machine learning & knowledge extraction

MDPI

Article

Automatic Feature Selection for Improved Interpretability on Whole Slide Imaging

Antoine Pirovano [1,2,*], Hippolyte Heuberger [1], Sylvain Berlemont [1], SaÏd Ladjal [2] and Isabelle Bloch [2,3]

[1] Keen Eye, 75012 Paris, France; hippolyte.heuberger@keeneye.ai (H.H.); sylvain.berlemont@keeneye.ai (S.B.)
[2] LTCI, Télécom Paris, Institut Polytechnique de Paris, 91120 Palaiseau, France; said.ladjal@telecom-paris.fr (S.L.); isabelle.bloch@telecom-paris.fr (I.B.)
[3] Centre National de la Recherche Scientifique, Laboratoire d'Informatique de Paris 6, Sorbonne Université, 75005 Paris, France
* Correspondence: antoine.pirovano@keeneye.ai

Abstract: Deep learning methods are widely used for medical applications to assist medical doctors in their daily routine. While performances reach expert's level, interpretability (highlighting how and what a trained model learned and why it makes a specific decision) is the next important challenge that deep learning methods need to answer to be fully integrated in the medical field. In this paper, we address the question of interpretability in the context of whole slide images (WSI) classification with the formalization of the design of WSI classification architectures and propose a piece-wise interpretability approach, relying on gradient-based methods, feature visualization and multiple instance learning context. After training two WSI classification architectures on Camelyon-16 WSI dataset, highlighting discriminative features learned, and validating our approach with pathologists, we propose a novel manner of computing interpretability slide-level heat-maps, based on the extracted features, that improves tile-level classification performances. We measure the improvement using the tile-level AUC that we called Localization AUC, and show an improvement of more than 0.2. We also validate our results with a RemOve And Retrain (ROAR) measure. Then, after studying the impact of the number of features used for heat-map computation, we propose a corrective approach, relying on activation colocalization of selected features, that improves the performances and the stability of our proposed method.

Keywords: histopathology; WSI classification; explainability; interpretability; heat-maps

check for updates

Citation: Pirovano, A.; Heuberger, H.; Berlemont, S.; Ladjal, S.; Bloch, I. Automatic Feature Selection for Improved Interpretability on Whole Slide Imaging. *Mach. Learn. Knowl. Extr.* **2021**, *3*, 243–262. https://doi.org/10.3390/make3010012

Academic Editor: Jaime Cardoso, Nicholas Heller, Pedro Henriques Abreu, Ivana Išgum and Diana Mateus

Received: 16 January 2021
Accepted: 13 February 2021
Published: 22 February 2021

1. Introduction

Since their successful application for image classification [1] on ImageNet [2], deep learning methods, especially Convolutional Neural Networks (CNN), have been extensively used and adapted to tackle efficiently a wide range of health issues [3,4].

Along with these new methods, the recent emergence of Whole Slide Imaging (WSI), microscopy slides digitized at a high resolution, represents a real opportunity for the development of efficient Computer-Aided Diagnosis (CAD) tools to assist pathologists in their work. Indeed, over the last three years, notably due to the WSI publicly available datasets, such as Camelyon-16 [5] and TCGA [6], and in spite of the very large size of these images (generally around 10 giga pixels per slide), deep learning architectures for WSI classification have been developed and proved to be really efficient.

In this work, we are interested in WSI classification architectures that use only the global label (e.g., diagnosis) to train and require no intermediate information such as cell labeling or tissue segmentation (which are time-consuming annotations). The training is regularized by introducing prior knowledge by design in the architectures which, in addition, makes the result interpretable. But the interpretability beyond the architectural design is still pretty shallow.

However, interpretability (the ability to provide explanations that are relevant and interpretable by experts in the field) for medical applications are critical in many ways. (i) For routine tools where useful features are well known and are subject to a consensus among experts, it is important to show that the same features are used by the trained model in order to gain the confidence of practitioners. (ii) A good explainability would enable us to get the most out of the architectural interpretability and thus assist more efficiently medical doctors in their slide reviews. (iii) The ability to train using only slide level supervision opens a new field we call discovery, which consists in predicting, based on easier access (e.g., less intrusive) data, outputs that generally require heavy processes or waiting such as surgery (e.g., prognosis, treatment response). In order to be able to guide experts towards new discoveries, the need for reliable interpretability is obviously high.

This work extends our previous publication in the Workshop on Interpretability of Machine Intelligence in Medical Image Computing (iMIMIC) at nternational Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI 2020) [7]. The new additional contributions consist of:

1. Further validation of heat-map improvement using a ROAR approach adapted for Multiple Instance Learning (MIL) context;
2. A study of the impact of the number of features selected to compute heat-maps;
3. A method relying on features acivation colocalization on slides, that enables us to filter out outlier features selected, thus improving results.

2. Related Work and Motivations

Most common WSI classification architectures deal with these very large images by cutting them into tiles, which is close to the workflow of pathologists who generally analyze these slides at levels of magnification between 5× and 40×.

Recently, as explained in Section 1, architectures that are able to learn using only global slide-level labels have been proposed. They rely on a context of MIL, i.e., slides are represented by bags of tiles with positive bags containing at least one positive tile and negative bags containing only negative tiles. In general, WSI classification methods relying on MIL require preprocessing steps to encode as efficiently as possible tiles from WSIs using tissue detection and normalization methods. The first step of preprocessing generally consists in detecting samples on the WSI since there are a lot of non-informative tiles that are just white background or artifacts. It can be performed using Otsu thresholding [8], color space transformation (e.g., thresholding on the saturation channel of HSV color space in [9]), or semantic segmentation using a U-Net [10] (e.g., in [11]). Once tissue is detected, another step that consists in stain normalization is generally carried out. The motivation behind this step is to improve the transferability to other datasets that come from a different hospital that might use a different scanner or staining to create their WSIs [12]. Color deconvolution [13] is the most popular approach [14–16] but other methods such as color transfer [17] or more recently cycle-GAN [18] can be applied as well. Finally, tiling and feature extraction (i.e., encoding) are performed to enter the MIL context and reduce the input size. Tiling consists in cutting the detected sample region into patches (called tiles) w.r.t. a grid defined by the tile size, the overlap size and the magnification. Feature extraction uses a CNN pre-trained on another task (almost systematically on ImageNet [2] in the literature) to compute descriptors. For example, among most popular works, encoding models used are ResNet [19] in [9,20–24], LeNet [25] in [26], Inception [27,28] in [29] and VGG [30] in [23,31]. There is no clear consensus on which feature extractor gives the best performances, we will give some insight regarding this in the discussions after the conclusion.

For example, CHOWDER [21] (that stands for "Classification of HistOpathology with Weak supervision via Deep fEature aggRegation") is an extension of WELDON [32] (that stands for "WEakly supervised Learning of Deep cOnvolutional neural Networks") and proposes a solution for WSI classification that uses min-max strategy to guide the training and make the decision. This approach reaches an AUC of 0.858 on Camelyon-16 and

0.915 on TCGA-lung (subset of TCGA dataset related to lung cancer). In [26], an attention module [33] is used instead of a min-max layer. AUC values of 0.775 for a breast cancer dataset and 0.968 for a colon cancer dataset were reported. Recently, more works on large datasets proposed architectures that follow the same design [22,24,29,31]. Heat-maps based on intermediate scores computed in these architectures are what we call architectural explainability that results from prior knowledge on WSI problems that is introduced by design in the architecture. They are of great interest and have proved to be really efficient to the point of being able to spot cancerous lesions that had been missed by experts (in [22]). However explanations are relying on a single "medical" score which might limit the interpretability regarding complex tissue structures that can be found on these slides.

While interpretability for deep learning CNN models is still at its beginning, some methods arise from the literature. "Feature Visualization", first introduced in [34], extended in [35] and then extensively developed in [36], consists of methods aiming at outputting visualizations to express in the most interpretable manner features associated with a single neuron or a group of neurons. It can be used to understand the general training of a model. For example, the question of transferring features learned from natural images (ImageNet) to medical images has only recently been investigated [37] while widely used and yet surprisingly good. It has also been used to measure how robust a learned feature is [38]. Another type of explainability methods consists of the attribution methods, i.e., methods that output values reflecting, for each component of the input (e.g., pixels), its contribution to the prediction. They are performed either through perturbation [39] or gradient computation (i.e., measure of the gradient of the output with respect to the input). This second group of methods is gaining more and more attention. In [34], the authors show that the gradient is a good approximation of the saliency of a model and even put forward a potential to perform weakly supervised localization. This work opened a new way of accessing explanations in deep neural networks and motivated a lot of interesting researches [40–42]. Mixed together, these explanation methods can provide meaningful and complementary interpretability.

As the quality of explanations improved, the usefulness to quantify these improvements grew with it, which pushed researches to question interpretability methods and to compare them by measuring their performances. For example, RemOve And Retrain (ROAR) [43] method consists in removing contributing items (features, pixels . . .) identified by an interpretability method from training samples to evaluate its completeness and relevance by measuring the impact of such an ablation on the learning and the performance of the model. Tests (cascade randomization, data randomization . . .) are designed in [44] to show whether interpretability methods are sensitive to model parameters or input. In [45], it is shown that some popular interpretability methods are doing a simple partial input image recovery that makes them model or class insensitivite and it is highlighted through adversarial examples.

To the best of our knowledge, a lot of explainability is still to be introduced in WSI classification architectures. In the next section, we present our approach to improve interpretability of a model trained for WSI classification in histopathology. We rely on gradient-based methods to identify and attribute the importance of features in intermediate descriptors, and on patch visualization for cell-level feature explanations. We also extend feature explanation to a slide level, thus drastically improving tumor localization and medical insights.

3. Proposed Methods

As introduced in Section 2, WSI classification architectures have a common design that is inspried from pathologist's workflow to classify a slide (screening the whole slide at a high magnification level, identifying informative regions and making a decision based on these regions).

We propose to formalize this common design here: Let i be the slide index that is divided into a bag of tiles w.r.t. a tissue detection relying on Otsu segmentation [8] and a non overlapping grid (see Figure 1). j is the tile index for each slide.

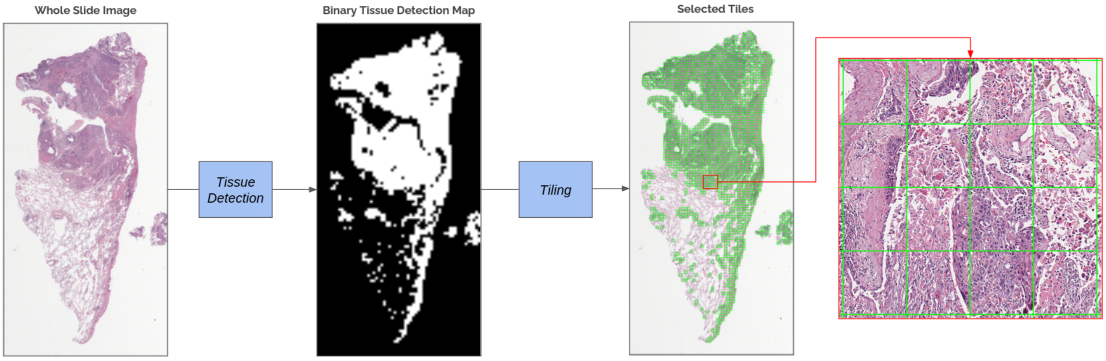

Figure 1. Illustration of tissue detection and tiling processes.

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

1. A feature extractor module f_e (typically a CNN architecture) that encodes each tile $x_{i,j}$ into a descriptor $d_{i,j} \in \mathbb{R}^N$ with N the descriptor size (depending on the feature extractor): $d_{i,j} = f_e(x_{i,j})$; Note that this block is part of pre-processing steps enabling to encode the slide into a bag of tile descriptors.
2. A tile scoring module f_s that, based on each tile descriptor $d_{i,j}$, assigns a single score per tile $s_{i,j} \in \mathbb{R}$: $s_{i,j} = f_s(d_{i,j})$;
3. An aggregation module f_a that, based on all tile scores $s_{i,j}$, and sometimes their tile descriptors $d_{i,j}$, computes a slide descriptor $D_i \in \mathbb{R}^M$ with M the slide descriptor size (depending on the aggregation module): $D_i = f_a(s_{i,j}, d_{i,j})$;
4. A decision module f_{cls} that, based on the slide descriptor D_i, makes a class prediction $P_i \in \mathbb{R}^C$ with C the number of classes: $P_i = f_{cls}(D_i)$.

Our approach (illustrated in Figure 2) consists in rewinding explanations from the decision module to tile information by applying interpretability methods and by answering successively the following three questions:

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

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

$$A_{c,m} = \sum_{i \in I_c} \mid \frac{\partial P_{i,c}}{\partial D_{i,m}} \mid = \sum_{i \in I_c} \mid \frac{\partial f_{cls}(D_{i,m})_c}{\partial D_{i,m}} \mid \tag{1}$$

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

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

K_c positions in slide descriptor. Each attribution $a_{c,n}$ at position n ($\in [0; N]$) of vector a_c is computed as:

$$a_{c,n} = \sum_{(i,j) \in J_c} \left| \frac{\partial s_{i,j}}{\partial d_{i,j,n}} \right| = \sum_{(i,j) \in J_c} \left| \frac{\partial f_s(d_{i,j,n})}{\partial d_{i,j,n}} \right| \tag{2}$$

with J_c the set of tile positions (i, j) that most activate K_c positions in slide descriptors (threshold empirically determined, see Section 4).

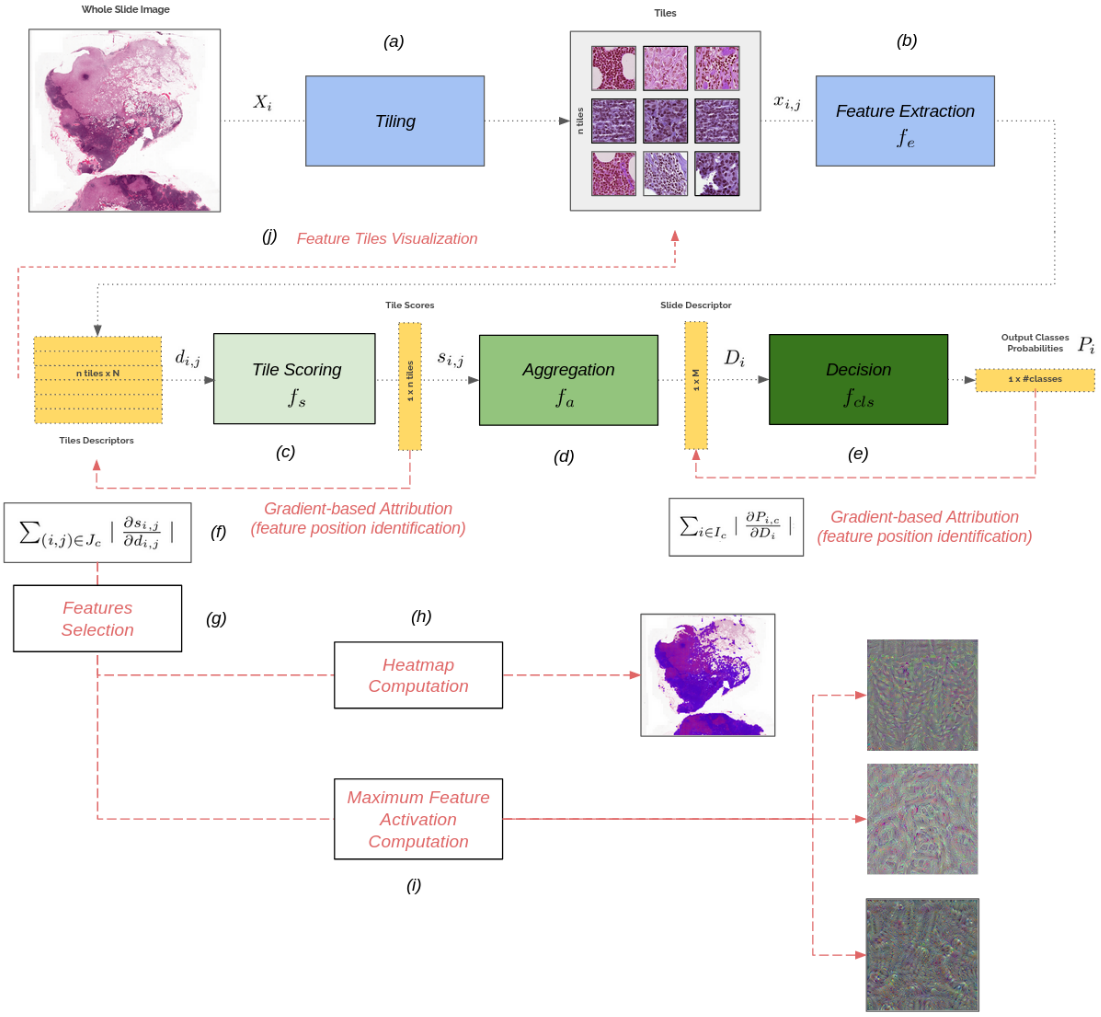
Whole Slide Image
X_i
(a)
Tiling
Tiles
n tiles
$x_{i,j}$
(b)
Feature Extraction
f_e
(j)
Feature Tiles Visualization
Tile Scores
Slide Descriptor
Output Classes Probabilities
P_i
n tiles x N
$d_{i,j}$
Tile Scoring
f_s
1 x n tiles
$s_{i,j}$
Aggregation
f_a
1 x M
D_i
Decision
f_{cls}
1 x #classes
Tiles Descriptors
(c)
(d)
(e)
$\sum_{(i,j)\in J_c} | \frac{\partial s_{i,j}}{\partial d_{i,j}} |$
(f)
Gradient-based Attribution (feature position identification)
$\sum_{i\in I_c} | \frac{\partial P_{i,c}}{\partial D_i} |$
Gradient-based Attribution (feature position identification)
Features Selection
(g)
(h)
Heatmap Computation
Maximum Feature Activation Computation
(i)

Figure 2. Overview of the proposed method. Whole Slide Imaging (WSI) classification: (**a**) Tiling from the slide; (**b**) Features extraction from the tiles; (**c**) Tile scoring from tile descriptors; (**d**) Aggregation; (**e**) Decision from the slide descriptor; Interpretability (in red): (**f**) Feature identification from gradient-based attributions; (**g**) Feature selection from feature colocalization; (**h**) Heat-map computation from activation of selected features; (**i**) Individual feature visualization through gradient ascent; (**j**) Individual feature visualization through tile feature activation.

To answer the third question, we rely on feature activation to highlight features identified as being discriminative to the task by selecting tiles $x_{i,j}$ that have the highest activation per feature in k_c identified over the whole test set. Along with these tiles, we display, for each position in tile descriptors $k \in k_c$, a maximum activation $\mathcal{X}^k$ image obtained by iteratively tuning pixels values to activate the feature by gradient ascent as follows: $\mathcal{X}^k$ is initialized as a uniformly distributed noise image $\mathcal{X}^k_0$; then while $f_e(\mathcal{X}^k_{n-1})_k$ (activation at position k) increases, iterate over $n > 0$:

$$\mathcal{X}_n^k = \mathcal{X}_{n-1}^k + \frac{\partial f_e(\mathcal{X}_{n-1}^k)_k}{\partial \mathcal{X}_{n-1}^k}. \tag{3}$$

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

$$H_{c,i,j} = \frac{1}{|k_c|} \cdot \sum_{k \in k_c} \frac{d_{i,j,k} - \min_k}{\max_k - \min_k} \tag{4}$$

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

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

We also validate results obtained with localization AUC by performing a ROAR analysis adapted to MIL context (defined in Section 2). Indeed, good heat-maps put forward discriminative tiles, thus removing these tiles from bags should prevent the model to learn. In this context, we propose to gradually (by thresholding on tile scores) removing tiles with a high tile score and to train a model with these new reduced bags. If heat-maps are relevant (i.e., if highlighted tiles represent the class of interest) and complete (i.e., if tiles representing the class of interest all have high tile scores) then slide classification performances should drop, while if heat-maps are not relevant or not complete the performances should remain stable through trainings.

Further and deeper analysis on the impact of the number of selected features on the quality of generated heat-maps presented in the next section enabled us to propose an additional *feature selection* block (in Figure 2) to filter out outliers selected. We will present this method after motivating it by our results.

In the next section, we present MIL architectures and the dataset on which we make our experiments and validate our approach. We also detail intermediate results that enable us to improve the overall interpretability and explanation heat-maps in particular.

4. Experiments and Results

4.1. Architectures

We validate our approach on two WSI classification trained architectures: CHOWDER and Attention-based classification.

CHOWDER [21] uses a 1×1 convolution layer to turn each tile descriptor into a single tile score. These scores are then aggregated using a min-max layer, that keeps the top-R and bottom-R scores (e.g., empirically $R = 5$ gives the best results), to give a slide descriptor ($M = 2 \times R$).

Attention-based architecture [26] uses an attention module (two 1×1 convolution layers with respectively 128 and 1 channels and a softmax layer) to compute competitive and normalized (sum to 1) tile scores from tile descriptors. Then, the slide descriptor is computed as the weighted (by tile scores) sum of tile descriptors ($M = N$).

Note that in our experiments the feature extractor is a ResNet-50 [19] ($N = 2048$) trained on ImageNet and is part of the preprocessing thus is not fine-tuned. The decision module is a two layers fully connected network with 200 and 100 hidden neurons, respectively.

4.2. Datasets

We validate our approach using Camelyon-16 dataset that contains 345 WSI divided into 209 “normal” cases and 136 “tumor” cases. This dataset contains slides digitized at 40× magnification from which we perform sample detection using Otsu thresholding [8] on a thumbnail of the slide downscaled by a factor 32 and keeping tiles that contain at least 50% of foreground pixels w.r.t. Otsu segmentation. Then, we extract, with regard to a non-overlapping grid, 224 × 224 pixels tiles at 20× magnification without stain normalization. Then, we pre-compute, for each tile, 2048-tile descriptors using a ResNet-50 model trained on ImageNet as it is done in [9,20–24]. 216 slides are used to train our models while 129 slides form the test set to evaluate performances of the different trained models.

4.3. Results on CHOWDER Model

CHOWDER model performs, at slide-level classification, with an AUC of 0.82. Let us now illustrate and detail the results of our approach on this CHOWDER model guided by the three questions raised in Section 3.

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

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

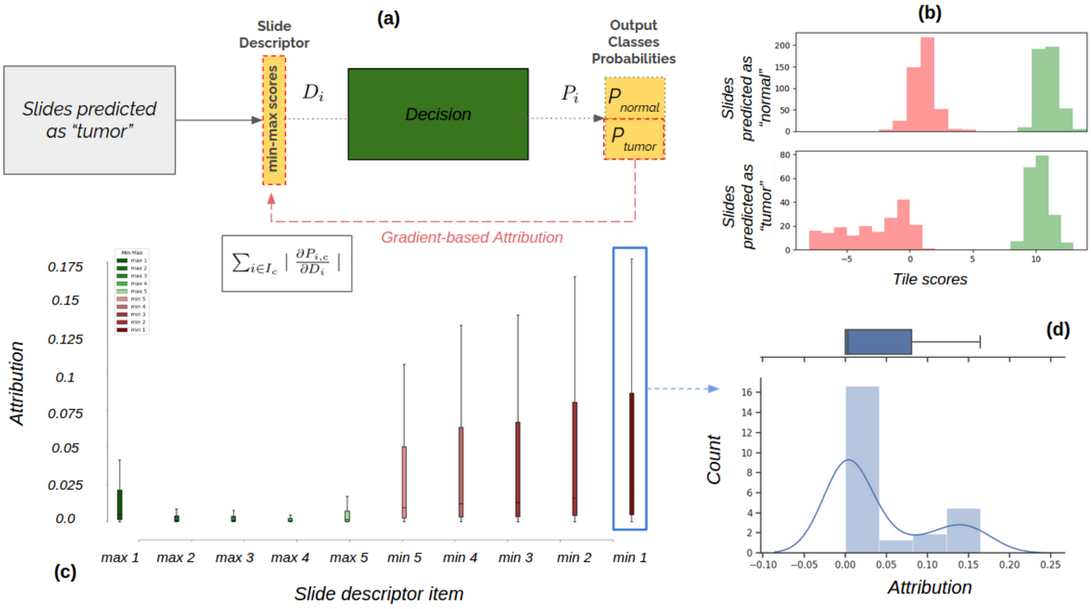

Figure 3. Slide descriptor attribution for the “tumor” class; (**a**) Illustration of gradient on the learned “Decision” block; (**b**) Distributions of min and max scores w.r.t. predicted class; (**c**) Distributions of attribution on each slide descriptor items (5-min scores and 5 max-scores); (**d**) Detail on the distribution of attribution for min 1 (lowest) tile score item and bimodal Gaussian approximation.

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

Most minimal tile scores are under −5 (and most maximal tile scores are above 11). For each of these groups of tiles, we compute the average attribution of each of the $N = 2048$ features in tile descriptors (extracted by a ResNet-50 trained on ImageNet). Figure 4 shows the distribution of features hence activated and allows us to identify which features are mostly responsible for min (and max tile scores), i.e., highest attribution for min (and max scored tiles).

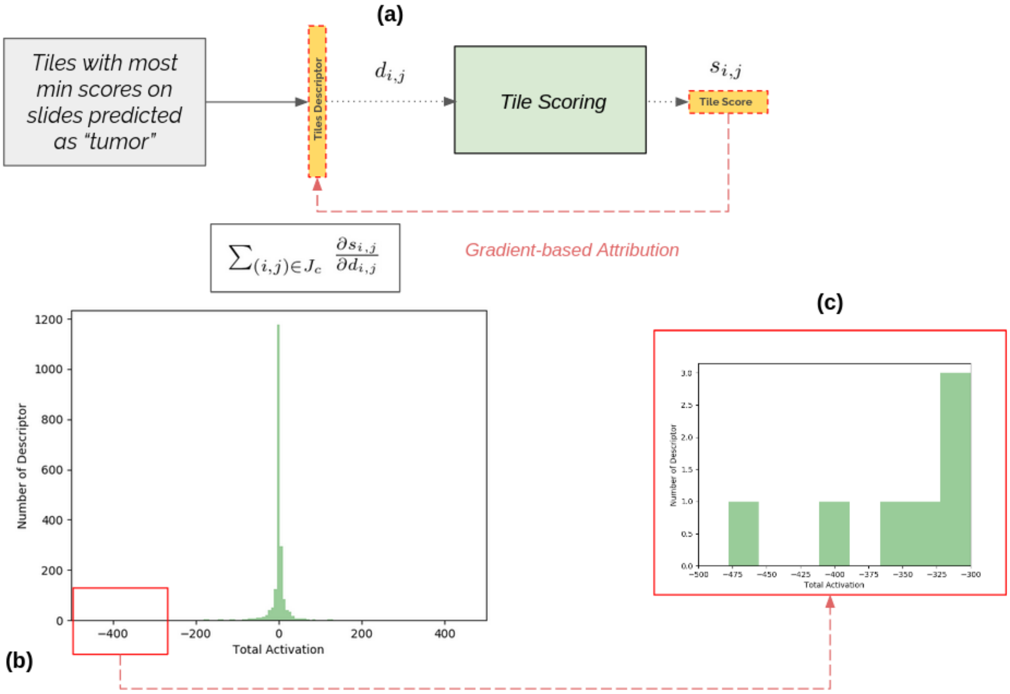

Figure 4. Attribution on tile descriptor items for the "tumor" class; (**a**) Illustration of gradient on the learned "Tile Scoring" block; (**b**) Distributions of attributions for tile descriptors with lowest score; (**c**) Zoom on the range of interest.

Thus, we are able to claim that features (defined by their position in the descriptor) that are mostly useful for the trained model for the "tumor" class are 242, 420, 602, 1154, 1644, 1652 and 1866. Following the same process, we identified features 565, 628, 647, 1158 and 1247 as being the most contributing features for the "normal" class according to CHOWDER model.

4.4. Results on Attention-Based Model

The attention-based model performs similarly to CHOWDER at slide-level classification with an AUC of 0.83. Here, tile scores are used to weight how much each tile is contributing to describe the slide w.r.t. the medical task the model has been trained on. As we understand that high tile scores should put forward tile descriptors that activate relevant features for the diagnosis, we also understand that, if the attention module makes its job well, relevant features should be used by both attention module and decision module. Thus, we propose to select features that have a high attribution in both tile descriptors and slide descriptors.

Using gradient-based attribution, we compute the histogram of attributions over the 2048 features of both slide descriptors and high scored tile descriptors of slides predicted as "tumor" (respectively w.r.t. the class prediction made and the predicted tile score). Figure 5 shows the selection of features for the "tumor" class, i.e., attribution of slide, and high (above 0.1) tile descriptors for slide predicted as "tumor".

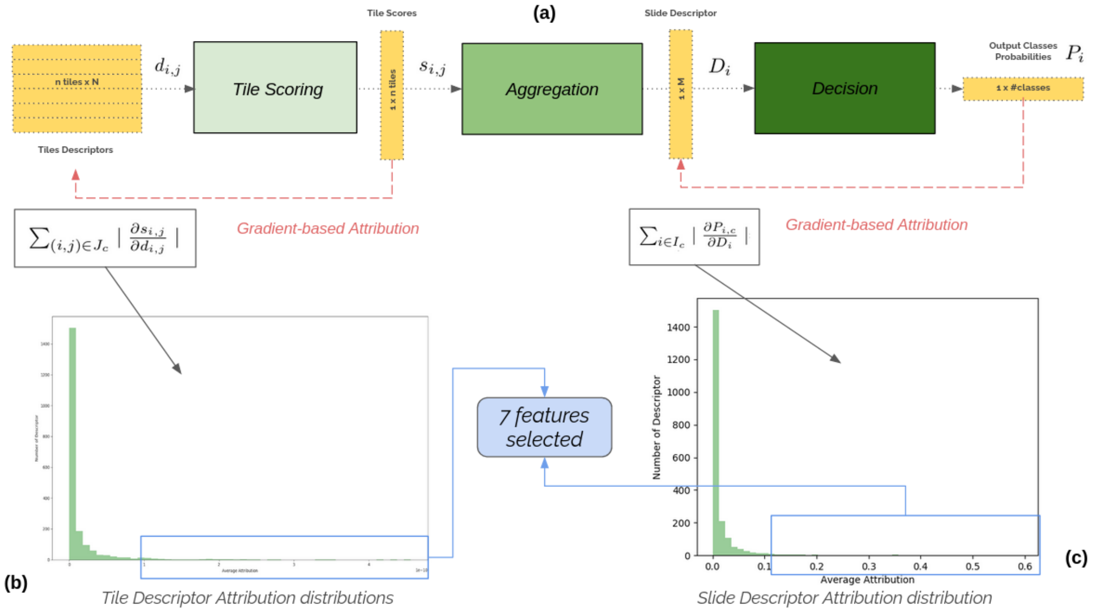
(a)
Tile Scores
Slide Descriptor
n tiles x N
Tiles Descriptors
$d_{i,j}$
Tile Scoring
1 x n tiles
$s_{i,j}$
Aggregation
1 x M
D_i
Decision
Output Classes Probabilities
P_i
1 x #classes
$\sum_{(i,j)\in J_c} \mid \frac{\partial s_{i,j}}{\partial d_{i,j}} \mid$
Gradient-based Attribution
$\sum_{i\in I_c} \mid \frac{\partial P_{i,c}}{\partial D_i} \mid$
Gradient-based Attribution
7 features selected
Number of Descriptor
1400
1200
1000
800
600
400
200
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Average Attribution
(b)
(c)
Tile Descriptor Attribution distributions
Slide Descriptor Attribution distribution

Figure 5. Feature selection for the "tumor" class using the attention-based model. (**a**) Illustration of the model and gradient-based explanation computation; (**b**) Distributions of attributions for tile descriptors of high scored tiles; (**c**) Distributions of attributions for slide descriptors of slides predicted as "tumor".

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

4.5. Tile-Level Interpretability

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

Figure 6 shows the 7 tiles that activate the most (over all tiles) each feature and the max activation image, that we expect to reveal what the feature means with regards to the histopathological problem it has been trained on.

Pathologists agreed that patch-based tile visualizations are highly interpretable and exhibit features that are indeed related to each class [46]. For example, feature 1652 tends to trigger spindle-shaped cells that indeed can be a metastasic tissue organization. For "normal" tissue features, feature 565 describes mainly clustered lymphocytes that are preponderant in normal tissues.

Appendix A shows more tile-level visualizations for more features.

4.6. Slide-Level Explanations: Heat-Maps

Coherence between patches exposed for a better interpretability led us to think about another way to present features to pathologists. Indeed, since tissues have a coherent and somehow organized structure, a relevant feature for histological problems would be activated in a coherent and somehow organized way over slides. Thus, along with patch-based visualization, we propose to access features activation heat-maps $H_{c,i}$ over slides, as presented in Section 3.

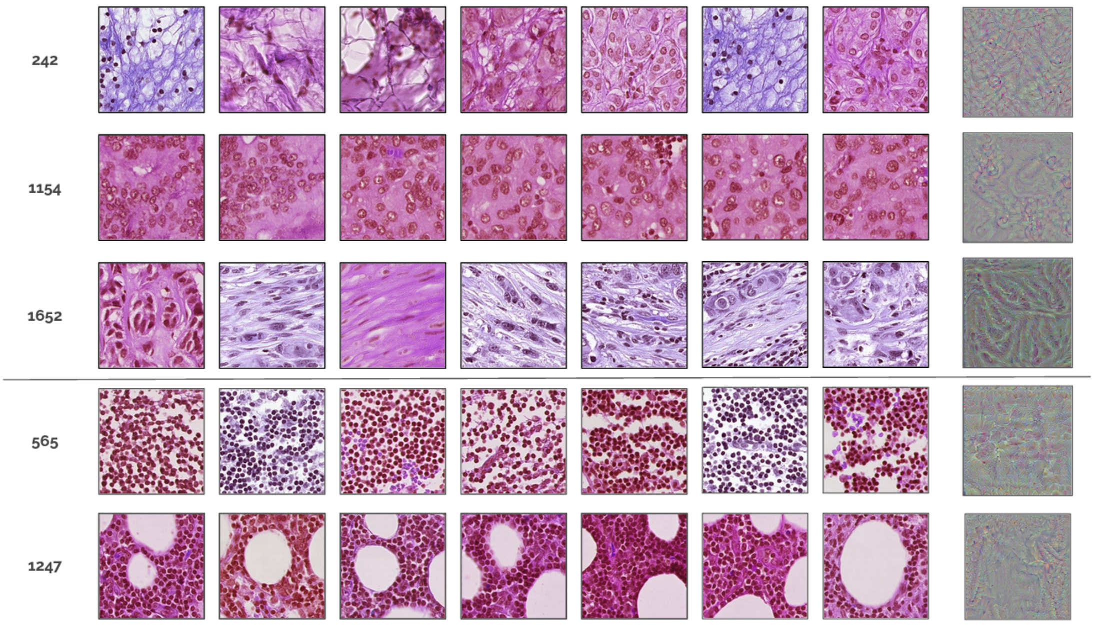

Figure 6. Patch-based visualizations obtained for features 242, 1154 and 1652 (for min-scores features); 565 and 1247 (for max-scores features); tiles and max activation images (right).

Figure 7 illustrates qualitative results. Quantitatively, we report a tile-level localization AUC of 0.884 for CHOWDER model and 0.739 for Attention-based model, using feature-based heat-map values (that are the average normalized feature activation over all features identified for the“tumor” class, see $H_{c,i,j}$ computation in Section 3) as a "tumor" prediction score and using lesion annotations provided by Camelyon-16 dataset to get the ground-truth label per tile. Both AUCs are significantly high, which validates our approach of identifying features that are relevant and of computing heat-maps for interpretation and explanation. Note that the AUC computed using tile scores is 0.684 for CHOWDER model and 0.421 for Attention-based model (see Table 1). We can also note that there is a gap in interpretability between CHOWDER model and Attention-based model while classification performances are similar (AUC of 0.82 for the CHOWDER model and 0.83 for the Attention-based model). This gap can be explained by the fact that, in the context of Camelyon-16, identifying one tumorous tile is enough to label a slide as "tumor", so implicit tile classification does not need to be exhaustive to provide meaningful information to the slide level decision module, however if so interpretability will decrease.

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

Figure 8 shows the performances of models retrained after the removal of tiles with different thresholds on the heat-maps obtained from the trained model on the full bags. We can observe that our feature-based heat-maps are the ones impacting the most the performances, which confirms the results in Table 1. Also it confirms that CHOWDER tile score heat-maps are complete and relevant while attention-based tile scores heat-maps are equivalent to random heat-maps due to the important number of positive tiles being scored with a low score by the attention module.

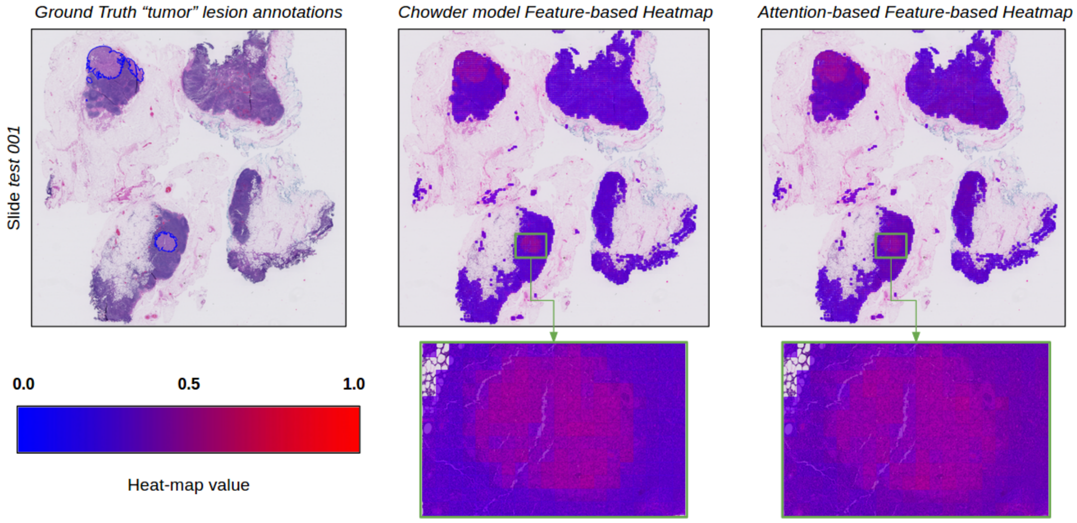
Ground Truth "tumor" lesion annotations
Chowder model Feature-based Heatmap
Attention-based Feature-based Heatmap
Slide test 001
0.0
0.5
1.0
Heat-map value

Figure 7. Slide-based visualizations: Heat-maps explaining the "tumor" class obtained by computing average normalized activation over identified features; ground-truth annotations for "tumor" tissue (left); CHOWDER model feature-based heat-maps (middle); attention-based model feature-based heat-maps (right).

Table 1. Results: classification and localization AUC using both methods (improvement of localization AUC by 0.200 for CHOWDER and 0.318 for Attention-based model).

Model	Classification AUC	Heat-Map Method	Localization AUC
CHOWDER	0.82	Tile scores Feature-based (ours)	0.684 0.884
Attention-based	0.83	Tile scores Feature-based (ours)	0.421 0.739

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

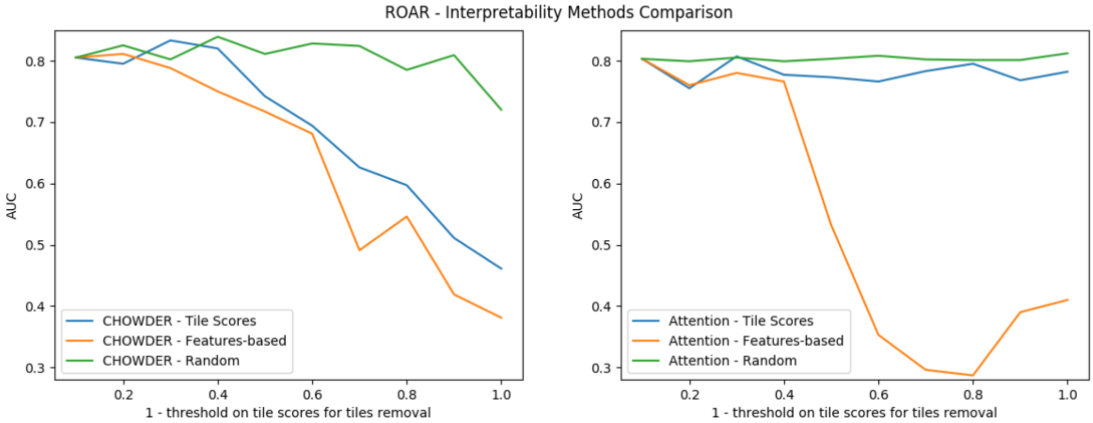

Figure 8. RemOve and Retrain experiment results: Impact of tile removal with regards to heat-maps from CHOWDER (**left**) and Attention-based model (**right**) on slide classification performances.

4.7. Study of the Impact of the Number of Selected Features

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

First, we measured the impact with a small number of features from using only the one most contributing feature up to the first 7 most contributing identified features. We can observe, in Figure 9, that there is an important variablily of localization AUC performances depending on the number of features with a variation between 0.903 and 0.82 (which are still great performances). We can also interpret that there are features of interest that make the localization AUC increase (such as feature 602 or feature 1866) and adversarial features that make the localization AUC decrease (such as feature 1644 or feature 420).

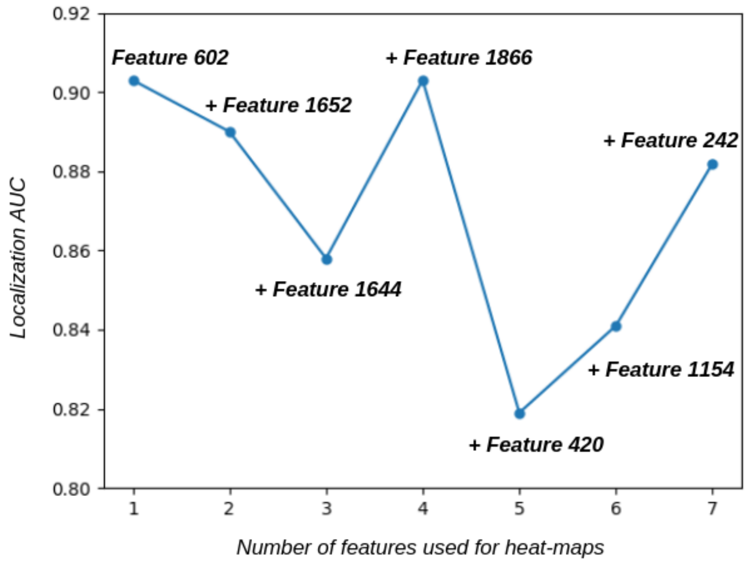

Figure 9. Impact of the number of selected features on localization AUC (between 0.80 and 0.92) for small numbers of features.

Thus, it seems critical to study more deeply this problem. So, as shown in Figure 10, by thresholding at different contribution scores, we select from 1 to (all) 2048 features (according to the distribution in Figure 4):

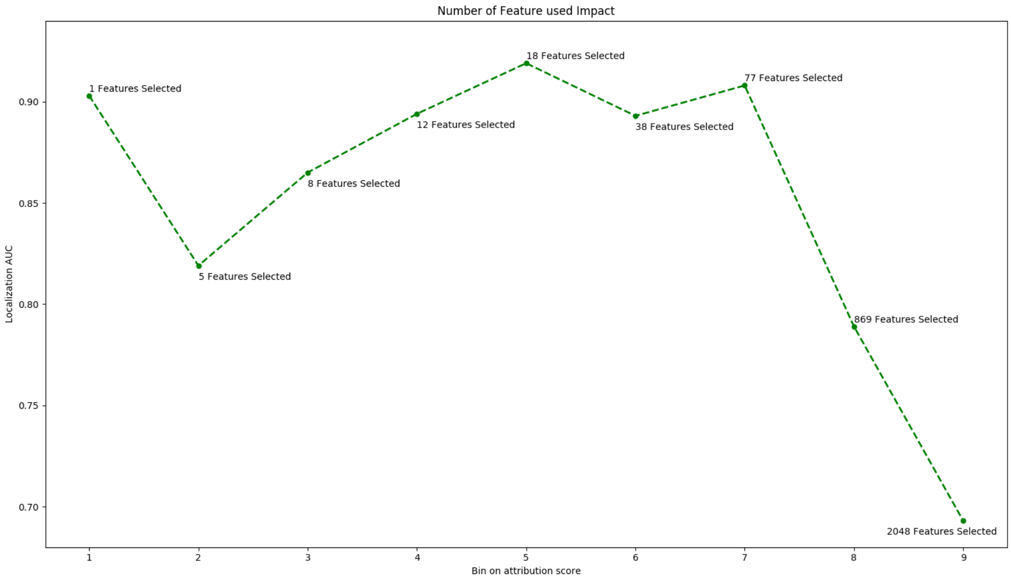

Figure 10. Impact of the number of selected features on localization AUC (between 0.7 and 0.9) for high numbers of features.

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

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

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

5. Feature Filtering Based on Colocalization

5.1. Approach

This study about the impact of the number of selected features on heat-map quality shows that three kinds of features stand out for Camelyon-16 dataset among Imagenet features: features of interest that activate homogeneously mostly in tumorous regions, adversarial features that activate homogeneously not only in tumorous regions, and unrelated features that either activate non homogeneously or almost do not activate over slides. Figure 11 illustrates the difference between features of interest and adversarial features. It also gives another way to think about the third question we put forward ("Are these features of tile descriptors relevant medically and representative of histopathological information?") by introducing a manner to measure the potential transfer of each individual feature to a given histopathological problem.

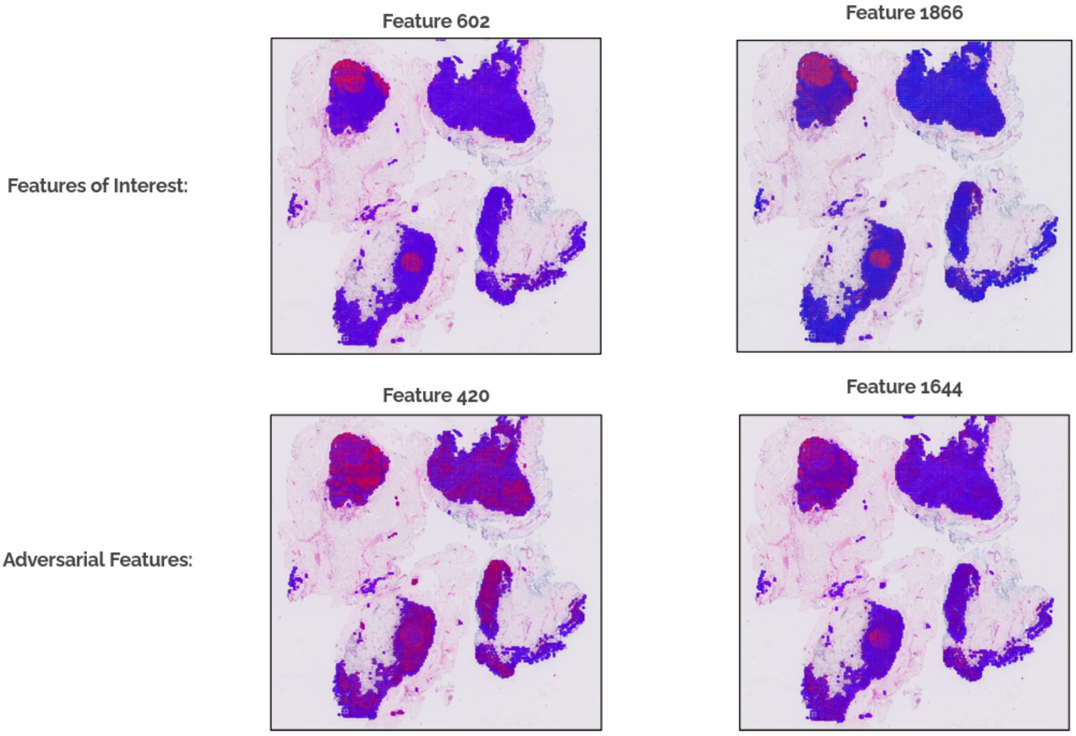

Figure 11. Contributing feature activation over slides.

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

To do so we propose to measure, for each selected feature individually, the mean absolute error (MAE) between the feature activation (normalized) and the feature-based heat-maps over whole slides. Thus, we obtain a distribution of MAE and we propose to keep only the ones that are on the lower half of the distribution. Figure 12 illustrates the method in the case of the 7 selected features for the CHOWDER model.

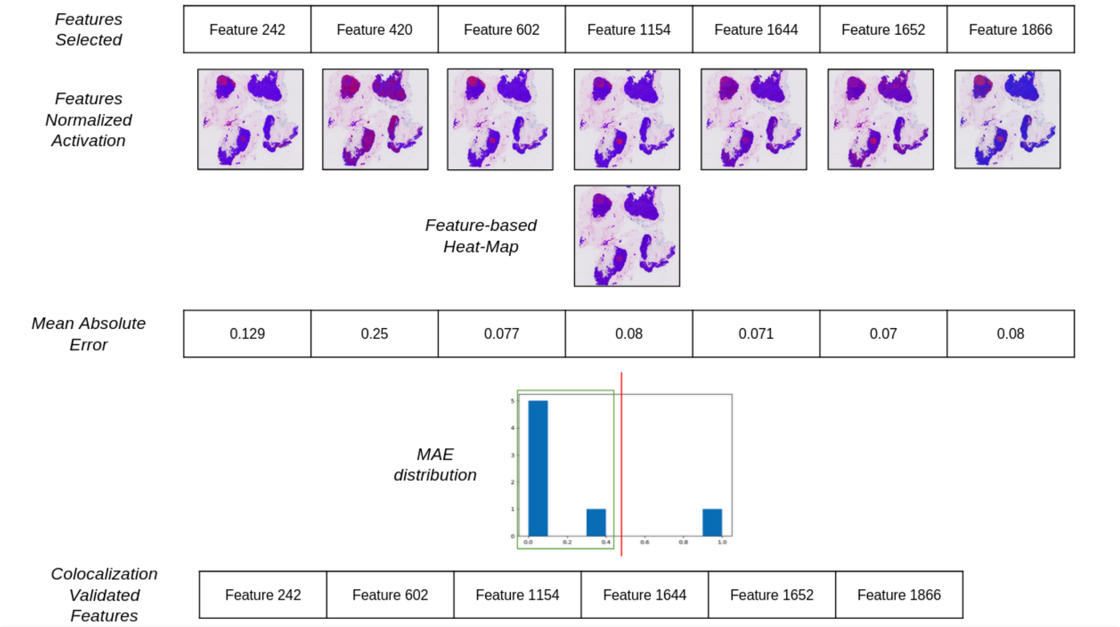

Figure 12. Illustration of the colocalization filtering method.

5.2. Results

We performed this filtering on every group of features previously studied for CHOWDER model. Figure 13 shows the obtained results, and confirms the usefulness of this filtering since the curve obtained with the additional colocalization filtering outperforms the feature-based heat-maps on localization AUC.

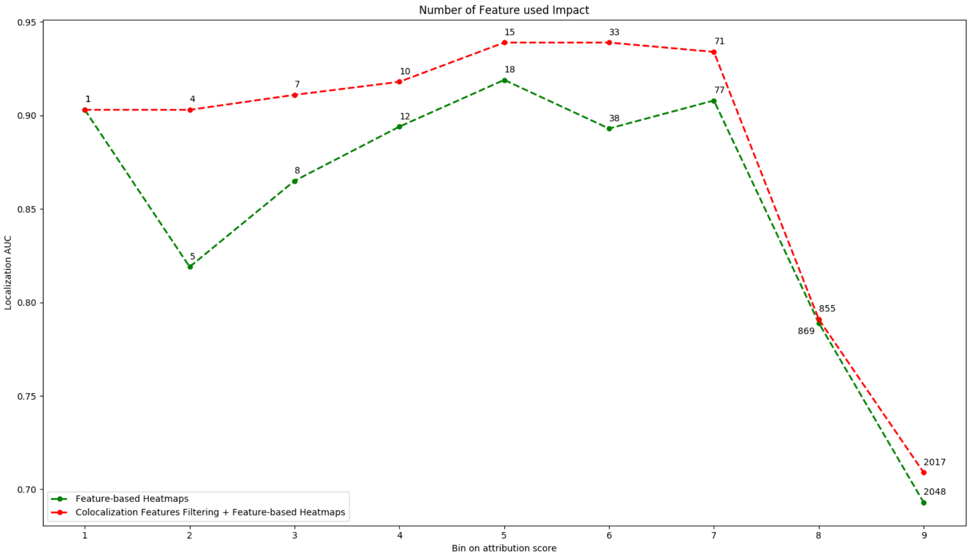

Figure 13. Colocalization filtering results for CHOWDER model.

We repeated this experiment for the attention-based model. Figure 14 confirms the three behaviors identified previously with an unsettled range of features where localization AUC highly varies, a stable range of features (around 5% of features), and a range of high number of features where performances drop. However, we can notice that due to the lack of interpretability at slide-level of the attention-based model, it can happen that in the first range of features the majority of features are adversarial, thus, colocalization filtering worsen performances. Still, in the vast majority of cases, it improves the performances and stabalizes them.

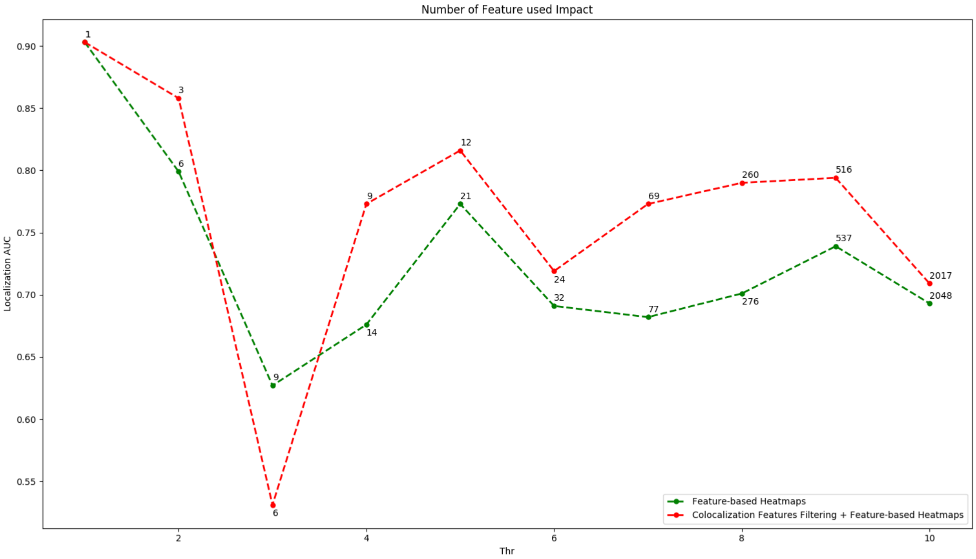

Figure 14. Colocalization filtering results for the attention-based model.

6. Conclusions

In this paper, we presented our interpretability approach and researches that apply to WSI classification architectures. We proposed a unified design that gather a vast majority of WSI classification methods relying on MIL learning. We motivated and applied a gradient-based attribution method to identify features that have been learned to be relevant in intermediate (tile and slide) descriptors. Then we showed the relevance of these features by visualization (with dataset patches and max activation) and validation by pathologists. These discussions made us consider measuring interpretability by computing explanability heat-maps over whole slides taking into account only identified features. Allying patch-based and slide-based visualization took interpretability to a next level for pathologists to understand histological meanings of features used by trained models. They confirmed that our approach gives explanations that are highly meaningful and interpretable, and enabled them to find out that characteristics used by the model are aligned with what pathologists have discovered over decades of researches. Our per-block approach can be used for all WSI classification pipelines trained on histpathological problems (and probably more, such as biomarker discoveries or treatment response) that follow the general design defined in this work, and brought the understanding of how WSI pipeline learns deeper. Validating our approach on two distinct architectures enabled us to claim the generalizability of our approach, and quantifying the improvement of heat-maps generated through two interpretability measures strengthen this point. Finally, our individual analysis of each feature selected at slide-level enabled us to filter out outlier features, to stabilize interpretability performances, and to automatically select the right number of features neededed for good heat-maps. This work dig deeper in the interpretability of WSI classification trained models.

7. Discussions

With this, we also open a lot of additional questions that could be of critical interest for the development of efficient CAD tools and to guide medical discoveries. We could start by adapting and/or validating the approach to a dataset with more than two classes and also trying fancier explanation methods such as SmoothGrad or Integrated Gradient. Measuring the interpretability could be performed by more pathologists to have quantitative evaluation of tile-level features. It would be also interesting to study the impact of stain normalization pre-processing on identified features and on the quality of feature-based heat-maps. We also want to discuss the fact that it seems that using and mixing together few numbers of features from ImageNet is good enough to describe tissues that can be found on histopathology slides, which may explain the choice of a specific feature extraction model, as long as it is only transfer from pre-training on ImageNet, is not a critical hyperparameter since most popular CNN architectures are able to learn complex features. Moreover, our work could lead to an improvement of classification performances through new regularizers for clustering constraint for example, or feature selection to reduce overfitting and improve generalizability. Finally, we have highlighted that most features used are texture classifiers, which is adapted to histopathological problems, but it could prevent these WSI classification pipelines to be applied directly to more challenging tasks such as liquid-based cytopathology problems that are not related to cell organizations and tissue structure but individual cell analysis.

Author Contributions: Conceptualization, A.P.; methodology, A.P., H.H., S.L. and I.B.; software, A.P. and H.H.; validation, S.B., S.L. and I.B.; formal analysis, A.P., H.H., S.L. and I.B.; investigation, A.P.; writing—original draft preparation, A.P.; writing—review and editing, H.H., S.L. and I.B.; visualization, A.P. and H.H.; supervision, H.H., S.B., S.L. and I.B.; project administration, S.B., S.L. and I.B.; funding acquisition, S.B., S.L. and I.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially funded by ANRT, Grant No. 2017/1076.

Acknowledgments: We would like to gratefully thank Jacques Bosq and Yaelle Harrar for their important medical inputs and feedbacks during this work, along with our colleagues at Keen Eye for their work and support.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CAD	Computer-Aided Diagnosis
MIL	Multiple Instance Learning
WSI	Whole Slide Image

Appendix A. Tile-Level Feature Visualizations

Appendix A.1. Features Selected by CHOWDER Model

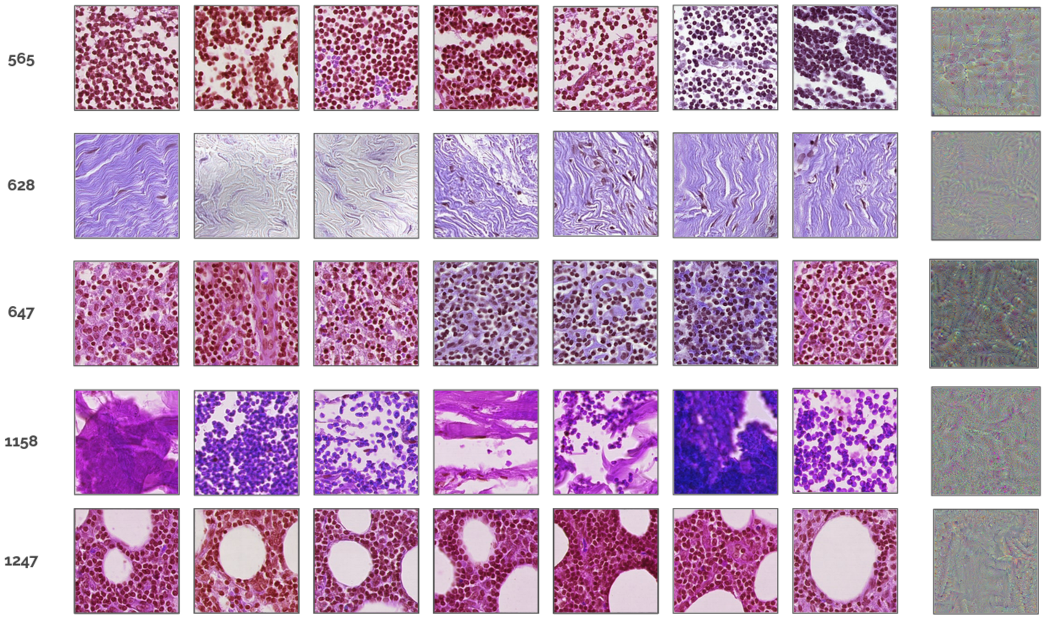

Figure A1. Features selected by CHOWDER model to describe "normal" class.

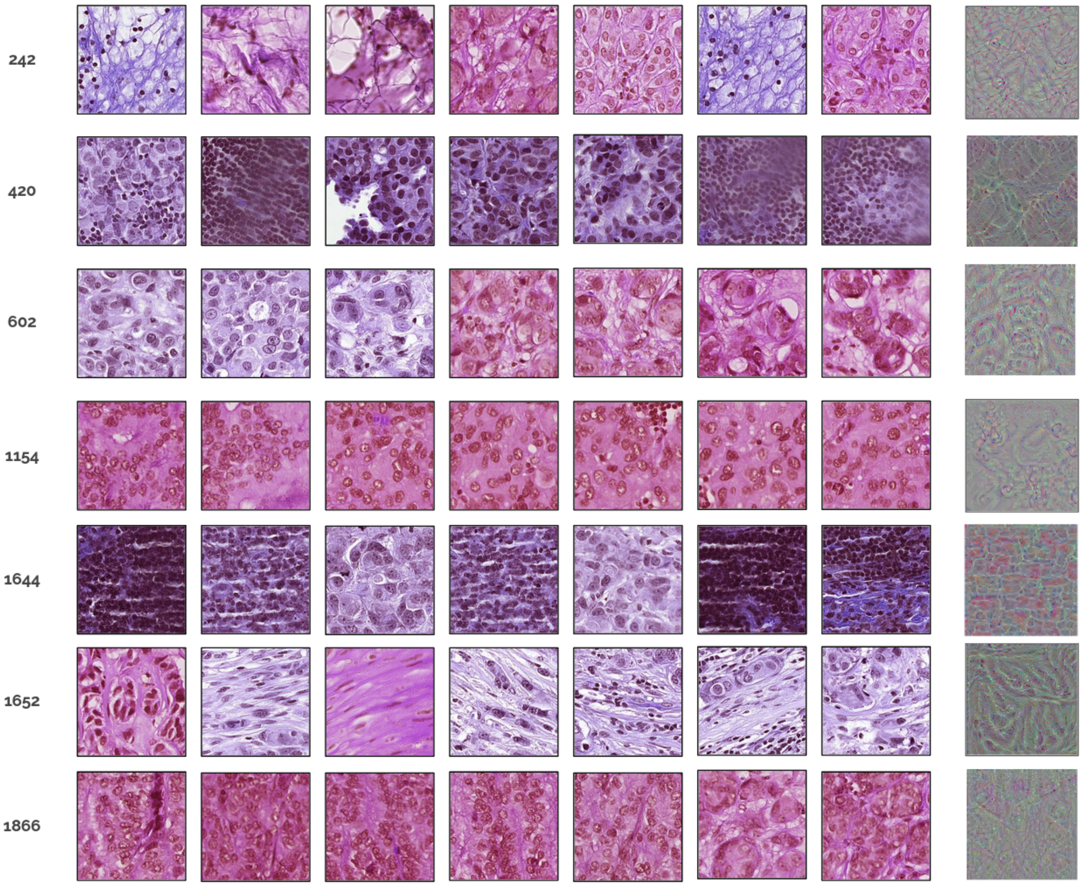

Figure A2. Features selected by CHOWDER model to describe "tumor" class.

Appendix A.2. Features Selected by Attention-Based Model

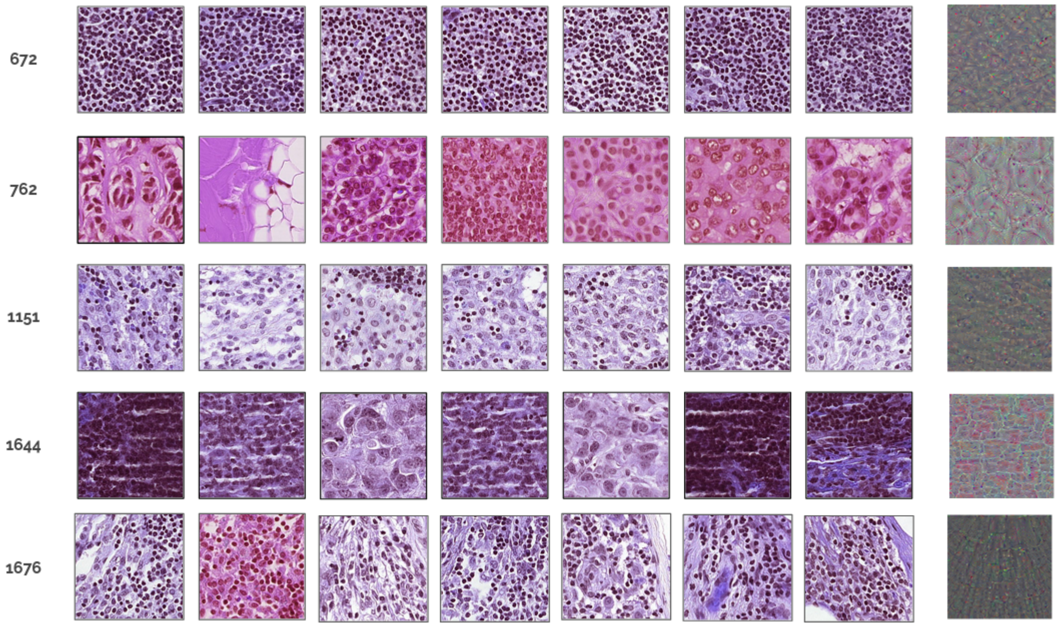

Figure A3. Features selected by Attention-based model to describe "normal" class.

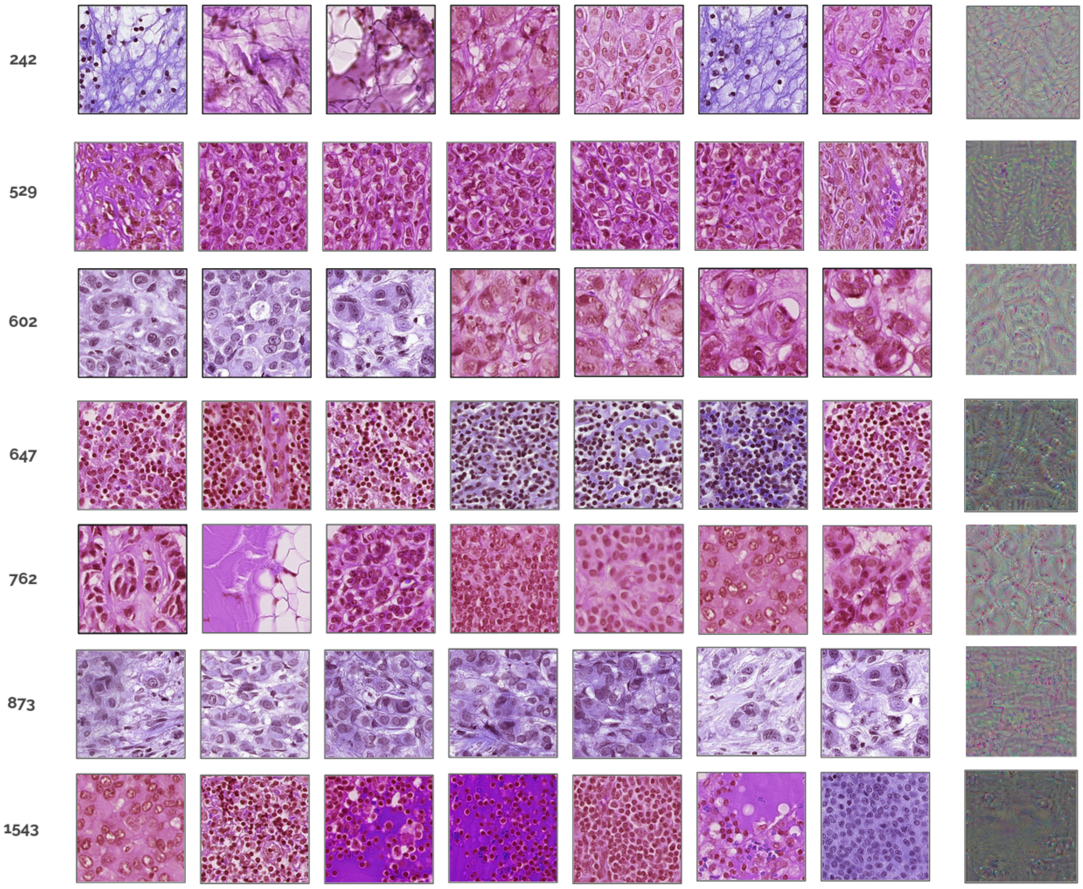

Figure A4. Features selected by Attention-based model to describe "tumor" class.

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing Systems 25 (NIPS 2012), Lake Tahoe, CA, USA, 3–6 December 2012; pp. 1097–1105.
2. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L.ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009.
3. Pasa, F.; Golkov, V.; Cremers, D.; Pfeiffer, F. Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. *Sci. Rep.* **2019**, *9*, 62–68. [CrossRef] [PubMed]
4. Pratt, H.; Coenen, F.; Broadbent, D.M.; Harding, S.P.; Zheng, Y. Convolutional Neural Networks for Diabetic Retinopathy. *Procedia Comput. Sci.* **2016**, *90*, 200–205. [CrossRef]
5. Bejnordi, B.E.; Veta, M.; Diest, P.J.V.; Ginneken, B.V.; Karssemeijer, N.; Litjens, G.; Laak, J.A.W.M.V.D.; Consortium, T.C. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA* **2017**, *312*, 2199–2210. [CrossRef] [PubMed]
6. Tomczak, K.; Czerwińska, P.; Wiznerowicz, M. The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* **2015**, *19*, A68–A77. [CrossRef] [PubMed]
7. Pirovano, A.; Heuberger, H.; Berlemont, S.; Ladjal, S.; Bloch, I. Improving Interpretability for Computer-aided Diagnosis tools on Whole Slide Imaging with Multiple Instance Learning and Gradient-based Explanations. *IMIMIC* **2020**, *12446*, 43–53.
8. Otsu, N. A threshold selection method from gray-level histogram. *IEEE Trans. Syst. Man Cybern.* **1979**, *99*, 62–66. [CrossRef] [CrossRef]
9. Lu, M.Y.; Williamson, D.F.K.; Chen, T.Y.; Chen, R.T.; Barbieri, M.; Mahmood, F. Data Efficient and Weakly Supervised Computational Pathology on Whole Slide Images. *arXiv* **2020**, arXiv:2004.09666.
10. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI* **2015**, *9351*, 234–241.
11. Ianni, J.D.; Soans, R.E.; Sankarapandian, S.; Chamarthi, R.V.; Ayyagari, D.; Olsen, T.G.; Bonham, M.J.; Stavish, C.C.; Motaparthi, K.; Cockerell, C.J.; et al. Tailored for Real-World: A Whole Slide Image Classification System Validated on Uncurated Multi-Site Data Emulating the Prospective Pathology Workload. *Sci. Rep.* **2020**, *10*, 3217. [CrossRef]
12. Ciompi, F.; Geessink, O.; Bejnordi, B.E.; Souza,G.S.d.; Baidoshvili, A.; Litjens, G.; Ginneken,B.v.; Nagtegaal, I.; Laak, J.v.d. The importance of stain normalization in colorectal tissue classification with convolutional networks. In Proceedings of the 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), Laak, Melbourne, Australia, 18–21 April 2017. [CrossRef]
13. Zhou, R.; Hammond, E.H.; Parker, D.L. A multiple wavelength algorithm in color image analysis and its applications in stain decomposition in microscopy images. *Med. Phys.* **1996**, *23*, 1977–1986. [CrossRef] [PubMed]
14. Macenko, M.; Niethammer, M.; Marron, J.; Borland, D.; Woosley, J.T.; Guan, X.; Schmitt, C.; Thomas, N.E. A method for normalizing histology slides for quantitative analysis. In Proceedings of the 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, Boston, MA, USA, 28 June–1 July 2009 . [CrossRef]
15. Khan, A.M.; Rajpoot, N.; Treanor, D.; Magee, D. A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. *IEEE Trans. Biomed. Eng.* **2014**, *61*, 1729–1738. [CrossRef] [PubMed]
16. Bejnordi, B.E.; Litjens, G.; Timofeeva, N.; Otte-Holler, I.; Homeyer, A.; Karssemeijer, N. Laak, J.A.v.d. Stain Specific Standardization of Whole-Slide Histopathological Images. *IEEE Trans. Med. Imaging* **2016**, *35*, 404–415. [CrossRef] [PubMed]
17. Reinhard, E.; Ashikhmin, M.; Gooch, B.; Shirley, P. Color transfer between images. *IEEE Comput. Graph. Appl.* **2001**, *21*, 34–41. [CrossRef]
18. Bel, T.D.; Hermsen, M.; Kers, J.; Laak, J.V.D.; Litjens, G. Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology. In Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning (MIDL), London, UK, 8–10 July 2019.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June 2016–1 July 2016.
20. Naylor, P.; Boyd, J.; Laé, M.; Reyal, F.; Walter, T. Predicting Residual Cancer Burden In A Triple Negative Breast Cancer Cohort. In Proceedings of the IEEE 16th International Symposium on Biomedical Imaging (ISBI), Venice, Italy, 8–11 April 2019; pp. 933–937.
21. Courtiol, P.; Tramel, E.W.; Sanselme, M.; Wainrib, G. Classification and Disease Localization in Histopathology Using Only Global Labels: A Weakly-Supervised Approach. *arXiv* **2018**, arXiv:1802.02212.
22. Campanella, G.; Hanna, M.G.; Geneslaw, L.; Miraflor, A.; Silva, V.W.K.; Busam, K.J.; Brogi, E.; Reuter, V.E.; Klimstra, D.S.; Fuchs, T.J. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **2019**, *25*, 1301–1309. [CrossRef] [PubMed]
23. Campanella, G.; Silva, V.W.K.; Fuchs, T.J. Terabyte-scale Deep Multiple Instance Learning for Classification and Localization in Pathology. *arXiv* **2018**, arXiv:1805.06983.
24. Li, B.; Li, Y.; Eliceiri, K.W. Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning. *arXiv* **2020**, arXiv:2011.08939.
25. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
26. Ilse, M.; Tomczak, J.M.; Welling, M. Attention-based deep multiple instance learning. In Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholmsmässan, Sweden, 10–15 July 2018.

27. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June 2016–1 July 2016.
28. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper With Convolutions. In Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
29. Zhang, Z.; Chen, P.; McGough, M.; Xing, F.; Wang, C.; Bui, M.; Xie, Y.; Sapkota, M.; Cui, L.; Dhillon, J.; et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nat. Mach. Intell.* **2019**, *1*, 236–245. [CrossRef]
30. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
31. Li, J.; Li, W.; Gertych, A.; Knudsen, B.S.; Speier, W.; Arnold, C.W. An attention-based multi-resolution model for prostate whole slide image classification and localization. *arXiv* **2019**, arXiv:1905.13208.
32. Durand, T.; Thome, N.; Cord, M. WELDON: Weakly Supervised Learning of Deep Convolutional Neural Networks. In Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June 2016–1 July 2016
33. Raffel, C.; Ellis, D.P.W. Feed-forward networks with attention can solve some long-term memory problems. *arXiv* **2015**, arXiv:1512.08756.
34. Simoyan, K.; Vedaldi, A.; Zissermn, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv* **2013**, arXiv:1312.6034.
35. Yosinski, J.; Clune, J.; Nguyen, A.M.; Fuchs, T.J.; Lipson, H. Understanding Neural Networks Through Deep Visualization. *arXiv* **2015**, arXiv:1506.06579.
36. Olah, C.; Satyanarayan, A.; Johnson, I.; Carter, S.; Schubert, L.; Ye, K.; Mordvintsev, A. The building blocks of interpretability. *Distill* **2018**, *3*, e10. [CrossRef]
37. Raghu, M.; Zhang, C.; Kleinberg, J.; Bengio, S. Transfusion: Understanding transfer learning for medical imaging. In Proceedings of the 33rd Annual Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; pp. 3342–3352.
38. Couteaux, V.; Nempont, O.; Pizaine, G.; Bloch, I. Towards Interpretability of Segmentation Networks by analyzing DeepDreams. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2019; pp. 56–63.
39. Fong, R.C.; Vedaldi, A. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3449–3457.
40. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 10–15 July 2018; pp. 3319–3328.
41. Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; Wattenberg, M. SmoothGrad: Removing noise by adding noise. *arXiv* **2017**, arXiv:1706.03825.
42. Goh, G.S.W.; Lapuschkin, S.; Weber, L.; Samek, W.; Binder, A. Understanding Integrated Gradients with SmoothTaylor for Deep Neural Network Attribution. *arXiv* **2020**, arXiv:2004.10484.
43. Hooker, S.; Erhan, D.; Kindermans, P.; Kim, B. Evaluating Feature Importance Estimates. *arXiv* **2018**, arXiv:1806.10758.
44. Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity Checks for Saliency Maps. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS), Montréal, QC, Canada, 3–8 December 2018; pp. 9525–9536.
45. Nie, W.; Zhang, Y.; Patel, A. A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations. In Proceedings of the 35th International Conference on Machine Learning (ICML), Stockholmsmässan, Sweden, 10–15 July 2018; pp. 3809–3818.
46. Tan, P.H.; Ellis, I.; Allison, K.; Brogi, E.; Fox, S.B.; Lakhani, S.; Lazar, A.J.; Morris, E.A.; Sahin, A.; Salgado, R.; et al. The 2019 World Health Organization classification of tumours of the breast. *Histopathology* **2020**, *77*, 181–185. [CrossRef] [PubMed]