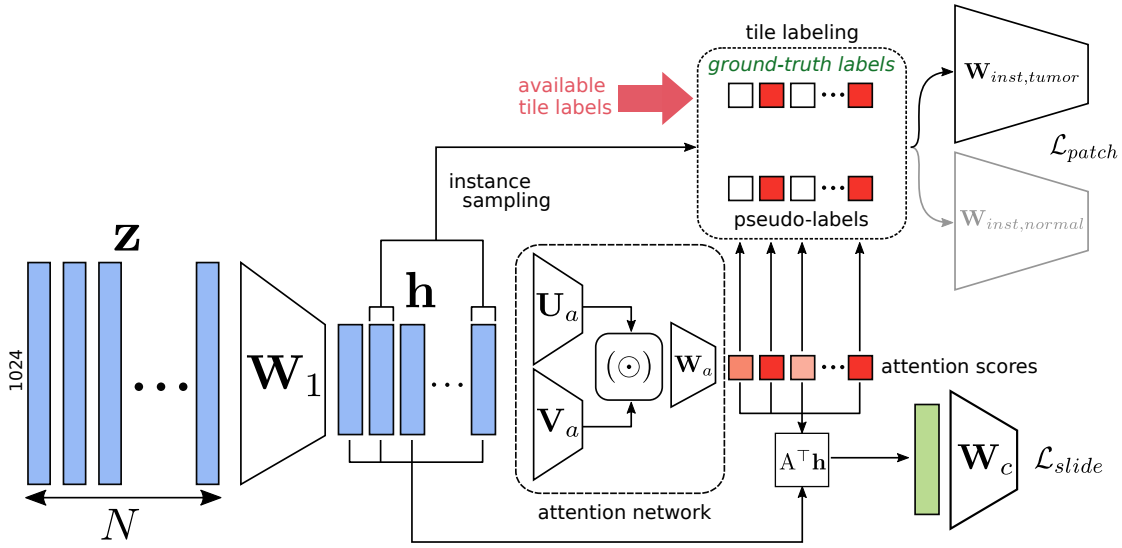


Fig. 2.1.: Overview of the CLAM model. Activation functions are not detailed in the interest of clarity. The original instance selection approach appears in black (pseudo-labels based on the attention scores), whereas our annotation-based instance selection approach appears in green in the "tile labeling" box.

- Finally, we show that we can improve the model's classification performance and the localization of tumorous tiles by adding a small amount of ground-truth tile-level labels along with slide-level labels.

2.2 Methods

2.2.1 Data preprocessing

For the preprocessing of the WSIs, we first transform a downsampled version of the image to the HSV space and apply Otsu thresholding [Otsu, 1979] on the hue and saturation channels to detect tissue. Then, we extract 256 x 256 non-overlapping tiles from the tissue region, at the highest magnification level. For the tiles feature extraction, we use the same model as Lu et al. [Lu, 2021], i.e. a ResNet50 architecture without the final classification layer, and an output dimension of 1024.

2.2.2 CLAM Algorithm

2.2.2.1 Description of the original model

Slide-level classification. The overview of the model can be seen in Figure 2.1. We consider a classification problem with n classes, and detail only the "multiple branches" version of CLAM (CLAM MB). A slide is represented as a feature matrix $\mathbf{z}$, of size 1024 x

N, where N is the number of tiles in the slide. A first fully-connected layer $\mathbf{W}_1$ transforms each 1024-dimensional feature vector $\mathbf{z}_k$ into a 512-dimensional feature vector $\mathbf{h}_k$, where k is the tile index. Then a gated-attention module computes the attention weights from $\mathbf{h}$ for each of the n classes, following:

$$a_{k,m} = \frac{\mathbf{W}_{a,m}(\tanh(\mathbf{V}_a\mathbf{h}_k^\top) \odot \mathrm{sigm}(\mathbf{U}_a\mathbf{h}_k^\top))}{\sum_{i=1}^{N} \exp\{\mathbf{W}_{a,m}(\tanh(\mathbf{V}_a\mathbf{h}_i^\top) \odot \mathrm{sigm}(\mathbf{U}_a\mathbf{h}_i^\top))\}} \tag{2.1}$$

where m is the class index, and $\mathbf{V}_a$, $\mathbf{U}_a$, and $\mathbf{W}_a$ are fully-connected layers. The attention weights are represented by the matrix $\mathrm{A} \in \mathbb{R}^{\mathrm{N}\times n}$. Before the final binary classification layers $\mathbf{W}_{c,m}$, each feature vector $\mathbf{h}_k$ is multiplied by its corresponding attention weight. This operation is represented by the matrix $\mathrm{M} = \mathrm{A}^\top\mathbf{h}$. Finally, each column of M (representing one of the n classes) is independently processed by one of the n classifiers to obtain the slide-level score for each class. A softmax activation function is eventually applied on the logits, and a cross-entropy loss $\mathcal{L}_{slide}$ is computed on the obtained scores.

Instance-level clustering. Based on the attention scores $a_{k,m}$, the tiles with the highest (resp. lowest) scores are retrieved for each class, and assigned a positive (resp. negative) pseudo-label. The assumption is that high-attention tiles are positive evidences of the class, whereas the low-attention tiles are negative evidences. For each class, a binary classifier $\mathbf{W}_{inst,m}$ classifies the gathered instances, and computes a tile-level smooth top-1 SVM loss $\mathcal{L}_{patch}$ which is added to $\mathcal{L}_{slide}$ to give the global loss term: $\mathcal{L} = c_1\mathcal{L}_{slide} + c_2\mathcal{L}_{patch}$ where c_1 and c_2 are hyperparameters.

The instance-level clustering does not directly affect the slide-level classification, rather, it encourages the learning of discriminative features by the model layer $\mathbf{W}_1$ to better separate classes. Note that the inference of slide-level classification does not use the instance level clustering.

2.2.2.2 Instance-level clustering for non-pathological slides

Given a tumor diagnosis task on a histopathological dataset, where some slides contain tumors, and others not, there is a caveat with the current formulation of the instance-level clustering: indeed, when assigning negative pseudo-labels to low-attention tiles, the model is actually learning that these tiles are *not* normal, and should therefore be classified as such. However, it is one of the hypotheses of MIL to consider all instances of a non-tumorous slide to be tumor-free. Therefore, all tiles originating from a non-tumorous slide should be classified as non-tumorous too. As a result, we decide to ignore the instance classifier $\mathbf{W}_{inst,normal}$, and instead use the tumor instance classifier $\mathbf{W}_{inst,tumor}$ to classify the B patches with the highest attention scores in normal slides as negative evidence of tumors. That way, we expect to reduce the number of false positives in the

final classification. Furthermore, we also ignore $\mathbf{W}_{inst,normal}$ to classify the patches from tumorous slides, and use only $\mathbf{W}_{inst,tumor}$. Indeed, without the need for histological subtyping, only a single binary classifier is required to perform the distinction between tumorous and non tumorous images. $\mathbf{W}_{inst,normal}$ remains therefore unused in our model.

2.2.2.3 Mixed supervision: Instance-level classification with ground-truth labels

Instead of selecting the tiles based on the highest and lowest attention scores, we introduce a mixed supervision formulation of the instance-level clustering which is now qualified as *instance-level classification*. On slides where tumorous regions were delineated by expert annotators, we propose to select tiles based on whether they belong to those tumorous regions. In doing so, we are sure that the instance-level classification layer learns with instances truly representative of two classes which is not the case when they are selected based on their attention score. This selection is only performed on slides for which annotations are available, which may be a small proportion of the histological training set. In fact, we hope to improve the model performance in classification and tumor localization (i.e. the tile-level classification of the slide) even with a small ratio of tile-level labels.

The training stage is therefore divided into two parts: first, only the subset of slides with pixel-wise annotations along with a subset of normal slides is used to train the model. This is to train the instance level classifier on true classes without the noise generated by the original pseudo-labeling procedure. We use various subset proportions (10, 50, 80% of the training slides) to measure the impact of the annotations on the results. For a tumorous slide containing N tiles such that $\mathrm{N} = \mathrm{N}_{tumor} + \mathrm{N}_{normal}$, we randomly sample $min(\mathrm{K}, \mathrm{N}_{tumor})$ tiles from the tumorous ones, and $min(\mathrm{K}, \mathrm{N}_{normal})$ among the normal ones, where K is a hyperparameter. Then, positive and negative labels are generated for the two sets of tiles, and a score is assigned to each tile t following: $l_t = \mathbf{W}_{inst,tumor}\mathbf{h}_t$. Second, the entire training set is used, this time without using any instance-level label, leveraging only slide-level labels. This procedure is summarized by Algorithm 1.

2.2.2.4 Implementation details

During training, we used a batch size of 1. We used the Adam optimizer [Kingma, 2014] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$, and a learning rate of 2.10^{-4}. All models were trained on a single NVIDIA GeForce GTX 1080 GPU. We used 70 epochs with early stopping after 20 epochs without improvement on the slide-level cross-entropy loss in the validation set.

During the first part of the training, we set $c_1 = c_2 = 0.5$ (when using tile-level labels) so that both loss terms $\mathcal{L}_{slide}$ and $\mathcal{L}_{patch}$ are weighted equally, and $c_1 = 0.7$ and $c_2 = 0.3$

Algorithm 1: Instance-level classification using tile-level labels

Data: $(\mathbf{h}_1, ..., \mathbf{h}_N)$, Y (the slide label), K, B
Result: l_Y
$l_Y \leftarrow \{\}$
if Y = *"tumor"* **then**
  $\mathrm{K}_{tumor} = min(\mathrm{K}, \mathrm{N}_{tumor})$
  $\mathrm{K}_{normal} = min(\mathrm{K}, \mathrm{N}_{normal})$
  Select $t_1, ..., t_{\mathrm{K}_{tumor}}$ // tumor tiles indexes
  Select $t'_1, ..., t'_{\mathrm{K}_{normal}}$ // normal tiles indexes
  for $t \leftarrow t_1, ..., t_{\mathrm{K}_{tumor}}$ **do**
    generate positive label $y_t = 1$
    $l_t = \mathbf{W}_{inst,tumor}\mathbf{h}_t$
  end
  $l_Y \leftarrow l_Y \cup \{l_t\}$
  for $t' \leftarrow t'_1, ..., t'_{\mathrm{K}_{normal}}$ **do**
    generate negative label $y_{t'} = 0$
    $l'_t = \mathbf{W}_{inst,tumor}\mathbf{h}'_t$
  end
  $l_Y \leftarrow l_Y \cup \{l'_t\}$
else
  Select $t_1, ..., t_B$
  for $t \leftarrow t_1, ..., t_B$ **do**
    generate negative label $y_t = 0$
    $l_t = \mathbf{W}_{inst,tumor}\mathbf{h}_t$
  end
  $l_Y \leftarrow l_Y \cup \{l_t\}$
end

during the second part, for these were the values used by [Lu, 2021] in their article. For hyperparameter K, we tested several increasing values (128, 256, 512, 1024, 5000), and we used the value $\mathrm{K} = 1024$, for which we reached the best performance. The value $B = 8$ was kept from the original model.

2.3 Results

2.3.1 Data description and experiments

2.3.1.1 The Camelyon16 dataset

The Camelyon16 challenge [Bejnordi, 2017] aimed at detecting metastases in H&E-stained WSIs of lymph node sections. The dataset contains 399 slides, split between a training set of 270 slides, and a test set of 129 slides. The slides were prepared and stained in two different medical centers. All slides that contain metastases (111

slides in the training set, 49 in the test set) have been exhaustively annotated (except for 20 of them in the training set, partially) by a group of expert pathologists. Annotations are available as XML files and can be converted to binary masks using the Automated Slide Analysis Platform (ASAP) open source software (https://github.com/computationalpathologygroup/ASAP). All slides were scanned at 40x magnification ($\simeq 0.25\mu$m/pixel).

2.3.1.2 Experimental setup

To measure the classification performance of our models, we decided to use the area under the receiver operating characteristic curve (AUC), as it was the metric used for the challenge, along with the F1-score. All models were evaluated on the challenge test set. The training set was further split into a training and a validation set. We used 5-fold cross validation to perform these splits in order to select the best performing model. In the training set, we randomly sampled k% ($k \in \{10, 50, 80\}$) of the slides to use with tile-level labels outside of the 20 with only partial annotations.

For the localization performance, we computed slide binary masks based on the outputs of the instance-clustering layer $\mathbf{W}_{inst,tumor}$ after having applied a threshold of 0.5. The masks were computed at the 5$^{\text{th}}$ resolution level. Furthermore, we used two different metrics for normal and tumorous slides: in the former, we computed the tile-wise specificity, i.e. the amount of tiles correctly classified as healthy divided by the total amount of healthy tiles. In the latter, we computed the Dice score in reference to the ground-truth mask. This metric was computed on 46 slides from the test set, as 3 metastatic slides out of the 49 were unavailable (one because the annotation file is absent from the dataset, the other two because of an error when computing the reference masks from the annotation file in ASAP).

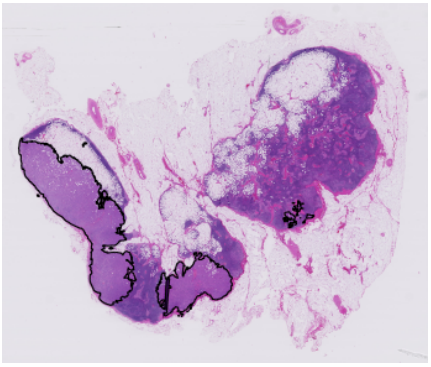
2.3.2 Evaluation

The test set classification results are shown in Table 2.1. For all models, we report the performance from the best fold, as it is usually done in challenges. We can see that the models with tile-level labels all outperform the CLAM algorithm in terms of slide-level classification, even when only 10% of the annotated slides were used. Moreover, we can see that the models trained with tile-level labels have also higher scores on localization tasks, both for tumorous (mean dice scores) and normal slides (mean specificity). In particular, models trained with tile-level labels tend to detect less falsely tumorous tiles than the reference model. Figure 2.2 shows an example of a tumorous slide from the test set and the masks computed using CLAM and two other annotation-based models. We can see that although the tumorous region is quite well detected in CLAM, there are many false positive tiles. When using tile-level labels however, these false positives tend

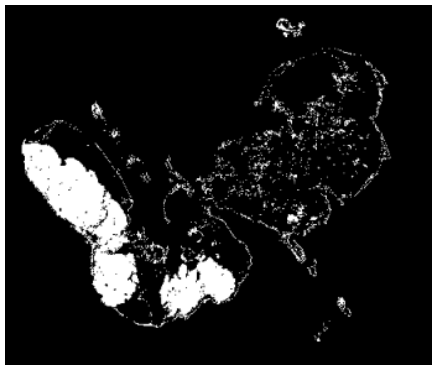
Method	AUC	F1-score	Mean Dice score (std)	Mean tile-level Specificity (std)
CLAM	0.895	0.8	0.215(0.28)	0.864(0.1)
CLAM w/ 10% annot.	0.924	0.835	0.35(0.263)	0.999(0.001)
CLAM w/ 50% annot.	0.939	**0.878**	0.375(0.279)	**0.999(0.001)**
CLAM w/ 80% annot.	**0.949**	0.873	**0.405(0.282)**	0.999(0.002)

Tab. 2.1.: Classification and localization metrics for the different methods.

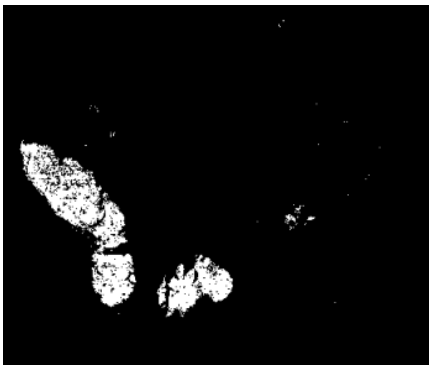
to disappear, and the detected region is closer to the ground truth. Figure 2.3 is another example from the test set of the difference between CLAM and annotation-based models. Figure 2.4 shows an example of a normal slide (also taken from the test set): here again, we can see that there are many more false positives when using CLAM. Using only 10% (resp. 80%) of the annotated slides, our method would have ranked 6th (resp. 4th) on the challenge open leaderboard.



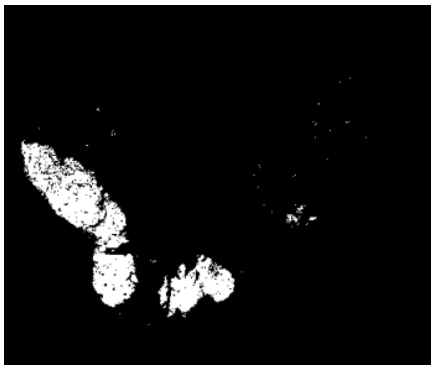
(a) Original slide



(b) Mask computed with CLAM.



(c) Using 10% annotations.

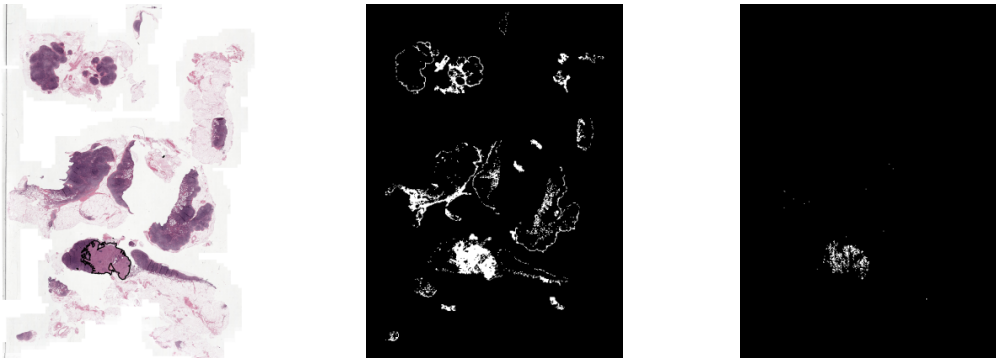


(d) Using 80% annotations.

Fig. 2.2.: Metastatic slide *test_016* from Camelyon16 (the metastasis region is delineated in black), next to binary masks computed using the different models. (b) displays the results of the CLAM algorithm, (c) and (d) show the results obtained using 10% and 80% of tile-level labels.

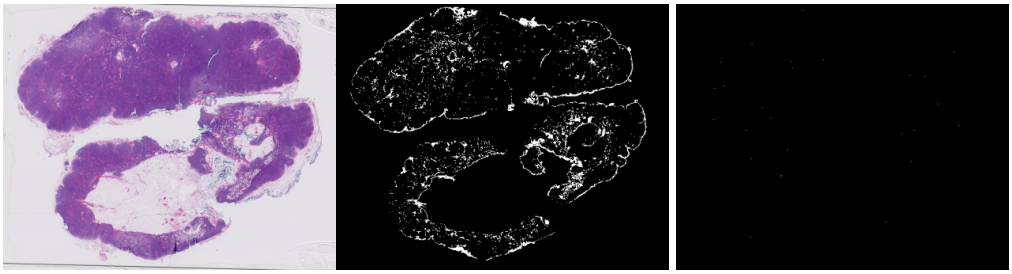
2.4 Discussion & Conclusion

In this work, we presented a mixed supervision approach for attention-based MIL. We proposed to add strong supervision in the tile classification branch of the CLAM algorithm for a subset of the training slides. We showed that even with a small amount



(a) Original slide **(b)** Mask computed with CLAM. **(c)** Using 10% annotations.

Fig. 2.3.: Metastatic slide *test_068* from Camelyon16 (the metastasis region is delineated in black), next to binary masks computed using CLAM, and the model with 10% of tile-level labels.



(a) Original slide **(b)** Mask computed with CLAM. **(c)** Using 10% annotations.

Fig. 2.4.: Slide *test_042* from Camelyon16 without tumor, next to binary masks computed using CLAM, and the model with 10% of tile-level labels.

(10%) of slides with pixel-wise annotations of tumors, we were able to obtain improved classification of slides and above all a better localization of the tiles with tumor tissue. In particular, we witnessed a sharp decrease in the number of false positive tiles, i.e. a better discrimination between tiles with normal and tumorous tissue. This better localization of tumors is crucial for WSI with only slide-level labels, since machine learning algorithms may provide the correct classification at the slide level but not necessarily with correct classifications at the tile level. This work shows that even with fairly limited pixel-wise annotation, it is possible to obtain more consistent and robust results at both local (tile) and global (slide) levels. Finally, the better localization of tumorous tiles and reduced rate of false positives in non-pathological slides also improves the interpretability of the provided slide classification algorithm which is key for the adoption of those algorithms in clinical practice.

The proposed approach may be improved in several ways. We have noticed that on the original CLAM and on our mixed supervised version, the localization performance of tumorous tiles may vary from one fold to the next. For our method, one or two folds (out of five) have significantly worse results than the other folds. This may be due to the lack of stain normalization performed since the slides originate from two distinct centers with

different acquisition equipment. Considering that only two centers were involved in the acquisition of the challenge data, we originally did not consider that stain normalization was essential. Moreover, in the method that won the challenge [Wang, 2016], some post-processing steps were applied to improve the localization accuracy. In our case, no post-processing was performed, and only a coarse localization map was computed, which could be refined for improved accuracy. Finally, the original CLAM paper considered the attention weights $a_{k,m}$ rather than the instance-level clustering as the source of information to localize pathologies. Therefore, it may also be interesting to supervise the attention mechanism with pixel-wise annotation similarly to what we proposed in this paper on the tile classification.

3

MS-CLAM: Mixed Supervision for the classification and localization of tumors in Whole Slide Images

Contents

Abstract
In this chapter, we extend the work introduced previously, by consolidating our take on mixed supervision for digital pathology with several additions to the model. Using the attention-based deep Multiple Instance Learning (MIL) model as our base weakly-supervised model, we propose once again to use mixed supervision – i.e., the use of both slide-level and patch-level labels – to improve both the classification and the localization performances of the original model, using only a limited amount of patch-level labeled slides. In addition to what we have already proposed, we define an attention loss term to regularize the attention between key instances, and a paired batch method to create balanced batches for the model. We also demonstrate the improvements of our method on not one but two different datasets: Camelyon16, which was already introduced in the previous chapter, and DigestPath2019, which we present here. First, we show that the changes made to the model already improve its performance and interpretability in the weakly-supervised setting. Furthermore, when using only between 12 and 62% of the total available patch-level annotations, we can reach performance close to fully-supervised models on Camelyon16 and DigestPath2019. This chapter was published in Medical Image Analysis [Tourniaire, 2023a].

3.1 Introduction

With the digitization of histological slides, deep learning algorithms have reached state-of-the-art performance on several tasks, e.g., cancer detection [Wang, 2016], tumor grading [Bulten, 2020] or mutation predictions and prognosis [Fu, 2020]. Nonetheless, Whole Slide Images (WSIs) still represent an atypical challenge in medical image analysis, as they often reach sizes of billions of pixels that are beyond the capacity of any current deep learning framework. For that matter, they are usually split into patches (or tiles) of smaller dimensions (e.g., 256x256 pixels), which are in turn processed by the models. As patch-level labels are usually unknown, because they are too time-consuming to obtain from expert pathologists, WSI analysis often falls under the Multiple Instance Learning (MIL) framework [Dietterich, 1997], where the slide is seen as a bag of which the tiles are instances. MIL often comes with weak supervision, meaning that only the slide-level label is known. Under this particular setting, two natural problems arise:

- Given the embeddings (or the probabilities) obtained at the instance level by a deep learning model, how can we recover the bag label? This is the MIL classification task.

- Given the bag label, is it possible to detect which are the key instances?

Regarding the latter, in a binary classification problem where one has to classify slides as containing tumorous tissue or not, the point is to be able to locate the tumor within the image, which is what we will refer to as tumor localization. This is a secondary task compared to the MIL classification one, but it is of great importance in histological image analysis, as it allows medical experts to confirm that the model's key instance selection *matches the slide prediction*. [Liu, 2012] refer to this as Key Instance Detection (KID), while putting forward that a good KID method allows for better bag classification.

3.1.1 Weakly-supervised classification

Under weak supervision, where only slide-level labels are known, several methods have been proposed to solve the binary tumor classification problem: in [Coudray, 2018], the authors simply follow the assumption that the tile labels are the same as the slide label, and proceed with this assumption to train a neural network to classify tiles, the outputs of which are averaged to recover the slide-level prediction. In [Campanella, 2019], the authors retrieve the S most suspicious tiles within the slide to feed a recurrent neural network (RNN) which in turn gives the slide-level classification. In [Courtiol, 2018], the authors adapt the WELDON method by [Durand, 2016] to MIL for binary classification of WSIs. However, these methods mainly focused on the bag-level classification. More recently, [Lerousseau, 2021] proposed a refined weak supervision approach by making use of the tumor cell percentage associated with each slide, instead of the sole binary label. With a training set of more than 18,000 slides from multiple cancer sites, they showed that they could outperform a fully-supervised model that was trained with tile-level labels on a binary tumor classification problem, while also producing convincing segmentation masks. For most of the aforementioned methods however, very large datasets were leveraged to obtain such good results, which are not necessarily available for every histological classification problem.

Self-supervision methods such as contrastive learning [Chopra, 2005], the aim of which is to minimize the distance between similar samples within a latent space, have been used as a training method to improve the feature extractors in histopathology. Using the Momentum Contrast v2 [Chen, 2020b] self-supervised framework to train a ResNet-50 neural network [He, 2016] as their tile-level feature extractor, [Dehaene, 2020], showed that they could reach slide-level classification scores close to the best fully-supervised method on the Camelyon16 challenge dataset [Bejnordi, 2017], but at the cost of intensive and time-consuming training, using many processing units.

3.1.2 Attention pooling

We already presented in the previous chapter the concept of attention pooling, introduced by [Ilse, 2018], the purpose of which is to compute attention scores to discriminate between low- and high-importance instances in the bag, and CLAM [Lu, 2021], based on the same architecture with an additional instance-specific classifier.

Self-attention [Vaswani, 2017], which can be used to model the interactions between instances within the bags, was used by [Rymarczyk, 2021] in addition to the classic attention-based MIL model (in which the instances are assumed to be independent), while exploring other MIL assumptions such as the presence-based or threshold-based assumptions. Unfortunately, [Rymarczyk, 2021] do not linger on the consequences of self-attention on the resulting attention scores, as bag-level classification is the main focus of their study. Self-attention was also used by [Li, 2021a], but this time as a distance measurement between the instance selected using max-pooling (denoted critical instance) and the other instances within the same bag, in a dual-stream model based on a self-supervised, multiscale feature extractor. Although their model integrates an instance classifier, it is mainly used for the critical instance selection, and serves no purpose in the localization of the tumor at inference time. [Shi, 2020] on the other hand, establish several theorems regarding attention-based MIL, notably showing how the instances' attention scores influence the bag-level prediction. They propose another method to compute the attention scores, called loss-based attention, and show on several datasets that it yields higher bag-level classification scores, and also boosts instance recall. However, the method was only tested on small MIL datasets, with few instances per bags (e.g. tens of instances) compared to what is commonly found in histopathology datasets (e.g. thousands of instances), and the method does not consider in particular the case of negative bags in binary MIL classification.

3.1.3 Mixed Supervision

[Shah, 2018] introduced mixed-supervision for image segmentation: strong supervision (i.e. pixel-level) and weak-supervision (bounding boxes, landmarks) were used together to improve segmentation while reducing the supervision cost. [Mlynarski, 2019] defined *mixed supervision* in their work as the joint use of image-level and pixel-level labels within an image. Compared to using only a few fully-annotated images, they showed that using the latter along with additional weakly-labeled images to train the same model also improved the segmentation results.

As for computational pathology, we already mentioned the work of [Ciga, 2021], where a two-headed patch-level ResNet-18 [He, 2016] was trained with both segmentation and classification labels, so as to reduce the segmentation labeling burden. In the case of WSI

analysis, we define mixed supervision as the joint use of instance- and bag-level labels. This kind of approach has previously been associated to semi supervision [Marini, 2021], where instance or patch-level labels are denoted “strong annotations", and bag or slide-level labels “weak annotations". Other works based on this joint approach have been published, especially on the topic of prostate cancer grading [Arvaniti, 2018; Bulten, 2020]. However, all these works either focused on WSI or instance classification, but did not try to perform both tasks simultaneously, using the labels from the two distinct levels of supervision. More recently, [Schmidt, 2022] proposed a model trained with both slide-level labels and a limited number of tile-level labels for tumor classification. The model is a tile-level classifier, that yields slide-level labels to generate pseudo labels for weakly-augmented tiles. After obtaining the pseudo labels, the same tiles are strongly augmented and classified by the same model using their respective true or pseudo labels all together. A loss function composed of both supervised and unsupervised terms is used to train the network. Like most of the previously mentioned frameworks, this one falls more in the semi-supervised learning category than in the mixedly-supervised one. Indeed, the purpose of their method lies in the training of the feature extractor, and not in a model that works at the slide-level. In [Lubrano di Scandalea, 2022], the authors devised a three-step approach, where a tile-level feature extractor is first trained in a self-supervised fashion using the SimCLR method [Chen, 2020a], then trained using both self-supervision and strong supervision on a small number of tiles. Once the feature extractor training is done, it is frozen, and used to extract features that will feed an attention-based deep MIL WSI classifier, in a similar way to [Lu, 2021]. This work, like the one by [Dehaene, 2020], aims at improving the performance of a WSI classifier by using a feature extractor specifically trained on histological data: mixed supervision is only used during the fine-tuning of the feature extractor, and is not fully integrated in the WSI classification process.

Although these previous methods made use of both tile and slide labels, it was always in the form of a combination of (1) semi supervision at the tile level and then (2) weak supervision at the slide level. Therefore, the joint use of tile and slide labels has been limited in prior work with usually separate statistical models to exploit both types of labels and distinct training process to perform mixed supervision.

3.1.4 Contributions

In this work, we propose to generalize the use of mixed supervision to both tile- and slide-level classification tasks in a joint framework to train a model more suited for histological slide analysis, with higher performance and interpretability. For that matter, we rely again on the CLAM architecture by [Lu, 2021], as the model is able to operate at both slide and tile levels, allowing for tumor localization in addition to slide classification. Moreover, we make use of a limited amount of tile-labeled slides, in the hope of reducing

the tedious work required from expert pathologists for the precise tumor delineation in histological datasets. Our contributions are listed as follows:

- The previously introduced strategy, where the tile-level classifier is trained on both true tile labels (when available) and pseudo-labels generated using the tiles' attention scores, is improved twofold: first, the two-step paradigm we initially developed now fits in a single, unique step. Second, to correct for the potential class imbalance between tumorous and non-tumorous tiles, we also propose a paired batch method that uses both kinds of slides at the same time at each training step.

- To better target the key instances responsible for the bag label and obtain a model less focused on few specific instances, we design a new loss function based on the attention scores of the slide-level classifier. The loss also enforces a uniform spread of the attention on the relevant tiles in the slide, which improves the slide-level classification as well as the interpretability of the model during inference. We propose an exponential weighted sampling strategy, designed to simplify the training procedure in a single step, using both annotated and unannotated slides at the same time.

- We evaluate our method on two different histology datasets tasked with binary tumor classification and localization, and show that it indeed improves the consistency of the model between the tile- and the slide-level predictions, i.e., classification and localization. Throughout the rest of the paper, our model is coined MS-CLAM, for *Mixedly Supervised*-CLAM.

3.2 Methods

3.2.1 CLAM

The CLAM framework [Lu, 2021] has already been presented in the previous chapter. For convenience reasons, we simply recall the main characteristics of the model. For each slide, a latent representation $\mathbf{z}$ is obtained by deriving a low-dimensional feature vector $\mathbf{h}_k$ for each tile $k \in \{1, ..., \mathrm{N}\}$, through a frozen deep convolutional neural network (e.g., a Resnet [He, 2016] pretrained on ImageNet). For each instance k, an attention score is computed:

$$a_k = \frac{\mathbf{W}_a(\tanh(\mathbf{V}_a\mathbf{h}_k^\top) \odot \mathrm{sigm}(\mathbf{U}_a\mathbf{h}_k^\top))}{\sum_{i=1}^{N} \exp\{\mathbf{W}_a(\tanh(\mathbf{V}_a\mathbf{h}_i^\top) \odot \mathrm{sigm}(\mathbf{U}_a\mathbf{h}_i^\top))\}} \tag{3.1}$$

where $\mathbf{U}_a$, $\mathbf{V}_a$, $\mathbf{W}_a$ are layers of a trainable neural network. The attention scores sum to 1, and the final representation of the slide is a sum of the instance feature vectors

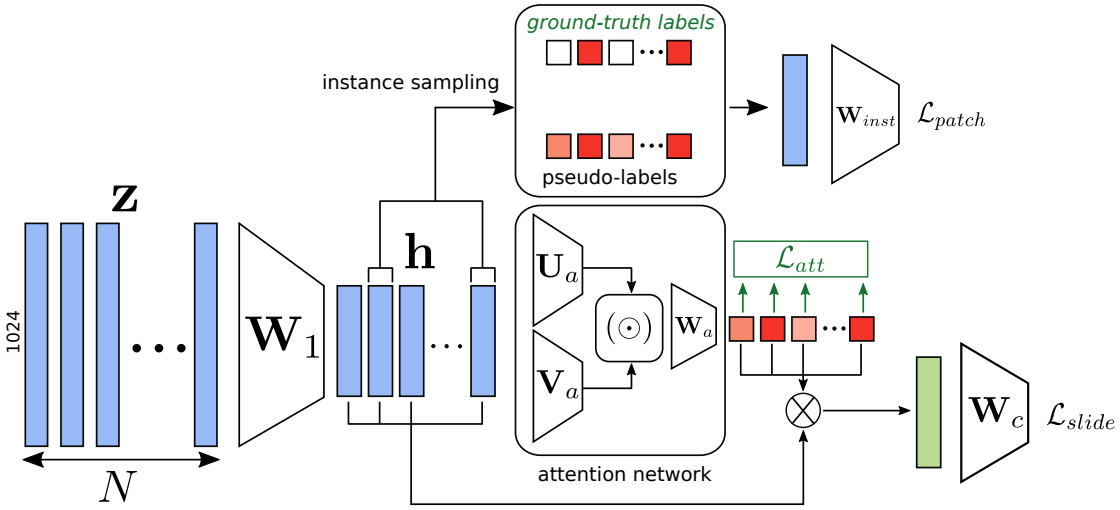


Fig. 3.1.: Overview of the MS-CLAM model. Regarding the attention scores, a faint (resp. bright) color represents a low (resp. high) score. For the ground-truth colors, red means the instance is positive (with respect to the bag label), while no color means the instance is negative. The light-green rectangle represents the attention-weighted average of the feature vectors. Our contributions to the original CLAM architecture are printed in dark green.

weighted by their respective attention scores, i.e. $\mathrm{A}^\top\mathbf{h}$ ($\mathbf{h} = [\mathbf{h}_1 ... \mathbf{h}_N]^\top$). Figure 3.1 recalls the graphical description of CLAM, along with our contributions that we will describe hereafter. In what follows, we refer to non-tumorous slides and tissue as “normal slides" and “normal tissue", regardless of the presence of artifacts in the tissue (tear, air bubble, etc...), given we tackle the problem of tumor vs. non-tumor classification of WSIs.

3.2.2 Instance-level classification supervision

In CLAM, under the weak supervision setting, we already saw that tiles are classified using pseudo labels generated based on the attention scores: the B tiles with the highest (resp. lowest) attention scores are labeled as positive (resp. negative) with respect to the slide-level label. We showed in chapter 2 that this assumption was wrong in the case of an all-negative bag (i.e., a normal slide), but another issue is that the parameter B is fixed, meaning that tiles are invariably sampled within slides, regardless of the actual number of tiles representative of the slide-level class. Therefore, only small values of B can help avoid sampling the wrong tiles, but in turn limit the number of training samples.

With MS-CLAM, we propose to solve some of these issues with the help of mixed supervision and the paired batch method. Like in the previous chapter, we once again make use of available tile-level labels, which are used instead of the pseudo labels on the corresponding slides, to train the tile-level classifier with both more numerous and

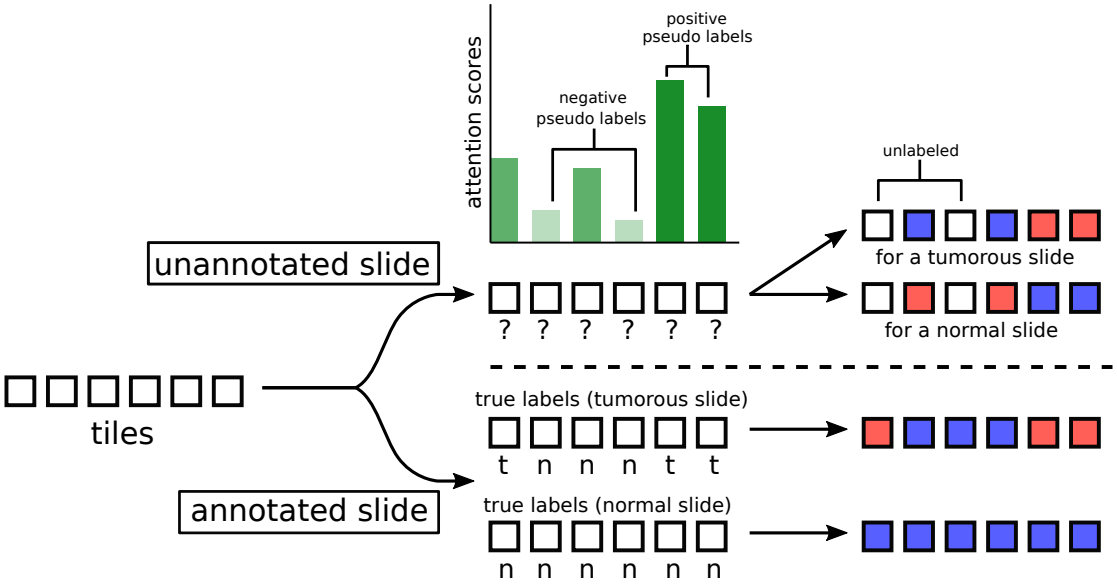


Fig. 3.2.: The two methods for labeling tiles (represented as colored squares) in WSIs: the top part corresponds to the weakly-supervised case (already used in CLAM), where attention scores are used to generate pseudo-labels for the tiles. The bottom part on the other hand corresponds to the case where the tile labels are available (only for MS-CLAM). The number of sampled tiles in the weakly-supervised setting here is $B = 2$. Red (resp. blue) squares represent tumorous (resp. normal) tiles.

accurate labels than in the original setting. To distinguish between the cases where tile labels are known or not, we use two different hyperparameters to sample the instances: B_+ when labels are available, and B_- when they are not ($B_+ > B_-$). For the case where B_+ might be greater than the actual number of tumorous tiles N_{tum} in the slide, we set $B_+ = N_{tum}$. Third, since in normal slides all tiles are normal (thanks to the MIL assumption), we use a different tile-sampling strategy, as the original method assigned wrong labels to the tiles with low attention scores. In our case, in normal slides, sampled tiles are only assigned the same label as the slide. Moreover, if we sample B tumorous tiles in tumorous slides then we only sample $B/2$ normal tiles in both tumorous and normal slides, to improve the balance between the two classes. A summary of the two tile labeling approaches is shown in Figure 3.2.

To train the original CLAM model, a single slide is sampled at each step, and a tile batch containing $2B$ instances is generated. With the modifications we made to the tile labeling, this approach is no longer recommended, because for normal slides, where only a single class is represented among the tiles, this would mean that the tile batch would correspond to a single label. Alternating between tumorous slides – with both labels represented at the tile level, and normal slides – with only a single label present in the tiles – could lead to unstable gradient computation for the tile classifier. To circumvent this issue, we propose to simultaneously process one tumorous and one normal slide (which we refer to as *Double Sampling*), and build a tile batch using instances from both

slides: we call this process the paired batch method. Figure 3.3 represents how the batch of tiles is produced.

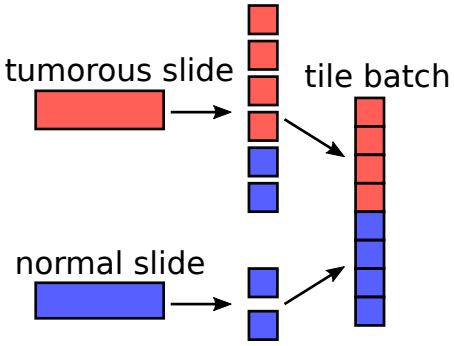


Fig. 3.3.: The paired batch creation process. The tumorous slide provides B tumorous and $B/2$ normal tiles, while the normal slide provides $B/2$ normal tiles to make a $2B$-sized tile batch ($B = 4$ in the figure).

3.2.3 Attention Loss

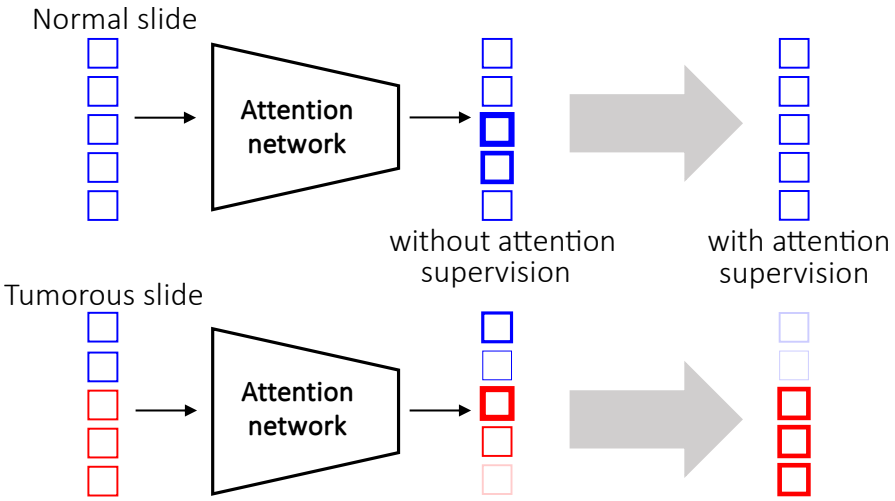


Fig. 3.4.: The goal of the supervision of the attention scores. Instances are represented as colored squares. The red color (resp. blue) represents instances with tumor (resp. without). A thick, bright square means the instance was given a high attention score, whereas a thin, faint square means the instance was given a low attention score. In normal slides, attention scores should be even, so as to weight each instance equally. In tumorous slides, tumorous patches' attention scores should be higher than the non-tumorous ones, but equally weighted between them. The attention loss is designed to guide the attention on the most relevant patches.

Until now, the mixed supervision was only designed for the instance-level classification, with a collateral impact on the slide-level classification. In CLAM, no constraint is applied on the attention scores, except that their sum must be equal to 1 (to be invariant to the bag size). In the weakly-supervised setting, we noticed that the attention scores associated to the patches were highly unbalanced, with only a few instances weighted much higher than the rest, whatever the slide label be. Although this effect is rather undetectable on small bag sizes, e.g., a few hundreds of instances, when facing bags with tens of thousands of instances, which is typical of WSIs, the attention tends to be

focused unevenly on a few instances only, or even sometimes on a single one. Here, we propose a new loss term based on the attention scores to orient the attention spread towards the most important patches, but also to equalize the attention among them (Figure 3.1, bottom). Figure 3.4 describes the purpose of the supervision. Oddly enough, the imbalance between attention scores is noticeable in both normal and tumorous slides. Ideally, we wish the attention scores to be distributed differently depending on the slide label. For normal slides, we want the attention scores to be all equal, i.e., $\forall(i,j) \in \{1,...,N\}^2, a_i = a_j = 1/N$. Seeing the attention scores as a probability distribution, this condition can be expressed through the means of the Shannon entropy [Shannon, 1948]. The entropy reaches a maximum when all of the outcomes are equally likely, and the maximum value is $\log N$ given there are N possible outcomes. For normal slides, the attention loss is therefore written as follows:

$$\mathcal{L}_{att} = \frac{1}{\log N} \sum_{k=1}^{N} a_k \log a_k \tag{3.2}$$

The entropy of the attention scores distribution should be maximum, meaning each instance is given equal importance in the weighted sum before the final classification layer. In other words, the attention pooling should mimic the mean pooling in the case of a normal slide. We take as a penalty term the negative entropy, normalized by $\log N$ to account for any number of instances within a WSI. This loss term does not require the availability of tile-level labels, since all tiles have the same label, which is the one of the slide. Such a regularization term on attention scores was also proposed by [Lu, 2019] and [Sharma, 2021], in the form of a KL-divergence with respect to a uniform distribution, which is equivalent to the entropy up to a constant. For tumorous slides on the other hand, the expression involves three terms, because we want to reach the three following objectives:

1. The attention scores of non-tumorous instances should be close to zero, as they have little to no impact on the slide label.

2. From the previous condition, we have that the tumorous instances' attention scores should be the only non-zero ones. Put differently, the sum of the attention scores of tumorous instances should be close to 1.

3. The entropy of the tumorous attention scores should be maximum, i.e., each tumorous instance should be weighted equally before the final classification. This is to ensure that all instances containing tumor contribute as equally as possible to the final prediction.

Therefore, the attention loss for tumorous slides is expressed as follows:

$$\mathcal{L}_{att} = \sum_{i=n_1}^{n_s} a_i + \frac{1}{\log m} \sum_{j=t_1}^{t_m} a_j \log a_j - \sum_{j=t_1}^{t_m} a_j \tag{3.3}$$

where m (resp. s) is the number of tumorous (resp. non-tumorous) instances, $t_1, ..., t_m$ are the indices of the tumorous attention scores, and $n_1, ..., n_s$ the indices of the non-tumorous ones. Contrary to the previous case, this loss term requires the knowledge of tile-level labels, hence why the mixed supervision is used during training. Table 3.1 summarizes the various losses computation depending on the slide's label.

Slide label	Inst. label availability	Inst. label	Att. loss	Inst. label nature	Inst. loss	Number of inst. in batch
Normal	yes	Normal	$-H(A)$	true label	CE	$B_+/2$ or $B_-/2$
Tumor	no	N/A	N/A	pseudo-label	CE	B_- (tum.) and $B_-/2$ (norm.)
	yes	Normal Tumor	$\|A_n\|_1$ $-H(A_t) - \|A_t\|_1$	true label true label	CE	$B_+/2$ B_+

Tab. 3.1.: Summary of the losses for each kind of slide label. The table also indicates how the losses are handled depending on the tile labels availability. The H function stands for the Shannon entropy, and A represents the vector of all attention scores. Similarly, A_t is the vector of tumorous tiles' attention scores, and A_n is the vector of normal tiles' attention scores. CE stands for cross-entropy.

3.2.4 Exponential Weighted Sampling

In the previous chapter, a two-step training procedure was designed to train the model where first only slides with tile-level labels were used, and then the entire training set in a second phase. This was done to ensure that the tile-level classifier was first trained on true labels, before being trained on pseudo-labels. In this paper, we introduce a sampling strategy devised to train the model in a single phase, where slides with tile-level labels are decreasingly more likely to be sampled during training, until all slides are sampled uniformly. Assuming we have $N_t = \hat{N}_t + \tilde{N}_t$ tumorous slides, where $\hat{N}_t$ is the number of tile-level labeled slides and $\tilde{N}_t$ the number of slides with unlabeled tiles: each slide i is assigned a sampling weight w_i, equal to a value $W > 1$ when the tile-level labels are known, or 1 otherwise. Then, these weights are converted into probabilities following $p_i = w_i / \sum_j w_j$. These probabilities serve as the parameters of a multinomial distribution used to sample the slides at each epoch. To progressively reduce the oversampling of annotated slides, their corresponding weights w_k are multiplied by a decay factor $\gamma < 1$ at the end of each epoch, until all slides' weights are equal to 1, resulting in equal sampling probabilities. Algorithm 2 summarizes the sampling strategy.

3.2.5 MS-CLAM without tile-level labels

The absence of tile-level annotated slides (weak supervision only) can be seen as a particular case, which requires several adjustments to the method. Concerning the

Algorithm 2: Tumorous slides' sampling strategy.

Data: Initial weight W, decay factor γ, number of training epochs E, the set of tumorous slides $\{S_1, ..., S_{N_t}\}$

for $e \leftarrow \{1, ..., E\}$ **do**
  for $i \leftarrow \{1, ..., N_t\}$ **do**
    if S_i *has tile-level labels* **then**
      $w_i \leftarrow W$
    else
      $w_i \leftarrow 1$
    end
    $p_i \leftarrow w_i / \sum_i w_i$
  end
  sample slides from Multinomial$(p_1, ..., p_{N_t})$
  $W \leftarrow \gamma W$
end

attention loss, we can only use Eq. 3.2, since Eq. 3.3 requires the knowledge of tile-level labels. Still, the paired batch method remains applicable, along with the other modifications we made to the model. Concerning the exponential weighted sampling, we fix $W = 1$ and $\gamma = 1$ so that all tumorous slides are sampled randomly (equal weights and no decay). This setting corresponds to the Double Sampling strategy we mention in section 3.2.2.

3.3 Materials

3.3.1 The Camelyon16 dataset

The Camelyon16 challenge [Bejnordi, 2017] was already introduced in the previous chapter. Here, we simply recall the class distribution of the dataset in Table 3.2, and add further comments on several aspects of the data. Indeed, this dataset is particularly challenging among histological datasets, as the metastasis size from one slide to the other greatly varies. From a MIL standpoint, this means that the number of positive instances per bag can differ significantly from one bag to the other, to the point where there can be only a few positive instances for tens of thousands of negative ones in one bag, while in the other there are nearly no negative instances. This variability is expressed in the form of a box plot in Figure 3.5 (notice the log scale on the horizontal axis). The mean percentage of tumorous tiles within tumorous slides is 6.3%, while the

Slide class	Train set	Test set
Normal	159	80
Metastatic	111	49

Tab. 3.2.: Summary of the Camelyon16 dataset class distribution.

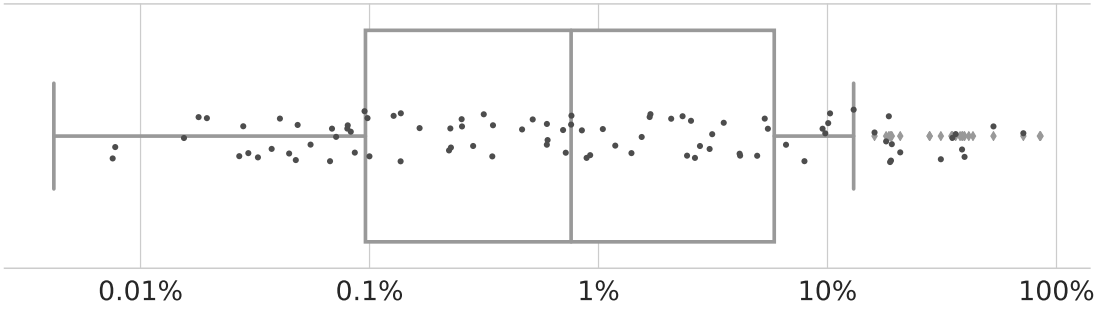


Fig. 3.5.: A box plot showing the percentage of tumorous tiles (log-scaled) in tumorous slides in the training set of Camelyon16. The grey, diamond-shaped points represent the outliers, while the black, circular points correspond to the data points themselves.

median percentage is only 0.76%. Therefore, we expect the attention loss to help the model coping with this variability among the positive bags. During the experiments, we split the training/validation set 5 times into a training (80%) and a validation (20%) set in a 5-fold cross-validation fashion, and report the average performance of the model on the competition test set

3.3.2 The DigestPath2019 dataset

The DigestPath2019 challenge [Li, 2019] was organized around two different gastric and colon histology datasets. In this work, we focused on the second dataset, for which the challenge task was to classify and segment tissue in colonoscopy images. It does not contain entire WSIs but regions selected within these colonoscopy slides. The resulting images have an average size of 5000x5000 pixels. As the competition test set is unavailable for download, we performed all of our experiments on the competition training set, which contains 660 images in total (from 324 patients, coming from 4 different centers), of which 250 (from 93 patients) display tumor regions. We perform a 5-fold cross-validation of the models on the competition training set (i.e., the available 660 images). For each fold, the training set is again divided between 80% training and 20% validation.

The challenge website mentions that some malignant glands were missed by pathologists, so the annotations are not exhaustive per se, but are considered as such during the experiments. As opposed to the ones in the Camelyon16 dataset, tumorous slides in DigestPath2019 display a much wider tumorous area with respect to the total tissue area: the mean percentage of tumorous tiles in tumorous images in this case is 31.8%, with a median value at 28.9%. The boxplot Figure 3.6 summarizes the ratio of tumorous tiles within tumorous images for DigestPath2019.

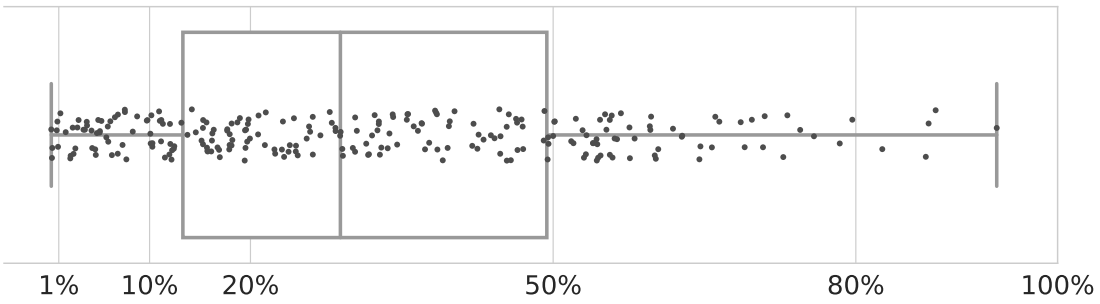


Fig. 3.6.: A box plot showing the percentage of tumorous tiles in tumorous images in the training set of DigestPath2019.

3.3.3 Data pre-processing

For both datasets, we followed the usual procedure for histological image analysis: first, the tissue region is filtered in the images using a threshold on the saturation channel in the HSV color space. For some images in the DigestPath2019 dataset, blurry regions were filtered out using a blur detector [Golestaneh, 2017]. Then, the images are split into squared tiles of dimensions 256x256 (for Camelyon16), 128x128 (for DigestPath2019). For DigestPath2019, the choice of a proper tile size is a more critical choice than for Camelyon16, since the original images have much smaller sizes (they are not original WSI files). Therefore, smaller tiles allow for a more accurate localization of the tumor within the images. However, since the feature extractor was initially trained with inputs of 224x224 pixels, using an input size too different from the initial dimensions would result in a performance decay. To assign a label to the tiles, we used a different approach for the two datasets since they both display very different tumorous area ratios as shown in Figures 3.5 and 3.6. For Camelyon16, we looked at the slide with the smallest tumorous region, centered a tile around this region, and computed the percentage of tumor inside the tile to obtain a threshold: the value we obtained was 20%. On DigestPath2019, as tumor regions were quite wide and similar from one slide to the other, we stuck to a 50% threshold. We used the same Imagenet pre-trained, Resnet-50 backbone as [Lu, 2021] to pre-extract the features from the tiles before training the attention layers.

As all images from each dataset were annotated, we randomly selected k% of the slides to be used with tile-level labels, with $k \in \{0, 6, 12, 25, 62, 100\}$, to evaluate the model's performance with an increasing percentage of available annotations.

3.3.4 Experimental setting

For all the experiments, we used the Adam optimizer [Kingma, 2014] during training, with the default values $\beta_1 = 0.9$ and $\beta_2 = 0.99$, with a learning rate of 2×10^{-4}. We took $B_+ = 1024$ for Camelyon16, $B_+ = 256$ for DigestPath2019, and $B_- = 8$ for the

tile sampling parameters. For the exponential weighted sampling, we kept $W = 90$ and $\gamma = 0.9$. All models were trained for either 50 epochs (DigestPath2019) or 90 epochs (Camelyon16) on a single NVIDIA GeForce 1080 GTX Ti GPU. A 5-fold training on Camelyon16 takes approximately 5 hours, while on DigestPath2019 it takes only 2 hours 30 minutes.

3.4 Results

To evaluate the results of the experiments, we use several classification and localization metrics. Although the model does not directly produce tumor masks, we use the outputs of the tile classifier to compute binary tile masks of the slides, which are then compared to the reference tumor masks using localization metrics. For the slide-level classification, we mainly look at the accuracy, the F1-score, and the AUC (Area Under the ROC Curve, which was the reference metric for both datasets). For both datasets, to evaluate the quality of the tumor masks, we first compute the reference tile-accurate masks based on the tumor delineations done by experts and using the assumptions made in 3.3.3 as to which tiles are considered tumorous. This is done to allow for a fairer comparison between the predicted and the reference masks, since none of the models are pixel-accurate. To compute the masks using the tile-level predictions, we use a threshold of 0.5 on the output probability of the tile classifier. For CLAM SB or MB, as two different tile-level classifiers coexist, the one corresponding to the tumor class is used. On Camelyon16, all masks are computed at the 5^{th} magnification level, the same magnification level used during the challenge for localization evaluation. Predicted masks are evaluated using the Dice score on tumorous slides. On normal slides on the other hand, we simply compute the tile-level specificity of each model.

3.4.1 Baselines

Aside from CLAM, we also compare the performance of our model with several baselines:

- **Weakly-supervised baselines.** We compare our model with other weakly-supervised models such as TransMIL [Shao, 2021b] and DS-MIL [Li, 2021a]. For the latter, an important part of the method is the training of a feature extractor in a self-supervised fashion. According to the authors, it took 2 weeks on 6 GPUs to complete this part for Camelyon16 only; therefore, we decided to use the pre-computed features they provide on github for this dataset in particular. We also show the performance of DS-MIL when using the same feature extractor as for CLAM, MS-CLAM and TransMIL (i.e., the ImageNet pre-trained ResNet-50). However, since neither TransMIL nor DS-MIL have a dedicated tile-level classifier, we only compare the

performances of these models on the slide-level classification task, but not on the tumor localization one.

- **Backbone fine-tuning.** We investigate the potential benefit of fine-tuning the feature extractor on the available tile labels before training CLAM and MS-CLAM. To this end, for each fold, we fine-tune the ImageNet pre-trained ResNet-50 backbone on tiles taken from the same slides used to supervise the training of MS-CLAM. The fine-tuning can be seen as supervised tile classification, after which we discard the final classification layer to recover the features. We use the Adam optimizer and cross-entropy loss, with a learning rate of 1×10^{-3}, divided by ten every time the loss plateaus for 15 epochs. The model is trained during 200 epochs for each fold, or until the loss stops decreasing during 20 epochs. We indicate *(FT)* in the tables next to models trained atop a fine-tuned backbone.

3.4.2 Slide-level classification

Model	% of annot. images	AUC (↑)	Acc. (↑)	F1-score (↑)
CLAM SB	0	0.973 ± 0.022	0.924 ± 0.039	0.901 ± 0.052
CLAM MB	0	0.953 ± 0.043	0.906 ± 0.048	0.879 ± 0.063
TransMIL	0	0.982 ± 0.015	0.932 ± 0.05	0.927 ± 0.053
DS-MIL (ImageNet)	0	0.615 ± 0.091	0.511 ± 0.138	0.212 ± 0.291
MS-CLAM	0	0.977 ± 0.014	0.941 ± 0.022	0.921 ± 0.026
MS-CLAM	6	0.982 ± 0.012	0.941 ± 0.026	0.922 ± 0.029
MS-CLAM	12	0.982 ± 0.013	**0.955 ± 0.034**	**0.940 ± 0.042**
MS-CLAM	25	0.980 ± 0.016	0.944 ± 0.038	0.927 ± 0.047
MS-CLAM	62	**0.984 ± 0.016**	0.947 ± 0.036	0.933 ± 0.041
MS-CLAM	100	0.981 ± 0.019	0.946 ± 0.035	0.931 ± 0.041
CLAM SB (FT)	12	0.987 ± 0.005	0.945 ± 0.014	0.928 ± 0.020
MS-CLAM (FT)	12	0.989 ± 0.009	0.953 ± 0.024	0.940 ± 0.027
CLAM SB (FT)	62	0.983 ± 0.013	0.953 ± 0.017	0.938 ± 0.022
MS-CLAM (FT)	62	**0.991 ± 0.008**	0.955 ± 0.029	0.941 ± 0.037
CLAM SB (FT)	100	0.990 ± 0.006	**0.964 ± 0.023**	**0.954 ± 0.025**
MS-CLAM (FT)	100	**0.991 ± 0.007**	0.950 ± 0.027	0.936 ± 0.033

Tab. 3.3.: Classification metrics over a 5-fold CV of the DigestPath2019 training set (± standard error is indicated for each experiment and metric).

Table 3.3 shows the results of the image classification for DigestPath2019. Without the use of any tile-level annotation, the addition of the attention loss on normal slides along with the paired batch method lead to an improvement of all the classification metrics, in particular the accuracy (improved by 1.7% compared to CLAM SB, 3.5% compared to CLAM MB) and the F1-score (improved by 2% compared to CLAM SB, 4.2% compared to CLAM MB). With an annotation burden as low as 12% of the available slides in the training set, we also notice an improvement of all the classification metrics for the MS-CLAM model compared to the CLAM baseline. While the gain in AUC is

rather marginal, there is a gradual improvement regarding the accuracy and the F1 score, meaning the model is less prone to classification errors.

Among the weakly-supervised baselines, TransMIL reaches the highest performance, on par with MS-CLAM using 6% of the annotated slides. However, the model does not provide tumor locations, therefore the slide-level predictions suffer from a lack of interpretability. For DS-MIL, without access to a dedicated backbone for this dataset, the model obtains poor performance compared to all the other methods.

Finally, when training CLAM on top of a fine-tuned feature extractor, its performance improves as the number of annotated slides increases. Nonetheless, when using only 12%, it still does not reach the performance of MS-CLAM (without FT) with the same amount of annotations in terms of accuracy and F1-score. Notwithstanding the case with 100% annotated slides, MS-CLAM (FT) still manages to improve the performance of CLAM (FT).

Classification results for Camelyon16 are detailed in Table 3.4. For the MS-CLAM models, there is again an improvement of all the classification metrics compared to CLAM when using 12% of annotated slides or more, which corresponds to only 11 tumorous slides in the Camelyon16 training set. We see similar effects to the ones observed on DigestPath2019: accuracy and F1-score largely profit from the added supervision (F1-score improved by 5% with only 6% annotated slides).

Just like for DigestPath2019, TransMIL is comparable to MS-CLAM trained with 6% annotated slides (with a 2% overhead in AUC), yet still without localization information. For DS-MIL, this time, we have access to a specifically trained backbone. The self-supervised version of the model reaches the highest AUC, but when looking at the accuracy and the F1-score, it falls behind TransMIL or MS-CLAM with 6% of annotated slides. Similar to what we observed on DigestPath2019, DS-MIL achieves poor results when using the ImageNet pre-trained backbone.

The results obtained by the FT models on Camelyon16 are similar to the ones obtained on DigestPath2019, although this time MS-CLAM (FT) is still superior to its CLAM (FT) counterpart when using 100% of the annotated samples. Interestingly enough, for the 12% and the 62% settings, CLAM (FT) and MS-CLAM (without FT) reach similar accuracy and F1-score, although CLAM (FT) obtains a higher AUC. Nonetheless, the computational burden is much lower for MS-CLAM, as no backbone fine-tuning is required.

Model	% of annot. images	AUC (↑)	Acc. (↑)	F1-score (↑)
CLAM SB	0	0.883 ± 0.033	0.863 ± 0.027	0.797 ± 0.049
CLAM MB	0	0.907 ± 0.010	0.870 ± 0.028	0.806 ± 0.049
TransMIL	0	0.910 ± 0.021	0.874 ± 0.030	0.857 ± 0.039
DS-MIL (SS)	0	**0.966 ± 0.021**	0.879 ± 0.066	0.849 ± 0.065
DS-MIL (ImageNet)	0	0.467 ± 0.185	0.478 ± 0.130	0.331 ± 0.302
MS-CLAM	0	0.884 ± 0.020	0.888 ± 0.026	0.830 ± 0.044
MS-CLAM	6	0.889 ± 0.017	0.898 ± 0.026	0.859 ± 0.022
MS-CLAM	12	0.908 ± 0.013	0.899 ± 0.028	0.861 ± 0.031
MS-CLAM	25	0.911 ± 0.016	0.902 ± 0.028	0.867 ± 0.035
MS-CLAM	62	0.932 ± 0.008	0.938 ± 0.009	0.913 ± 0.013
MS-CLAM	100	0.939 ± 0.008	**0.938 ± 0.012**	**0.916 ± 0.017**
CLAM SB (FT)	12	0.932 ± 0.029	0.898 ± 0.030	0.860 ± 0.039
MS-CLAM (FT)	12	0.921 ± 0.036	0.910 ± 0.018	0.877 ± 0.029
CLAM SB (FT)	62	0.967 ± 0.011	0.936 ± 0.025	0.915 ± 0.033
MS-CLAM (FT)	62	0.970 ± 0.012	0.935 ± 0.015	0.916 ± 0.018
CLAM SB (FT)	100	0.980 ± 0.008	0.946 ± 0.016	0.928 ± 0.020
MS-CLAM (FT)	100	**0.982 ± 0.013**	**0.950 ± 0.018**	**0.935 ± 0.022**

Tab. 3.4.: Classification metrics on the Camelyon16 test set. All metrics are averaged on a 5-fold cross validation split of the training set (± a standard error reported).

Model	% of annot. slides	Dice score (tum)	Specificity (norm)
CLAM SB	0	0.520 ± 0.075	0.525 ± 0.075
CLAM MB	0	0.443 ± 0.087	0.443 ± 0.101
MS-CLAM	0	0.310 ± 0.051	**0.998 ± 0.001**
MS-CLAM	6	0.460 ± 0.073	0.998 ± 0.002
MS-CLAM	12	0.530 ± 0.063	0.997 ± 0.003
MS-CLAM	25	0.595 ± 0.061	0.993 ± 0.005
MS-CLAM	62	**0.677 ± 0.026**	0.978 ± 0.009
MS-CLAM	100	0.676 ± 0.027	0.960 ± 0.016
CLAM SB (FT)	12	0.598 ± 0.069	0.695 ± 0.233
MS-CLAM (FT)	12	0.453 ± 0.060	**0.997 ± 0.001**
CLAM SB (FT)	62	0.582 ± 0.123	0.590 ± 0.306
MS-CLAM (FT)	62	**0.715 ± 0.027**	0.976 ± 0.011
CLAM SB (FT)	100	0.596 ± 0.094	0.609 ± 0.289
MS-CLAM (FT)	100	0.714 ± 0.014	0.967 ± 0.014

Tab. 3.5.: Localization metrics on DigestPath2019. All metrics are averaged on a 5-fold cross validation split of the training set (± a standard error reported).

3.4.3 Localization of tumor regions

Localization results (i.e., the tumor masks derived from the tile classifier) for DigestPath2019 (resp. Camelyon16) are detailed in Table 3.5 (resp. Table 3.6). For both datasets, with the exception of MS-CLAM with 0 and 6% annotations on DigestPath2019, we obtain both a higher mean Dice score on tumorous images, and a higher mean specificity for normal images, close to 1, indicating very few false positives. In both

cases, the baseline models lack specificity and fail to point at the tumor region accurately, triggering many false positives in the case of normal images. Yet, for DigestPath2019 in particular, the tumor region in tumorous slides spans a relatively wide area (see Figure 3.6), sometimes covering nearly all of the tissue. Therefore, the Dice score for CLAM is fairly high (0.520) . Conversely, MS-CLAM models tend to have a much higher specificity, at the cost of a lower sensitivity, which gradually improves with the percentage of tile-level labeled slides. This, combined with the previous observation on the ratio between tumor and healthy tissue in the case of DigestPath2019, explains why the Dice score of MS-CLAM with 0 and 6% annotations is lower for this dataset in particular compared to CLAM SB. It is in the images that contain smaller tumor regions regarding the total tissue area that the difference is perceptible. An example of the latter case is visible Figure 3.7, where the CLAM model has classified all tissue as tumorous, whereas it is in fact limited to sub-parts of the image. With MS-CLAM on the other hand, the tumorous tissue is correctly located within all the annotated regions. Furthermore, this image is a likely example of incomplete annotations, as several small unmarked regions in the image, in particular at the bottom right, are likely to be tumorous. Nonetheless, these regions are still correctly picked by the models, as shown on the two right-most masks in Figure 3.7. We can also notice that adding more supervision in MS-CLAM improves the recall of tumorous instances, which is expected since the model trains on more tumorous tile samples. In terms of localization, although fine-tuning the feature extractor improves the performance of CLAM in terms of both Dice score and specificity, the latter still remains far below what is achieved with MS-CLAM, regardless of the backbone used. When using more annotated samples (62% onward), the performance of the model improves with fine-tuning.

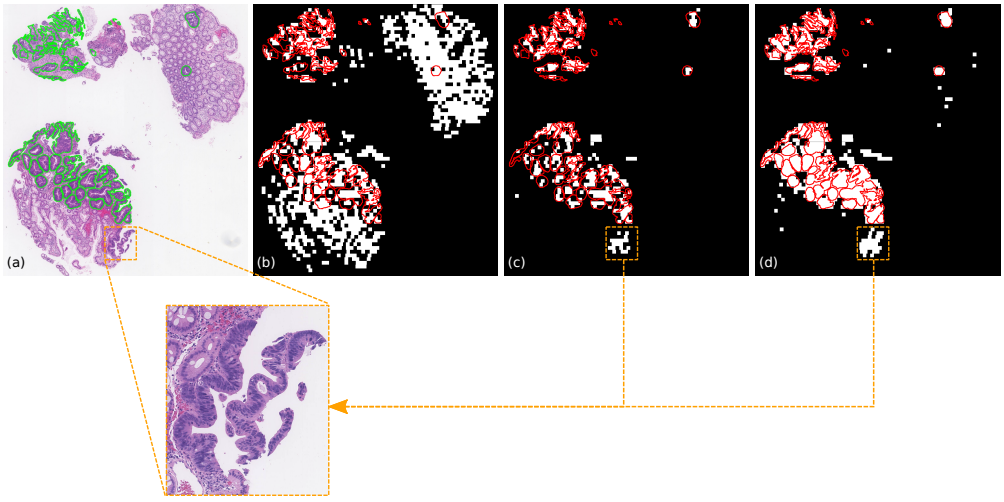


Fig. 3.7.: Examples of tumor masks obtained on a tumorous image from the DigestPath2019 dataset. **(a)** The slide region with the tumorous tissue delineated in green on the left image, and in red on the four binary masks. **(b)-(d)** The tile-level masks computed by the models' tile-level classifier with various amounts of supervision. **(b)** CLAM SB (0%). **(c)** MS-CLAM (6%). **(d)** MS-CLAM (62%). The orange square contains a tissue region, likely tumorous, that is not delineated in the ground truth annotations.

Model	% of annot. slides	Dice score (tum)	Specificity (norm)
CLAM SB	0	0.212 ± 0.005	0.740 ± 0.034
CLAM MB	0	0.223 ± 0.031	0.755 ± 0.029
MS-CLAM	0	0.331 ± 0.015	**1.000 ± 0.000**
MS-CLAM	6	0.425 ± 0.052	**1.000 ± 0.000**
MS-CLAM	12	0.473 ± 0.023	**1.000 ± 0.000**
MS-CLAM	25	0.503 ± 0.039	0.999 ± 0.001
MS-CLAM	62	**0.513 ± 0.029**	0.996 ± 0.002
MS-CLAM	100	0.475 ± 0.023	0.991 ± 0.003
CLAM SB (FT)	12	0.210 ± 0.024	0.691 ± 0.125
MS-CLAM (FT)	12	0.425 ± 0.043	**1.000 ± 0.000**
CLAM SB (FT)	62	0.287 ± 0.031	0.872 ± 0.041
MS-CLAM (FT)	62	**0.533 ± 0.033**	0.998 ± 0.001
CLAM SB (FT)	100	0.270 ± 0.032	0.840 ± 0.082
MS-CLAM (FT)	100	0.442 ± 0.048	0.984 ± 0.008

Tab. 3.6.: Localization metrics on the Camelyon16 test set. All metrics are averaged on a 5-fold cross validation split of the training set (± a standard error reported).

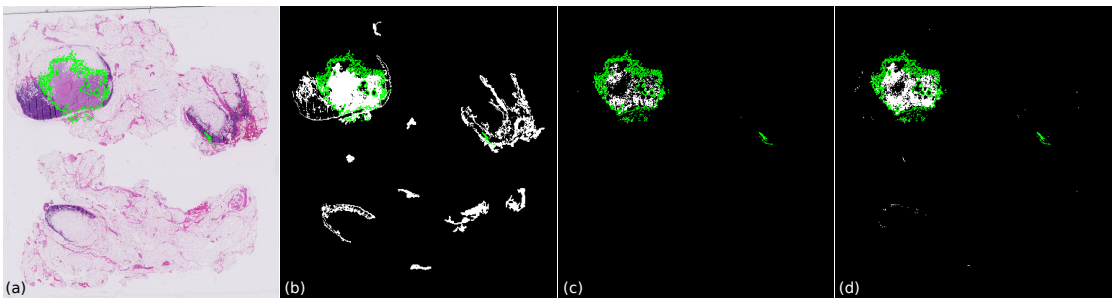


Fig. 3.8.: Slide #26 from the test set of Camelyon16, along with the tile-level tumor mask computed by each model using the tile-level classifier. **(a)** The slide thumbnail (metastasis delineated in green). **(b)-(d)** The tile-level masks computed by the models with various amounts of supervision. **(b)** CLAM SB (0%). **(c)** MS-CLAM (6%). **(d)** MS-CLAM (62%).

On Camelyon16, contrary to what was observed on DigestPath2019, all of the MS-CLAM models outperform CLAM regardless of the percentage of annotated slides used, in terms of both Dice score and specificity. Again, part of the explanation lies in the ratio between tumor and healthy tissue in tumorous slides: with many false positives, CLAM models tend to overestimate far more the tumorous regions than in the previous dataset. One noticeable result in Table 3.6 is the slight decrease of specificity and Dice score for the MS-CLAM models with 100% of annotated slides. This is because this model has a much higher recall at the cost of more false positives, and is therefore penalized by the relatively small tumorous regions in Camelyon16. However, it offers a much higher tile-level AUC and Average Precision (AP) than its counterpart with few tile-level labels (MS-CLAM with 100% of annotated slides reaches a mean AUC of 0.950 and a mean AP of 0.763, against an AUC of 0.736 and an AP of 0.429 for MS-CLAM with 6%). In the same way, although it seems like MS-CLAM with 62% of annotated slides performs better than the one with

100%, these two models have in fact very close performance. With 62%, the mean tile-level AUC is 0.948, but the mean recall is 0.551 (against 0.605 with 100%). Given the small number of tumorous tiles per slide in Camelyon16, false positives are more hurtful to the Dice score than false negatives. Figure 3.8 shows an example of a tumorous slide from Camelyon16, where the mask computed using the weakly-supervised model lacks specificity, which is higher for the MS-CLAM models, while recall increases with the annotation percentage. When dealing with minute tumorous regions, which is the case for the slides in Camelyon16, a high specificity is essential to accurately pick the tumor region: with many false positives, the inspection of the slide becomes tedious, whereas a high specificity guarantees that the user can rapidly check the regions raised suspicious by the model. The same observations made on DigestPath2019 regarding the effect of backbone fine-tuning on the localization performance can be made on Camelyon16, although this time MS-CLAM (FT) systematically obtains higher Dice score and specificity compared to CLAM (FT). When comparing MS-CLAM with and without fine-tuning, it is only in the case of 62% annotated samples that we obtain a performance gain, while all the other amounts of annotation are negatively affected by the fine-tuning procedure. Fine-tuning alone seems insufficient to increase the localization performance of the model, whereas our framework does bring significant improvements.

3.5 Ablation studies

In this section, we show the contributions of each main module of our model to the global performance, both in terms of WSI classification and localization. We also highlight the impact of the attention loss on the attention scores and the visualized attention maps. All ablation experiments are performed on the Camelyon16 dataset.

3.5.1 Attention loss

Model	% of annot. images	AUC (↑)	Acc. (↑)	F1-score (↑)	$\mathcal{L}_{att}$ (norm) (↓)	$\mathcal{L}_{att}$ (tum) (↓)
CLAM SB	0	0.883 ± 0.033	0.863 ± 0.027	0.797 ± 0.049	-0.664 ± 0.219	-0.441 ± 0.100
MS-CLAM (no att. loss)	12	0.910 ± 0.018	0.902 ± 0.019	0.865 ± 0.027	-0.537 ± 0.086	-0.780 ± 0.081
MS-CLAM	12	0.908 ± 0.013	0.899 ± 0.028	0.861 ± 0.031	-0.963 ± 0.042	-0.954 ± 0.182
MS-CLAM (no att. loss)	25	0.901 ± 0.031	0.895 ± 0.020	0.848 ± 0.036	-0.545 ± 0.087	-0.762 ± 0.145
MS-CLAM	25	0.911 ± 0.016	0.902 ± 0.028	0.867 ± 0.035	-0.966 ± 0.023	-1.170 ± 0.105
MS-CLAM (no att. loss)	62	0.914 ± 0.018	0.905 ± 0.013	0.866 ± 0.016	-0.497 ± 0.164	-0.777 ± 0.066
MS-CLAM	62	0.932 ± 0.008	0.938 ± 0.009	0.913 ± 0.013	-0.946 ± 0.017	-1.271 ± 0.107
MS-CLAM (no att. loss)	100	0.919 ± 0.006	0.907 ± 0.019	0.874 ± 0.023	-0.397 ± 0.147	-0.831 ± 0.065
MS-CLAM	100	0.939 ± 0.008	0.938 ± 0.012	0.916 ± 0.017	-0.915 ± 0.032	-1.318 ± 0.085

Tab. 3.7.: Impact of the attention loss on the slide-level classification performance. All metrics are averaged on a 5-fold cross validation split of the training set (± a standard error reported).

Slide-level classification. Table 3.7 shows the impact of the attention loss on the slide classification task. Our attention loss improves the classification results from 25%

annotated slides onward, although the performance is only slightly superior (< 0.5% difference) without the loss for 12% annotated slides. Regarding the impact on the attention scores: for normal slides, an attention loss $\mathcal{L}_{att}$ (norm) close to -1 means that the attention scores are nearly all equal. In this case, the model was able to give equal importance to each tile instead of focusing on just a few. For tumorous slides on the other hand, $\mathcal{L}_{att}$ (tum) has two main effects: first, the attention scores of tumorous tiles are higher than the ones of normal tiles; second, the attention scores of tumorous tiles have similar values, which means the tiles contribute equally to the final attention-based mean of the features. These observations regarding the attention scores are visible in the coarse attention maps represented in Figures 3.9 and 3.10, where each tile is colored according to its score. In the first one, the tumorous region is much better described by the attention scores when using the attention loss. In the second, although not perfectly uniform, the coarse attention map shows a more widespread attention distribution on the slide.

Localization. Table 3.8 shows the impact of the attention loss on the tumor localization task. Both the Dice score and the specificity are strongly affected: when using 62% of annotated slides, the presence of the attention loss increases the Dice score by 16%, while also preserving a very high specificity of 0.996. In general, the model's tendency to make more false positives (decrease in specificity) as the percentage of annotated slides increases is better contained thanks to the attention loss, as the specificity does not fall below 0.991 (against 0.927 without it).

Overall, the attention loss improves the slide-level classification performance from 25% annotated slides onward, and yields remarkable improvements for tumor localization for every annotation ratio.

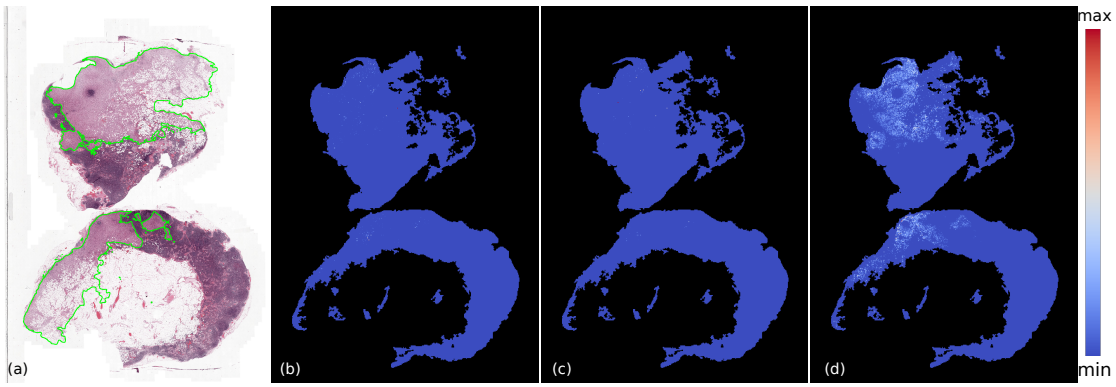


Fig. 3.9.: Slide #90 (tumorous) from the test set of Camelyon16, along with the tile-level coarse attention map computed by each model using the attention scores. The color scale on the right indicates the mapping between colors and attention scores. The former have been rescaled following $a'_k = (a_k - \min(\mathrm{a}))/(\max(\mathrm{a}) - \min(\mathrm{a}))$. **(a)** The slide thumbnail (metastasis delineated in green). **(b)-(d)** The coarse attention maps computed by the models with various amounts of supervision. **(b)** CLAM SB (0%). **(c)** MS-CLAM (12%, no attention loss). **(d)** MS-CLAM (12%)

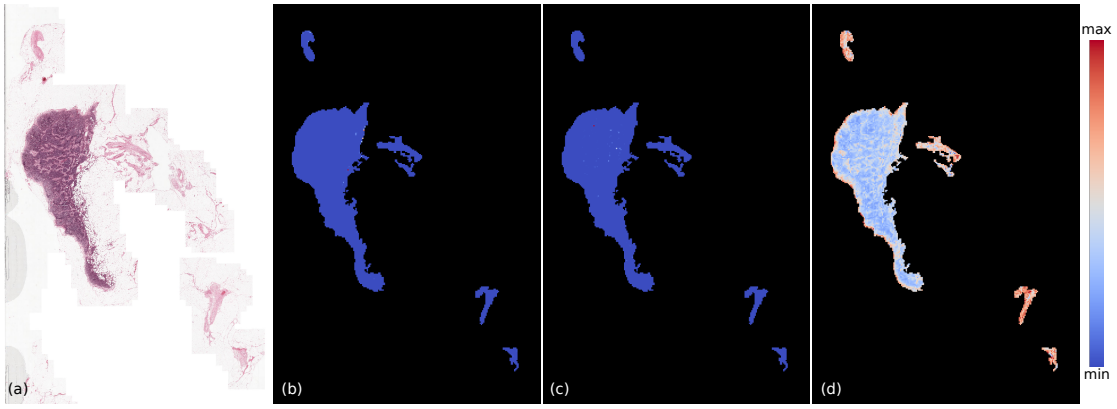


Fig. 3.10.: Slide #119 (normal) from the test set of Camelyon16, along with the tile-level coarse attention map computed by each model using the attention scores. The attention scores have been rescaled and matched to colors following the same procedure as in Figure 3.9. **(a)** The slide thumbnail. **(b)-(d)** The coarse attention maps computed by the models with various amounts of supervision. **(b)** CLAM SB (0%). **(c)** MS-CLAM (12%, no attention loss). **(d)** MS-CLAM (12%).

Model	% of annot. slides	Dice score (tum)	Specificity (norm)
CLAM SB	0	0.212 ± 0.005	0.740 ± 0.034
MS-CLAM (no att. loss)	12	0.437 ± 0.015	0.989 ± 0.003
MS-CLAM	12	0.473 ± 0.023	1.000 ± 0.000
MS-CLAM (no att. loss)	25	0.423 ± 0.034	0.981 ± 0.008
MS-CLAM	25	0.503 ± 0.039	0.999 ± 0.001
MS-CLAM (no att. loss)	62	0.351 ± 0.031	0.946 ± 0.035
MS-CLAM	62	0.513 ± 0.029	0.996 ± 0.002
MS-CLAM (no att. loss)	100	0.323 ± 0.019	0.927 ± 0.021
MS-CLAM	100	0.475 ± 0.023	0.991 ± 0.003

Tab. 3.8.: Impact of the attention loss on localization. All metrics are averaged on a 5-fold cross validation split of the training set ($\pm$ a standard error reported).

3.5.2 Exponential Weighted Sampling

In this section, we compare the exponential weighted sampling strategy with 2 other sampling strategies:

- RandSamp: The same slide sampling strategy as CLAM, i.e. sample randomly a single slide at each iteration.

- DoubleSamp: We use the Double Sampling strategy introduced in section 3.2.5 regardless of the amount of annotated slides used. In this setting, a normal slide and a tumorous slide are both sampled at each step without any specific weight for the annotated samples.

- EWSamp: The Exponential Weighted Sampling strategy: a normal slide and a tumorous slide are sampled simultaneously like in DoubleSamp, but annotated tumorous slides are more likely to be sampled than non-annotated ones, as detailed in Algorithm 1.

Slide-level classification. Table 3.9 shows the impact of the sampling strategy on the slide classification task. When the amount of annotated slides is $> 25\%$, the EWSamp strategy yields the best results in terms of F1-score and accuracy, while being comparable to the DoubleSamp strategy in terms of AUC. It is only when there are a few annotated slides (e.g., 12%) that the RandSamp strategy reaches higher accuracy and F1-score, albeit with a slightly lower AUC than the other two.

Localization. Table 3.10 shows the impact of the sampling strategy on the tumor localization task. This time, the RandSamp strategy is far behind the other two: the tile-level paired batch method we present in section 2.2 greatly improves the localization performance of the model. When the number of annotated slides is small (12%), the EWSamp strategy is advantageous compared to the DoubleSamp one, yielding a higher Dice score. When the amount of annotated slides increases, the exponential weighting of the annotated samples is likely to become less important, since the probability that an annotated slide is sampled at each step is higher.

Like the attention loss, the two sampling strategies we proposed bring much higher localization performance compared to the standard sampling method. When only a few annotated slides are available, the exponential weighted sampling is preferable.

Model	% of annot. images	AUC (↑)	Acc. (↑)	F1-score (↑)
CLAM SB	0	0.883 ± 0.033	0.863 ± 0.027	0.797 ± 0.049
MS-CLAM (RandSamp)	12	0.898 ± 0.013	0.916 ± 0.003	0.877 ± 0.007
MS-CLAM (DoubleSamp)	12	0.904 ± 0.006	0.902 ± 0.007	0.866 ± 0.009
MS-CLAM (EWSamp)	12	0.908 ± 0.013	0.899 ± 0.028	0.861 ± 0.031
MS-CLAM (RandSamp)	25	0.902 ± 0.020	0.916 ± 0.019	0.878 ± 0.028
MS-CLAM (DoubleSamp)	25	0.906 ± 0.013	0.899 ± 0.015	0.861 ± 0.019
MS-CLAM (EWSamp)	25	0.911 ± 0.016	0.902 ± 0.028	0.867 ± 0.035
MS-CLAM (RandSamp)	62	0.926 ± 0.010	0.930 ± 0.016	0.900 ± 0.025
MS-CLAM (DoubleSamp)	62	0.933 ± 0.007	0.929 ± 0.003	0.901 ± 0.005
MS-CLAM (EWSamp)	62	0.932 ± 0.008	0.938 ± 0.009	0.913 ± 0.013
MS-CLAM (RandSamp)	100	0.937 ± 0.008	0.933 ± 0.015	0.904 ± 0.025
MS-CLAM (DoubleSamp)	100	0.943 ± 0.004	0.936 ± 0.013	0.913 ± 0.019
MS-CLAM (EWSamp)	100	0.939 ± 0.008	0.938 ± 0.012	0.916 ± 0.017

Tab. 3.9.: Impact of the exponential weighted sampling on the slide-level classification performance. ($\pm$ a standard error reported).

Model	% of annot. slides	Dice score (tum)	Specificity (norm)
CLAM SB	0	0.212 ± 0.005	0.740 ± 0.034
MS-CLAM (RandSamp)	12	0.318 ± 0.027	1.000 ± 0.000
MS-CLAM (DoubleSamp)	12	0.456 ± 0.038	1.000 ± 0.000
MS-CLAM (EWSamp)	12	0.473 ± 0.023	1.000 ± 0.000
MS-CLAM (RandSamp)	25	0.337 ± 0.020	1.000 ± 0.000
MS-CLAM (DoubleSamp)	25	0.507 ± 0.033	1.000 ± 0.000
MS-CLAM (EWSamp)	25	0.503 ± 0.039	0.999 ± 0.001
MS-CLAM (RandSamp)	62	0.351 ± 0.025	1.000 ± 0.000
MS-CLAM (DoubleSamp)	62	0.510 ± 0.028	0.996 ± 0.001
MS-CLAM (EWSamp)	62	0.513 ± 0.029	0.996 ± 0.002
MS-CLAM (RandSamp)	100	0.329 ± 0.016	1.000 ± 0.000
MS-CLAM (DoubleSamp)	100	0.482 ± 0.022	0.993 ± 0.003
MS-CLAM (EWSamp)	100	0.475 ± 0.023	0.991 ± 0.003

Tab. 3.10.: Evaluation of the impact of the slide exponential weighted sampling strategy. All metrics are averaged on a 5-fold cross validation split of the training set ($\pm$ a standard error reported).

3.6 Discussion

With MS-CLAM, we showed the benefits of using a few slides with tile-level labels in addition to the slide-level ones on both DigestPath2019 and Camelyon16. On the latter, using only 62% of the tile-level labeled slides would have been enough to reach the 1st position on the challenge leaderboard based on the AUC results (5th position on the final leaderboard). On DigestPath2019 on the other hand, using 6% of the tile-level labeled slides would have reached the 5th rank in terms of AUC on the second task of the challenge [Da, 2022]. For both of the challenges, the best results were obtained with the help of deep neural networks trained from scratch on the challenge data, with additional post-processing steps, and sometimes using an ensemble of various heavy architectures (ensemble of networks) to reach the highest possible score. Furthermore, each challenge had its unique best methods, while here we presented a model that reaches near top performance without any post-processing steps, on both datasets. On the DigestPath2019 data, the AUC improvement with respect to the CLAM baseline is rather modest, but MS-CLAM clearly reduces the number of classification errors compared to CLAM (see accuracy and F1-score in Table 3.3). What is more, MS-CLAM trains in a matter of hours, and reaches state-of-the-art performance with an out-of-domain pre-trained feature extractor, proving the efficiency of such a model. With both higher classification scores, and lower attention losses, the MS-CLAM models provide better performance thanks to a higher key instance recall, that the attention loss promotes. Coupled with paired batch sampling, it allows MS-CLAM to outperform CLAM even without tile-level labels, using weak supervision only. Mixed supervision also allows to sample more tiles within annotated slides, and provides ground-truth labels instead of pseudo-labels for these

samples in particular. In turn, MS-CLAM models achieve higher localization performance than their weakly-supervised counterparts.

To profit even further from the annotations, we evaluated the impact of fine-tuning the feature extractor in CLAM and MS-CLAM. Although fine-tuning on its own was sufficient to reach higher results in terms of slide-level classification, its effect on localization was far from what we could achieve with MS-CLAM alone. Furthermore, fine-tuning is a long and costly process (16 hours per fold over 2 GPUs for Camelyon16 when using 62% of the annotated slides). Given this dataset contains only 270 samples, fine-tuning could be very expensive to scale to bigger datasets. MS-CLAM on the other hand needs nearly no additional time compared to CLAM, and is far superior in terms of localization, while even being competitive with fine-tuning in terms of slide-level classification.

There are still several limitations with this implementation of mixed supervision for attention-based MIL, the most critical one being that the tumor region must be exhaustively located by annotations within the annotated set. Missing tumorous regions could induce erroneous tile labels and hamper the tile-level classification. However, this need is limited to only a few slides as shown for both datasets (in Camelyon16, only 11 slides suffice for a performance improvement). Furthemore, we showed that the model was still robust to partial annotations, as some DigestPath2019 tumorous slides exhibit unannotated tumor tissue which was correctly classified (Figure 3.7). Moreover, although the ground-truth segmentation masks for the Camelyon16 challenge are particularly meticulous, coarser segmentation masks could suffice for our models as a tiling approach is used. This however, brings us to the second limitation of the model: the tile-level localization suffers from inaccuracy, due to square tiles only approximately fitting the tissue parts. It is therefore impossible to obtain more subtle tumor localization using tiles only, although they still offer a good first approximation of the tumor location. In the case of DigestPath2019, where tiles are often overlapping with benign tissue or background, predicted tumorous tiles tend to overestimate the tumor region. The labeling of tiles is also imperfect for the very same reasons. It would be interesting to supervise the attention of the model with finer tile-level labels, accounting for instance for the ratio of tumor within the tile, instead of its mere presence obtained after a hard threshold. Finally, the model presented here was only designed for binary classification, where tile- and slide-level labels coincide: it could be extended to multi-label classification, with different labels at the tile and the slide levels. A good example for this is the Gleason grading of prostate cancer, where tile-level Gleason patterns are insufficient to qualify the entire slide, without the knowledge of the area spanned by the patterns, which is typically accessible via the slide-level label. The tile-to-slide cooperation offered by this kind of model, along with mixed supervision could potentially be of great interest in this scenario.

3.7 Conclusion

In this chapter, we presented a new loss function, coined attention loss, that leverages partially available tile-level labels to constrain the attention distribution in CLAM, an attention-based, weakly-supervised MIL model. Using mixed supervision to exploit both slide- and tile-level labels, we were able to improve the performances of the model for the classification of both entities. With greater coherence between classification and localization, these newly trained models offer better interpretability and fewer false positives among the suspicious regions, furthering their usability in a clinical setting. The framework was built atop an already cost-efficient architecture [Lu, 2021], that required few slide-level labels, and limited computational resource, and extended in a similar fashion this effectiveness to the mixed supervision setting, narrowing the amount of required labels to improve upon the baseline. Although for the moment limited to binary classification, with local and global label coherence, we aim to extend the application of mixed supervision to multi-class classification, with fewer constraints on the relations between the labels at different scales.

Acknowledgments

The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support. This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

The authors would like to thank Hind Dadoun for her suggestions on the paired batch method.

4

Lung-IO: A new dataset for the analysis of immunotherapy outcome in lung cancer patients

Contents

Abstract
The purpose of this chapter is two-fold: first, we introduce Lung-IO, a new digital histopathology dataset of lung cancer patients treated with immune checkpoint blockade, collected through the collaboration of five different French medical centers. We show the cohort statistics, as a whole or separate by center, and discuss the available clinical information, the observed discrepancies between centers, and the meaning of what is a positive or negative response to the treatment, based on the RECIST criteria. Finally, we apply several WSI classification models to our dataset, including MS-CLAM (introduced in previous chapters), and show that these models generally fail at accurately predicting the treatment response, and suffer from generalization issues once a dedicated center-wise split of the data is done. This allows us to motivate the different approach we present in the next chapter.

4.1 Introduction

There are currently several digitized histopathological datasets available online, either coming from past computational challenges, or public health services. The purpose and the organ of interest of these datasets are diverse, ranging from metastasis detection in lymph node [Bejnordi, 2017], to Gleason grading of prostate cancer [Bulten, 2022], but also cell segmentation in breast slides [Amgad, 2022]. Usually, these datasets are built with a specific and histologically identifiable purpose in mind. Ground truth annotations are often carefully collected with consensus-based approaches, and among multiple participating centers. However, since these datasets are task-dependent, the only available information is the ground-truth needed to evaluate one's ability to perform the expected objective. Thus, it is often impossible to leverage the available data for any application other than the one for which it was collected. On the other hand, there exist datasets which serve a more general purpose, such as The Cancer Genome Atlas (TCGA[1]). Through the TCGA portal, it is possible to access the information of thousands of pan-cancer cases, which includes digital slides, but also clinical information (such as age, gender, etc...) and sequencing transcripts. Therefore, it becomes possible to perform multi-omics and multi-purpose analysis on various types of cancer and different organs. Although very detailed and complete, the lung cases registered in the TCGA are nearly all treated with chemotherapy, and the treatment response is rarely documented, even though patient follow-up is generally accessible. This is expected given that the TCGA was designed much before ICIs were approved by the FDA as first or second line treatment. Therefore, a dataset that contains lung slides and corresponding clinical information must be built to conduct immunotherapy outcome prediction.

[1] https://www.cancer.gov/ccg/research/genome-sequencing/tcga

4.2 Dataset and definition of the task

4.2.1 Case selection process

The purpose of the dataset we introduce is to gather a retrospective cohort of ICI-treated lung cancer patients, with both histology slides and clinical information available. The primary objective of this dataset is the treatment response prediction, evaluated thanks to radiological examination of the patients. Therefore, the ones who were enrolled had lung cancer, received immune checkpoint blockade, and had follow-up examinations to assess the treatment effect. Other than these requirements, no selection was performed based on age, histological type, tumor stage or sample origin. The slides which were used for this study are all diagnostic slides, used routinely in clinical practice for cancer diagnosis; they are not purposely crafted samples. This is both an asset and a weakness of this dataset: the variability observed among the slides is representative of the samples found in clinical practice, but it can also hold back the development of efficient proof-of-concept methods. Five different University Hospitals were involved in this project: Caen, Dijon, Rouen, Toulouse, and Nice. Concerning the tissue samples, every center provided:

- A Formalin-Fixed, Parafin Embedded (FFPE) block of pathological tissue
- Two unstained tissue slides
- A spreadsheet describing the clinical information of the patients.

All of the tissue slides were stained with Hematoxylin, Eosin and Saffron (HES) and scanned at the Laboratory of Clinical and Experimental Pathology, Nice University Hospital, FHU OncoAge using a NanoZoomer scanner (Hamamatsu Photonics, Hamamatsu, Japan). Because of missing information or poor quality, some samples were left out of the final dataset. The flow chart in Figure 4.1 describes the subsequent steps that lead to the final dataset. “Good quality” in the flow chart is associated to four exclusion criteria:

1. the tissue remains blurry, in spite of several scanning attempts
2. the tissue is missing from the slide, or only a few cells are visible
3. the slide is a cytology sample
4. there is no tumor in the tissue (healthy sample)

Beside these four rules, every slide was considered a valid sample, even though some of them included very little or sparse tissue. Moreover, we did not filter based on the origin of the tissue, nor on the way it was acquired (i.e., resection or biopsy).

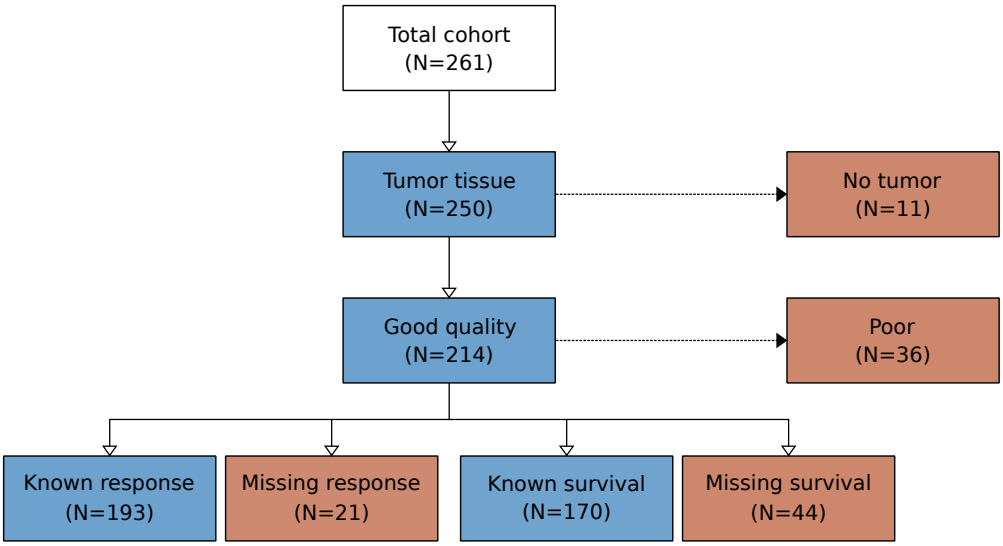


Fig. 4.1.: A flow chart representing the selection process we used to build our dataset.

Concerning the clinical information, the only exclusion criterion was the absence of treatment response evaluation. For several cases from the external contributing centers (all but Nice), the treatment response was given as a yes/no answer instead of the RECIST evaluation. Since we could not retrieve the RECIST label corresponding to these cases, and had established our own rule for converting RECIST labels into binary treatment response (see Section 4.2.2), we decided to ignore these patients to avoid potential confusions. For the cases from Nice however, the classical CT-scan RECIST evaluation was followed by a PET-scan one for most of the patients, so as to confirm the measurements obtained with the CT-scan only. This is nonetheless in compliance with the RECIST baseline instructions [Eisenhauer, 2009], and only constitutes a more thorough examination of the treatment effect. Of note, histology was not considered a selection criterion given that nearly all of the cases were either adenocarcinomas or squamous cell carcinomas, with only a few rare subtypes; therefore, we included each one of them. Table 4.1 summarizes the clinical information of the cohort.

In light of this table, there are several striking observations to make. Age, sex, histology and smoking history (especially the number of nonsmokers) all show similar trends across the various centers. However, it is not the case for the other variables. First, Nice is the only center with cancer stages below III, i.e., resectable tumors for most of them. Consequently, the slides collected at the Nice University Hospital are much richer in terms of tissue quantity, since most of them were from resection specimens. Then, the PD-L1 expression levels differ significantly between centers: Rouen and Caen only include cases with a TPS >50% (except for three of them). Accordingly, the treatment response in both these cohorts is often positive, with a 50% ratio of responders in Caen,

Characteristic		All (N=193)	Nice (N=60)	Caen (N=15)	Dijon (N=42)	Rouen (N=26)	Toulouse (N=50)
Age, years, median (range)		63 (30-90)	63 (38-83)	69 (54-83)	64 (47-79)	63 (45-90)	62 (30-79)
Sex, no.	Female	60	15	4	16	8	17
	Male	133	45	11	26	18	33
Stage, no.	<3	15	15	0	0	0	0
	3	35	21	0	0	2	12
	4	141	23	14	42	24	38
Histology, no.	ADK	141	41	9	35	19	37
	SCC	42	12	6	5	7	12
	Other	10	7	0	2	0	1
Smoking, no.	yes	86	30	14	18	11	13
	no	12	4	1	0	2	5
	former	88	26	0	17	13	32
	unknown	7	0	0	7	0	0
TPS, no.	<1%	46	12	0	6	0	28
	1 – 49%	35	11	1	11	2	10
	>50%	75	11	14	20	23	7
	unknown	37	26	0	5	1	5
Response, no.	no	108	41	8	26	7	26
	yes	85	19	7	16	19	24

Tab. 4.1.: The clinical information of the entire cohort and for each center. ADK stands for adenocarcinoma, while SCC stands for squamous cell carcinoma."Other" means any other histology, i.e., Large Cell Neuro-endocrine Carcinoma (LCNC), sarcomatoid carcinoma, adenosquamous, and undifferentiated. TPS expression is reported following intervals based on the thresholds commonly found in the literature.

and an unusual ratio of 73% of responders in Rouen. Surprisingly, the patients from Toulouse are also balanced between responders and nonresponders (52% vs. 48%), even though the TPS values are more in line with the ones observed in Nice. Despite these discrepancies between the cohorts, we used all of the cases to conduct our analyses.

The differences observed in response rates and PD-L1 expression are also reflected in the survival period of the patients between centers. As Figure 4.2 shows, there are significant differences in both overall and progression-free survival (OS, PFS) between the centers. This is confirmed by a log-rank statistical test, which indicates that the OS in Rouen is statistically different from all of the other centers ($p<0.05$, except from Nice). The significance threshold is also crossed between Caen and Nice. For PFS, the differences are not as numerous as for OS, but there is yet again a statistically significant difference between Rouen and Caen.

4.2.2 Treatment response definition

To simplify the objective of the study, we decided to convert the four different RECIST labels into two distinct outcomes: responder or nonresponder. For three out of the four original labels, the conversion suffers little to no doubt: either complete or partial responses (CR, PR) correspond unequivocally to a positive response, while progressive disease (PD) indicates more than likely an absence of response. On the other hand, the definition of the stable disease (SD) label is by essence unclear, since it translates to "neither sufficient shrinkage to qualify for PR, nor sufficient increase to qualify for PD" [Eisenhauer, 2009]. In some previous studies on treatment response, patients with a SD

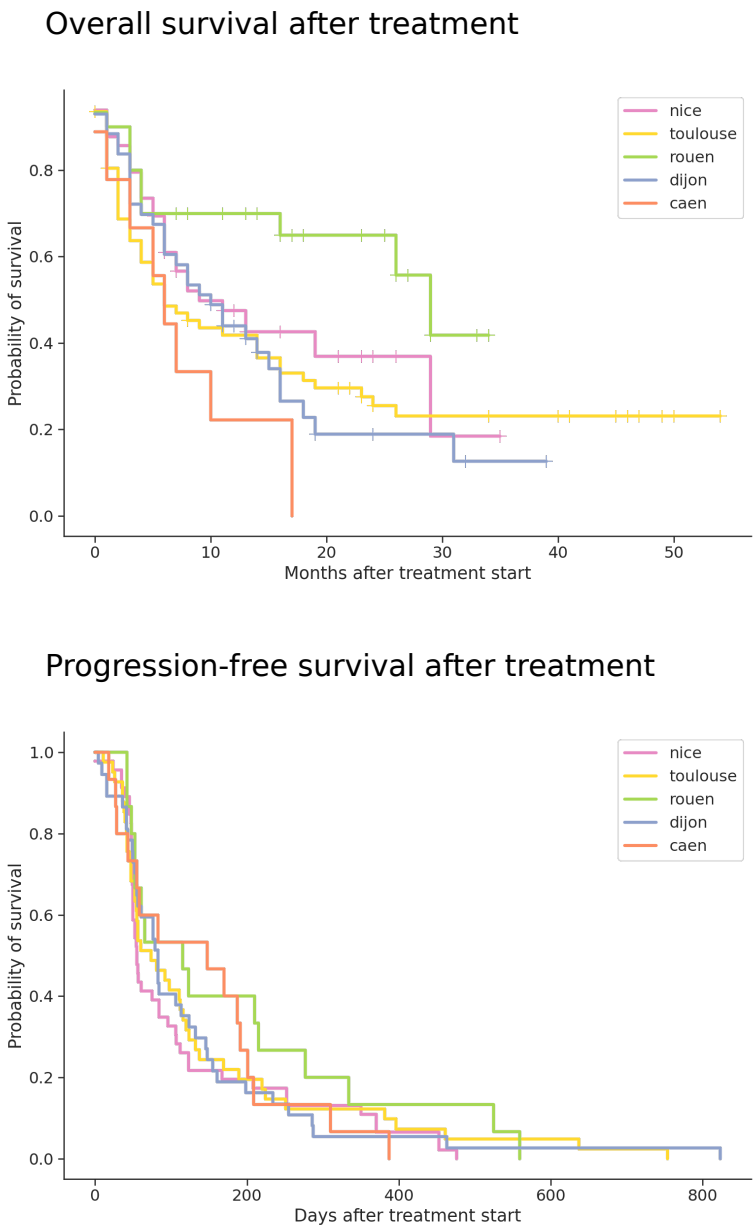


Fig. 4.2.: The Kaplan-Meier estimates of the patient overall and progression-free survival probability for each center.

RECIST were not included to keep only extreme-most labels [Johannet, 2021]. Given the uncertainty around the actual effect of the treatment in this case, the patients from the Nice hospital were submitted to a PET-scan examination to either confirm or refute the positive response to the treatment. In some cases, SD was associated to a negative treatment response. For the other participating centers, the only available information was the CT-scan evaluation. Therefore, SD was considered a positive response. Figure 4.3 illustrates with boxplots the discrepancies between the RECIST-labelled patients in terms of survival. Considering the very small p-value magnitude between SD and PD survivals (both for OS and PFS), it is reasonable to consider them to be associated to opposite treatment responses.

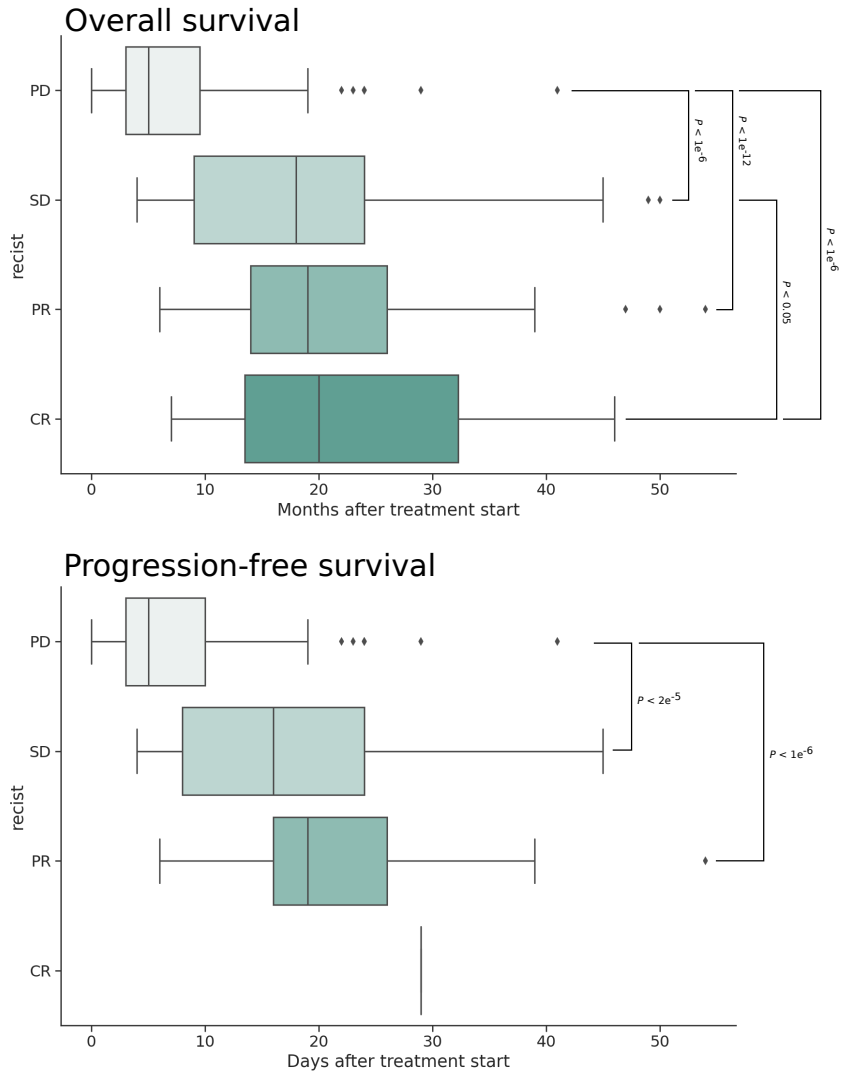


Fig. 4.3.: Boxplots showing overall and progression-free survival among patients for each RECIST label (all centers).

4.3 Treatment response prediction

4.3.1 Problem definition

Given the nature of the task that we consider, i.e., predict the binary outcome of the ICI on lung cancer patients using histological slides, we resort to the multiple instance learning or MIL framework, already mentioned in the introduction of this thesis. In the case of binary classification, the problem can be formulated the following way: let us consider a dataset $\mathcal{X} = \{x_1, ..., x_N\}$ of N slides and associated treatment response $\mathcal{Y} = \{y_1, ..., y_N\}$ where $\forall i \in \{1...N\}, y_i \in \{0, 1\}$. Each slide x_i of $\mathcal{X}$ is seen as a bag containing K_i instances such that $x_i = \{x_{i,1}, ..., x_{i,K_i}\}$, where each instance corresponds

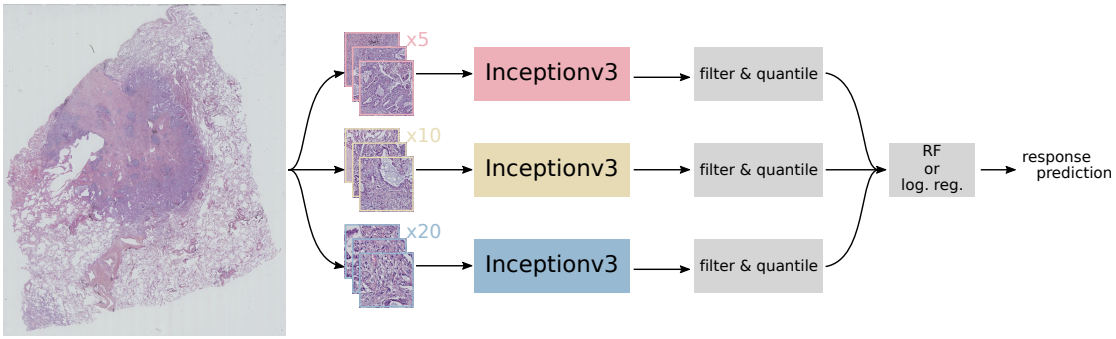


Fig. 4.4.: An illustration of the model developed by [Jain, 2020] and that we slightly adapt to perform treatment response classification.

to a sub-region of the slide, typically a 256×256 pixels square. In the case of treatment response prediction, the label of each instance $x_{i,j}$ is unknown, since only the patient-level label is accessible. Therefore, this problem is said to be weakly-supervised. The objective is, as for every supervised learning task, to find f such that $f(\mathcal{X}) = \mathcal{Y}$. In our case, however, this problem cannot be tackled directly, given the size of the slides in the dataset. Thus, we write $f = g \circ h$ where h is a "feature extractor", the purpose of which is to extract a low-dimensional representation of the instances within each slide, or instance scores directly, and g is an *aggregator*, which is there to recover the label y_i of sample x_i given the instance embeddings (or scores) obtained through h. It is also customary to call g a *pooling* operator, since in most cases g pools the instances and applies some kind of selection to infer the bag (or slide) label. There are two most-basic pooling operations which usually serve as a baseline to the others, namely mean-pooling and max-pooling. For both of these, we assume that h computes scores from instances, i.e. $\forall (i,j) \in \{1...N\} \times \{1...K_i\}, h(x_{i,j}) = r_{i,j} \in \mathbb{R}$. Then, the bag-level prediction $\tilde{y}_i$ is obtained through $\tilde{y}_i = g(\{r_{i,j} \mid j \in \{1...K_i\}\})$, where:

$$g(\{r_{i,j}\}_{j\in\{1...K\}}) = \begin{cases} \max_j(r_{i,j}) & \text{for max pooling} \\ \frac{1}{K}\sum_{j=1}^{K} r_{i,k} & \text{for mean-pooling} \end{cases} \tag{4.1}$$

In what follows, h will typically be a deep convolutional neural network, such as a residual network (ResNet, [He, 2016]), and g can either be a fixed or a trainable pooling operator (again, a neural network for instance). Both end-to-end or multiple-step training schemes exist, but in most cases the latter is considered, since training both a model that computes both g and h can be resource-intensive, and cause problem in case of nondifferentiable pooling operations.

4.3.2 Models

There exist numerous models designed for weakly-supervised, binary MIL, and in particular in the context of histopathology, since classification of tumors is one such task, that is

very frequently addressed in the literature [Courtiol, 2018; Campanella, 2019; Dehaene, 2020; Lu, 2019]. Other histology-related tasks are simplified as binary classification problems [Sha, 2019; Jain, 2020]. Among all of them, we tested three in particular, given they showed excellent performance on closely-related tasks.

TileNet The first model comes from [Coudray, 2018], where it was originally used to classify and predict mutations in NSCLC slides taken from the TCGA. In this work, a straightforward slide-to-tile correspondence is applied to generate tile labels for each case. An Inceptionv3 CNN [Szegedy, 2016], pretrained on the Imagenet dataset [Deng, 2009] is used to classify the tiles. To recover the slide label, either mean-pooling or max-pooling is used (Equation 4.1). We refer to this model as *TileNet*, and specify which pooling method was used to obtain the results.

MultiMag The second model comes from [Jain, 2020]. This time, the original task is the prediction of the TMB status of NSCLC patients (again from the TCGA), either high or low given a predefined threshold. Knowing the close relationship between TMB and treatment response, the transition to the second task is well motivated. The method is similar to the previous one, in that we use again an Inceptionv3 CNN as a tile classifier where the tiles are given the same label as the slide. One of the major differences is that one tile classifier is used for three distinct magnification levels ($\times 5$, $\times 10$ and $\times 20$) instead of a single one. The other principal difference lies in the pooling method that is used: after training each tile classifier, the scores obtained at each magnification level are first filtered to eliminate the probabilities inside the interval $[0.3, 0.7]$. This is done to ensure that low-confidence tiles are removed in the first place. Then, the median probability is computed for each magnification, and a Random Forest classifier [Breiman, 2001] is used to classify the slides using the set of medians. If there are multiple slides per patient, then all the tile scores from all the slides are considered before computing medians and predictions. We tried to make some modifications to the initial model, especially on the aggregation part. Instead of taking the median of the probabilities, which is a rather limited representation, we computed the deciles of the probabilities, and used these for each magnification level. We also evaluated logistic regression instead of the random forest to study the effect of the final classifier on the performance. This model and the small changes we propose can be seen in Figure 4.4. We refer to it as *MultiMag*, and specify if necessary the chosen hyperparameters (e.g., final classification algorithm).

ABMIL and CLAM Finally, we also use the attention-based deep MIL (ABMIL) model [Ilse, 2018] and CLAM [Lu, 2021], which have already been introduced and discussed in chapters 2 and 3. As a reminder, for these models, h is an Imagenet-pretrained ResNet50 CNN that is only used to extract low-dimensional tile representations. g however is a trainable two-layer neural network, that computes attention scores a_k for each tile $x_{i,k}$

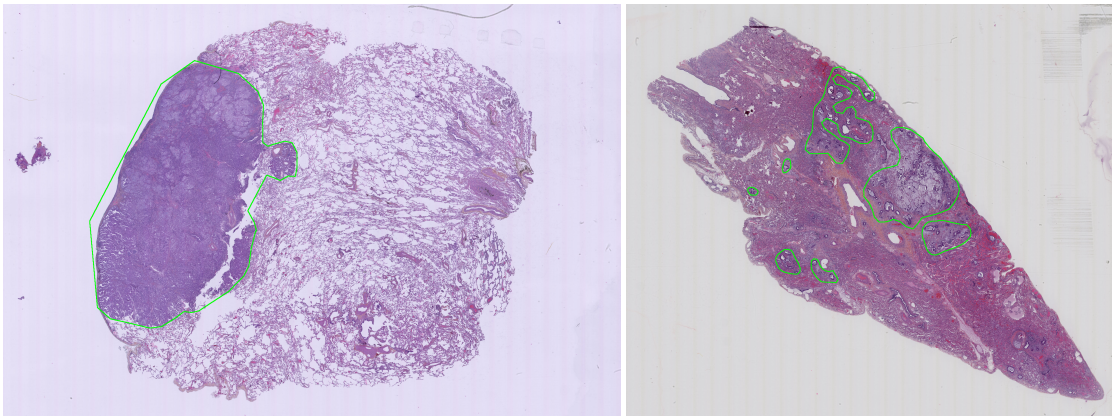


Fig. 4.5.: Two examples of tumor region annotations (green). The slide on the left exhibits a large and sparse parenchymal region on the right, which has probably no impact on the response classification.

of slide x_i, which are used to compute a weighted sum of the embeddings that better represents the slide in the latent space.

4.3.3 Tumor region annotations for MS-CLAM

To be able to leverage the potential of MS-CLAM, annotations are required to help the model determine which tiles should be given high attention scores, and apply the loss functions defined in chapter 3. Yet, contrary to the task of tumor classification, treatment response does not correspond to identifiable regions of interest in WSIs. Nonetheless, one could assume that among the multitude of tissue types that are present in the slide, some of them are more than likely unrelated to the outcome of ICI, such as, for instance, parenchyma and other benign or usual histological structures of the lung. Based on these assumptions, the tumor regions within the slides were annotated, so that tile labels could be used for MS-CLAM. The tumor regions were targeted as being most likely the primary source of information concerning the treatment response, compared to other surrounding regions. Figure 4.5 shows two examples of slides and their annotated tumor region.

4.3.4 Experiments

To test the presented models, a systematic 5-fold cross-validation (CV) was conducted using all the available samples. The CV was performed regardless of the centers: the patients from each center were split randomly in both training and test sets.

Model	AUC	Accuracy	Precision	Recall
TileNet (mean pooling)	0.571 (0.114)	0.477 (0.121)	0.53 (0.271)	0.455 (0.383)
TileNet (max pooling)	0.576 (0.144)	0.533 (0.122)	0.429 (0.315)	0.431 (0.310)
MultiMag (RF)	0.577 (0.099)	0.585 (0.084)	0.537 (0.191)	0.458 (0.204)
MultiMag (logistic regression)	0.550 (0.105)	0.585 (0.069)	0.451 (0.320)	0.337 (0.219)
ABMIL	0.574 (0.098)	0.584 (0.079)	0.504 (0.110)	0.485 (0.288)
CLAM	0.572 (0.098)	0.554 (0.067)	0.509 (0.169)	0.470 (0.048)
MS-CLAM	0.583 (0.155)	0.559 (0.137)	0.560 (0.171)	0.478 (0.116)

Tab. 4.2.: The 5-fold CV results obtained by the various model for treatment response prediction.

4.3.5 Results

Table 4.2 shows the results obtained by the various models on the treatment response prediction task. Overall, the performance of the models is rather poor, even though there are significant design gaps between them. For the attention-based model, MS-CLAM was unfortunately unable to improve upon the results of either ABMIL or CLAM, and even falls behind the models based on tile classification.

4.3.6 Discussion and Conclusion

The results yielded by various models on the treatment response prediction do not reach satisfactory levels although each of them showcased high performance on other histological tasks, such as tumor classification or histological subtyping [Lu, 2021; Coudray, 2018], and TMB level prediction [Jain, 2020]. There may be several reasons why all of these models, despite the differences in their architecture, fail to deliver convincing predictions.

First, as already stated in section 4.2.2, the conversion between the RECIST and the binary labels is not perfectly clear. Although by looking at the survival times it seems like the PD RECIST is different from SD, this difference might not be well measurable from the WSIs only. The uncertainty surrounding the labels could also be explained by the short period that separates the start of the treatment and the CT evaluation (3 months), which is insufficient to fully characterize unequivocal progression or regression of the tumor in some cases. The survival data, on the other hand, is a less noisy source of labels, since it does not rely on specific time points, but is rather a continuous follow-up of the patients.

Second, the models we have seen so far necessarily focus on a specific type of instance or tile within the slide (for all of them, the dependencies between the tiles are not taken into account), which may not be relevant to spot differences accurately between responders and nonresponders, given the nature of the data (diagnostic slides). Since we cannot claim that a specific phenotype is sufficient to indicate a link with a positive or a negative

response, there may be further explanations to be found when looking at neighboring regions within the slides. Single tiles are limited fields of view, and therefore do not necessarily represent a viable source of information alone.

5 WhARIO: Whole-slide image-based survival Analysis for patients tReated with ImmunOtherapy

Contents

Abstract
Immune checkpoint inhibitors (ICIs) are now one of the standards of care for patients with lung cancer, and have greatly improved both progression-free and overall survival. However, less than 20% of the patients treated with ICIs actually respond to the treatment, when some suffer from acute adverse events. Although a few biomarkers have integrated the clinical workflow to help in patient selection, they often require additional modalities on top of diagnostic Whole-slide Images (WSIs), and can lack efficiency or robustness. In this work, we propose a new biomarker derived solely from the analysis of histology slides. We develop a 3-step framework, combining contrastive learning and nonparametric clustering to distinguish tissue patterns within the slides, before exploiting the adjacencies of previously defined regions to train a proportional hazards model for survival analysis. Based on a cohort of 193 patients from 5 different centers, we show that our newly designed set of features is an efficient predictor of survival for lung cancer patients who received ICI treatment. We achieve similar performance to the current gold standard biomarker, without the need to access other imaging modalities, and show that both can be used together to reach even better results. Finally, we provide a histological interpretation of the most significant clusters, and highlight the correlation between them and the literature with respect to the signs of ICI effectiveness. This chapter was submitted to a journal [Tourniaire, 2023b]

5.1 Introduction

Immune checkpoint inhibitors have been one of the major recent breakthroughs in cancer therapy. In particular, several studies showed that lung cancer, the deadliest kind of cancer globally [Sung, 2021], faced significant improvements in terms of survival, with the introduction of Programmed cell death protein 1 (PD-1) and Programmed Death-Ligand 1 (PD-L1) inhibitors [Horn, 2017; Reck, 2019]. Other types of ICIs that target CTLA-4 protein receptors have also been shown to be efficient when combined with anti-PD-1 or anti-PD-L1 treatments [Hellmann, 2018]. However, one common problem with this treatment is the usual low response rate, which is slightly below 20% for non-small cell lung cancer (NSCLC), its most common form [Mazieres, 2019; Berghmans, 2020]. Another main issue is, as with every other treatment, the occurrence of adverse effects such as rash, diarrhea, or even severe allergic and inflammatory reactions which can potentially be fatal [Martins, 2019; Wang, 2018]. To better select patients eligible to this kind of therapy, several biomarkers have been devised. The current gold standard is the measure of PD-L1 expression in tumor cells through immunohistochemistry (IHC), or Tumor Proportion Score (TPS), for which two different thresholds (1% and 50% respectively) have been identified as relevant criteria to select patients with higher

response rates (27% and 39% respectively) [Reck, 2019; Mok, 2019]. Yet, the efficacy of such a biomarker remains limited, with additional concerns regarding the robustness of its assessment and the variability between observers [Grigg, 2016; Ilie, 2017; Cooper, 2017]. Another recent biomarker is the tumor mutational burden or TMB, which corresponds to the number of somatic mutations per megabase in the DNA of cancer cells. Patients with high TMB (i.e. $\geq$10 mutations per megabase) were shown to have higher progression-free and overall survival, as well as higher response rates (up to 45%) than others [Hellmann, 2018; Marabelle, 2020; Klein, 2021]. TMB is not yet routinely used in a clinical setting because it primarily requires Whole Exome Sequencing, a method that is currently not available in many hospitals due to its high cost and complexity.

To compensate for the current lack of available biomarkers, several works have proposed to use deep learning for the analysis of Hematoxylin and Eosin (H&E) stained whole-slide images to either recover existing biomarkers, or to develop new ones. [Sha, 2019] proposed a multi-field-of-view analysis of lung H&E WSIs to predict the PD-L1 status (i.e., TPS $>$ 1%). In this work, an IHC analysis of the slides is first conducted to label regions based on PD-L1 positivity (above threshold). Then, a deep residual network (ResNet-18) – modified to process different fields of view in small patches – is used to classify the patches between PD-L1$^+$ and PD-L1$^-$. During inference, the ratio of PD-L1$^+$ patches is computed for each slide to derive the PD-L1 status of each patient. [Jain, 2020] use three Inceptionv3 networks [Szegedy, 2016] at three different magnification levels ($\times$5, $\times$10, $\times$20) to classify the TMB status of lung H&E slide patches. During inference, low confidence patches are discarded, and a random forest classifier predicts the TMB status from the median probabilities of each magnification level. These two works address proxies to treatment outcome prediction through intermediate biomarkers, that could be obtained using cheaper modalities (i.e., H&E), but do not go beyond their limits, and in particular their limited prognostic power.

On the topic of straightforward treatment response prediction, a few methods have been proposed to classify melanoma patients between responders and nonresponders. [Harder, 2019] use both IHC and H&E images to extract features which are then used to train small classification models such as random forests, support vector machine or logistic regression. The feature extraction leverages a deep learning-based detection of lymphocytes thanks to multimodal registration and pathologist annotations of cells and tissue types. [Johannet, 2021] use 2 different deep neural networks to segment the tumor regions in melanoma slides and classify patches. Here, the patch labels are the same as the slides'. During inference, the average of the probabilities of the patches is used to get the slide-level score, which is either used directly, or through a logistic regression with clinical variables to output the response. Although these works propose to overstep previous markers and predictions, they nonetheless require careful expert annotations of the tissue, if not additional modalities (such as IHC) to select specific regions in the tissue, and guide their analysis.

Lately, automated approaches for the assessment of Tumor Infiltrating Lymphocytes (TILs) have been proposed to help in the prediction of survival of ICI-treated NSCLC patients. [Park, 2022] first train a deep neural network to segment tumor and stroma and detect TILs in lung WSIs based on a consensus of pathologists' annotations, before defining three different phenotypes based on the ratios of TIL-invaded stroma and tumor regions in slides. The authors show that one phenotype in particular, which they refer to as the inflamed immune phenotype, shows survival trends which are significantly better than the other two phenotypes. This phenotype also correlates positively with high PD-L1 expression and TMB. [Wang, 2022] use a very similar approach at the start, using a U-net-like network to segment cells, and another one to segment tumor and stroma using a small set of annotated regions. However, instead of defining phenotypes, the authors manually build a feature set of over 700 features based on TILs and tumor cells interactions, as well as geometric characteristics of TILs. The feature set is pruned during the training of the survival model by the means of elastic-net regularization. The authors show that a Cox Proportional Hazards (PH) [Cox, 1972] model trained on the final feature set is able to correctly rank and stratify patients in low- and high-risk groups on three other lung cancer cohorts, as well as a gynecological cancer one. These two approaches use deep learning to detect lymphocytes and tumor or stroma within WSIs, before features are manually constructed to feed a survival prediction model, such as the Cox PH model. Therefore, the quality of the TIL assessment can be controlled by pathologists before features are extracted and used as predictors for survival. This type of method offers a clearer interpretation of the results, as the deep learning model does not intervene directly in the decision process. However, it requires the introduction of domain-specific, prior information that is considered to define the features that will be used for the survival regression ; here, the focus on TILs. Although the choice of such prior information is legitimated by literature [Tumeh, 2014], there are potentially other unknown factors which could be associated to favorable prognosis, and that are deliberately ignored in this kind of method.

With these drawbacks in mind, we develop an approach, called WhARIO (Whole-slide image-based survival Analysis for patients tReated with ImmunOtherapy), that does not rely on any histological prior at all, but harvests unsupervised mechanisms to extract features that are then used for survival analysis. In particular, we introduce the following contributions:

- We develop a three-step pipeline, that allows to cluster low-dimensional representations of the tissue in WSIs, and use the cluster interactions to build a feature matrix for each patient. The feature extraction is based on contrastive learning, while the clustering approach is nonparametric, making the entire feature extraction process unsupervised.

- We propose a feature selection method to select the most relevant ones for survival in the aforementioned matrices, using the concordance index and the log-rank test in a cross-validation of a Cox PH model.

- Using an in-house, multicentric dataset of 149 patients, we show that the features we crafted from the unsupervised tissue analysis in WSIs are prognostic of survival for lung cancer patients treated with ICI, and are on par with the current gold standard PD-L1 biomarker, which requires additional IHC analysis.

- Finally, we discuss the histological interpretation of the clusters that are most correlated to longer survival, thus establishing further interpretability of our pipeline.

5.2 Methods

In this section, we describe the various steps needed to leverage Hematoxylin, Eosin and Saffron (HES) slides for survival analysis. Figure 5.1 shows an overview of our method. Our framework involves three steps: first, contrastive learning is used to extract low-dimensional features from patches taken in WSIs. Once this is done, these low-dimensional projections of patches are used to perform deep nonparametric clustering. Finally, after training the clustering model, the obtained clusters are projected back to the slides, and adjacency between clusters within the slides are used to build patient-wise feature matrices, which serve as inputs to a survival regression model. Each step is detailed in the following sections.

5.2.1 Contrastive learning

For the clustering to work, we need to have the input data lie in a low-dimensional space (i.e., $d \leq 10$), to avoid the curse of dimensionality, which prevents the Euclidean distance between samples from being discriminative. To this end, a low-dimensional latent representation of each tile in every WSI should be derived before clustering can happen. This is why we chose to perform the unsupervised training of a deep neural network to create low-dimensional representations of the tiles that we can then use for the DeepDPM clustering method. To achieve this, we use the SimCLR contrastive learning method [Chen, 2020a], which has already been proven efficient for histopathology [Ciga, 2022]. The purpose of this method is to learn a mapping from a high- to a low-dimensional space that is invariant to a set of geometric transformations and color distortions. This is achieved by maximizing the similarity between two different projections $\hat{z}$ and $\tilde{z}$ of

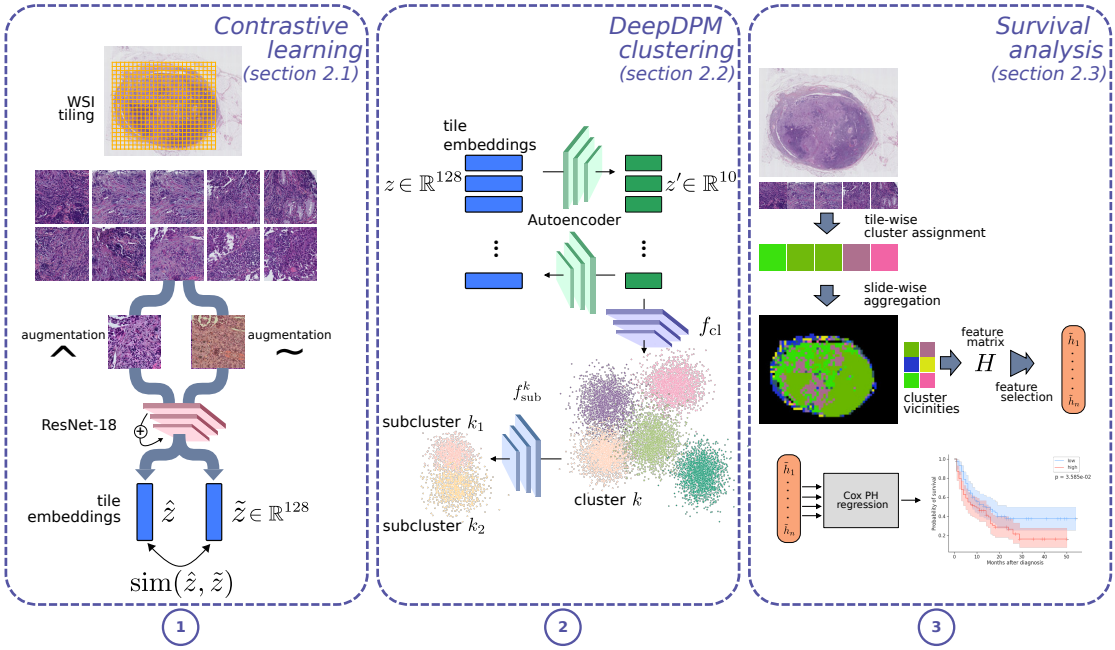


Fig. 5.1.: Overview of the WhARIO three-step workflow we use in this chapter. The method requires first contrastive pretraining, then clustering the tissue in lung slides, before feature matrices are derived from cluster vicinities and selected for the final survival analysis.

the same image augmented in two different ways, i.e., by minimizing the Normalized Temperature-scaled cross-entropy (NT-Xent):

$$\ell = -\log \frac{\exp(\mathrm{sim}(\hat{z}, \tilde{z})/\tau)}{\sum_{t \neq \hat{z}} \exp(\mathrm{sim}(\hat{z}, t)/\tau)} \tag{5.1}$$

in which $\mathrm{sim}(\cdot)$ is the cosine similarity function and τ is a temperature parameter. For the set of transformations, we follow the same protocol as [Ciga, 2022], that is random resized cropping, horizontal or vertical flipping, rotations, color jittering and Gaussian blur. Another advantage of contrastive learning is that it already enforces similar tiles to be closer in the latent space, which can help the following clustering algorithm.

5.2.2 DeepDPM Clustering

Given that we want to devise a data-driven approach without using any prior knowledge on histological patterns associated to the treatment response, we start our approach by clustering the tissue within each slide. However, most of the clustering methods – even among the most recent ones – require to define a number of clusters beforehand. There are nonetheless a few clustering algorithms which overcome this difficulty, such as DBSCAN [Ester, 1996]. More recently, [Ronen, 2022] introduced DeepDPM, a deep clustering method (i.e., based on a deep learning model) that uses a Dirichlet Process Gaussian Mixture Model (DPGMM) to remove the need to predefine a fixed number of

clusters. The method is based on two different models that are trained alternatively: a clustering network, that infers a number of clusters and assigns each point to them, and an autoencoder, which not only reduces yet again the dimension of the latent space for clustering, but also projects input data closer to the cluster centroids. We start by describing the principles of the clustering model, before introducing the autoencoder. Let $\mathcal{X} = (\boldsymbol{x}_i)_{i=1}^N$ denote a dataset of N points in $\mathbb{R}^d$. The mixture can be written:

$$p(\boldsymbol{x} \mid (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k)_{k=1}^{\infty}) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{5.2}$$

where $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is a Gaussian density function parameterized by $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and π_k a strictly positive real number such that $\sum_{k=1}^{\infty} \pi_k = 1$. Two different prior distributions are defined: for the components $\boldsymbol{\theta} = (\boldsymbol{\theta}_k)_{k=1}^{\infty}$, it is the Normal-Inverse Wishart (NIW) distribution, whereas for the weights $\boldsymbol{\pi} = (\pi_k)_{k=1}^{\infty}$, it is a Griffiths-Engen-McCloskey stick-breaking process (GEM) with concentration parameter α, the expected number of clusters.

DeepDPM adopts a Metropolis-Hastings inspired split/merge framework to automatically handle the total number of clusters, where the split of a cluster is accepted with probability $\min(1, H_s)$, where:

$$H_s = \frac{\alpha \Gamma(N_{k,1}) f_{\boldsymbol{x}}(\mathcal{X}_{k,1}; \lambda) \Gamma(N_{k,2}) f_{\boldsymbol{x}}(\mathcal{X}_{k,2}; \lambda)}{\Gamma(N_k) f_{\boldsymbol{x}}(\mathcal{X}_k; \lambda)} \tag{5.3}$$

where $\Gamma(\cdot)$ is the Gamma function, $\mathcal{X}_k$, $\mathcal{X}_{k,1}$ and $\mathcal{X}_{k,2}$ represent the sets of points in cluster k, and its subclusters k_1 and k_2 respectively (with $|\mathcal{X}_\bullet| = N_\bullet$), $f_{\boldsymbol{x}}$ is the marginal data likelihood with respect to the NIW distribution and its parameters λ. Consequently, the merging of two clusters is accepted with probability $\min(1, H_m)$ where $H_m = 1/H_s$.

The (soft) cluster and subcluster assignments of the data are obtained using single hidden layer perceptrons: f_{cl} computes for each data point a vector that contains the membership probabilities for each cluster, i.e. $f_{\text{cl}}(\mathcal{X}) = \boldsymbol{P} \in \mathbb{R}^{N \times K}$ where K is the number of clusters. For each current cluster k, a subcluster network f_{sub}^k computes a vector of membership probabilities for the two subclusters, $f_{\text{sub}}^k(\mathcal{X}_k) = \widetilde{\boldsymbol{P}}_k \in \mathbb{R}^{N_k \times 2}$. Each kind of network has its own loss function. For f_{cl}, it is:

$$\mathcal{L}_{\text{cl}} = \sum_{i=1}^{N} \text{KL}(\boldsymbol{p}_i || \boldsymbol{p}_i^{\text{E}}) \tag{5.4}$$

where KL is the Kullback-Leibler divergence, $\boldsymbol{p}_i^{\text{E}} = (p_{i,k}^{\text{E}})_{k=1}^K$ are the expected cluster membership probabilities obtained during the E-step of the Expectation-Maximisation algorithm (EM) [Dempster, 1977], following:

$$p_{i,k}^{\text{E}} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})} \tag{5.5}$$

An isotropic loss is used for f_{sub}, i.e.:

$$\mathcal{L}_{\text{sub}} = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \sum_{j=1}^{2} \tilde{p}_{i,j} \|\boldsymbol{x}_i - \tilde{\boldsymbol{\mu}}_{k,j}\|_2^2 \tag{5.6}$$

where $\tilde{\boldsymbol{\mu}}_{k,j}$ is the mean of subcluster j in cluster k.

On top of the previously detailed mechanisms, the authors of DeepDPM propose to alternate between pure clustering and feature learning, by the means of an autoencoder (AE) $\boldsymbol{g} \circ \boldsymbol{f}$ initialized beforehand by minimizing a reconstruction loss:

$$\mathcal{L}_{\text{rec}} = \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}_i)) - \boldsymbol{x}_i\|_2^2 \tag{5.7}$$

and then trained to minimize the mean-square error between embeddings $\boldsymbol{f}(\boldsymbol{x}_i)$ and cluster centers $\boldsymbol{\mu}_{z_i}$:

$$\mathcal{L}_{\text{MSE}} = \|\boldsymbol{f}(\boldsymbol{x}_i) - \boldsymbol{\mu}_{z_i}\|_2^2 \qquad z_i = \text{argmax}_k \, p_{i,k} \tag{5.8}$$

The entire model alternates between training the clustering and subclustering networks while the AE is frozen, and training the AE through $\mathcal{L}_{\text{AE}} = \mathcal{L}_{\text{rec}} + \gamma \mathcal{L}_{\text{MSE}}$ ($\gamma \in \mathbb{R}^+$) while the clusters are fixed. The number of alternations is a fixed hyperparameter, as well as the number of epochs to train each part of the model. When training the clustering and subclustering networks, the total number of clusters changes following the trigger of split or merge operations.

5.2.3 Feature selection and survival analysis

After clustering the tiles of the slides in the training set, we assume that cluster adjacency within the slides hold useful prognostic information. To capture the interactions between clusters in slides, we count for each tile the different clusters represented in its 8-neighborhood, i.e. in adjacent tiles. We also include the background as an extra cluster for tiles on the edge of the tissue region. Therefore, for each slide, we build a matrix $H \in \mathbb{R}^{K \times (K+1)}$ where K is the final number of clusters obtained. When there are several slides for a single patient, the matrices are summed together to obtain a single matrix per patient. As the slides can include various amounts of tiles, we also normalize the matrix H by the column-wise sum of its elements. Figure 5.2 illustrates how the matrix H is filled.

To select the features predictive of survival, and that are used by the Cox PH model [Cox, 1972], we first apply iterative forward variable selection to the entire set of features H. One of the most well established methods for this is MRMR [Ding, 2005], or

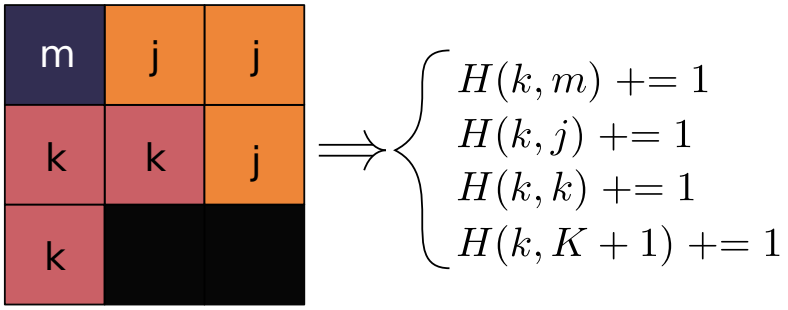


Fig. 5.2.: Construction of the feature matrix H based on cluster neighborhoods. Here, we assume a center tile in a slide belonging to cluster k, and the tiles in its 8-neighborhood. For each different cluster k' touching it, the matrix entry $H(k, k')$ is incremented by 1. The background (black region on the image) corresponds to index $K + 1$.

minimum redundancy, maximum relevance. In MRMR, the feature set is progressively filled by taking the feature that has the highest correlation with the outcome while being the least correlated to the other ones. However, MRMR does not have a stopping criterion to prevent the addition of undesired features, and was primarily designed for classification tasks. Boruta [Kursa, 2010] is another popular method, but it relies on permutation-sensitive models such as Random Forests, which is not the case of the Cox PH model. Therefore, we propose our own feature selection method specific to survival analysis, using two survival-related metrics. First, the concordance index (c-index), which evaluates the model's ability to rank correctly the survival times:

$$C_{\text{index}} = \frac{\sum_{i,j} \mathbb{1}_{T_j<T_i} \cdot \mathbb{1}_{\eta_j>\eta_i} \cdot \Delta_j}{\sum_{i,j} \mathbb{1}_{T_j<T_i} \cdot \Delta_j} \tag{5.9}$$

where T_i and T_j are survival times, η_i and η_j the predicted risks and $\Delta_j = 1 - \delta_j$, with δ_j the censorship indicator, which is equal to one if the survival time is censored (i.e., the patient was still alive at the end of the observation period). The c-index stands between 0.5 (random ranking) and 1 (perfect ranking). The second metric is the p-value associated to the log-rank test [Mantel, 1966] between the low and the high risk groups of patients. At each step, a single feature is added to the feature set used for regression, and a mean c-index $\bar{c}$ is computed on a M-fold cross-validation of the dataset. The patients are separated in low- and high-risk groups based on the inferred risk scores, so that the p-value p_{lr} of the log-rank test is used to check the difference. The results are ranked according to the ratio $z = -\bar{c}/\log_{10}(p_{\text{lr}})$, and the feature that decreases this ratio is added to the selected ones, until the improvement plateaus. The intuition behind this ratio is the following: the mean c-index provides a raw performance metric of the model, that we wish to improve. However, to discard only minor improvements, and make sure the average performance is not caused by some result peak on a specific fold, we use the magnitude of the p-value of the log-rank test on the entire training set to check that the patient stratification also improves significantly. Algorithm 3 summarizes the procedure.

Algorithm 3: The feature forward selection algorithm.

Data: A dataset $\mathcal{D} = (\boldsymbol{H}, \boldsymbol{T}, \boldsymbol{\delta})$ // H = feature matrix, T = survival times, δ = censorship

Result: A feature set $\mathcal{S}_H$

Initialization:

$\mathcal{S}_H \leftarrow \{\}$

Divide $\mathcal{D}$ in M subsets $\mathcal{D}_m$, $m \in \{1, ..., M\}$ // M-fold CV

$z_{\text{old}} \leftarrow \infty$ // Initialize the best baseline score

do

  for $h \leftarrow \{\boldsymbol{H}(0,0), ..., \boldsymbol{H}(K, K+1)\} \setminus \mathcal{S}_H$ **do**

    $\widetilde{\mathcal{S}_H} \leftarrow \mathcal{S}_H \cup \{h\}$

    $\mathcal{C} \leftarrow \{\}$

    $\mathcal{P} \leftarrow \{\}$

    $\boldsymbol{\eta} \leftarrow \{\}$

    $c \leftarrow 0$

    for $m \leftarrow \{1, ..., M\}$ **do**

      fit $\text{CPH}_{\widetilde{\mathcal{S}_H}}$ on $\mathcal{D} \setminus \mathcal{D}_m$ // CPH = Cox PH

      $\boldsymbol{\eta} \leftarrow \boldsymbol{\eta} \cup \{\text{CPH}_{\widetilde{\mathcal{S}_H}}(\mathcal{D}_m)\}$ // Test the Cox model on each validation set

      $c \leftarrow c + C_{\text{index}}(\mathcal{D}, \boldsymbol{\eta})$

    end

    $\overline{c} \leftarrow c/M$ // average performance on the M folds

    $\mathcal{C} \leftarrow \mathcal{C} \cup \{\overline{c}\}$

    Split $\mathcal{D}$ in $\mathcal{D}_{\text{low}}$ and $\mathcal{D}_{\text{high}}$ based on $\text{median}(\boldsymbol{\eta})$

    $\mathcal{P} \leftarrow \mathcal{P} \cup \{p_{\text{lr}}(\mathcal{D}_{\text{low}}, \mathcal{D}_{\text{high}})\}$

  end

  $z_{\text{new}} \leftarrow \min(\{z \,|\, z = -\frac{c}{\log_{10}(p_{\text{lr}})}, c \in \mathcal{C}, p \in \mathcal{P}\})$

  $\mathcal{S}_H \leftarrow \widetilde{\mathcal{S}_H}$

  $z_{\text{old}} \leftarrow z_{\text{new}}$

while $z_{\text{new}} < z_{\text{old}}$

One last important aspect of the Cox PH model is to check that the computed regression coefficients are significantly different from zero. We use the usual Wald statistical test on the regression coefficients to do backward elimination of the features by removing the ones which have a p-value > 0.05 associated to their coefficient.

5.3 Materials

5.3.1 Dataset

For this work, we use an in-house dataset of 193 lung cancer patients, collected through 5 different medical centers in France. The PD-L1 expression and overall survival information was available for 149 patients out of the 193. One or several cancerous tissue slides were collected for each patient, either through biopsy (all the centers except for

Variable name		All (N=149)	Nice (N=22)	Caen (N=8)	Dijon (N=36)	Rouen (N=27)	Toulouse (N=56)
Age, years, median (range)		62 (30-90)	61 (38-82)	70 (62-83)	63 (47-79)	61 (45-90)	61 (30-82)
Sex, no.	Female	44	2	0	13	9	20
	Male	105	20	8	23	18	36
Stage, no.	<3	5	5	0	0	0	0
	3	23	7	0	0	3	13
	4	121	10	8	36	24	43
Histology, no.	ADK	107	12	4	30	19	42
	SCC	37	8	4	4	8	13
	Other	5	2	0	2	0	1
Smoking, no.	yes	61	13	8	17	10	13
	no	8	0	0	0	2	6
	former	74	9	0	13	15	37
	unknown	6	0	0	6	0	0
TPS, no.	<1%	47	8	0	5	0	34
	1 – 49%	40	9	1	12	3	15
	>50%	62	5	7	19	24	7
Survival, months, median (range)		10 (1-54)	9 (2-35)	7 (1-17)	11 (1-39)	14 (1-34)	8 (1-54)
Tiles per slide, no., median (range)		425 (11-7355)	2885 (60-7355)	139 (22-1289)	83 (11-7197)	371 (37-4274)	604 (14-7333)

Tab. 5.1.: The clinical information of the patients in the cohort. ADK stands for adenocarcinoma, while SCC stands for squamous cell carcinoma."Other" means other rare histological subtypes, either sarcomatoid carcinoma or undifferentiated. TPS expression is reported following intervals based on the thresholds commonly found in the literature. For categorical variables, the number of patients is given. For continuous ones, we provide the median and the range.

Nice) or resection (Nice, and a single slide per center for the other ones). All of the slides were scanned at the Nice Pasteur hospital using a Hamamatsu NanoZoomer scanner with $\times 20$ magnification. The cohort is rather heterogeneous, as it contains several histological subtypes of non-small cell lung cancer (NSCLC), at different stages and with various levels of PD-L1 expression. Table 5.1 summarizes the statistics of the cohort. We use the slides from the 193 patients for the first two unsupervised steps of the method (i.e., contrastive learning and DeepDPM), but restrict to the set of 149 patients for the survival analysis.

5.3.2 Experimental setting

5.3.2.1 Tile extraction and pretraining

From the available WSIs, 256×256 tiles were extracted at $\times 20$ magnification within the tissue regions after thresholding the saturation channel in the HSV color space, and removing artifacts and blurry regions thanks to Gaussian filtering and the coverslip edge detector from the HistoQC package [Janowczyk, 2019]. Some regions in the slides such as blood stains were manually removed based on their unlikely correlation with survival. After preprocessing and tile extraction, we obtained approximately 350,000 tiles. As table 5.1 shows, the number of extracted tiles per slide greatly varies between the centers, (the median number of tiles per slide is 82 for the cases of Dijon, 2885 for the cases of Nice). Since the cases from Nice mostly correspond to resection samples, it explains why the numbers are so high for this center in particular. For pretraining, we used a ResNet-18 [He, 2016], with a batch size of 1024. We used the LARS optimizer [You, 2017] with

cosine annealing and initial learning rate set to 0.3 $\times$ *batch size* / 256 following the recommendations of the authors of SimCLR. Concerning the augmentations, the color jitter was applied with probability 0.8 to brightness, contrast, saturation and hue channels with factors 0.8 for the first three, and 0.2 for the last one. Rotations and flips were applied with probability 0.5. The Gaussian blur kernel size was taken as 1% of the patch size with sigmas ranging from 0.1 to 2.0, and the random crop was cut from 8 to 100 % of the patch. The network was trained for 200 epochs with early stopping triggered by validation loss plateau, on two NVIDIA A40 GPUs (40 hours). At the end of pretraining, all tiles were projected in the final 128-dimensional space of the network for the next part. The entire implementation was done using *pytorch v1.12.1* [Paszke, 2019].

5.3.2.2 DeepDPM clustering

From the set of 350,000 tiles, we decided to apply a sampling limit of 1000 tiles per slide given the high variability in tissue quantity between the slides. This led to a dataset of approximately 120,000 tiles for clustering. As stated in Section 5.2.2, we use the setting where clustering alternates with the training of an AE. For the encoder, we use the same architecture as the authors of the original paper for their Imagenet experiment, i.e. a 128-500-500-2000-10 MLP. The AE is pretrained for 50 epochs at the start (using only $\mathcal{L}_{\text{rec}}$), before the alternation scheme begins. We use 150 epochs for clustering, 100 for the AE, and 3 alternations in total. The total training time was 15 hours on a single NVIDIA RTX 2080 Ti.

5.3.2.3 Survival Analysis

After the training of the clustering model is over, each tile in the dataset is associated to a cluster. The slide-level feature matrices are computed by aggregating all of the tiles within each slide and following the description in Figure 5.2, and summed in case there are several for one patient. The lifelines package (v0.27.3, [Davidson-Pilon, 2019]) is used to conduct all of the subsequent survival analyses. We use the Cox PH model to output risk scores which allow for c-index computation and risk group separation. Unless otherwise specified, the experiments are conducted on a 5-fold CV of the dataset. The mean c-index is computed based on the c-indexes obtained for each fold, and the results of all 5 validation groups were gathered to perform risk group separation. We also compute the Kaplan-Meier estimates [Kaplan, 1958] of the survival function for each group.

5.4 Results

5.4.1 Clustering

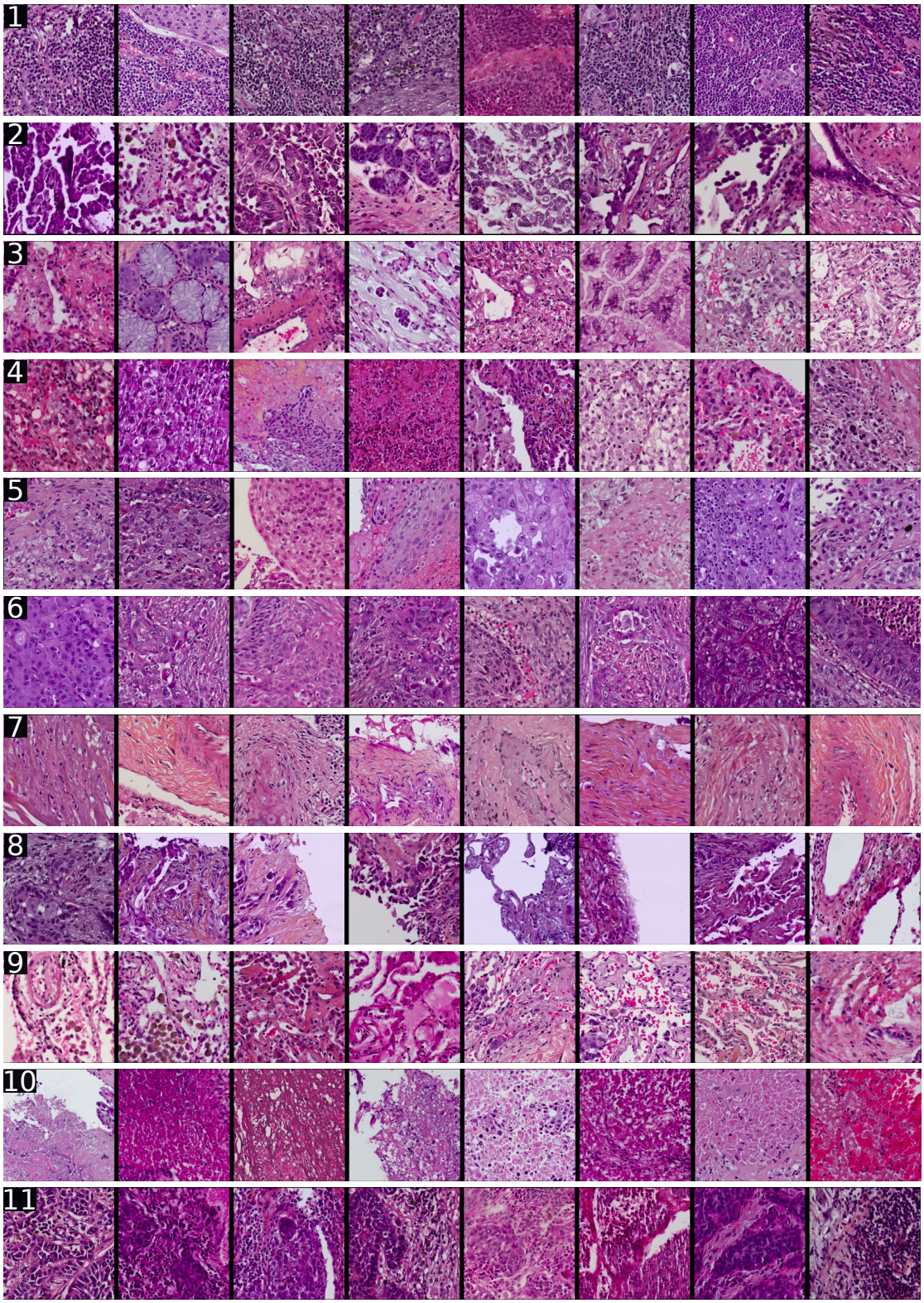


Fig. 5.3.: Tile samples corresponding to each discovered cluster.

At the end of the DeepDPM training phase, we obtained 11 clusters. To make sure the clusters did not correspond to trivial groups within the dataset (such as the center of

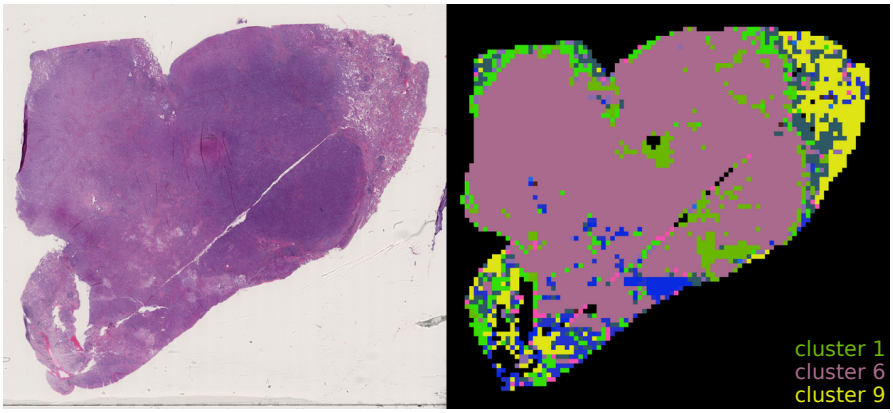


Fig. 5.4.: Example of a WSI in the dataset next to its tile-wise representation as clusters. The pink, green and yellow colors correspond to clusters 6, 1, and 9 respectively, which clearly identify the center tumor bulk (6) and surrounding lymphocytes (1), with normal lung parenchyma on the right (9).

origin, or the histological subtype), we computed the point-biserial correlations (which is the equivalent of the Pearson correlation coefficient between a continuous variable and a categorical one) and observed little to no correlation between them ($R<0.3$). Figure 5.3 shows tile samples corresponding to each cluster, while Figure 5.4 shows an example of a WSI next to its cluster representation. To examine the nature of the tissue within the clusters, we created mosaics of tiles sampled within each one of them (similar to what is shown Figure 5.3), and submitted them to the pathologist's inspection. For all of them, specific and identifiable histological patterns were found, with certain clusters harboring particularly homogeneous tissue (e.g., cluster 7 containing only fibrosis). Although there is partial overlap between clusters, or intra-cluster variability, the overall cluster assignment translates a certain histological logic. Table 5.2 summarizes the comments made by the expert pathologist on all of the clusters.

Cluster ID	Comments
1	Mostly tumor and/or inflammation regions with lymphocytes
2	Mainly papillary adenocarcinoma and fibrosis
3	Mainly mucinous adenocarcinoma
4	Mostly inflammation regions again, but with more diversity among the cells: lymphocytes, macrophages and neutrophils. Some of the tiles sport fibrosis
5	A mixture of different tissues: mostly squamous cell carcinoma, and some with fibrosis or inflammation
6	A lot of solid tumor areas, and a bit of stroma or fibrous tissue. Some tiles come from bronchial cartilage tissue
7	Nearly only fibrosis, with no tumor at all
8	There is more background in these tiles than in any other cluster, with mainly fibrosis and inflammation
9	Normal alveolar parenchyma mostly, and some fibrosis or necrosis
10	Mostly necrosis and hemorrhage, and some normal alveolar tissue.
11	Mainly inflammation with lymphocytes again, with some tiles displaying tumor

Tab. 5.2.: Summary of the pathologist's comments on the different clusters.

5.4.2 Feature selection

From the 11 discovered clusters, we obtain for each slide a feature matrix of dimension $11\times(11+1)$. After running Algorithm 3, we obtain a set of 6 features. Then, these 6 features are filtered to recover only the features with a p-value < 0.05 in the Cox PH model: only three features are kept after this process: $h_{1,2}$, $h_{6,4}$ and $h_{6,7}$. Going back to Table 5.2, we see that the interactions correspond to either tumor/inflammation or tumor/fibrosis neighborhoods. The former is very coherent with the nature of immunotherapy and previous findings on the role of the inflammatory response with respect to survival [Saltz, 2018; Thorsson, 2018], thus it is a reassuring result with respect to the considered cohort. The following section illustrates how prognostic these three features are in terms of survival.

5.4.3 Survival Analysis

5.4.3.1 Cross-validation on the entire dataset

features	c-index (95% CI)	p_{lr}	HR (95% CI)
TPS (1% threshold)	N/A	0.003	1.81 (1.21-2.69)
TPS	0.617 (0.558-0.676)	0.06	1.46 (0.98-2.17)
WhARIO	0.638 (0.603-0.673)	1×10^{-4}	2.29 (1.48-3.56)
WhARIO + TPS	**0.697 (0.650-0.744)**	$\mathbf{3\times10^{-6}}$	**2.60 (1.72-3.94)**

Tab. 5.3.: C-indexes, HRs and p-values of the log-rank test comparison between the various feature sets on the cross-validation. The best metrics appear in bold.

To evaluate the results of our method, we first check what is the prognostic power of the TPS on our dataset. First, the threshold of 1% [Reck, 2019] is used to separate patients in low- and high-risk groups (without any model). Then, we also evaluate the performance of a Cox PH model fitted on the TPS only. Finally, we train a Cox PH model using the WhARIO features we selected in Section 5.4.2, but also the combination of these with TPS. Figure 5.5 shows the obtained survival curves with the associated log-rank p-value and c-indexes, while Table 5.3 summarizes the results. When using the 1% threshold, low- and high-risk groups have statistically different survival (p = 0.003), which is coherent with the literature. However, this is not verified for each center individually, as only a

Center	p_{lr}	HR (95% CI)
Caen	N/A	N/A
Dijon	0.002	4.14 (1.56-11.02)
Nice	0.31	1.68 (0.60-4.66)
Rouen	N/A	N/A
Toulouse	0.15	1.68 (0.82-3.45)
All	0.003	1.81 (1.21-2.69)

Tab. 5.4.: Hazard Ratios and p-value of the log-rank test when using the 1% threshold of TPS to split risk groups.

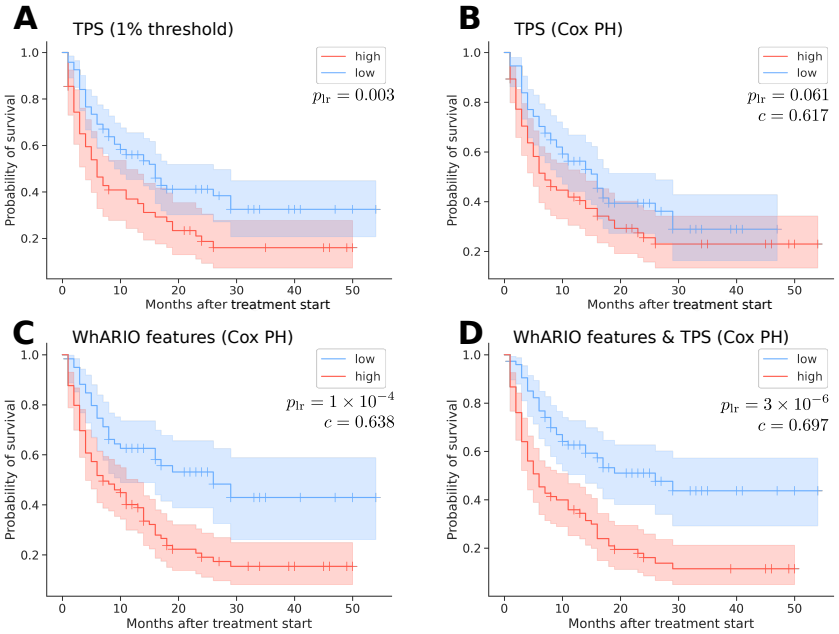


Fig. 5.5.: Low- and high-risk group survival curves based on (A) the 1% TPS threshold, (B) a Cox PH regression on TPS values, (C) a Cox PH regression on WhARIO features, and (D) a Cox PH regression on WhARIO features and TPS combined.

single center, Dijon, has significantly different survival for each group (cf. Table 5.4). What is more, two centers out of the five (Caen and Dijon) do not include low-TPS patients, making a threshold-based grouping impossible.

On the other hand, a Cox PH regression on all continuous values of the TPS does not lead to a statistically significant difference between the groups ($p = 0.06$). On the contrary, our features yield two groups with statistically significant difference in terms of overall survival ($p = 1 \times 10^{-4}$). The mean c-index is 0.638, which is comparable, and even slightly superior to the one obtained with TPS alone. Another remarkable result can be achieved when we add the TPS to the set of selected features: the p-value of the log-rank test gets smaller by a factor 100 ($p = 3 \times 10^{-6}$), while the HR increases from 2.29 to 2.60 and the c-index from 0.638 to 0.697. Combining WhARIO features and TPS scores in a Cox PH model allows for more distinguishable risk groups than the reference 1% threshold.

5.4.3.2 Leave one center out

To further validate the prognostic power of our selected features, we conducted additional experiments where one center was completely left out to be used as a test set. In this setting, we use 4 centers for 5-fold CV, report the CV performance as in the previous section, and select the best performing model (based on the metric we introduced in Section 5.2.3) to make predictions on the left-out center. Based on the center characteristics in Table 5.1 however, we chose not to select Caen and Rouen as the test set for these experiments. For Caen, it is due to the lack of a sufficiently large cohort (8 patients

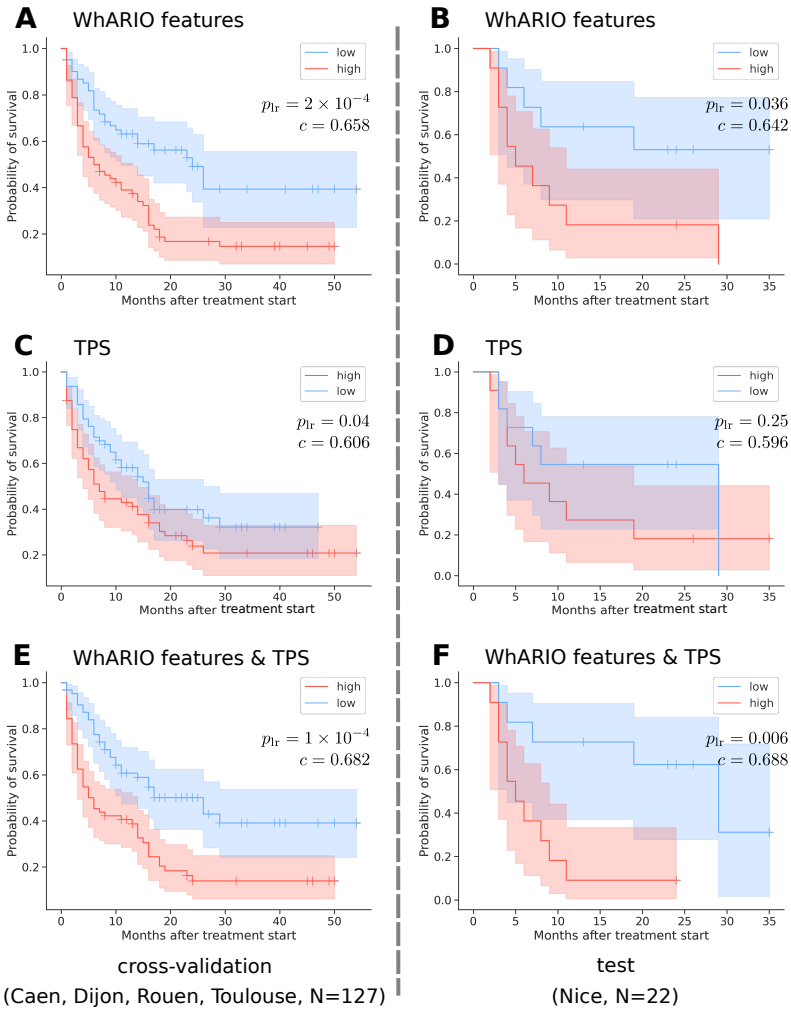


Fig. 5.6.: Low- and high-risk group survival curves based on a Cox PH regression using (A,B) WhARIO features only, (C,D) TPS only and (E,F) a combination of the two when Nice is the left-out test set. The left column corresponds to the CV, and the right column to the test set.

only). For Rouen, the main reason is that it is an outlier compared to the other centers in terms of both TPS and median survival. For the remaining centers, we present and discuss the outcomes of our experiments on Nice (N=22) and Toulouse (N=53) in what follows. Again, we compare three settings: using TPS only, using WhARIO features only, and finally combining the two. Tables 5.5 and 5.6 and Figures 5.6 and 5.7 show the corresponding metrics and survival curves when using the Nice center and the Toulouse center as the test sets. With the WhARIO features, we obtain results consistent with what we obtained in section 5.4.3.1: a mean c-index of 0.658 (resp. 0.628), a HR of 2.31 (resp. 2.01), with a comparable log-rank test p-value (2×10^{-4}) in the first experiment. In the second one, although higher (p=0.01), the p-value stays reasonably under 0.05. The test set yields a c-index of 0.642 for Nice (resp. 0.678 for Toulouse), a HR of 3.01 (resp. 2.77) and statistically significant survival difference between low- and high-risk groups ($p_{\text{lr}} = 0.036$ and $p_{\text{lr}} = 0.008$). With the TPS alone, however, although the results on the CV are very close to the ones obtained on the entire dataset, the model only yields

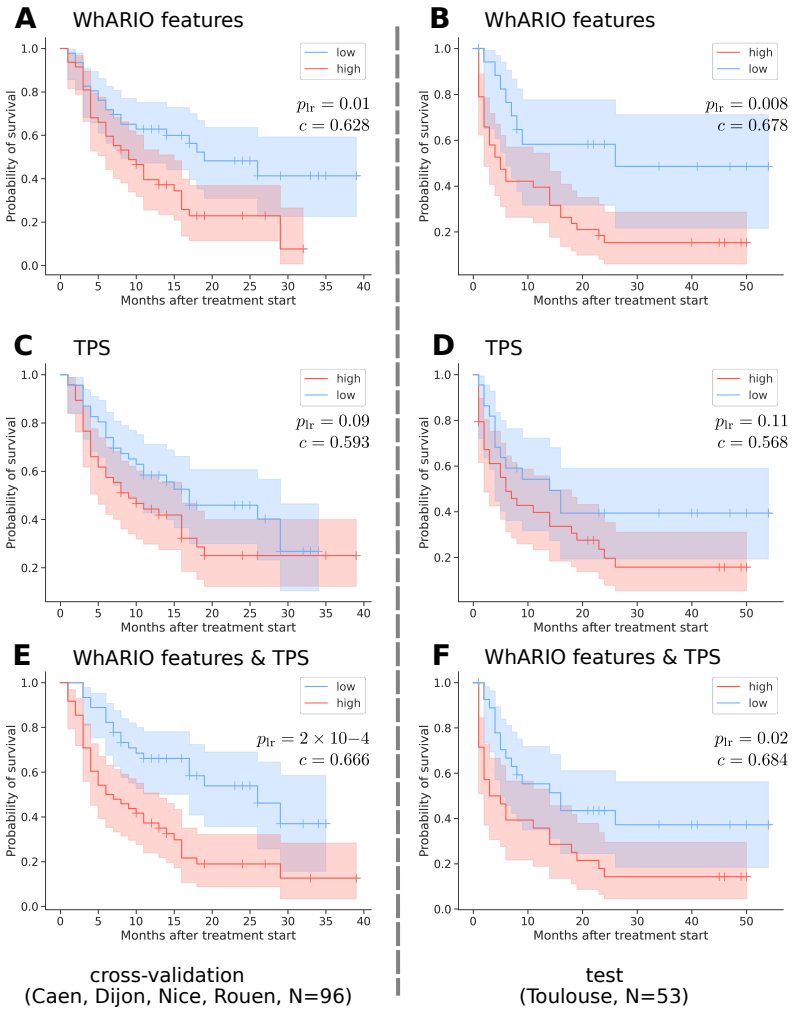


Fig. 5.7.: Low- and high-risk group survival curves based on a Cox PH regression using (A,B) WhARIO features only, (C,D) TPS only and (E,F) a combination of the two when Toulouse is the test set. The left column corresponds to the CV, and the right column to the test set. Following the comments in Section 5.4.3.2, although there is a clear difference in the stratification between (A) and (E), it is much less visible between (B) and (F).

a c-index of 0.596 on the test set in the first experiment (0.568 in the second), with no statistically significant difference between low- and high-risk groups of patients. When combining PD-L1 with WhARIO features, the results on the cross-validation improve once again. Regarding the test set, the combination also yields a more accurate prognosis for Nice, both in terms of c-index and log-rank test p-value. For Toulouse on the other hand, the c-index is indeed slightly higher with the combination, but so is the p-value: the benefit of combining the two is not as explicit this time.

The Dijon cohort is a special case for this set of experiments, since it is the center that has the lowest amount of tiles per slide, with a median at 82 (against >2000 tiles per slide in Nice see Table 5.1), which has a strong impact on the presence of the clusters of interest in the slides: in 28 out of the 36 slides of this center (78%), none of the

	features	c-index (95% CI)	p_{lr}	HR (95% CI)
CV (Caen, Dijon, Rouen, Toulouse)	WhARIO	0.658 (0.608-0.707)	2×10^{-4}	2.31 (1.47-3.63)
	TPS	0.606 (0.489-0.722)	0.04	1.57 (1.01-2.43)
	WhARIO + TPS	**0.682 (0.654-0.709)**	$\mathbf{1\times10^{-4}}$	**2.39 (1.53-3.71)**
test (Nice)	WhARIO	0.642	0.036	3.01 (1.02-8.90)
	TPS	0.596	0.25	1.79 (0.64-5.06)
	WhARIO + TPS	**0.688**	**0.006**	**4.67 (1.41-15.50)**

Tab. 5.5.: C-indexes, HRs and p-values of the log-rank test comparison between the various feature sets on the CV and the left out test set (Nice). The best metrics appear in bold.

	features	c-index (95% CI)	p_{lr}	HR (95% CI)
CV (Caen, Dijon, Nice, Rouen)	WhARIO	0.628 (0.545-0.710)	0.01	2.01 (1.18-3.43)
	TPS	0.593 (0.476-0.709)	0.09	1.56 (0.92-2.62)
	WhARIO + TPS	**0.666 (0.588-0.743)**	$\mathbf{2\times10^{-4}}$	**2.66 (1.55-4.59)**
test (Toulouse)	WhARIO	0.678	**0.008**	**2.77 (1.80-6.03)**
	TPS	0.568	0.11	1.71 (0.88-3.31)
	WhARIO + TPS	**0.684**	0.02	2.15 (1.14-4.05)

Tab. 5.6.: C-indexes, HRs and p-values of the log-rank test comparison between the various feature sets on the CV and the left out test set (Toulouse). The best metrics appear in bold.

clusters whose neighborhoods are of interest to us are present, which severely hinders the potential of our approach. Nonetheless, we still performed the same kind of experiment we have conducted above on Nice and Toulouse, the results of which can be found in the Appendix A. We obtain similar trends regarding the metrics and the curves, but without the same level of statistical significance, mainly because of the lack of representativeness of the relevant clusters.

5.5 Discussion

With our approach, we showed that it was possible to perform survival prediction of lung cancer patients treated with ICIs, from the sole analysis of HES WSIs. The prognostic power of our method showed similar performances to the gold standard PD-L1 evaluation done on IHC slides, without having to resort to such modality. Another advantage of our pipeline compared to previous ones is that we did not use any annotation nor pathology prior information to select features or process tissue: our framework is entirely unsupervised, and purely data-driven. Nonetheless, we were still able to validate our histological findings a posteriori, by submitting the cluster samples to the gaze of an expert pathologist. From the different uncovered clusters, the interactions between two pairs in particular had significant impact on the distinction between low- and high-risk patients. When looking at their histological characteristics, it appeared that they were representing tumor/inflammation and tumor/immune adjacencies (Table 5.2, in particular pairs 1-2 and 6-4). Figure 5.8 illustrates this by showing two samples of slides

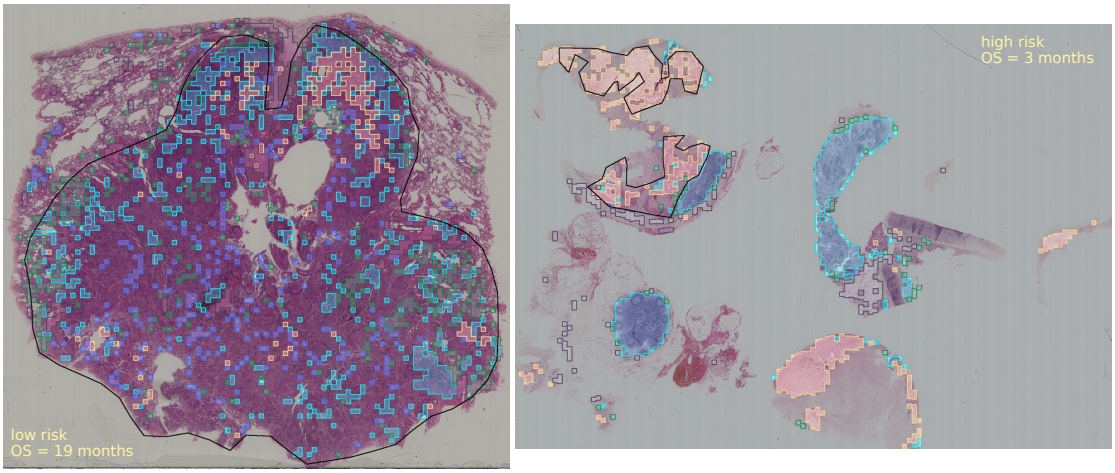


Fig. 5.8.: Low- and high-risk examples of slides from the Nice cohort, with the superposition of tile cluster assignments with respect to the selected ones for survival prediction. Cyan and green correspond to clusters 1 and 4 (mostly inflammation/lymphocytes), blue and yellow correspond to clusters 2 and 6 (mostly tumor). Cluster 7 appears in grey. The black line defines the contours of the tumor region. The unassigned tissue in the tumor area on the left has been assigned to another tumor-related cluster (cluster 11, cf Table 5.2).

classified as low and high risk. On the left, we can see that clusters 1 and 3 (cyan and green), i.e., lymphocyte-rich areas and inflammatory tissue, are largely present within the bounds of the tumor. On the right, however, the tumor region remains rather unscathed, with most of the immune-related tissue outside of its boundaries. These findings are well correlated to the nature of immunotherapy, as the interactions between immune and tumor cells are suspected to be associated to a positive ICI response [Tumeh, 2014]. We managed to uncover these interactions without any prior, only through nonparametric clustering. We also showed that the interaction between these regions yielded statistically significant risk group separations, in a cross-validation setting as well as for a left-out test set. Our approach is an interpretable way to exploit H&E slides directly to predict survival of lung cancer patients who received ICI, without having to rely on pathologists' annotations.

There are nonetheless some improvements to make, and further validation to perform. Regarding the clustering method used, there are several points we intend to address in the future. First, each new experiment with a different seed is susceptible to generate a different number of clusters. Yet, most of the experiments we ran ended with a stable final number of clusters between 11 and 15. For this article in particular, we decided to pick the experiment that generated the fewest clusters as possible, so as to reduce the feature exploration space that grows with K^2. Moreover, the clustering part also depends on the latent space that is learned with SimCLR. Since contrastive learning relies on the robustness of the representations with respect to transformations, and in particular, color jittering, more experiments involving pathology-specific augmentations [Tellez, 2019] or

color normalization [Macenko, 2009; Vahadane, 2016], would be needed to measure their impact on the obtained clusters.

Second, the inclusion of slide-level spatial information in the clustering process would be a welcome addition, as the spatial dependency between the tiles of a single slide are likely to bear important information. Yet, even without the tile dependency taken into account, the current clustering method already yields coherent slide-level clusters, which shows its ability to find common patterns without additional supervision (see Figure 5.4 for instance). Third, the spatial distribution of the clusters within the slides could also be an interesting information to consider when predicting patient survival. For now, we have only addressed this by summarizing direct cluster interactions in a matrix (i.e., tissues in contact), but perhaps a more global approach could yield further results. Given the sometimes random tissue distribution on slides, however, this should be studied with caution to avoid any undesired bias from artificial tissue proximity due to laboratory manipulation.

Fourth, concerning the feature extraction method, we only used in this chapter the strong cluster assignments as a way to describe the nature of the tissue. The reality is likely to be different, since tiles cover different histological structures at the same time. We could also look at how the cluster assignments vary when the tiles are shifted in different directions with a small step. That way, we could obtain a smoother representation of the slides, taking into account the superposition of cluster assignments, instead of a single one. This would probably also help correct some misassignments between clusters, as we observed that sometimes, tiles were given a cluster label different from their neighbors in spite of their resemblance.

Fifth, regarding the data itself, our analysis of the tissue only partially takes into account the spatial information relating to tissue organization in WSIs. However, as the dataset contains both resection and biopsy samples, in which the tissue arrangement can sometimes be somewhat artificial, one should be cautious as to which interactions are represented. The presence of biopsies with very little amount of tissue is in fact one of the drawbacks of the dataset we used. Among the slides with the lowest amount of tissue, there probably were missing elements to fully characterize them. Our comments concerning the Dijon cohort in Section 5.4.3.2 support this claim. To conclude on the dataset, the sheer amount of patients could be higher, so as to validate our findings. Another external dataset could confirm the results we obtained here, but also help us get stronger significance evaluations of the features we obtained. Although we decided to stick to a specific, common p-value threshold of 0.05, it is also likely that some features among the ones we pruned are nonetheless relevant to the outcome prediction. As an example of this phenomenon, we observed on the left-out-center experiments that TPS itself was sometimes above this particular threshold, although by a small margin. To

confirm the importance of other features unfortunately, we would need more samples to obtain a better estimate.

5.6 Conclusion

In this chapter, we have presented a 3-step pipeline to automatically, and without supervision, analyze tissue from WSIs of lung cancer patients who received ICI treatment to train a model for survival prognosis. We showed that our model yielded risk groups with statistically significant differences in survival in a multicentric population, both in a global cross-validation setting, as well as in 2 leave-one-center-out experiments. The histological interpretation of the most significant clusters pointed at the interactions between tumor and inflammation regions, a finding concordant with the literature and that we were able to recover without prior tissue selection.

Acknowledgments

The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support. This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

The authors would like to thank Hamila Marame, Julien Fayada, Marine Pedro, Cyrielle Falduzza, Olivier Carruggi, and Pascal Grier for their contribution to the preparation and digitization of the whole-slide images at the Nice hospital. The authors would also like to thank Amina Oyuntogos for her help in the analysis of the tissue observed in the clusters.

6 Conclusion

In this thesis, we proposed a new method for both classification and localization within WSIs, and another one for unsupervised feature extraction from tissue slides for survival prediction of patients treated with immune-checkpoint blockades. In particular, each of these methods represents our proposed answer to the challenges introduced at the beginning of this thesis: simultaneous classification and localization for digital pathology, annotation-efficient methods, and finally, the discovery of new biomarkers prognostic of treatment outcome. We summarize each contribution and discuss future directions in the remainder of this section.

6.1 Main Contributions

Mixed supervision for whole-slide image classification
In chapter 2, we introduced the concept of mixed supervision in digital pathology. Based on a pre-existing weakly-supervised, attention-based, multiple instance learning model, we showed that a few instance labels only could be used in order to simultaneously classify and localize tumors in whole-slide images. We showed that a few annotated slides were sufficient to improve the classification performance of the model, while increasing tremendously the localization accuracy. Even though the initial weakly-supervised model had already good slide-level classification results, its localization performance, i.e., tile-level classification, was rather poor. This is a major drawback which hinders the adoption of such models in clinical practice. With our approach, we promoted a more coherent model, that is able to establish a better correspondence between the local (tile) and the global (slide) predictions.

MS-CLAM: a more robust and unified approach to mixed supervision
In chapter 3, we redesigned more thoroughly the model that was used in chapter 2 and made several additions that truly embedded mixed supervision within the method. With the help of a newly designed loss function to better distribute the weights between the attention scores, and a slide sampling strategy that reduced the training process to a single step, we were able to correct and improve our first contributions to the model. Without sacrificing the efficiency of the original model, and without changes to its architecture, the coherence between slide-level and tile-level predictions was improved. Indeed, with attention scores more evenly distributed on key instances, i.e. tumor tiles

in tumorous slides, and nontumorous tiles in healthy slides, the attention-weighted sum of the features better represents the slide label. With our sampling strategy, we simplified our original approach to have the model trained with both annotated and unannotated slides at the same time. In turn, this reduced the training time, which is another advantage of our method compared to others based on the pre-training of a feature extractor such as [Li, 2021a]. On this note, we showed that pre-training the feature extractor was insufficient to really bridge the localization gap regardless of the percentage of annotations considered.

A new dataset for the study of lung cancer immunotherapy response

In chapter 4, we introduced a new histology-focused dataset, Lung-IO, built thanks to the contributions of 5 different medical centers in France, with nearly 200 lung cancer patients treated with ICIs. The dataset includes several clinical characteristics, such as age, gender, histological subtype, tumor stage, smoking status, PD-L1 expression levels, progression-free and overall survival, as well as the response to the treatment assessed using the RECIST criteria. The available diagnostic HES slides contain both biopsy and resection samples, thus representing a large variety of tissue that are routinely observed in clinical practice. We discussed the interpretation that could be made of the response criteria with respect to the patient overall and progression-free survival in order to design a binary classification problem. We also produced a set of baselines on response prediction using various models with different architectures and working principles. We showed that standard WSI classification models were not able to reach relevant performance thresholds on K-fold cross validations, and provided several explanations as to why these models failed, before motivating the need for a different approach on the subject.

Unsupervised discovery of biomarkers associated to immunotherapy outcome

Given that direct binary treatment outcome prediction (responders vs. nonresponders) failed in several attempts and with various models, we resorted to a different approach concerning treatment outcome by targeting survival analysis. In chapter 5, we have described how a three-step pipeline entirely data-driven was successful in stratifying patients in low- and high-risk groups after immunotherapy. To distinguish between several phenotypes without any histological guidance, we used a nonparametric clustering approach, i.e., that does not require the user to indicate a specific number of clusters. To achieve this, the method relies on Bayesian Dirichlet process mixtures to automatically infer the final number of clusters during training. These clusters were then used to build a neighboring matrix for each slide, that would represent the interactions between phenotypes for each patient. After applying forward feature selection from these matrices, we identified two specific clusters that were prognostic of survival. The posterior inspection of the clusters by a pathologist revealed that the most important ones contained solid tumor on the one hand, and immune cell-rich inflammatory tissue on the other. Our findings concur with recent studies on the relationship between the presence

of lymphocytes in the tumor area and the treatment response [Tumeh, 2014]. In our case, though, the discovery of these phenotype interactions as potential outcome predictors was done without any targeted focus or hints based on histological knowledge.

6.2 Future Research

In this thesis, we have presented several contributions to the field of digital pathology, through both whole-slide image classification and treatment outcome prediction, which nonetheless call for further inquiries.

Developing a large-scale digitized WSI dataset for the study of immunotherapy response in lung cancer
Although the dataset presented in chapter 4 is a very good starting point for the study of ICI response in lung cancer, this first draft could be improved in many different ways. First, more samples would obviously help consolidate it, as a little under 200 cases limits the possibility to both train large-scale models, but also to evaluate them on a larger test set. In comparison, a dataset such as Camelyon16 [Bejnordi, 2017], focused on the much more simple task of classifying and identifying tumors, contains as much as 400 cases, while TCGA hosts slightly more than 1000 lung cancer cases. Having more cases also offers the opportunity to select suitable samples more properly, especially in terms of slide quality. Regarding the latter, perhaps it would be also interesting to focus on similar samples, for instance on resection first, since they usually hold much more tissue than biopsy slides, and larger regions in the environment of the tumor. Yet, the presence of biopsy slides is nonetheless something to wish for, since they are representative of the slides that are generated in clinical practice, and should ensure that the models can leverage these slides too. The acquisition and digitization of the slides also demands more investigation, with this time a different process for each contributing center, involving other technicians and scanners. This would generate more diversity in the observed colors of the slides, which is crucial since it has a major impact on the performance of the models, as several studies have shown [Khan, 2014; Ciompi, 2017].

Second, aside from the sole histological point of view, there are additions to make regarding complementary sources of data. Throughout this work, we have never used genomics and have given justifications for it, but it is yet an inevitable source of information that should reinforce such a dataset in the future. The mutational profiles are very rich and detailed pieces of intelligence, and they are at the origin of several other promising biomarkers, such as the TMB. Used in conjunction with histological data, they could allow for the development of stronger predictors. However, mutational profiles are not necessarily standard in clinical practice, therefore additional efforts are required to obtain them. On the topic of modalities, the IHC slides, used until today in only a

few computational studies, also bear their lot of information that could be included for the treatment response prediction. They are themselves the source of a biomarker identification, namely the PD-L1 expression level, and reveal new features in the tissue compared to common HES samples.

Third, the evaluation of the treatment response should be discussed with care, since in many cases the treatment effect seems equivocal. In what we have seen, a PET-scan evaluation consecutive to the classical CT-scan examination allowed for a more precise sanction of the effect of immunotherapy on several patients. This procedure should perhaps be generalized to every new entry of the dataset, since it reduces the noise surrounding the labels. Finally, other factors such as the progression-free survival could be considered to strengthen the observations. Distinguishing properly responders and nonresponders is crucial for the pursuit of this study.

Improving and generalizing MS-CLAM to a broader range of tasks

With MS-CLAM, we have shown that mixed supervision could successfully be applied to straightforward tumor classification tasks. However, there are several ways in which we could improve this model. First, there are a few issues that still need to be solved. We have noticed throughout the experiments several behaviors of the model that would need to be changed. For instance, we have pointed in section 3.4.3 at a seemingly worst performance in localization when more annotations are used on Camelyon16. It is not exactly true, as table 6.1 reminds.

model	**AUC**	**AP**	*Precision*	**Recall**	**F1-score**
MS-CLAM (6%)	0.805 (0.036)	0.540 (0.065)	0.921 (0.020)	0.241 (0.055)	0.379 (0.065)
MS-CLAM (12%)	0.851 (0.045)	0.618 (0.078)	0.921 (0.008)	0.296 (0.026)	0.448 (0.029)
MS-CLAM (25%)	0.907 (0.027)	0.698 (0.057)	0.908 (0.024)	0.372 (0.079)	0.522 (0.077)
MS-CLAM (62%)	0.948 (0.005)	0.765 (0.019)	0.864 (0.030)	0.550 (0.035)	0.671 (0.021)
MS-CLAM (100%)	0.950 (0.004)	0.763 (0.014)	0.814 (0.022)	0.604 (0.026)	0.693 (0.016)

Tab. 6.1.: Tile classification metrics on the Camelyon16 test set. All metrics are averaged on a 5-fold cross validation split of the training set ($\pm$ a standard error reported). AUC, Average Precision (AP), recall and F1-score, marked in **bold**, all increase with the amount of annotations used. On the other hand, precision, marked in *italic*, steadily decreases.

However, one bothering trend is the constant decrease in the precision of the tile classifier, that is to say the increase of false positives, at the benefit of a higher recall, which means more true positives. This is desirable is some sense, since we expect from the model to be as sensitive as possible to the presence of tumor, but it also damages the legibility of the localization maps, especially when the tumor region is extremely small, or when there is not tumor on the slide (see e.g., Figure 6.1). To control the trade-off between the two, a minimum level of precision should be set, and the recall of the model could be adjusted

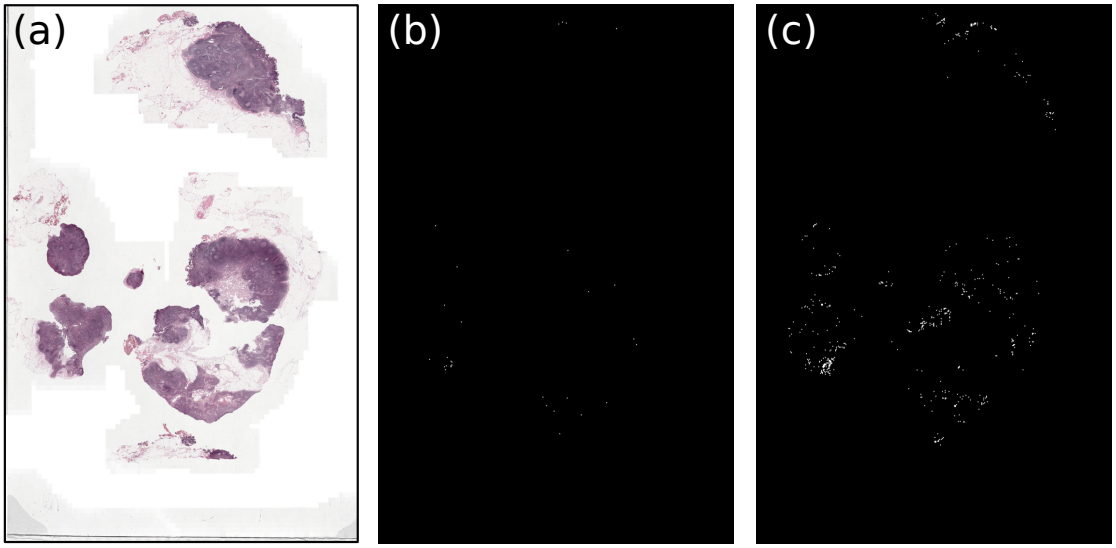


Fig. 6.1.: An example of a *healthy* WSI from the Camelyon16 next to tumor masks computed using MS-CLAM with two different percentages of annotations. (a) is the H&E slide, (b) corresponds to the tumor mask obtained from MS-CLAM with 6% annotations, (c) with 100%. There are many more false positives on the right-most image, making it less convenient for a pathologist to inspect.

with respect to this level. This would induce a more consistent progression between the models with increasing annotations.

Second, the model should be able to generalize to multi-class classification problems. In the original CLAM paper, [Lu, 2021] already propose what they refer to as a "multi-branch" version of the model, or CLAM-MB. In this version, instead of scalar attention scores, there is an attention matrix $A \in \mathbb{R}^{N \times C}$ where N is the number of tiles in the slide and C is the total number of classes. A weighted bag representation is computed using the attention scores of each column in A. With mixed supervision, we could enforce several constraints on this matrix, on top of the ones we have already introduced. If the classes are mutually exclusive, we could have an additional loss term on the columns of the matrix, that would penalize a cross-column attention distribution. For the cases where there is not a straightforward correspondence between the tile labels and the slide label, such as the Gleason grading problem, the model should also be modified to handle such contingency.

Third, the architecture of the model could be simplified and enhanced. The current two-branch structure is perhaps not the optimal design we could think of, and the attention network could serve directly as the instance classifier, in a manner that is similar to what [Shi, 2020] propose. Moreover, many improvements of the attention-based MIL model have been proposed since [Ilse, 2018], with for instance the introduction of self-supervision to factor in the dependency between the tiles [Rymarczyk, 2021; Li, 2021a]. Therefore, the added value of mixed supervision on top of these new features should be evaluated. Finally, the current localization precision of MS-CLAM is limited to the

tile size, which is a bit coarse and lacks subtlety. Several works like [Li, 2023] have introduced methods for weakly-supervised segmentation in whole-slide images, thus it would be interesting to push future research around MS-CLAM in this direction.

Immunotherapy outcome prediction: towards higher efficiency and better interpretation

The results we have obtained on the survival prediction for lung cancer patients treated with immunotherapy are encouraging, but these are only first steps towards more efficient models, with identifiable and practical biomarkers to be used in the clinical setting. Within the pipeline we presented in chapter 5, several elements require further exploration. The clustering method in itself should in the future incorporate domain knowledge, i.e., the spatial proximity that exists between the patches, since it is more likely that neighbor patches belong to the same cluster rather than distant ones. Perhaps a starting point for this would be to use contrastive predictive coding [Oord, 2018], as it has already been used for histopathology [Lu, 2019]. As an additional modification to the pipeline, there could be a way to group steps 1 and 2, i.e., contrastive learning and clustering, into a single step. There are several examples of deep clustering methods which learn image embeddings and clusters in a concurrent manner [Caron, 2018; Huang, 2020]. Recently, contrastive clustering [Li, 2021b] was proposed, and was applied successfully to digital pathology for tissue clustering [Yan, 2022]. The combination of contrastive learning with nonparametric clustering could help reduce the final number of clusters, and in turn identify robust common patterns across slides.

The method in chapter 5 is based on a texture analysis of the tissue. However, similar individual histological elements such as lymphocytes reside in various areas of the tissue, but are not specifically identified by our method. Rather, it is prone to assigning patches with the same kind of cells to different clusters depending on the tissue they lie on, or their concentration. Combining our current approach with fine-grained information such as nucleus positions could be an interesting way to incorporate multi-level information.

Regarding the treatment outcome prediction in general, a first interesting step would be to evaluate the method on data collected and scanned in a different center. Even though the dataset we have used is already multicentric, all the slides were digitized in the same center. With a different scanner, the colors produced by the stains would probably be different, thus causing a generalization barrier for the model. To prevent this from happening, the model should be trained with an efficient color normalization scheme to increase its robustness to stain variations. A second step would be to test it on other kinds of cancer, such as melanoma. In [Wang, 2022], the authors show that their methods is able to generalize to a small set of gynecological cancer cases, from a lung cancer training set. The same process could be applied to our method, but straightforward generalization would probably fail, since skin or cervix tissue is very different from the lung's, and we have already pointed at the limits of a texture-based analysis in the previous point. In this

case, another training set specific to the organ of interest would be necessary. Finally, to channel as much information as possible, multiple modalities ought to be used together to obtain accurate predictions. Works like the one of [Vanguri, 2022] seem to support such a claim.

A

Appendix: WhARIO – Leave-one-center-out experiment with Dijon as the test set

As Figure A.1 shows, the results on the Dijon cohort are more mixed than on the Nice or the Toulouse cohorts. Since 78% of the slides do not include any of the clusters that we leverage to predict survival, this is unsurprising. It also goes with the observation on the amount of tissue per slide, which is the lowest for the Dijon cohort (median number of tiles per slide: 82). However, similar trends appear yet again in this situation as for the other two cohorts: a regression on PD-L1 values alone is insufficient to correctly stratify risk groups in the test set (here, the two curves are nearly identical), and the combination of WhARIO features with the TPS does yield a higher c-index for both CV and test, and better separated risk groups than with the TPS alone. Table A.1 summarizes the results from this experiment.

	features	c-index (95% CI)	p_{lr}	HR (95% CI)
CV (Caen, Nice, Rouen, Toulouse)	WhARIO	0.665 (0.628-0.702)	2×10^{-5}	2.86 (1.72-4.77)
	TPS	0.603 (0.540-0.666)	0.022	1.71 (1.07-2.72)
	WhARIO + TPS	**0.699 (0.611-0.786)**	**2×10^{-6}**	**3.12 (1.91-5.07)**
test (Dijon)	WhARIO	0.541	**0.198**	**2.77 (1.27-6.03)**
	TPS	0.579	0.95	1.03 (0.45-2.36)
	WhARIO + TPS	**0.618**	0.59	1.25 (0.55-2.87)

Tab. A.1.: C-indexes, HRs and p-values of the log-rank test comparison between the various feature sets on the CV and the left out test set (Toulouse). The best metrics appear in bold.

Figure A.2 shows the results of the risk group stratification when a threshold different from the median is used to separate the cases in the last experiment (WhARIO + TPS): raising the threshold to distinguish patients using the 65$^{\text{th}}$ quantile of the risk score allows for a statistically significant difference between the survival periods to appear ($p_{\text{lr}} = 0.026$). The two groups have nonetheless a comparable number of samples, since there are 19 patients in the low-risk group, and 17 in the high-risk one. The same observation cannot be made on any of the other experiments, and especially in the experiment with TPS as the sole feature.

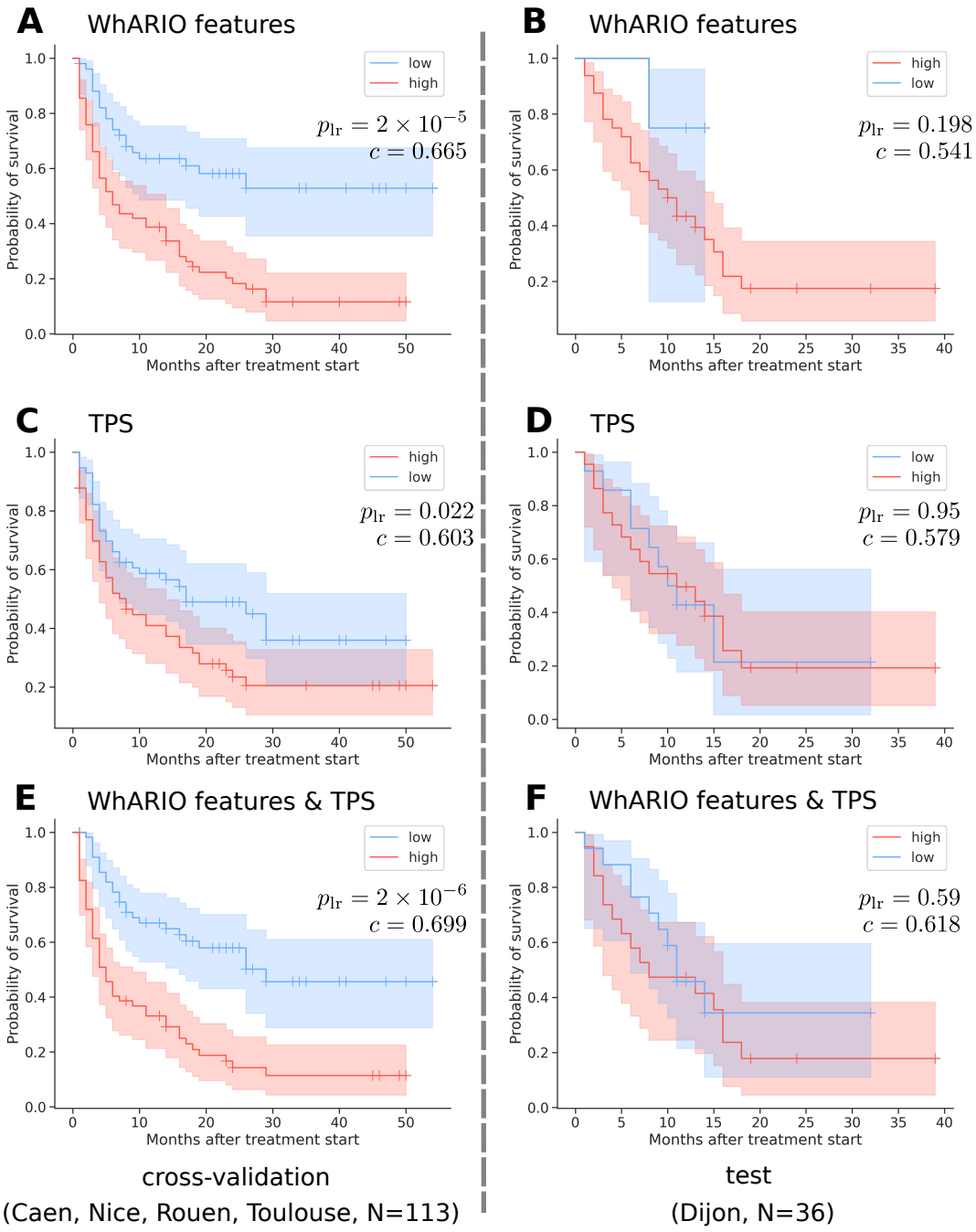


Fig. A.1.: Low- and high-risk group survival curves based on a Cox PH regression using (A,B) WhARIO features only, (C,D) TPS only and (E,F) a combination of the two when Dijon is the test set. The left column corresponds to the CV, and the right column to the test set. The absence of clusters of interest in most of the slides (78%) prevents any obvious and efficient patient stratification when using the clusters only. The combination of PD-L1 and WhARIO still yields a better c-index however.

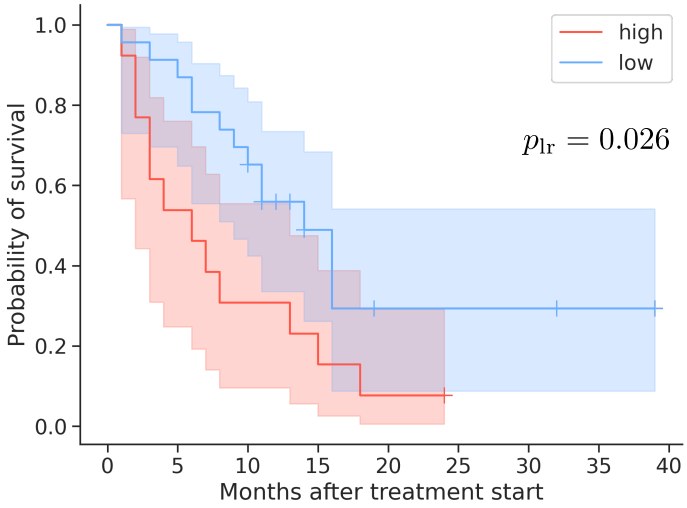


Fig. A.2.: Risk group stratification of the patients in the Dijon cohort using the 65$^{\text{th}}$ percentile of the risk scores instead of the 50$^{\text{th}}$ (i.e., the median).

Bibliography

[Amgad, 2022] Mohamed Amgad, Lamees A Atteya, Hagar Hussein, Kareem Hosny Mohammed, Ehab Hafiz, Maha A T Elsebaie, Ahmed M Alhusseiny, Mohamed Atef AlMoslemany, Abdelmagid M Elmatboly, Philip A Pappalardo, Rokia Adel Sakr, Pooya Mobadersany, Ahmad Rachid, Anas M Saad, Ahmad M Alkashash, Inas A Ruhban, Anas Alrefai, Nada M Elgazar, Ali Abdulkarim, Abo-Alela Farag, Amira Etman, Ahmed G Elsaeed, Yahya Alagha, Yomna A Amer, Ahmed M Raslan, Menatalla K Nadim, Mai A T Elsebaie, Ahmed Ayad, Liza E Hanna, Ahmed Gadallah, Mohamed Elkady, Bradley Drumheller, David Jaye, David Manthey, David A Gutman, Habiba Elfandy, and Lee A D Cooper. "NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer". In: *GigaScience* 11 (2022). giac037 (cit. on p. 52).

[Arvaniti, 2018] Eirini Arvaniti and Manfred Claassen. "Coupling weak and strong supervision for classification of prostate cancer histopathology images". In: *arXiv preprint arXiv:1811.07013* (2018) (cit. on p. 27).

[Bejnordi, 2017] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer". In: *Jama* 318.22 (2017), pp. 2199–2210 (cit. on pp. 12, 13, 17, 25, 34, 52, 87).

[Berghmans, 2020] Thierry Berghmans, Valérie Durieux, Lizza E. L. Hendriks, and Anne-Marie Dingemans. "Immunotherapy: From Advanced NSCLC to Early Stages, an Evolving Concept". In: *Frontiers in Medicine* 7 (2020) (cit. on pp. 3, 64).

[Berrada, 2018] Leonard Berrada, Andrew Zisserman, and M Pawan Kumar. "Smooth Loss Functions for Deep Top-k Classification". In: *International Conference on Learning Representations* (2018) (cit. on p. 13).

[Breiman, 2001] Leo Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32 (cit. on p. 59).

[Bulten, 2020] Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, and Geert Litjens. "Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study". In: *The Lancet Oncology* 21.2 (2020), pp. 233–241 (cit. on pp. 24, 27).

[Bulten, 2022] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, et al. "Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge". In: *Nature Medicine* 28.1 (2022), pp. 154–163 (cit. on p. 52).

[Campanella, 2019] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images". In: *Nature medicine* 25.8 (2019), pp. 1301–1309 (cit. on pp. 5, 12, 25, 59).

[Caron, 2018] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. "Deep Clustering for Unsupervised Learning of Visual Features". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018 (cit. on p. 90).

[Chen, 2020a] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. "A Simple Framework for Contrastive Learning of Visual Representations". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 1597–1607 (cit. on pp. 27, 67).

[Chen, 2020b] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. "Improved baselines with momentum contrastive learning". In: *arXiv preprint arXiv:2003.04297* (2020) (cit. on p. 25).

[Chen, 2022] Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. "Scaling Vision Transformers to Gigapixel Images via Hierarchical Self-Supervised Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 16144–16155 (cit. on p. 5).

[Chopra, 2005] S. Chopra, R. Hadsell, and Y. LeCun. "Learning a similarity metric discriminatively, with application to face verification". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. 2005, 539–546 vol. 1 (cit. on p. 25).

[Ciga, 2021] Ozan Ciga and Anne L. Martel. "Learning to segment images with classification labels". In: *Medical Image Analysis* 68 (2021), p. 101912 (cit. on pp. 13, 26).

[Ciga, 2022] Ozan Ciga, Tony Xu, and Anne Louise Martel. “Self supervised contrastive learning for digital histopathology”. In: *Machine Learning with Applications* 7 (2022), p. 100198 (cit. on pp. 67, 68).

[Ciompi, 2017] Francesco Ciompi, Oscar Geessink, Babak Ehteshami Bejnordi, Gabriel Silva de Souza, Alexi Baidoshvili, Geert Litjens, Bram van Ginneken, Iris Nagtegaal, and Jeroen van der Laak. “The importance of stain normalization in colorectal tissue classification with convolutional networks”. In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. 2017, pp. 160–163 (cit. on p. 87).

[Cooper, 2017] Wendy A. Cooper, Prudence A. Russell, Maya Cherian, Edwina E. Duhig, David Godbolt, Peter J. Jessup, Christine Khoo, Connull Leslie, Annabelle Mahar, David F. Moffat, Vanathi Sivasubramaniam, Celine Faure, Alena Reznichenko, Amanda Grattan, and Stephen B. Fox. “Intra- and Interobserver Reproducibility Assessment of PD-L1 Biomarker in Non–Small Cell Lung Cancer”. In: *Clinical Cancer Research* 23.16 (Aug. 2017), pp. 4569–4577 (cit. on pp. 4, 65).

[Coudray, 2018] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. “Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning”. In: *Nature medicine* 24.10 (2018), pp. 1559–1567 (cit. on pp. 5, 25, 59, 61).

[Courtiol, 2018] Pierre Courtiol, Eric W Tramel, Marc Sanselme, and Gilles Wainrib. “Classification and disease localization in histopathology using only global labels: A weakly-supervised approach”. In: *arXiv preprint arXiv:1802.02212* (2018) (cit. on pp. 5, 12, 25, 59).

[Cox, 1972] D. R. Cox. “Regression Models and Life-Tables”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202 (cit. on pp. 66, 70).

[Da, 2022] Qian Da, Xiaodi Huang, Zhongyu Li, Yanfei Zuo, Chenbin Zhang, Jingxin Liu, Wen Chen, Jiahui Li, Dou Xu, Zhiqiang Hu, et al. “DigestPath: a Benchmark Dataset with Challenge Review for the Pathological Detection and Segmentation of Digestive-System”. In: *Medical Image Analysis* (2022), p. 102485 (cit. on p. 47).

[Davidson-Pilon, 2019] Cameron Davidson-Pilon. “lifelines: survival analysis in Python”. In: *Journal of Open Source Software* 4.40 (2019), p. 1317 (cit. on p. 74).

[Dehaene, 2020] Olivier Dehaene, Axel Camara, Olivier Moindrot, Axel de Lavergne, and Pierre Courtiol. “Self-supervision closes the gap between weak and strong supervision in histology”. In: *arXiv preprint arXiv:2012.03583* (2020) (cit. on pp. 5, 13, 25, 27, 59).

[Dempster, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum Likelihood from Incomplete Data Via the EM Algorithm”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22 (cit. on p. 69).

[Deng, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255 (cit. on p. 59).

[Dietterich, 1997] Thomas G. Dietterich, Richard H. Lathrop, and Tomás Lozano-Pérez. “Solving the multiple instance problem with axis-parallel rectangles”. In: *Artificial Intelligence* 89.1 (1997), pp. 31–71 (cit. on pp. 5, 24).

[Ding, 2005] Chris Ding and Hanchuan Peng. “Minimum redundancy feature selection from microarray gene expression data”. en. In: *J. Bioinform. Comput. Biol.* 3.2 (Apr. 2005), pp. 185–205 (cit. on p. 70).

[Durand, 2016] Thibaut Durand, Nicolas Thome, and Matthieu Cord. “WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016 (cit. on p. 25).

[Eisenhauer, 2009] E.A. Eisenhauer, P. Therasse, J. Bogaerts, L.H. Schwartz, D. Sargent, R. Ford, J. Dancey, S. Arbuck, S. Gwyther, M. Mooney, L. Rubinstein, L. Shankar, L. Dodd, R. Kaplan, D. Lacombe, and J. Verweij. “New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1)”. In: *European Journal of Cancer* 45.2 (2009). Response assessment in solid tumours (RECIST): Version 1.1 and supporting papers, pp. 228–247 (cit. on pp. 3, 54, 55).

[Ester, 1996] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. “A density-based algorithm for discovering clusters in large spatial databases with noise.” In: *kdd*. Vol. 96. 34. 1996, pp. 226–231 (cit. on p. 68).

[Forde, 2022] Patrick M. Forde, Jonathan Spicer, Shun Lu, Mariano Provencio, Tetsuya Mitsudomi, Mark M. Awad, Enriqueta Felip, Stephen R. Broderick, Julie R. Brahmer, Scott J. Swanson, Keith Kerr, Changli Wang, Tudor-Eliade Ciuleanu, Gene B. Saylors, Fumihiro Tanaka, Hiroyuki Ito, Ke-Neng Chen, Moishe Liberman, Everett E. Vokes, Janis M. Taube, Cecile Dorange, Junliang Cai, Joseph Fiore, Anthony Jarkowski, David Balli, Mark Sausen, Dimple Pandya, Christophe Y. Calvet, and Nicolas Girard. “Neoadjuvant Nivolumab plus Chemotherapy in Resectable Lung Cancer”. In: *New England Journal of Medicine* 386.21 (2022). PMID: 35403841, pp. 1973–1985 (cit. on p. 3).

[Fu, 2020] Yu Fu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. "Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis". In: *Nature Cancer* 1.8 (2020), pp. 800–810 (cit. on p. 24).

[Garon, 2015] Edward B. Garon, Naiyer A. Rizvi, Rina Hui, Natasha Leighl, Ani S. Balmanoukian, Joseph Paul Eder, Amita Patnaik, Charu Aggarwal, Matthew Gubens, Leora Horn, Enric Carcereny, Myung-Ju Ahn, Enriqueta Felip, Jong-Seok Lee, Matthew D. Hellmann, Omid Hamid, Jonathan W. Goldman, Jean-Charles Soria, Marisa Dolled-Filhart, Ruth Z. Rutledge, Jin Zhang, Jared K. Lunceford, Reshma Rangwala, Gregory M. Lubiniecki, Charlotte Roach, Kenneth Emancipator, and Leena Gandhi. "Pembrolizumab for the Treatment of Non–Small-Cell Lung Cancer". In: *New England Journal of Medicine* 372.21 (2015). PMID: 25891174, pp. 2018–2028 (cit. on p. 3).

[Golestaneh, 2017] S Alireza Golestaneh and Lina J Karam. "Spatially-Varying Blur Detection Based on Multiscale Fused and Sorted Transform Coefficients of Gradient Magnitudes." In: *CVPR*. 2017, pp. 596–605 (cit. on p. 36).

[Grigg, 2016] Claud Grigg and Naiyer A. Rizvi. "PD-L1 biomarker testing for non-small cell lung cancer: truth or fiction?" In: *Journal for ImmunoTherapy of Cancer* 4.1 (2016) (cit. on pp. 3, 65).

[Habis, 2022] Antoine Habis, Vannary Meas-Yedid, Daniel Felipe González Obando, Jean-Christophe Olivo-Marin, and Elsa D. Angelini. "Smart Learning of Click and Refine for Nuclei Segmentation on Histology Images". In: *2022 IEEE International Conference on Image Processing (ICIP)*. 2022, pp. 2281–2285 (cit. on p. 6).

[Harder, 2019] Nathalie Harder, Ralf Schönmeyer, Katharina Nekolla, Armin Meier, Nicolas Brieu, Carolina Vanegas, Gabriele Madonna, Mariaelena Capone, Gerardo Botti, Paolo A. Ascierto, and Günter Schmidt. "Automatic discovery of image-based signatures for ipilimumab response prediction in malignant melanoma". In: *Scientific Reports* 9.1 (May 2019), p. 7449 (cit. on pp. 6, 65).

[He, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016 (cit. on pp. 13, 25, 26, 28, 58, 73).

[He, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738 (cit. on p. 13).

[Heeke, 2020] Simon Heeke, Jonathan Benzaquen, Véronique Hofman, Elodie Long-Mira, Virginie Lespinet, Olivier Bordone, Charles-Hugo Marquette, Hervé Delingette, Marius Ilié, and Paul Hofman. "Comparison of Three Sequencing Panels Used for the Assessment of Tumor Mutational Burden in NSCLC Reveals Low Comparability". In: *Journal of Thoracic Oncology* 15.9 (2020), pp. 1535–1540 (cit. on p. 4).

[Hellmann, 2018] Matthew D. Hellmann, Tudor-Eliade Ciuleanu, Adam Pluzanski, Jong Seok Lee, Gregory A. Otterson, Clarisse Audigier-Valette, Elisa Minenza, Helena Linardou, Sjaak Burgers, Pamela Salman, Hossein Borghaei, Suresh S. Ramalingam, Julie Brahmer, Martin Reck, Kenneth J. O'Byrne, William J. Geese, George Green, Han Chang, Joseph Szustakowski, Prabhu Bhagavatheeswaran, Diane Healey, Yali Fu, Faith Nathan, and Luis Paz-Ares. "Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden". In: *New England Journal of Medicine* 378.22 (2018). PMID: 29658845, pp. 2093–2104 (cit. on pp. 4, 64, 65).

[Horn, 2017] Leora Horn, David R. Spigel, Everett E. Vokes, Esther Holgado, Neal Ready, Martin Steins, Elena Poddubskaya, Hossein Borghaei, Enriqueta Felip, Luis Paz-Ares, Adam Pluzanski, Karen L. Reckamp, Marco A. Burgio, Martin Kohlhäeufl, David Waterhouse, Fabrice Barlesi, Scott Antonia, Oscar Arrieta, Jérôme Fayette, Lucio Crinò, Naiyer Rizvi, Martin Reck, Matthew D. Hellmann, William J. Geese, Ang Li, Anne Blackwood-Chirchir, Diane Healey, Julie Brahmer, and Wilfried E.E. Eberhardt. "Nivolumab Versus Docetaxel in Previously Treated Patients With Advanced Non–Small-Cell Lung Cancer: Two-Year Outcomes From Two Randomized, Open-Label, Phase III Trials (CheckMate 017 and CheckMate 057)". In: *Journal of Clinical Oncology* 35.35 (2017). PMID: 29023213, pp. 3924–3933 (cit. on pp. 3, 64).

[Huang, 2020] Jiabo Huang, Shaogang Gong, and Xiatian Zhu. "Deep semantic clustering by partition confidence maximisation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8849–8858 (cit. on p. 90).

[Ilie, 2017] Marius Ilie and Paul Hofman. "Reproducibility of PD-L1 assessment in non-small cell lung cancer—know your limits but never stop trying to exceed them". In: *Translational Lung Cancer Research* 6.Suppl 1 (2017) (cit. on pp. 4, 65).

[Ilié, 2022] Marius Ilié, Jonathan Benzaquen, Paul Tourniaire, Simon Heeke, Nicholas Ayache, Hervé Delingette, Elodie Long-Mira, Sandra Lassalle, Marame Hamila, Julien Fayada, et al. "Deep learning facilitates distinguishing histologic subtypes of pulmonary neuroendocrine tumors on digital whole-slide images". In: *Cancers* 14.7 (2022), p. 1740 (cit. on p. 9).

[Ilse, 2018] Maximilian Ilse, Jakub Tomczak, and Max Welling. "Attention-based deep multiple instance learning". In: *International conference on machine learning*. PMLR. 2018, pp. 2127–2136 (cit. on pp. 13, 26, 59, 89).

[Jain, 2020] Mika S. Jain and Tarik F. Massoud. “Predicting tumour mutational burden from histopathological images using multiscale deep learning”. In: *Nature Machine Intelligence* 2.6 (June 2020), pp. 356–362 (cit. on pp. 6, 58, 59, 61, 65).

[Janowczyk, 2019] Andrew Janowczyk, Ren Zuo, Hannah Gilmore, Michael Feldman, and Anant Madabhushi. “HistoQC: An Open-Source Quality Control Tool for Digital Pathology Slides”. In: *JCO Clinical Cancer Informatics* 3 (2019). PMID: 30990737, pp. 1–7 (cit. on p. 73).

[Johannet, 2021] Paul Johannet, Nicolas Coudray, Douglas M. Donnelly, George Jour, Irineu Illa-Bochaca, Yuhe Xia, Douglas B. Johnson, Lee Wheless, James R. Patrinely, Sofia Nomikou, David L. Rimm, Anna C. Pavlick, Jeffrey S. Weber, Judy Zhong, Aristotelis Tsirigos, and Iman Osman. “Using Machine Learning Algorithms to Predict Immunotherapy Response in Patients with Advanced Melanoma”. In: *Clinical Cancer Research* 27.1 (Jan. 2021), pp. 131–140 (cit. on pp. 6, 56, 65).

[Kaplan, 1958] E. L. Kaplan and Paul Meier. “Nonparametric Estimation from Incomplete Observations”. In: *Journal of the American Statistical Association* 53.282 (1958), pp. 457–481 (cit. on p. 74).

[Kather, 2020] Jakob Nikolas Kather, Lara R. Heij, Heike I. Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M. Niehues, Kai A. J. Sommer, Peter Bankhead, Loes F. S. Kooreman, Jefree J. Schulte, Nicole A. Cipriani, Roman D. Buelow, Peter Boor, Nadina Ortiz-Brüchle, Andrew M. Hanby, Valerie Speirs, Sara Kochanny, Akash Patnaik, Andrew Srisuwananukorn, Hermann Brenner, Michael Hoffmeister, Piet A. van den Brandt, Dirk Jäger, Christian Trautwein, Alexander T. Pearson, and Tom Luedde. “Pan-cancer image-based detection of clinically actionable genetic alterations”. In: *Nature Cancer* 1.8 (Aug. 2020), pp. 789–799 (cit. on p. 5).

[Khan, 2014] Adnan Mujahid Khan, Nasir Rajpoot, Darren Treanor, and Derek Magee. “A Nonlinear Mapping Approach to Stain Normalization in Digital Histopathology Images Using Image-Specific Color Deconvolution”. In: *IEEE Transactions on Biomedical Engineering* 61.6 (2014), pp. 1729–1738 (cit. on p. 87).

[Kingma, 2014] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on pp. 16, 36).

[Klein, 2021] Oliver Klein, Damien Kee, Ben Markman, Matteo S. Carlino, Craig Underhill, Jodie Palmer, Daniel Power, Jonathan Cebon, and Andreas Behren. “Evaluation of TMB as a predictive biomarker in patients with solid cancers treated with anti-PD-1/CTLA-4 combination immunotherapy”. In: *Cancer Cell* 39.5 (2021), pp. 592–593 (cit. on pp. 4, 65).

[Kursa, 2010] Miron B. Kursa and Witold R. Rudnicki. “Feature Selection with the Boruta Package”. In: *Journal of Statistical Software* 36.11 (2010), pp. 1–13 (cit. on p. 71).

[Le Bescond, 2022] Loïc Le Bescond, Marvin Lerousseau, Ingrid Garberis, Fabrice André, Stergios Christodoulidis, Maria Vakalopoulou, and Hugues Talbot. “Unsupervised Nuclei Segmentation Using Spatial Organization Priors”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Ed. by Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li. Cham: Springer Nature Switzerland, 2022, pp. 325–335 (cit. on p. 6).

[Lerousseau, 2021] Marvin Lerousseau, Marion Classe, Enzo Battistella, Théo Estienne, Théophraste Henry, Amaury Leroy, Roger Sun, Maria Vakalopoulou, Jean-Yves Scoazec, Eric Deutsch, et al. “Weakly supervised pan-cancer segmentation tool”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2021, pp. 248–256 (cit. on p. 25).

[Li, 2018] Ruoyu Li, Jiawen Yao, Xinliang Zhu, Yeqing Li, and Junzhou Huang. “Graph CNN for Survival Analysis on Whole Slide Pathological Images”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Ed. by Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger. Cham: Springer International Publishing, 2018, pp. 174–182 (cit. on p. 6).

[Li, 2019] Jiahui Li, Shuang Yang, Xiaodi Huang, Qian Da, Xiaoqun Yang, Zhiqiang Hu, Qi Duan, Chaofu Wang, and Hongsheng Li. “Signet ring cell detection with a semi-supervised learning framework”. In: *International Conference on Information Processing in Medical Imaging*. Springer. 2019, pp. 842–854 (cit. on p. 35).

[Li, 2021a] Bin Li, Yin Li, and Kevin W. Eliceiri. “Dual-Stream Multiple Instance Learning Network for Whole Slide Image Classification With Self-Supervised Contrastive Learning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 14318–14328 (cit. on pp. 26, 37, 86, 89).

[Li, 2021b] Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. “Contrastive Clustering”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.10 (May 2021), pp. 8547–8555 (cit. on p. 90).

[Li, 2023] Kailu Li, Ziniu Qian, Yingnan Han, Eric I-Chao Chang, Bingzheng Wei, Maode Lai, Jing Liao, Yubo Fan, and Yan Xu. “Weakly supervised histopathology image segmentation with self-attention”. In: *Medical Image Analysis* 86 (2023), p. 102791 (cit. on p. 90).

[Liu, 2012] Guoqing Liu, Jianxin Wu, and Zhi-Hua Zhou. “Key Instance Detection in Multi-Instance Learning”. In: *Proceedings of the Asian Conference on Machine Learning*. Ed. by Steven C. H. Hoi and Wray Buntine. Vol. 25. Proceedings of Machine Learning Research. Singapore Management University, Singapore: PMLR, Nov. 2012, pp. 253–268 (cit. on p. 25).

[Lu, 2019] Ming Y Lu, Richard J Chen, Jingwen Wang, Debora Dillon, and Faisal Mahmood. “Semi-supervised histology classification using deep multiple instance learning and contrastive predictive coding”. In: *arXiv preprint arXiv:1910.10825* (2019) (cit. on pp. 32, 59, 90).

[Lu, 2021] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. “Data-efficient and weakly supervised computational pathology on whole-slide images”. In: *Nature Biomedical Engineering* 5.6 (2021), pp. 555–570 (cit. on pp. 5, 7, 13, 14, 17, 26–28, 36, 49, 59, 61, 89).

[Lubrano di Scandalea, 2022] Melanie Lubrano di Scandalea, Tristan Lazard, Guillaume Balezo, Yaëlle Bellahsen-Harrar, Cécile Badoual, Sylvain Berlemont, and Thomas Walter. “Automatic grading of cervical biopsies by combining full and self-supervision”. working paper or preprint. Jan. 2022 (cit. on p. 27).

[Macenko, 2009] Marc Macenko, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. “A method for normalizing histology slides for quantitative analysis”. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. 2009, pp. 1107–1110 (cit. on pp. 5, 83).

[Malhotra, 2017] Jyoti Malhotra, Salma K Jabbour, and Joseph Aisner. “Current state of immunotherapy for non-small cell lung cancer”. In: *Translational lung cancer research* 6.2 (Apr. 2017), pp. 196–211 (cit. on p. 3).

[Mantel, 1966] N Mantel. “Evaluation of survival data and two new rank order statistics arising in its consideration”. en. In: *Cancer Chemother. Rep.* 50.3 (Mar. 1966), pp. 163–170 (cit. on p. 71).

[Marabelle, 2020] Aurélien Marabelle, Marwan Fakih, Juanita Lopez, Manisha Shah, Ronnie Shapira-Frommer, Kazuhiko Nakagawa, Hyun Cheol Chung, Hedy L. Kindler, Jose A. Lopez-Martin, Wilson H Miller Jr., Antoine Italiano, Steven Kao, Sarina A. Piha-Paul, Jean-Pierre Delord, Robert R. McWilliams, David A. Fabrizio, Deepti Aurora-Garg, Lei Xu, Fan Jin, Kevin Norwood, and Yung-Jue Bang. “Association of tumour mutational burden with outcomes in patients with advanced solid tumours treated with pembrolizumab: prospective biomarker analysis of the multicohort, open-label, phase 2 KEYNOTE-158 study”. In: *The Lancet Oncology* 21.10 (Oct. 2020), pp. 1353–1365 (cit. on pp. 4, 65).

[Marini, 2021] Niccolò Marini, Sebastian Otálora, Henning Müller, and Manfredo Atzori. “Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification”. In: *Medical Image Analysis* 73 (2021), p. 102165 (cit. on p. 27).

[Martins, 2019] Filipe Martins, Latifyan Sofiya, Gerasimos P Sykiotis, Faiza Lamine, Michel Maillard, Montserrat Fraga, Keyvan Shabafrouz, Camillo Ribi, Anne Cairoli, Yan Guex-Crosier, et al. “Adverse effects of immune-checkpoint inhibitors: epidemiology, management and surveillance”. In: *Nature reviews Clinical oncology* 16.9 (2019), pp. 563–580 (cit. on pp. 3, 64).

[Mazieres, 2019] J. Mazieres, A. Drilon, A. Lusque, L. Mhanna, A.B. Cortot, L. Mezquita, A.A. Thai, C. Mascaux, S. Couraud, R. Veillon, M. Van den Heuvel, J. Neal, N. Peled, M. Früh, T.L. Ng, V. Gounant, S. Popat, J. Diebold, J. Sabari, V.W. Zhu, S.I. Rothschild, P. Bironzo, A. Martinez-Marti, A. Curioni-Fontecedro, R. Rosell, M. Lattuca-Truc, M. Wiesweg, B. Besse, B. Solomon, F. Barlesi, R.D. Schouten, H. Wakelee, D.R. Camidge, G. Zalcman, S. Novello, S.I. Ou, J. Milia, and O. Gautschi. “Immune checkpoint inhibitors for patients with advanced lung cancer and oncogenic driver alterations: results from the IMMUNOTARGET registry”. In: *Annals of Oncology* 30.8 (2019). Triple-negative breast cancer - clinical results and biomarker analysis of GeparNuevo study, pp. 1321–1328 (cit. on pp. 3, 64).

[Mlynarski, 2019] Pawel Mlynarski, Hervé Delingette, Antonio Criminisi, and Nicholas Ayache. “Deep learning with mixed supervision for brain tumor segmentation”. In: *Journal of Medical Imaging* 6.3 (2019), p. 034002 (cit. on pp. 13, 26).

[Mok, 2019] Tony S. K. Mok, Yi-Long Wu, Iveta Kudaba, et al. “Pembrolizumab versus chemotherapy for previously untreated, PD-L1-expressing, locally advanced or metastatic non-small-cell lung cancer (KEYNOTE-042): a randomised, open-label, controlled, phase 3 trial”. In: *The Lancet* 393.10183 (May 2019), pp. 1819–1830 (cit. on pp. 3, 65).

[Oord, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. “Representation learning with contrastive predictive coding”. In: *arXiv preprint arXiv:1807.03748* (2018) (cit. on p. 90).

[Otsu, 1979] Nobuyuki Otsu. “A threshold selection method from gray-level histograms”. In: *IEEE transactions on systems, man, and cybernetics* 9.1 (1979), pp. 62–66 (cit. on p. 14).

[Park, 2022] Sehhoon Park, Chan-Young Ock, Hyojin Kim, Sergio Pereira, Seonwook Park, Minuk Ma, Sangjoon Choi, Seokhwi Kim, Seunghwan Shin, Brian Jaehong Aum, Kyunghyun Paeng, Donggeun Yoo, Hongui Cha, Sunyoung Park, Koung Jin Suh, Hyun Ae Jung, Se Hyun Kim, Yu Jung Kim, Jong-Mu Sun, Jin-Haeng Chung, Jin Seok Ahn, Myung-Ju Ahn, Jong Seok Lee, Keunchil Park, Sang Yong Song, Yung-Jue Bang, Yoon-La Choi, Tony S. Mok, and Se-Hoon Lee. "Artificial Intelligence–Powered Spatial Analysis of Tumor-Infiltrating Lymphocytes as Complementary Biomarker for Immune Checkpoint Inhibition in Non–Small-Cell Lung Cancer". In: *Journal of Clinical Oncology* 40.17 (2022). PMID: 35271299, pp. 1916–1928 (cit. on pp. 6, 66).

[Paszke, 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035 (cit. on p. 74).

[Paz-Ares, 2021] Luis Paz-Ares, Tudor-Eliade Ciuleanu, Manuel Cobo, Michael Schenker, Bogdan Zurawski, Juliana Menezes, Eduardo Richardet, Jaafar Bennouna, Enriqueta Felip, Oscar Juan-Vidal, Aurelia Alexandru, Hiroshi Sakai, Alejo Lingua, Pamela Salman, Pierre-Jean Souquet, Pedro De Marchi, Claudio Martin, Maurice Pérol, Arnaud Scherpereel, Shun Lu, Thomas John, David P. Carbone, Stephanie Meadows-Shropshire, Shruti Agrawal, Abderrahim Oukessou, Jinchun Yan, and Martin Reck. "First-line nivolumab plus ipilimumab combined with two cycles of chemotherapy in patients with non-small-cell lung cancer (CheckMate 9LA): an international, randomised, open-label, phase 3 trial". In: *The Lancet Oncology* 22.2 (Feb. 2021), pp. 198–211 (cit. on p. 3).

[Reck, 2019] Martin Reck, Delvys Rodríguez–Abreu, Andrew G. Robinson, Rina Hui, Tibor Csőszi, Andrea Fülöp, Maya Gottfried, Nir Peled, Ali Tafreshi, Sinead Cuffe, Mary O´Brien, Suman Rao, Katsuyuki Hotta, Kristel Vandormael, Antonio Riccio, Jing Yang, M. Catherine Pietanza, and Julie R. Brahmer. "Updated Analysis of KEYNOTE-024: Pembrolizumab Versus Platinum-Based Chemotherapy for Advanced Non–Small-Cell Lung Cancer With PD-L1 Tumor Proportion Score of 50% or Greater". In: *Journal of Clinical Oncology* 37.7 (2019). PMID: 30620668, pp. 537–546 (cit. on pp. 3, 64, 65, 77).

[Ronen, 2022] Meitar Ronen, Shahaf E. Finder, and Oren Freifeld. "DeepDPM: Deep Clustering With an Unknown Number of Clusters". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 9861–9870 (cit. on p. 68).

[Rymarczyk, 2021] Dawid Rymarczyk, Adriana Borowa, Jacek Tabor, and Bartosz Zielinski. “Kernel Self-Attention for Weakly-Supervised Image Classification Using Deep Multiple Instance Learning”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Jan. 2021, pp. 1721–1730 (cit. on pp. 26, 89).

[Saillard, 2021] Charlie Saillard, Olivier Dehaene, Tanguy Marchand, Olivier Moindrot, Aurélie Kamoun, Benoit Schmauch, and Simon Jegou. “Self-supervised learning improves dMMR/MSI detection from histology slides across multiple cancers”. In: *Proceedings of the MICCAI Workshop on Computational Pathology*. Ed. by Manfredo Atzori, Nikolay Burlutskiy, Francesco Ciompi, Zhang Li, Fayyaz Minhas, Henning Müller, Tingying Peng, Nasir Rajpoot, Ben Torben-Nielsen, Jeroen van der Laak, Mitko Veta, Yinyin Yuan, and Inti Zlobec. Vol. 156. Proceedings of Machine Learning Research. PMLR, 2021, pp. 191–205 (cit. on p. 5).

[Saltz, 2018] Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R. Shroyer, Tianhao Zhao, Rebecca Batiste, John Van Arnam, The Cancer Genome Atlas Research Network, Ilya Shmulevich, Arvind U.K. Rao, Alexander J. Lazar, Ashish Sharma, and Vésteinn Thorsson. “Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images”. In: *Cell Reports* 23.1 (2018), 181–193.e7 (cit. on p. 77).

[Schmauch, 2020] Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, Thomas Clozel, Matahi Moarii, Pierre Courtiol, and Gilles Wainrib. “A deep learning model to predict RNA-Seq expression of tumours from whole slide images”. In: *Nature Communications* 11.1 (Aug. 2020), p. 3877 (cit. on p. 5).

[Schmidt, 2022] Arne Schmidt, Julio Silva-Rodríguez, Rafael Molina, and Valery Naranjo. “Efficient Cancer Classification by Coupling Semi Supervised and Multiple Instance Learning”. In: *IEEE Access* 10 (2022), pp. 9763–9773 (cit. on p. 27).

[Sha, 2019] Lingdao Sha, Boleslaw L. Osinski, Irvin Y. Ho, Timothy L. Tan, Caleb Willis, Hannah Weiss, Nike Beaubier, Brett M. Mahon, Tim J. Taxter, and Stephen S. F Yip. “Multi-Field-of-View Deep Learning Model Predicts Nonsmall Cell Lung Cancer Programmed Death-Ligand 1 Status from Whole-Slide Hematoxylin and Eosin Images”. In: *Journal of Pathology Informatics* 10.1 (2019), p. 24 (cit. on pp. 6, 59, 65).

[Shah, 2018] Meet P. Shah, S. N. Merchant, and Suyash P. Awate. "MS-Net: Mixed-Supervision Fully-Convolutional Networks for Full-Resolution Segmentation". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*. Ed. by Alejandro F. Frangi, Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-López, and Gabor Fichtinger. Cham: Springer International Publishing, 2018, pp. 379–387 (cit. on p. 26).

[Shannon, 1948] Claude Elwood Shannon. "A mathematical theory of communication". In: *The Bell system technical journal* 27.3 (1948), pp. 379–423 (cit. on p. 32).

[Shao, 2021a] Wei Shao, Tongxin Wang, Zhi Huang, Zhi Han, Jie Zhang, and Kun Huang. "Weakly Supervised Deep Ordinal Cox Model for Survival Prediction From Whole-Slide Pathological Images". In: *IEEE Transactions on Medical Imaging* 40.12 (2021), pp. 3739–3747 (cit. on p. 6).

[Shao, 2021b] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and yongbing zhang yongbing. "TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 2136–2147 (cit. on p. 37).

[Sharma, 2021] Yash Sharma, Aman Shrivastava, Lubaina Ehsan, Christopher A Moskaluk, Sana Syed, and Donald Brown. "Cluster-to-conquer: A framework for end-to-end multi-instance learning for whole slide image classification". In: *Medical Imaging with Deep Learning*. PMLR. 2021, pp. 682–698 (cit. on p. 32).

[Shi, 2020] Xiaoshuang Shi, Fuyong Xing, Yuanpu Xie, Zizhao Zhang, Lei Cui, and Lin Yang. "Loss-based attention for deep multiple instance learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 04. 2020, pp. 5742–5749 (cit. on pp. 26, 89).

[Sirinukunwattana, 2016] Korsuk Sirinukunwattana, Shan E Ahmed Raza, Yee-Wah Tsang, David RJ Snead, Ian A Cree, and Nasir M Rajpoot. "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images". In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1196–1206 (cit. on p. 13).

[Srinidhi, 2021] Chetan L. Srinidhi, Ozan Ciga, and Anne L. Martel. "Deep neural network models for computational histopathology: A survey". In: *Medical Image Analysis* 67 (2021), p. 101813 (cit. on p. 5).

[Stenzinger, 2019] Albrecht Stenzinger, Jeffrey D. Allen, Jörg Maas, Mark D. Stewart, Diana M. Merino, Madison M. Wempe, and Manfred Dietel. "Tumor mutational burden standardization initiatives: Recommendations for consistent tumor mutational burden assessment in clinical samples to guide immunotherapy treatment decisions". In: *Genes, Chromosomes and Cancer* 58.8 (2019), pp. 578–588 (cit. on p. 4).

[Sumi, 2022] Toshiyuki Sumi, Haruhiko Michimata, Daiki Nagayama, Yuta Koshino, Hiroki Watanabe, Yuichi Yamada, and Hirofumi Chiba. "Tumor-Associated Raynaud’s Phenomenon Exacerbated by Administration of Immune Checkpoint Inhibitors". In: *Journal of Thoracic Oncology* 17.10 (Oct. 2022), pp. 1233–1234 (cit. on p. 3).

[Sung, 2021] Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Mathieu Laversanne, Isabelle Soerjomataram, Ahmedin Jemal, and Freddie Bray. "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries". In: *CA: A Cancer Journal for Clinicians* 71.3 (2021), pp. 209–249 (cit. on pp. 1, 64).

[Szegedy, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the Inception Architecture for Computer Vision". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016 (cit. on pp. 59, 65).

[Tellez, 2019] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology". In: *Medical Image Analysis* 58 (2019), p. 101544 (cit. on p. 82).

[Thorsson, 2018] Vésteinn Thorsson, David L. Gibbs, Scott D. Brown, et al. "The Immune Landscape of Cancer". In: *Immunity* 48.4 (2018), 812–830.e14 (cit. on p. 77).

[Tourniaire, 2021] Paul Tourniaire, Marius Ilie, Paul Hofman, Nicholas Ayache, and Herve Delingette. "Attention-based Multiple Instance Learning with Mixed Supervision on the Camelyon16 Dataset". In: *Proceedings of the MICCAI Workshop on Computational Pathology*. Ed. by Manfredo Atzori, Nikolay Burlutskiy, Francesco Ciompi, Zhang Li, Fayyaz Minhas, Henning Müller, Tingying Peng, Nasir Rajpoot, Ben Torben-Nielsen, Jeroen van der Laak, Mitko Veta, Yinyin Yuan, and Inti Zlobec. Vol. 156. Proceedings of Machine Learning Research. PMLR, Sept. 2021, pp. 216–226 (cit. on pp. 7, 9, 12).

[Tourniaire, 2022] Paul Tourniaire, Marius Ilie, Paul Hofman, Nicholas Ayache, and Hervé Delingette. "Mixed supervision to improve the classification and localization: Coherence of tumors in histological slides". In: *Cancer Research* 82.12_Supplement (2022), pp. 461–461 (cit. on p. 9).

[Tourniaire, 2023a] Paul Tourniaire, Marius Ilie, Paul Hofman, Nicholas Ayache, and Hervé Delingette. "MS-CLAM: Mixed supervision for the classification and localization of tumors in Whole Slide Images". In: *Medical Image Analysis* 85 (2023), p. 102763 (cit. on pp. 7, 9, 24).

[Tourniaire, 2023b] Paul Tourniaire, Marius Ilie, Julien Mazi'eres, Anna Vigier, François Ghiringhelli, Nicolas Piton, Jean-Christophe Sabourin, Frédéric Bibeau, Paul Hofman, Nicholas Ayache, and Hervé Delingette. "WhARIO: Whole-slide image-based survival Analysis for patients tReated with ImmunOtherapy". preprint submitted to a journal. 2023 (cit. on pp. 8, 9, 64).

[Tumeh, 2014] Paul C. Tumeh, Christina L. Harview, Jennifer H. Yearley, I. Peter Shintaku, Emma J. M. Taylor, Lidia Robert, Bartosz Chmielowski, Marko Spasic, Gina Henry, Voicu Ciobanu, Alisha N. West, Manuel Carmona, Christine Kivork, Elizabeth Seja, Grace Cherry, Antonio J. Gutierrez, Tristan R. Grogan, Christine Mateus, Gorana Tomasic, John A. Glaspy, Ryan O. Emerson, Harlan Robins, Robert H. Pierce, David A. Elashoff, Caroline Robert, and Antoni Ribas. "PD-1 blockade induces responses by inhibiting adaptive immune resistance". In: *Nature* 515.7528 (Nov. 2014), pp. 568–571 (cit. on pp. 6, 66, 82, 87).

[Vahadane, 2016] Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. "Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images". In: *IEEE Transactions on Medical Imaging* 35.8 (2016), pp. 1962–1971 (cit. on pp. 5, 83).

[van Rijthoven, 2021] Mart van Rijthoven, Maschenka Balkenhol, Karina Siliņa, Jeroen van der Laak, and Francesco Ciompi. "HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images". In: *Medical Image Analysis* 68 (2021), p. 101890 (cit. on p. 5).

[Vanguri, 2022] Rami S. Vanguri, Jia Luo, Andrew T. Aukerman, Jacklynn V. Egger, Christopher J. Fong, Natally Horvat, Andrew Pagano, Jose de Arimateia Batista Araujo-Filho, Luke Geneslaw, Hira Rizvi, Ramon Sosa, Kevin M. Boehm, Soo-Ryum Yang, Francis M. Bodd, Katia Ventura, Travis J. Hollmann, Michelle S. Ginsberg, Jianjiong Gao, Rami Vanguri, Matthew D. Hellmann, Jennifer L. Sauter, Sohrab P. Shah, and M. S. K. M. I. N. D. Consortium. "Multimodal integration of radiology, pathology and genomics for prediction of response to PD-(L)1 blockade in patients with non-small cell lung cancer". In: *Nature Cancer* 3.10 (Oct. 2022), pp. 1151–1164 (cit. on p. 91).

[Vaswani, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017) (cit. on p. 26).

[Viswanathan, 2022] Vidya Sankar Viswanathan, Paula Toro, Germán Corredor, Sanjay Mukhopadhyay, and Anant Madabhushi. "The state of the art for artificial intelligence in lung digital pathology". In: *The Journal of Pathology* 257.4 (2022), pp. 413–429 (cit. on p. 5).

[Wang, 2016] Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H Beck. “Deep learning for identifying metastatic breast cancer”. In: *arXiv preprint arXiv:1606.05718* (2016) (cit. on pp. 13, 21, 24).

[Wang, 2018] Daniel Y. Wang, Joe-Elie Salem, Justine V. Cohen, Sunandana Chandra, Christian Menzer, Fei Ye, Shilin Zhao, Satya Das, Kathryn E. Beckermann, Lisa Ha, W. Kimryn Rathmell, Kristin K. Ancell, Justin M. Balko, Caitlin Bowman, Elizabeth J. Davis, David D. Chism, Leora Horn, Georgina V. Long, Matteo S. Carlino, Benedicte Lebrun-Vignes, Zeynep Eroglu, Jessica C. Hassel, Alexander M. Menzies, Jeffrey A. Sosman, Ryan J. Sullivan, Javid J. Moslehi, and Douglas B. Johnson. “Fatal Toxic Effects Associated With Immune Checkpoint Inhibitors: A Systematic Review and Meta-analysis”. In: *JAMA Oncology* 4.12 (Dec. 2018), pp. 1721–1728 (cit. on p. 64).

[Wang, 2019] Xi Wang, Hao Chen, Caixia Gan, Huangjing Lin, Qi Dou, Efstratios Tsougenis, Qitao Huang, Muyan Cai, and Pheng-Ann Heng. “Weakly supervised deep learning for whole slide lung cancer image analysis”. In: *IEEE transactions on cybernetics* 50.9 (2019), pp. 3950–3962 (cit. on p. 12).

[Wang, 2022] Xiangxue Wang, Cristian Barrera, Kaustav Bera, Vidya Sankar Viswanathan, Sepideh Azarianpour-Esfahani, Can Koyuncu, Priya Velu, Michael D. Feldman, Michael Yang, Pingfu Fu, Kurt A. Schalper, Haider Mahdi, Cheng Lu, Vamsidhar Velcheti, and Anant Madabhushi. “Spatial interplay patterns of cancer nuclei and tumor-infiltrating lymphocytes (TILs) predict clinical benefit for immune checkpoint inhibitors”. In: *Science Advances* 8.22 (2022), eabn3966 (cit. on pp. 6, 66, 90).

[Yan, 2022] Jiangpeng Yan, Hanbo Chen, Xiu Li, and Jianhua Yao. “Deep contrastive learning based tissue clustering for annotation-free histopathology image analysis”. In: *Computerized Medical Imaging and Graphics* 97 (2022), p. 102053 (cit. on p. 90).

[Yao, 2020] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. “Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks”. In: *Medical Image Analysis* 65 (2020), p. 101789 (cit. on p. 6).

[You, 2017] Yang You, Igor Gitman, and Boris Ginsburg. “Scaling SGD Batch Size to 32K for ImageNet Training”. In: *CoRR* abs/1708.03888 (2017). arXiv: 1708.03888 (cit. on p. 73).

List of Figures

List of Tables