WhARIO: whole-slide-image-based survival analysis for patients treated with immunotherapy

Paul Tourniaire, Marius Ilie, Julien Mazières, Anna Vigier, François Ghiringhelli, Nicolas Piton, Jean-Christophe Sabourin, Frédéric Bibeau, Paul Hofman, Nicholas Ayache, et al.

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

Paul Tourniaire[a,*], **Marius Ilie**[b,c,d], **Julien Mazières**[e], **Anna Vigier**[f], **François Ghiringhelli**[g], **Nicolas Piton**[h], **Jean-Christophe Sabourin**[h], **Frédéric Bibeau**[i], **Paul Hofman**[b,c,d], **Nicholas Ayache**[a,d], **Hervé Delingette**[a,d]

[a]Université Côte d'Azur, Inria, Epione project-team, Sophia Antipolis, France
[b]Laboratory of Clinical and Experimental Pathology, Pasteur Hospital, Université Côte d'Azur Nice, France
[c]Hospital-Related Biobank BB-0033-00025
[d]FHU OncoAge
[e]Department of Pneumology, CHU Toulouse-Hôpital Larrey, Université Paul Sabatier, Toulouse, France
[f]Department of Pathology, IUCT-Oncopole, Toulouse, France
[g]Department of Medical Oncology, Centre Georges-François Leclerc, Dijon, France
[h]Department of Pathology, Rouen University Hospital, France and Normandie University, UNIROUEN, Inserm U124, Rouen, France
[i]Departement of Pathology, Centre Hospitalier Universitaire de Besançon, Besançon, France

Abstract.
Purpose: 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, although less than 20% of the patients respond to the treatment, while some face acute adverse events. Although a few predictive biomarkers have integrated the clinical workflow, they require additional modalities on top of Whole-slide Images (WSIs), and lack efficiency or robustness. In this work, we propose a new biomarker of immunotherapy outcome derived solely from the analysis of histology slides.
Approach: 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 derive features and train a proportional hazards model for survival analysis. We test our approach on an in-house dataset of 193 patients from 5 medical centers, and compare it with the gold standard Tumor Proportion Score (TPS) biomarker.
Results: On a 5-fold cross-validation (CV) of the entire dataset, the WhARIO features are able to separate a low- and a high-risk group of patients with a Hazard Ratio (HR) of 2.29 (CI_{95}=1.48 to 3.56), while the TPS 1% reference threshold only reaches a HR of 1.81 (CI_{95}=1.21 to 2.69). Combining the two yields a higher HR of 2.60 (CI_{95}=1.72 to 3.94). Additional experiments on the same dataset, where 1 out of 5 centers is excluded from the CV and used as a test set confirm these trends.
Conclusions: Our newly designed WhARIO features are 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.

Keywords: digital pathology, deep learning, nonparametric clustering, immunotherapy, lung cancer, survival analysis.

*Paul Tourniaire, paul.tourniaire@inria.fr

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,[1] 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.[2,3] 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.[4] 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.[5,6] 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.[7,8] To select patients eligible to this kind of therapy better, 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).[3,9] Yet, the efficacy of such a biomarker remains limited, with additional concerns regarding the robustness of its assessment and the variability between observers.[10–12] 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.[4,13,14] 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. [15] 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.[16] use three Inceptionv3 networks[17] 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.[18] 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.[19] 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.[20] 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.[21] 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)[22] 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,[23] 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 193 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.

2 Methods

In this section, we describe the various steps needed to leverage H&E slides for survival analysis. Figure 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.

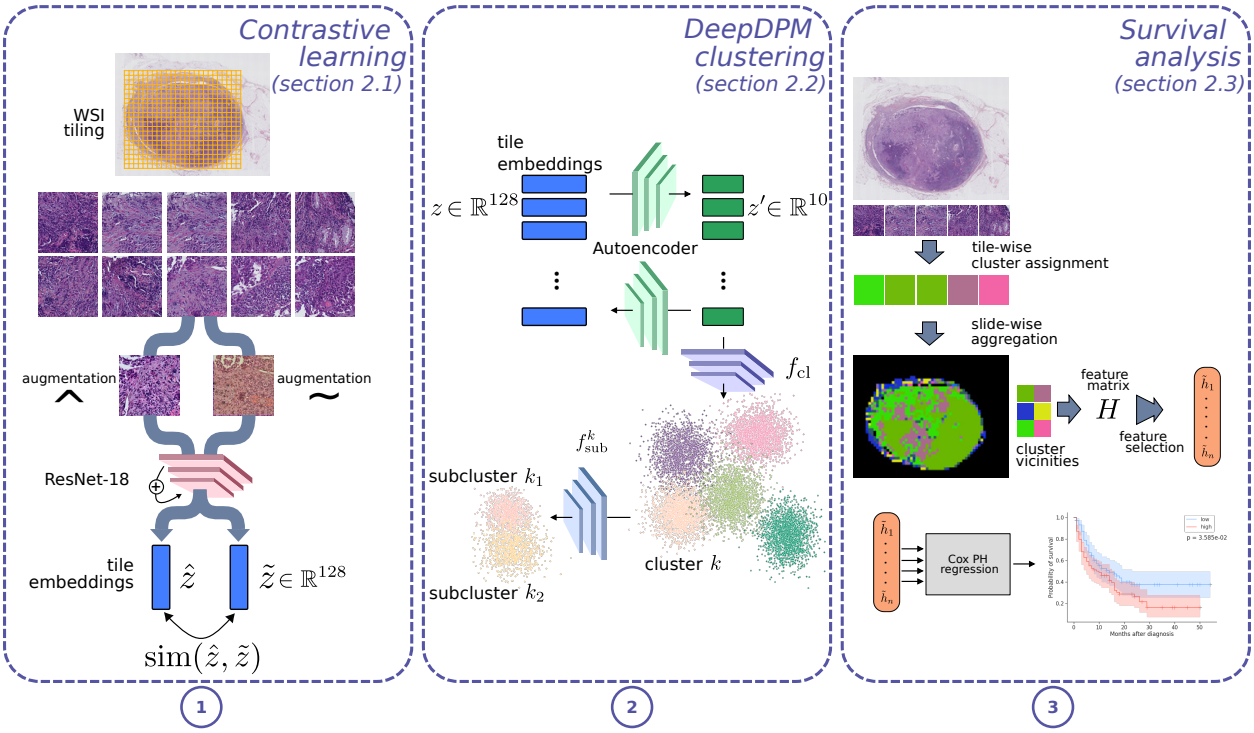

Fig 1 Overview of the WhARIO three-step workflow we use in this paper. The method requires first contrastive pretraining, then clustering the tissue in lung slides, before feature matrices are derived from cluster vicinities and selected for the final survival analysis.

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,[24] which has already been proven efficient for histopathology.[25] 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 the same image augmented in two different ways, i.e., by minimizing the Normalized Temperature-scaled cross-entropy (NT-Xent):

$$\ell = -\log \frac{\exp(\mathrm{sim}(\hat{z}, \tilde{z})/\tau)}{\sum_{t \neq \hat{z}} \exp(\mathrm{sim}(\hat{z}, t)/\tau)} \tag{1}$$

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,[25] 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.

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.[26] More recently,[27] 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:

$$p(\boldsymbol{x} \mid (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k)_{k=1}^{\infty}) = \sum_{k=1}^{\infty} \pi_k \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2}$$

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:

$$H_s = \frac{\alpha \Gamma(N_{k,1}) f_{\boldsymbol{x}}(\mathcal{X}_{k,1}; \lambda) \Gamma(N_{k,2}) f_{\boldsymbol{x}}(\mathcal{X}_{k,2}; \lambda)}{\Gamma(N_k) f_{\boldsymbol{x}}(\mathcal{X}_k; \lambda)} \tag{3}$$

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:

$$\mathcal{L}_{\text{cl}} = \sum_{i=1}^{N} \text{KL}(\boldsymbol{p}_i || \boldsymbol{p}_i^{\text{E}}) \tag{4}$$

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),[28] following:

$$p_{i,k}^{\text{E}} = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma}_{k'})} \tag{5}$$

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

$$\mathcal{L}_{\text{sub}} = \sum_{k=1}^{K} \sum_{i=1}^{N_k} \sum_{j=1}^{2} \tilde{p}_{i,j} \|\boldsymbol{x}_i - \tilde{\boldsymbol{\mu}}_{k,j}\|_2^2 \tag{6}$$

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:

$$\mathcal{L}_{\text{rec}} = \frac{1}{N} \sum_{i=1}^{N} \|\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}_i)) - \boldsymbol{x}_i\|_2^2 \tag{7}$$

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

$\boldsymbol{\mu}_{z_i}$:

$$\mathcal{L}_{\text{MSE}} = \|\boldsymbol{f}(\boldsymbol{x}_i) - \boldsymbol{\mu}_{z_i}\|_2^2 \qquad z_i = \text{argmax}_k\, p_{i,k} \tag{8}$$

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.

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 2 illustrates how the matrix H is filled.

To select the features predictive of survival, and that are used by the Cox PH model,[22] 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,[29] or 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

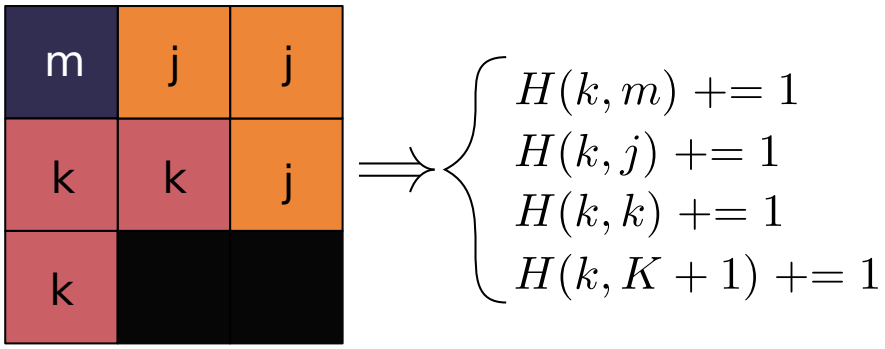

Fig 2 Construction of the feature matrix H based on cluster neighborhoods. Here, we assume a center tile in a slide belonging to cluster k, and the tiles in its 8-neighborhood. For each different cluster k' touching it, the matrix entry $H(k,k')$ is incremented by 1. The background (black region on the image) corresponds to index $K+1$.

for classification tasks. Boruta[30] 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 ability of the model to rank the survival times correctly:

$$C_{\text{index}} = \frac{\sum_{i,j} \mathbb{1}_{T_j<T_i} \cdot \mathbb{1}_{\eta_j>\eta_i} \cdot \delta_j}{\sum_{i,j} \mathbb{1}_{T_j<T_i} \cdot \delta_j} \tag{9}$$

where T_i and T_j are survival times, η_i and η_j the predicted risks and 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[31] 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 $\overline{c}$ 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 $z = -\overline{c}/\log_{10}(p_{\text{lr}})$, 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 1 summarizes the procedure.

Algorithm 1: 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
    $\bar{c} \leftarrow c/M$ `// average performance on the M folds`
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{\bar{c}\}$
    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 coeffi-

Table 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.

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)	82 (11-7197)	371 (37-4274)	604 (14-7333)

cients 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.

3 Materials

3.1 Dataset

This study involved the five following French university hospitals: Caen, Dijon, Nice, Rouen and Toulouse. All data and materials were collected by the Laboratory of Clinical and Experimental Pathology (LPCE), within the Nice university hospital. The following criteria were set:

- All patients had to be diagnosed with Non-Small Cell Lung Cancer.
- All patients had to be treated with immunotherapy.
- For each patient, there had to be two unstained histological slides available.

- The histological slides had to contain at least 10% of tumorous cells.
- The clinical and follow-up information had to be available (in particular, the treatment response assessed with the RECIST 1.1 criteria[32] and the survival).

The tissue slides were stained with Hematoxylin, Eosin and Saffron (HES) and scanned at the LPCE using a NanoZoomer scanner (Hamamatsu Photonics, Hamamatsu, Japan). It is also where the case selection was carried out. On top of the previously mentioned criteria, some slides were unusable after scanning due to persistent blur in the image, and had to be removed from the study. Past the selection phase, every slide was considered a valid sample, even though some of them included very little or sparse tissue. Moreover, we included both biopsy and resection samples, although the latter was only present for one center only (Nice). The flow chart in Figure 3 describes the subsequent steps that lead to the final dataset.

The PD-L1 expression and overall survival information was available for 149 patients out of the 193 with known response. 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 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.

3.2 Experimental Setting

3.2.1 Tile extraction and pretraining

From the available WSIs, 256×256 tiles were extracted at ×20 magnification within the tissue regions after thresholding the saturation channel in the HSV color space, and removing artifacts

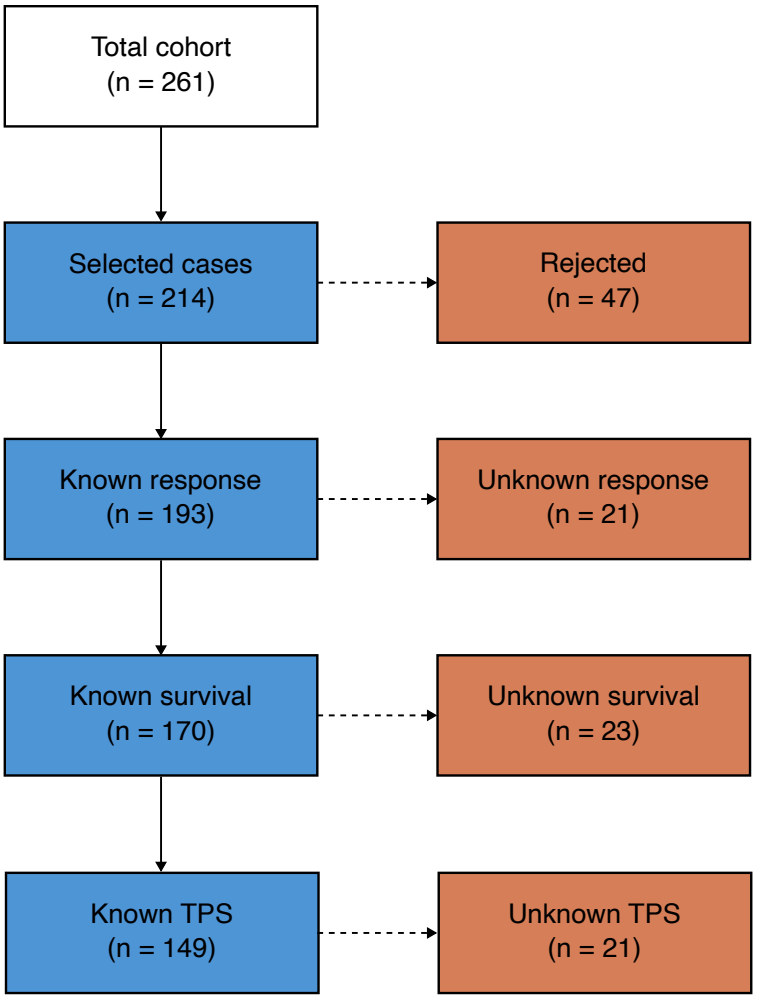

Fig 3 Flow chart of the case selection process

and blurry regions thanks to Gaussian filtering and the coverslip edge detector from the HistoQC package.[33] 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 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,[34] with a batch size of 1024. We used the LARS optimizer[35] with cosine annealing and initial learning rate set to $0.3 \times \textit{batch size} / 256$ following the recommendations of the authors of SimCLR. Con-

cerning 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*.[36]

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 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.

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 2, and summed in case there are several for one patient. The

lifelines package (v0.27.3,[37]) 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[38] of the survival function for each group.

4 Results

4.1 Clustering

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 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 4 shows tile samples corresponding to each cluster, while Figure 5 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 4). For each cluster, the tiles were first sorted by decreasing membership probability. Then, the sorted list was split in 10 different sublists, each corresponding to a decrease of 10% in probability (from 1.0 to 0.1). Tiles from the sublists corresponding to high probabilities were sampled to appear in the mosaic corresponding to the cluster. Once these mosaics were ready, they were submitted to the Nice hospital for inspection, which was performed by a senior thoracic pathologist. For all of the mosaics, specific and identifiable histological patterns were found, with certain clusters harboring particularly homogeneous tissue (e.g., cluster 7

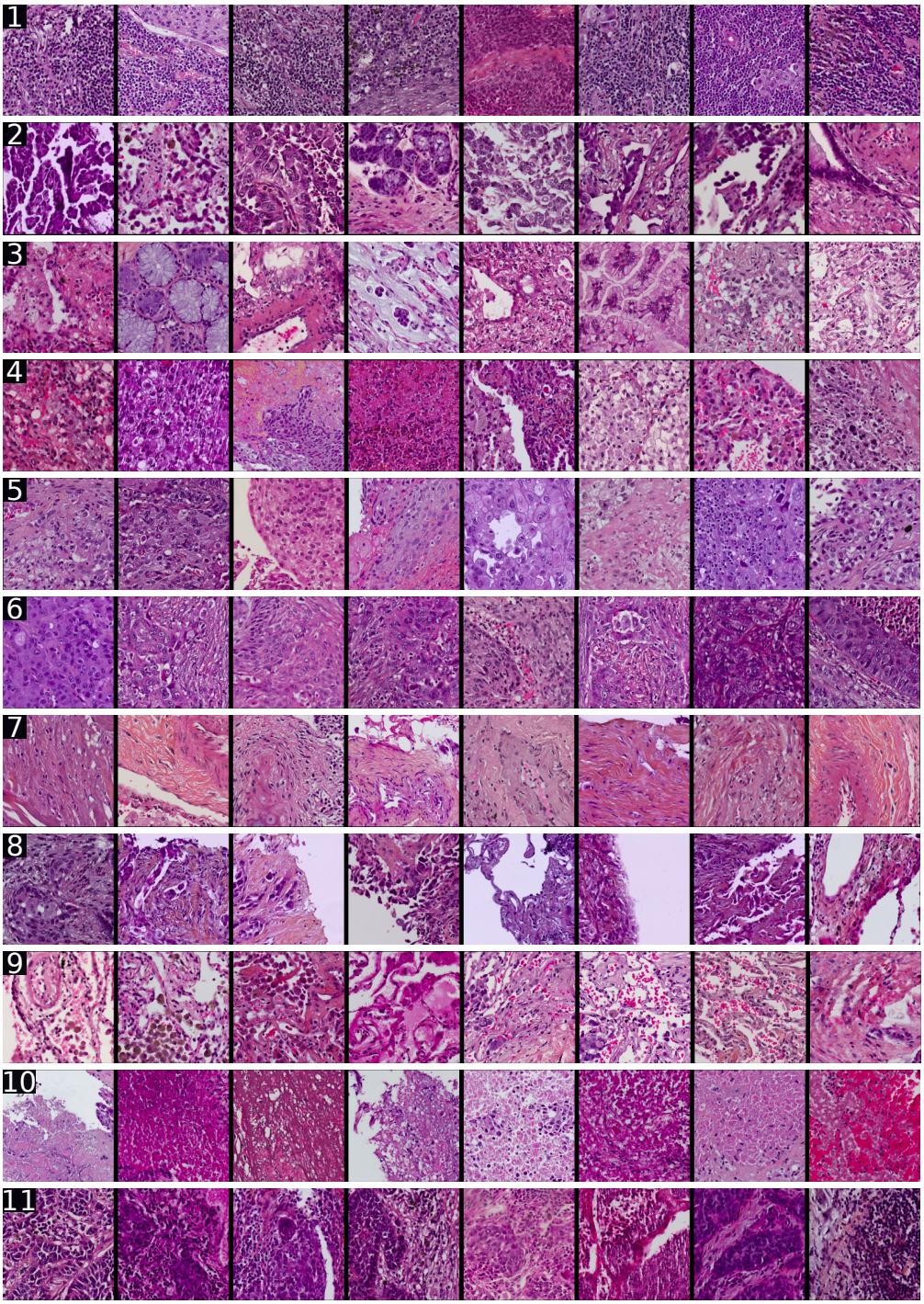

Fig 4 Tile samples corresponding to each discovered cluster.

containing only fibrosis). Although there is partial overlap between clusters, or intra-cluster variability, the overall cluster assignment translates a certain histological logic. Table 2 summarizes the comments made by the expert pathologist on all of the clusters.

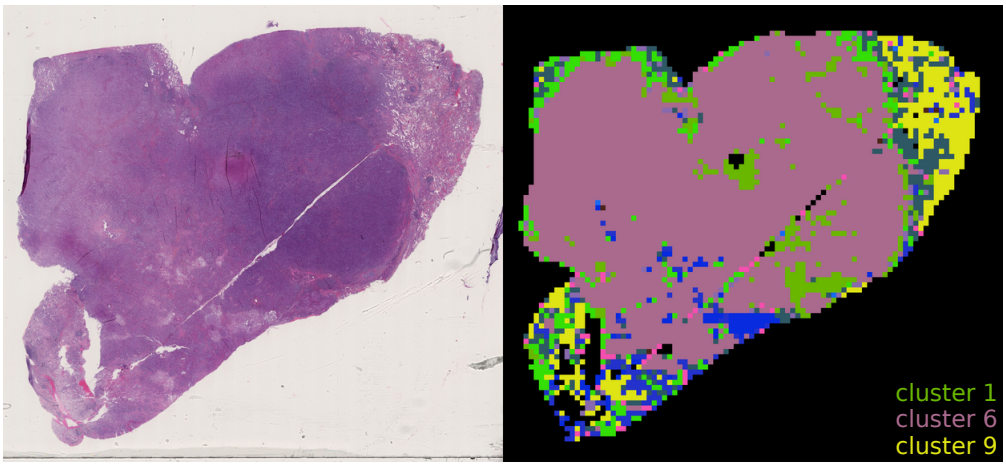

Fig 5 Example of a WSI in the dataset next to its tile-wise representation as clusters. The pink, green and yellow colors correspond to clusters 6, 1, and 9 respectively, which clearly identify the center tumor bulk (6) and surrounding lymphocytes (1), with normal lung parenchyma on the right (9).

Table 2 Summary of the pathologist's comments on the different 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

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 1, 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 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,[39,40] 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.

4.3 Survival Analysis

4.3.1 Cross-validation on the entire dataset

Table 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.

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)**

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%[3] 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 4.2, but also the combination of these with TPS. Figure 6 shows the obtained survival curves with the associated log-rank p-value and c-indexes, while Table 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 single center, Dijon, has significantly different survival for each group (cf. Table 4). What is more, two centers out of the five (Caen and Rouen) do not include low-TPS patients, making a

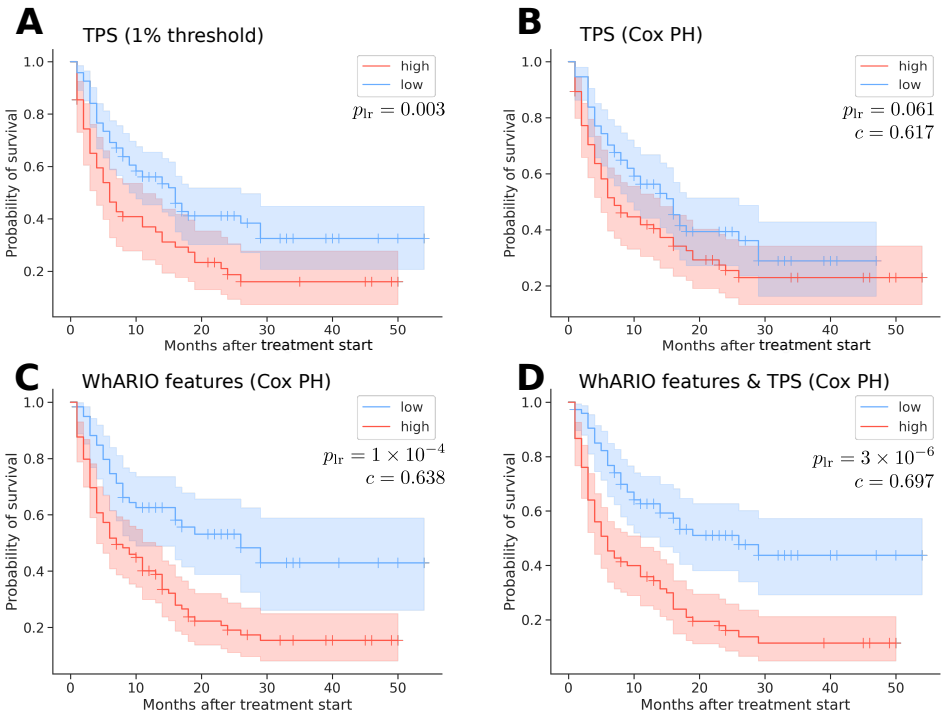

Fig 6 Low- and high-risk group survival curves (with the 95% CI) based on (A) the 1% TPS threshold, (B) a Cox PH regression on TPS values, (C) a Cox PH regression on WhARIO features, and (D) a Cox PH regression on WhARIO features and TPS combined.

threshold-based grouping impossible.

Table 4 Hazard Ratios and p-value of the log-rank test when using the 1% threshold of TPS to split risk groups.

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)

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 = 1×10^{-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 ($p = 3\times10^{-6}$), 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.

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 2.3) to make predictions on the left-out center. Based on the center characteristics in Table 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 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 and 6 and Figures 7 and 8 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 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

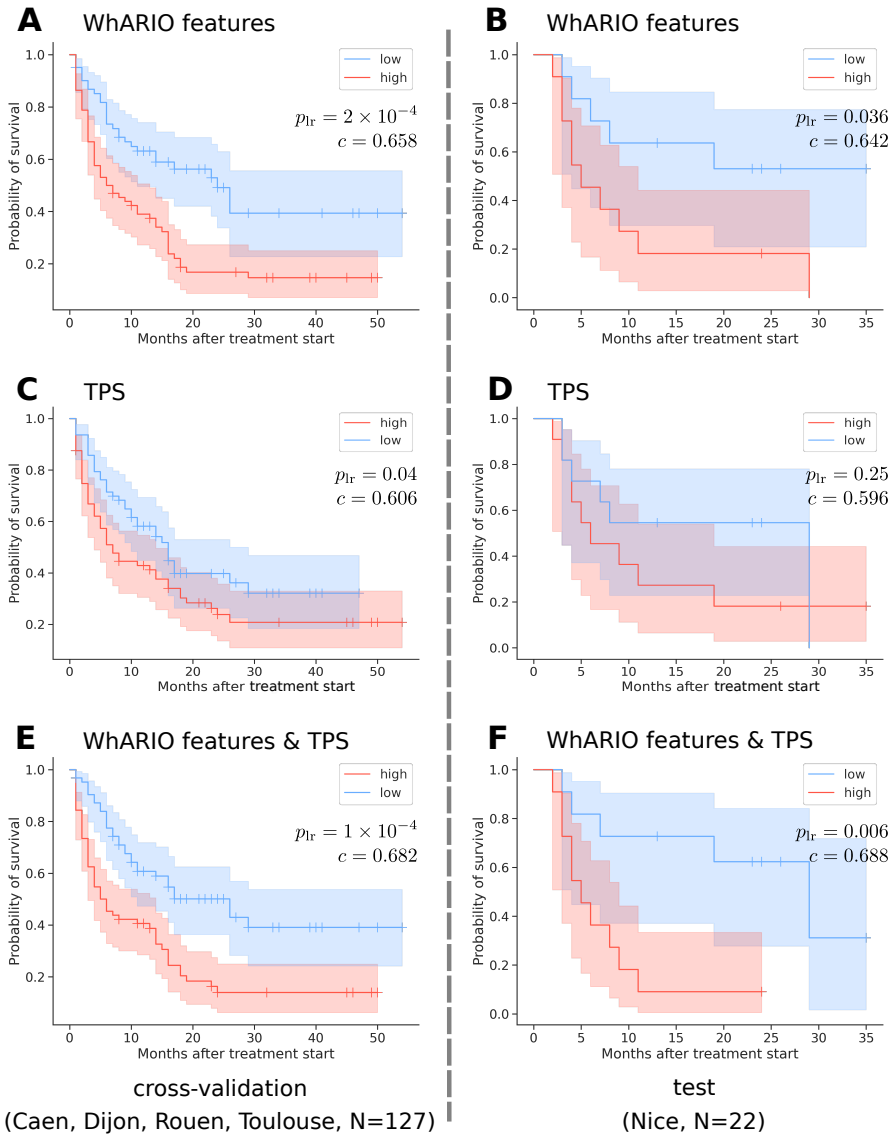

Fig 7 Low- and high-risk group survival curves (with the 95% CI) based on a Cox PH regression using (A,B) WhARIO features only, (C,D) TPS only and (E,F) a combination of the two when Nice is the left-out test set. The left column corresponds to the CV, and the right column to the test set.

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 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.

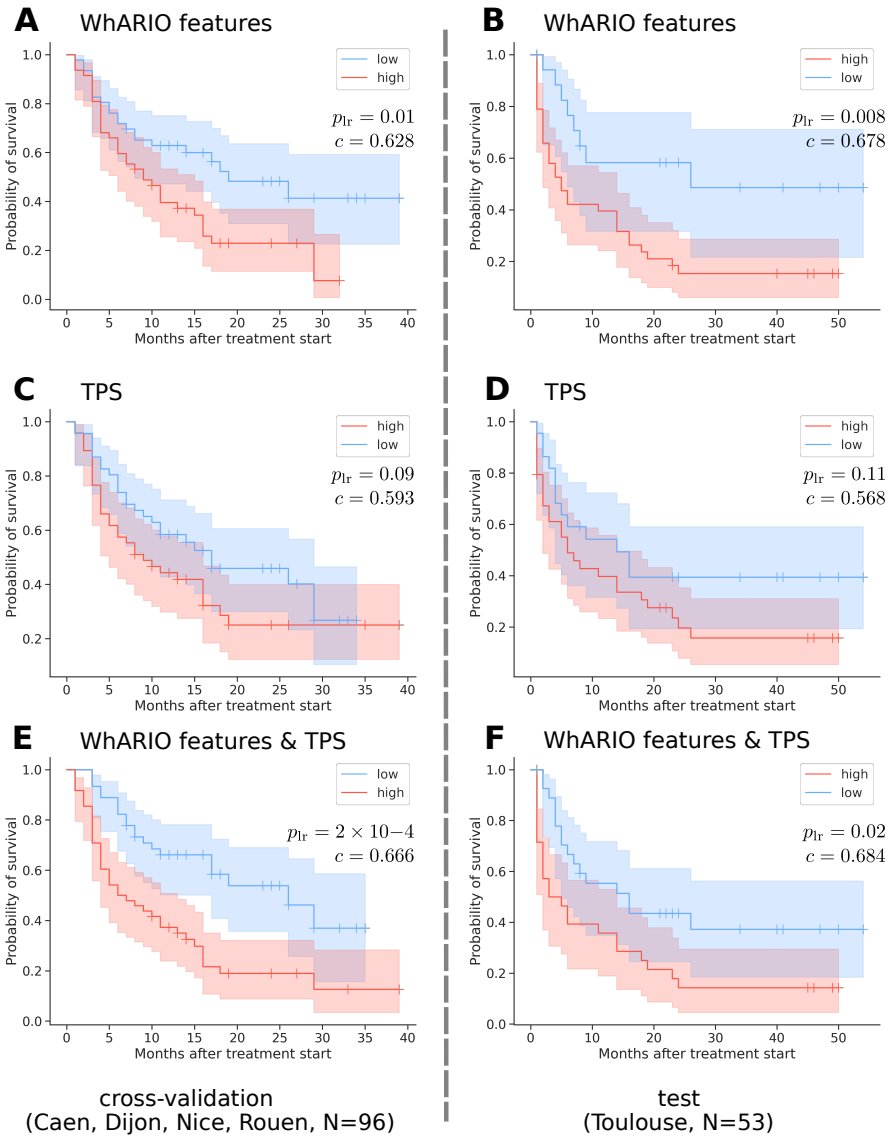

Fig 8 Low- and high-risk group survival curves (with the 95% CI) based on a Cox PH regression using (A,B) WhARIO features only, (C,D) TPS only and (E,F) a combination of the two when Toulouse is the test set. The left column corresponds to the CV, and the right column to the test set. Following the comments in Section 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).

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

Table 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, 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)**

Table 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.

	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)

the lowest amount of tiles per slide, with a median at 82 (against >2000 tiles per slide in Nice, see Table 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 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 supplementary material. 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 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 H&E WSIs. The prognostic power of our

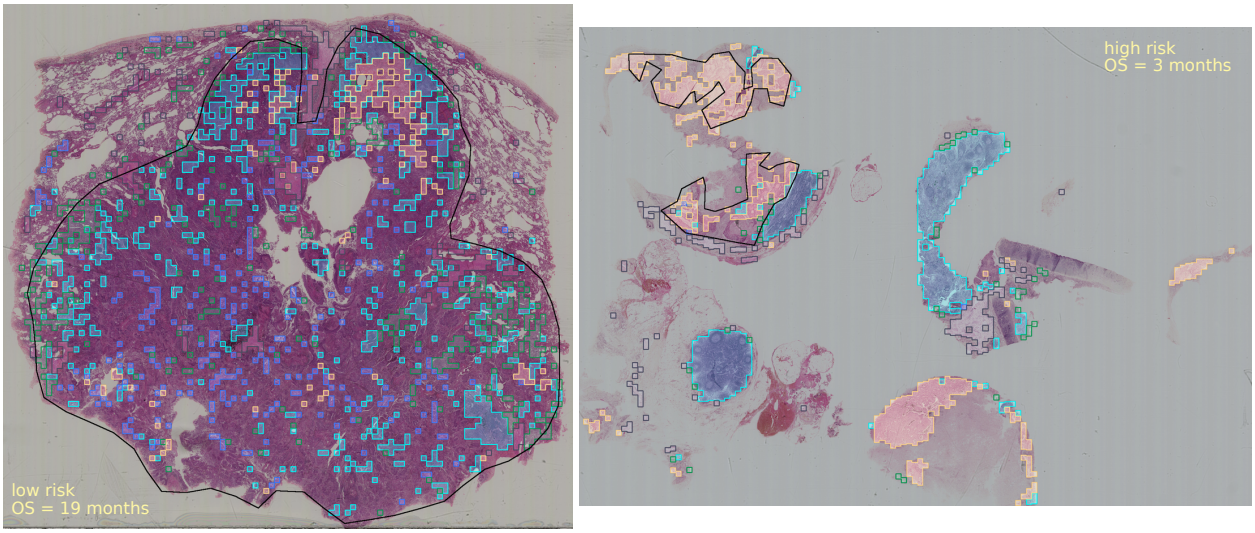

Fig 9 Low- and high-risk examples of slides from the Nice cohort, with the superposition of tile cluster assignments with respect to the selected ones for survival prediction. Cyan and green correspond to clusters 1 and 4 (mostly inflammation/lymphocytes), blue and yellow correspond to clusters 2 and 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 2).

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 2, in particular pairs 1-2 and 6-4). Figure 9 illustrates this by showing two samples of slides classified as low and high risk. On the left, we can see that clusters 1 and 4 (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.[23] 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 10 and 15. For this article in particular, we decided to pick the experiment that generated the fewest clusters possible, so as to reduce the feature exploration space that grows with K. 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[41] or color normalization,[42,43] 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 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 paper 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 characterize them fully. Our comments concerning the Dijon cohort in Section 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.

6 Conclusion

In this paper, 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.

Disclosures

The authors declare no conflicts of interest.

Code, Data, and Materials Availability

The code will be made available upon publication.

References

1 H. Sung, J. Ferlay, R. L. Siegel, *et al.*, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians* **71**(3), 209–249 (2021).

2 L. Horn, D. R. Spigel, E. E. Vokes, *et al.*, "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)," *Journal of Clinical Oncology* **35**(35), 3924–3933 (2017). PMID: 29023213.

3 M. Reck, D. Rodríguez–Abreu, A. G. Robinson, *et al.*, "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," *Journal of Clinical Oncology* **37**(7), 537–546 (2019). PMID: 30620668.

4 M. D. Hellmann, T.-E. Ciuleanu, A. Pluzanski, *et al.*, "Nivolumab plus ipilimumab in lung cancer with a high tumor mutational burden," *New England Journal of Medicine* **378**(22), 2093–2104 (2018). PMID: 29658845.

5 J. Mazieres, A. Drilon, A. Lusque, *et al.*, "Immune checkpoint inhibitors for patients with advanced lung cancer and oncogenic driver alterations: results from the immunotarget reg-

istry," *Annals of Oncology* **30**(8), 1321–1328 (2019). Triple-negative breast cancer - clinical results and biomarker analysis of GeparNuevo study.

6 T. Berghmans, V. Durieux, L. E. L. Hendriks, *et al.*, "Immunotherapy: From advanced nsclc to early stages, an evolving concept," *Frontiers in Medicine* **7** (2020).

7 F. Martins, L. Sofiya, G. P. Sykiotis, *et al.*, "Adverse effects of immune-checkpoint inhibitors: epidemiology, management and surveillance," *Nature reviews Clinical oncology* **16**(9), 563–580 (2019).

8 D. Y. Wang, J.-E. Salem, J. V. Cohen, *et al.*, "Fatal Toxic Effects Associated With Immune Checkpoint Inhibitors: A Systematic Review and Meta-analysis," *JAMA Oncology* **4**, 1721–1728 (2018).

9 T. S. K. Mok, Y.-L. Wu, I. 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," *The Lancet* **393**, 1819–1830 (2019).

10 C. Grigg and N. A. Rizvi, "Pd-l1 biomarker testing for non-small cell lung cancer: truth or fiction?," *Journal for ImmunoTherapy of Cancer* **4**(1) (2016).

11 M. Ilie and P. Hofman, "Reproducibility of pd-l1 assessment in non-small cell lung cancer—know your limits but never stop trying to exceed them," *Translational Lung Cancer Research* **6**(Suppl 1) (2017).

12 W. A. Cooper, P. A. Russell, M. Cherian, *et al.*, "Intra- and interobserver reproducibility assessment of pd-l1 biomarker in non–small cell lung cancer," *Clinical Cancer Research* **23**, 4569–4577 (2017).

13 A. Marabelle, M. Fakih, J. Lopez, *et al.*, "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," *The Lancet Oncology* **21**, 1353–1365 (2020).

14 O. Klein, D. Kee, B. Markman, *et al.*, "Evaluation of tmb as a predictive biomarker in patients with solid cancers treated with anti-pd-1/ctla-4 combination immunotherapy," *Cancer Cell* **39**(5), 592–593 (2021).

15 L. Sha, B. L. Osinski, I. Y. Ho, *et al.*, "Multi-field-of-view deep learning model predicts nonsmall cell lung cancer programmed death-ligand 1 status from whole-slide hematoxylin and eosin images," *Journal of Pathology Informatics* **10**(1), 24 (2019).

16 M. S. Jain and T. F. Massoud, "Predicting tumour mutational burden from histopathological images using multiscale deep learning," *Nature Machine Intelligence* **2**, 356–362 (2020).

17 C. Szegedy, V. Vanhoucke, S. Ioffe, *et al.*, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2818–2826 (2016).

18 N. Harder, R. Schönmeyer, K. Nekolla, *et al.*, "Automatic discovery of image-based signatures for ipilimumab response prediction in malignant melanoma," *Scientific Reports* **9**, 7449 (2019).

19 P. Johannet, N. Coudray, D. M. Donnelly, *et al.*, "Using machine learning algorithms to predict immunotherapy response in patients with advanced melanoma," *Clinical Cancer Research* **27**, 131–140 (2021).

20 S. Park, C.-Y. Ock, H. Kim, *et al.*, "Artificial intelligence–powered spatial analysis of tumor-

infiltrating lymphocytes as complementary biomarker for immune checkpoint inhibition in non–small-cell lung cancer," *Journal of Clinical Oncology* **40**(17), 1916–1928 (2022). PMID: 35271299.

21 X. Wang, C. Barrera, K. Bera, *et al.*, "Spatial interplay patterns of cancer nuclei and tumor-infiltrating lymphocytes (tils) predict clinical benefit for immune checkpoint inhibitors," *Science Advances* **8**(22), eabn3966 (2022).

22 D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)* **34**(2), 187–202 (1972).

23 P. C. Tumeh, C. L. Harview, J. H. Yearley, *et al.*, "Pd-1 blockade induces responses by inhibiting adaptive immune resistance," *Nature* **515**, 568–571 (2014).

24 T. Chen, S. Kornblith, M. Norouzi, *et al.*, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., *Proceedings of Machine Learning Research* **119**, 1597–1607, PMLR (2020).

25 O. Ciga, T. Xu, and A. L. Martel, "Self supervised contrastive learning for digital histopathology," *Machine Learning with Applications* **7**, 100198 (2022).

26 M. Ester, H.-P. Kriegel, J. Sander, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *kdd*, **96**(34), 226–231 (1996).

27 M. Ronen, S. E. Finder, and O. Freifeld, "Deepdpm: Deep clustering with an unknown number of clusters," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9861–9870 (2022).

28 A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data

via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)* **39**(1), 1–22 (1977).

29 C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinform. Comput. Biol.* **3**, 185–205 (2005).

30 M. B. Kursa and W. R. Rudnicki, "Feature selection with the boruta package," *Journal of Statistical Software* **36**(11), 1–13 (2010).

31 N. Mantel, "Evaluation of survival data and two new rank order statistics arising in its consideration," *Cancer Chemother. Rep.* **50**, 163–170 (1966).

32 E. A. Eisenhauer, P. Therasse, J. Bogaerts, *et al.*, "New response evaluation criteria in solid tumours: revised recist guideline (version 1.1)," *European journal of cancer* **45**(2), 228–247 (2009).

33 A. Janowczyk, R. Zuo, H. Gilmore, *et al.*, "Histoqc: An open-source quality control tool for digital pathology slides," *JCO Clinical Cancer Informatics* (3), 1–7 (2019). PMID: 30990737.

34 K. He, X. Zhang, S. Ren, *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (2016).

35 Y. You, I. Gitman, and B. Ginsburg, "Scaling SGD batch size to 32k for imagenet training," *CoRR* **abs/1708.03888** (2017).

36 A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, 8024–8035, Curran Associates, Inc. (2019).

37 C. Davidson-Pilon, "lifelines: survival analysis in python," *Journal of Open Source Software* **4**(40), 1317 (2019).

38 E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association* **53**(282), 457–481 (1958).

39 J. Saltz, R. Gupta, L. Hou, *et al.*, "Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images," *Cell Reports* **23**(1), 181–193.e7 (2018).

40 V. Thorsson, D. L. Gibbs, S. D. Brown, *et al.*, "The immune landscape of cancer," *Immunity* **48**(4), 812–830.e14 (2018).

41 D. Tellez, G. Litjens, P. Bándi, *et al.*, "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Medical Image Analysis* **58**, 101544 (2019).

42 M. Macenko, M. Niethammer, J. S. Marron, *et al.*, "A method for normalizing histology slides for quantitative analysis," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 1107–1110 (2009).

43 A. Vahadane, T. Peng, A. Sethi, *et al.*, "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE Transactions on Medical Imaging* **35**(8), 1962–1971 (2016).

List of Figures

1 Overview of the WhARIO three-step workflow we use in this paper. 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.

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 $K + 1$.

3 Flow chart of the case selection process

4 Tile samples corresponding to each discovered cluster.

5 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).

6 Low- and high-risk group survival curves (with the 95% CI) 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.

7 Low- and high-risk group survival curves (with the 95% CI) 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.

8 Low- and high-risk group survival curves (with the 95% CI) 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 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).

9 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 2).

List of Tables

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.

2 Summary of the pathologist's comments on the different clusters.

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.

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

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.

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.