UNIVERSITÉ DE GENÈVE
Département d'informatique

FACULTÉ DES SCIENCES
Professeur Dr. Stéphane Marchand–Maillet

Service de radiologie

FACULTÉ DE MÉDECINE
Professeur Dr. Henning Müller

Interpretability of Deep Learning for Medical Image Classification: Improved Understandability and Generalization

THÈSE

présentée à la Faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur dès sciences, mention informatique

par

Mara Graziani

de
Nettuno (Italie)

Thèse N° 5622

GENÈVE
2021

La Faculté des sciences, sur le préavis de Monsieur S. MARCHAND-MAILLET, professeur associé et directeur de thèse (Département d'informatique), Monsieur H. MÜLLER, professeur titulaire et codirecteur de thèse (Faculté de médecine, Département de radiologie et informatique et Haute Ecole Spécialisée de Suisse Occidentale, Sierre, Valais, Suisse), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 27 April 2021

Thèse - XXXX -

Le Doyen

Index

Nomenclature

Roman Symbols

$\boldsymbol{A}$	Matrix
$\boldsymbol{a}$	Vector
a	Scalar
$\boldsymbol{W}$	Weight matrix
$\mathcal{D}$	Dataset
$\boldsymbol{X}$	Dataset inputs (matrix with N rows, one for each datapoint)
$\boldsymbol{y}$	Dataset outputs (vector with N elements, one for each datapoint)
$\boldsymbol{x_i}$	Input consisting of the i-th data point
$\hat{\boldsymbol{y}}_{\boldsymbol{i}}$	Model output of the i-th data point
$\boldsymbol{y_i}$	Ground-truth label of the i-th data point
$w \times h \times p$	Dimensionality of the feature maps with width w, height h, and p channels
L	Number of network layers
λ	Learning rate in Gradient Descent (GD)

Number Sets

$\Re$	Real numbers

Other Symbols

$\mathcal{N}$	The Gaussian distribution

Abbreviations

e.g.	Exempli gratia ("for the sake of an example")
i.e.	Id est ("it is")
w.r.t.	With respect to

Notation

The following notation is used throughout the work. Bold lower-case letters denote vectors (e.g. an input image $\boldsymbol{x}$), and standard-weight letters (e.g. its ground-truth label y) denote scalar quantities. We use subscripts to denote either entire rows or columns (with bold letters, $\boldsymbol{x}_i$), or specific elements (x_{ij}). A dataset of input images is identified by standard-weight upper-case letters (e.g. the set of input images X). The model parameters are identified as $\boldsymbol{\theta}$. The feature extraction obtained by a forward pass of an input image $\boldsymbol{x}$ to the network layers is denoted as the transformation function $\phi(\cdot)$, where $\phi^l(\cdot)$ represents the representation obtained at the l-th layer.

Abstract

The application of deep learning to medical imaging tasks has led to exceptional results in several contexts, including the analysis of human tissue samples. Convolutional neural networks (CNNs) constitute a highly performant model, that can almost perfectly detect even the smallest tumor cells in tissue biopsies. These models may have a great potential to support physicians if introduced in the clinical routines.

Despite their impeccable performance on the test sets, CNNs fail in the real-world settings of the clinical workflow, lacking generalization capabilities to unseen data coming from diverse domains. New approaches shall be researched to evaluate whether a model has learned to detect correct patterns and can provide a reliable outcome. Particularly in the medical domain, understanding what are the limitations of a model is a compelling task, to ensure physicians that the model predictions are in line with the standards of clinical practice and can thus be considered in clinical routines. This thesis investigates this task by developing new interpretability techniques, with the aim of making the inner working of deep learning classifiers understandable to physicians and applicable to new inputs.

By narrowing the focus onto microscopy images of breast cancer, my work starts by demonstrating that prior knowledge is a valuable source of input for explaining the model behavior. I introduce information about where the nuclei are located in the images to generate visual explanations that demonstrate that the model predictions are based on the pixels inside the nuclei contours. I then propose a method called Regression Concept Vectors (RCVs) to produce explanations based on the representation of arbitrary concepts that can be obtained as measures directly extracted from the images or annotated by experts. This approach demonstrates that variations of the texture in the images are relevant to the model.

Going beyond the purpose of generating explanations, I directly tackle the generalization deficiencies of existing models. I propose a pruning system that uses RCVs to remove from the model's learning process the undesired behavior of capturing content about unwanted features. As an example, I analyze the removal of the implicitly learned invariance to object scale in models that are pre-trained on natural images, since scale is instead a relevant measure for the analysis of medical images. I then guide the training of CNNs to learn morphology features while discarding the confounding information about data provenance, demonstrating that the resulting model has increased generalization capabilities.

Résumé

La recherche sur le *deep learning* appliqué aux images médicales a conduit à des résultats d'analyse presque parfaits dans de multiples contextes tels que les images de biopsie d'organe. Les réseaux d'apprentissage convolutifs, appelés Convolutional Neural Networks (CNN), constituent un modèle extrêmement performant, qui peut être appliqué pour identifier les plus petites cellules tumorales dans les biopsies d'organes avec une précision exceptionnelle. S'ils étaient introduits dans les routines cliniques, les CNN offriraient un énorme potentiel pour soutenir le personnel médical.

Néanmoins, les capacités des CNN ne sont parfaites que sur les données de tests expérimentaux et échouent dans l'environnement de flux de travail clinique du monde réel. Ces modèles ne s'avèrent pas suffisamment capables d'étendre leur capacité d'analyse à des données nouvells, et donc de généraliser leurs performances sur des données provenant de nouvelles sources, rendant impossible leur application dans les routines de laboratoire. Il existe un besoin de générer de nouvelles approches qui permettent de vérifier si un modèle a réellement appris les motifs corrects dans les données pour donner une réponse fiable. Comprendre les limites des CNN et y remédier est une tâche extrêmement importante pour rassurer le personnel médical que les réponses du processus automatisé sont conformes aux normes de la pratique clinique. Le but de ce travail de thèse est de remplir cette tâche à travers l'investigation et la conception de nouvelles technologies d'interpretabilité, qui peuvent clarifier et expliquer le mécanisme d'apprentissage ainsi que permettre la compréhension des résultats du deep learning par le personnel médical et expliquer, en les documentant, les limites.

En restreignant le champ d'application aux images microscopiques du cancer du sein, mon travail de thèse commence par démontrer que les connaissances préalables sont une source d'information importante pour expliquer ce que le modèle a appris. En introduisant des informations sur les positions des noyaux dans les images, j'ai développé une méthode pour générer des explications visuelles, qui illustrent que la réponse du modèle est basée sur les pixels dans les noyaux. J'ai ensuite développé une approche appelée Regression Concept Vectors (RCV) pour expliquer le modèle à l'aide de mesures représentant un concept spécifique (par exemple la taille des noyaux) pouvant être directement mesuré sur les images ou annoté par des experts. Cette approche démontre que les variations d'apparence, et précisément de texture, sont prépondérantes dans la décision des CNN.

J'ai alors directement analysé le problème de la généralisation. À l'aide de la méthode RCV, j'ai conçu un système de *pruning* qui supprime l'apprentissage indésirable de certaines caractéristiques, telles que l'invariance d'échelle d'un objet, ce qui est important dans l'analyse d'images médicales. J'ai guidé l'apprentissage des CNN pour incorporer des informations sur les variations morphologiques des noyaux et obtenir l'invariance vers l'origine des données. Le modèle résultant augmente ses capacités de généralisation vers des données inconnues.

Riassunto

La ricerca sul *deep learning* applicato ad immagini mediche ha portato verso risultati quasi perfetti in molteplici contesti come, ad esempio, l'analisi microscopica di tessuti patologici. Le reti di apprendimento a base convoluzionale, le cosiddette Convolutional Neural Networks (CNNs), costituiscono un modello estremamente performante, che puó essere applicato per identificare con accuratezza eccezionale le piú piccole cellule tumorali in biopsie di organi. Se introdotte nelle routine cliniche, le CNNs avrebbero un potenziale enorme nel supportare il personale medico.

Ciononostante, le capacitá delle CNNs sono impeccabili soltanto sui dati sperimentali di test e falliscono nell'ambiente reale del flusso di lavoro clinico. Questi modelli non si dimostrano sufficientemente in grado di estendere, e quindi generalizzare, le loro performance su dati provenienti da nuove fonti, rendendo impossibile la loro applicazione nelle routine di laboratorio. Sussiste il bisogno di generare nuovi approcci che permettano di verificare se un modello abbia appreso effettivamente i pattern corretti nei dati per donare una risposta affidabile. Capire le limitazioni delle CNNs é un compito estremamente importante per assicurare al personale medico che le risposte del processo automatico sono in linea con gli standard della pratica clinica. Lo scopo di questo lavoro di tesi é di adempiere a tal compito attraverso lo studio e l'ideazione di nuove tecnologie di *interpretability*, che possano chiarire e spiegare il meccanismo appreso dai modelli di deep learning al personale medico.

Restringendo l'ambito applicativo alle immagini microscopiche di cancro al seno, il mio lavoro di tesi inizia dal dimostrare che la conoscenza pregressa é una fonte d'informazioni importante per spiegare le *features* apprese dal modello. Introducendo informazioni sulle posizioni dei nuclei nelle immagini, ho sviluppato un metodo per generare delle spiegazioni visive, le quali illustrano che la risposta del modello si basa sui pixel all'interno dei nuclei. Ho poi sviluppato un approccio chiamato Regression Concept Vectors (RCV) per spiegare il modello utilizzando delle misure rappresentanti un concetto di tipo arbitrario (ad esempio la dimensione dei nuclei) che possano essere direttamente misurati sulle immagini o annotate da esperti. Questo approccio dimostra che variazioni nell'aspetto, e precisamente nella texture, sono di rilievo nella decisione delle CNNs.

Ho poi analizzato direttamente il problema della generalizzazione. Utilizzando il metodo RCV, ho ideato un sistema di *pruning* che rimuove l'apprendimento indesiderato di alcune caratteristiche, ad esempio l'invarianza alla scala di un oggetto, che e' importante nell'analisi d'immagini mediche. Ho guidato l'apprendimento delle CNNs per inglobare informazioni sulle variazioni di tipo morfologico dei nuclei e per ottenere invarianza verso la provenienza dei dati. Il modello risultante aumenta le sue capacitá di generalizzazione verso dati non noti.

Acknowledgments

None of this work would have been possible without the help of Prof. Henning Müller and Prof. Stéphane Marchand-Maillet. I thank them for their spotless availability, their trust in my capacities and their guidance. For being the busiest and most available person I have ever met, Henning has always found the time to give me advice and support, for which I express my deepest appreciation. By surrounding me with opportunities to visit new research labs, he allowed me to exchange ideas (and ideals) with acknowledged experts and students, which is one of the most beautiful aspects of being an academic.

I may now thank the external committee member, Prof. Mauricio Reyes. His excitement about my work kept me attentive to the needs of real-world clinical problems. I feel lucky for having his influence on my work.

I am also grateful to Prof. Manfredo Aztori and Prof. Barbara Caputo. Their words in a particular time of my life encouraged me to follow the path of scientific research. I still remember Babara's excitement, being my professor at the time, when she lent me her book on Support Vector Machines. She did introduce me, with her knowledge and passion, to this beautiful field.

I could not thank enough Dr. Vincent Andrearczyk for his patient supervision and advice over these years. Our discussions about the methods went on for days in the most outlandish settings: bending our necks under a rock cliff (together with Dr. Cognolato, who may remember some of these occasions), walking in knee-deep snow, hiking amidst the clouds, kayaking over cold splashes of river water. Merciless criticizer and interested listener, he helped me to develop as a researcher, as a better person. All that is yet to come for us gives me unbounded happiness.

It would be ungrateful not to thank Prof. Adrien Depeursinge and Valentin Oreiller, for their help with making my published papers statistically more significant than the initial drafts.

I thank the warm family of Technopôle (that includes several doctors, whose titles are omitted hereafter for simplicity of the notation), including Davide, Sebas, Gaeta, Roger, Oscar, Yashin, Anjani: you colored my days with tropical shades. Grazie Niccolo', for always adding more kilometers and more experiments to run. Gracias Cristina y Alejandra! You are both a source of inspiration about life, relationships, and nutrition. Merci Marie-Helene, for healing my mind and my posture with the ancient tradition of enchaining asanas.

And with all my heart and soul, thanks to mum, dad and Ale. You three have been here with me all the time, despite the distance, despite the difficulties of these years. Your support has given me strength and determination over this long journey.

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation: Deep Learning Classifiers of Medical Images Require Interpretability

"All models are wrong, but some are useful" is a quote that I have often heard in statistics and ML classes. The quote refers to George Box's statement that "all models are wrong" in the Journal of the American Statistical Association of 1976 (Box 1976). Box had the intention of clarifying that the models are approximations based on assumptions, either implicit or explicit, that are never exactly true. The adoption of a model, even if wrong, is justified by its usefulness in describing the properties of a given phenomenon.

I find it rather compelling to understand when a model is good enough to be useful for a given application. Working on medical tasks where mistakes come at a high cost, understanding the limitations of the current models has a high priority. If pitfalls are uncovered, new models can be built to be more reliable and trustworthy than the already existing ones. The model's quality can be evaluated by its performance and generalization to unseen input data, namely the expected value of the model's error on new inputs. Under the simplified conditions of training and testing data being sampled from very similar underlying distributions (i.e. with little domain shift), near-perfect performance was shown by Deep Learning (DL) models in various applications (Gulshan et al. 2016, Giusti et al. 2014), with Convolutional Neural Networks (CNNs) becoming the backbone of numerous state-of-the-art approaches in medical image classification for diagnostic support (Ertosun & Rubin 2015, Wang et al. 2014, Ehteshami Bejnordi et al. 2017, Brown et al. 2018). As retrospective studies have shown, these techniques are not yet ready to generalize to real-world data and clinical workflows (Nagendran et al. 2020, van der Laak et al. 2021, Arvidsson et al. 2018, Kelly et al. 2019), hence this is where existing models need to be improved the most. My concerns about model reliability are, in fact, a relevant and debated issue in this field (Doshi-Velez & Kim 2017, Caruana et al. 2015, Babic et al. 2021). The performance drop is most often due to the poor availability of well-curated, multi-institutional datasets resembling real-world scenarios where data originates from different hospitals, are acquired with multiple protocols and devices. This shift between the real-world conditions of the clinical setting and the simplified ones of the existing training and testing datasets reduces the pertinence of performance as a way to evaluate the usefulness of a model.

The applicability of DL models to clinical settings is consequently surrounded by uncertainty on whether the model performance will be sufficiently reliable for trusting the

algorithm output on the real-world tasks. This particularly worries physicians, who restrain from relying on opaque automated tools and question about when and why is the model expected to fail on their data (Graziani, Marini, Otálora, Ciompi, Aztori, Fragetta & Müller 2021, Tonekaboni et al. 2019). As remarked by Doshi-Velez & Kim (2017), the answer to these types of questions should be sought in a "different approach to evaluating model performance", where the reliability of the automated outcomes is evaluated not only by their testing performance but also by the understandability of the model mechanisms and priorities. Physicians often ignore the processes of feature extraction, model selection and training that is involved in the generation of automated outcomes. If only they were explained on the basis of which features the model can predict a certain output, they would then be able to predict eventual model failures and justify unexpected outcomes. Equating the "correct functioning" of a DL system to high performance on a test set is, therefore, an insufficient definition of the system's purpose and design objective, particularly when this system interacts with, makes decisions about, or has an impact on human lives (Graziani, Dutkiewiczk, Calvaresi, Pereira Amorima, Yordanova, Vered, Nair, Abreu, Blanke, Pulignano, O. Prior, Lauwaert, Reijers, Depeursinge, Andrearczyk & Müller 2021). Doshi-Velez & Kim (2017) defined this fundamental under-specification of the system evaluation as the main limitation towards providing reliable models, highlighting the necessity of new evaluation criteria that include measuring model interpretability.

In this context, interpretability, as formally defined in Section 2.2.1, represents a way to gain an understanding of the underlying mechanics that drive the predictions. The central point of this thesis work is the use of interpretability as a human-centric tool: for the physicians, to improve their understanding of the model and the features used for the prediction; for ML developers, to address the generalization drop; for patients, to improve their acceptance and trust in DL-based tools for diagnostic support.

Figure 1.1 shows the high-level organization of the thesis, which is further described in Section 1.5. Sections 1.2 and 1.3 introduce the main notions of ML and interpretability that will be used in the other chapters. Section 1.4 presents the main research question of this work and the thesis objectives. Section 1.7 reports the entire list of my publications.

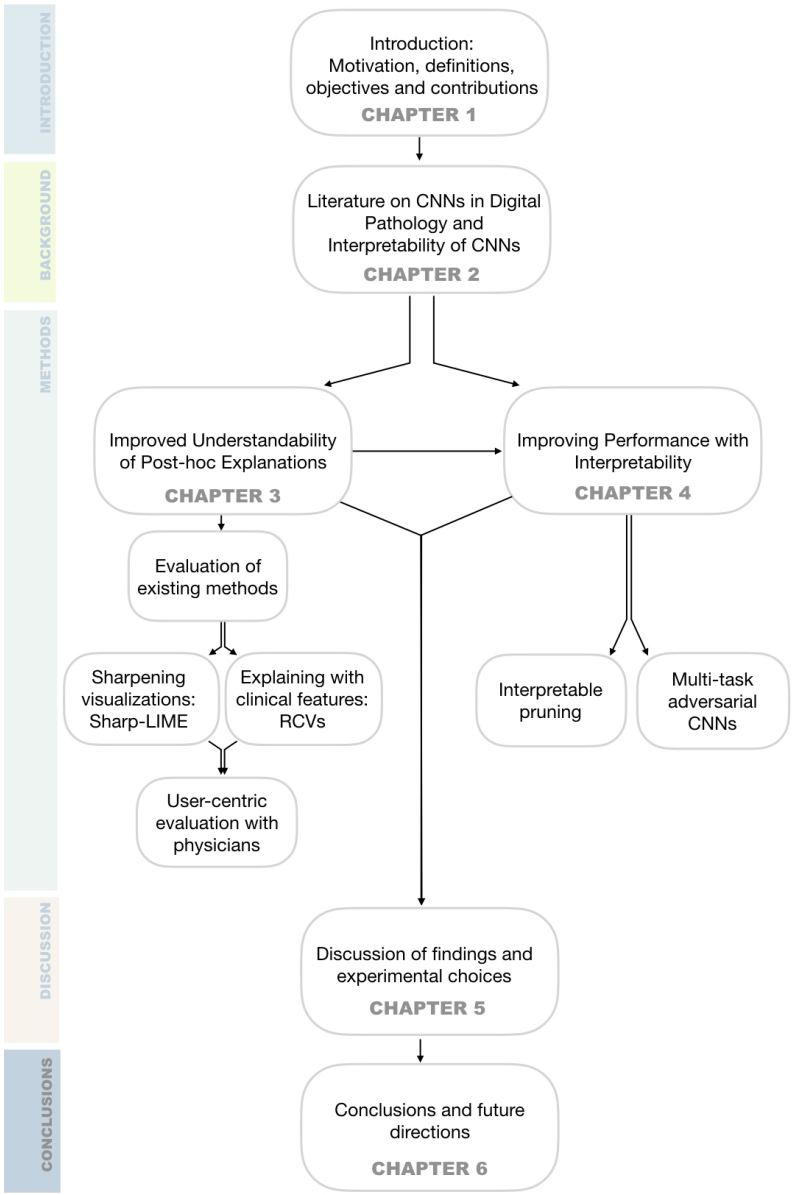


Figure 1.1: Organization of the manuscript. An arrow from one chapter to the other indicates work that was built on top of the considerations in the previous chapters. Among the proposed methods, this work presents the techniques of Sharp Local Interpretable Model-agnostic Explanations (Sharp-LIME) (Graziani, Palatnik de Sousa, B. R. Vellasco, Costa da Silva, Müller & Andrearczyk 2021) and Regression Concept Vectors (RCVs) (Graziani et al. 2018)

1.2 Machine Learning and Deep Neural Networks

This thesis assumes that the reader has experience with ML and DL. To provide self-contained content, however, this section presents the basic concepts in ML and DL that are used throughout the thesis. The many improvements to the methodology, network training and architecture design that have been proposed in recent years, despite being relevant, are not reported in this section because they are not necessary to understand the content of this work. Most of the content in this section summarizes the concepts presented in Goodfellow et al. (2016). Readers willing to deepen their knowledge on these topics may refer to the reference book.

1.2.1 From Statistics to Machine Learning: Linear Regression

Developed in the field of statistics as a way to understand the relationship between two numerical variables, *linear regression* has then been borrowed by ML and is now a prominent approach in this field. Given N input-output pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, we assume that there is a linear function that maps each of the $\boldsymbol{x}_i$ to the labels y_i. The linear regression model is a function of the type:

$$f(\boldsymbol{x}_i) = \boldsymbol{x}_i \boldsymbol{W} + b. \tag{1.1}$$

Where $\boldsymbol{W} \in \Re^{n \times d}$ is the weight matrix, $b \in \Re$ is the bias, and $\boldsymbol{x}_i \in \Re^n$, $y_i \in \Re$. The aim of learning a linear regression model is finding parameters $\boldsymbol{W}, b$ that minimize the error over our observed data, wich can be computed as the average sum of squares $\mathcal{L}(\boldsymbol{x}, y) = \frac{1}{N} \sum_{i=1}^{N} [y_i - (\boldsymbol{x}_i \boldsymbol{W} + b)]^2$.

1.2.2 Deep Feed-forward Neural Networks

Deep Feed-forward Neural Networks (DNNs) are a family of DL models in which the input data flow in a *feed-forward* way. Given a mapping of some input $\boldsymbol{x}$ to a label y (for simplicity, we consider $y \in \{0, 1\}$, although this can easily be scaled to $\boldsymbol{y} \in \Re^d$, with $d \geq 1$), we approximate the mapping with a function $f(\cdot)$ by learning some parameters $\boldsymbol{\theta}$. These networks can have multiple internal layers which introduce a sequence of non-linearities that are typically not observed, the reason for which they are called *hidden layers*. The DNN elaborates the data flowing through each layer in cascading order, with the intermediate computations in the hidden layers being used to compute the prediction $\hat{y}_i = f(\boldsymbol{x_i})$. In these architectures, there are no connections that feed backward $\hat{y}$ into the intermediate layers.

In a network with L hidden layers, an intermediate layer l (with $l < L$) computes a function of the output of the $(l-1)$-th layer. This function computes what is referred to as the pre-activated output as follows.

$$z^l(\boldsymbol{x}; \boldsymbol{\theta}) = \boldsymbol{W}^l h^{l-1}(\boldsymbol{x}; \boldsymbol{\theta}) + \boldsymbol{b}^l, \tag{1.2}$$

where the matrix $\boldsymbol{W}^l$ is the weight matrix of layer l and $\boldsymbol{b}^l$ is the bias. The value $z^l(\boldsymbol{x})$ is then passed through an activation function $\sigma(\cdot)$ to obtain the layer's output:

$$h^l(\boldsymbol{x}; \boldsymbol{\theta}) = \sigma(z^l(\boldsymbol{x}; \boldsymbol{\theta})). \tag{1.3}$$

For a binary classification task, the output layer of the network is a logistic regression function, defined for an input $\boldsymbol{z}^{(l-1)}$ as:

$$f(\boldsymbol{z}^{(l-1)}) = \frac{1}{1 + e^{-\boldsymbol{z}^{(l-1)}}}. \tag{1.4}$$

Training a neural network consists of learning the parameters $\boldsymbol{\theta}$ that minimize a loss function $\mathcal{L}_y$, for example, the Binary Cross-Entropy (BCE) loss:

$$\mathcal{L}_y = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log \hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)) \tag{1.5}$$

The non-linearity introduced by the composition of multiple intermediate layers causes the loss function to become non-convex. The optimization, in this case, needs to be solved by an iterative optimizer that aims at reducing the loss as much as possible. This approach, while making it possible to work in a non-convex landscape, offers no guarantee of convergence, and may be sensitive to parameter initialization. Gradient Descent (GD) is one of the simplest and most used optimization algorithms for DL. The key idea of GD is to follow the gradient of $\mathcal{L}_y$ for the entire training set on a downhill path. The computation of one GD iteration on all input points is, however, very expensive since it requires evaluating the model on the entire dataset. In practice, Stochastic Gradient Descent (SGD) is used to evaluate the gradient on a mini-batch of m input samples drawn from the input data. The small batches provide a regularizing effect and have lower memory requirements than the computation on the full dataset used in GD. The parameter update at iteration τ is given by the following equation:

$$\boldsymbol{\theta}^{\tau} = \boldsymbol{\theta}^{\tau-1} - \lambda \nabla \mathcal{L}_y(\boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^{\tau-1}} \tag{1.6}$$

where $\lambda \in \Re_{\geq 0}$ is the *learning rate* determining the size of the downhill step of the gradient. It is a common practice to gradually decrease the learning rate over time, hence having a value λ_τ that also changes depending on the iteration τ. Several other considerations can be made on the learning rate, for example, the addition of momentum to accelerate the learning processes (Goodfellow et al. 2016). These are not reported in this section for brevity. Where not clearly stated otherwise, SGD with linear weight decay will be used for the experiments.

1.2.3 Convolutional Neural Networks

CNNs (LeCun et al. 1999) are specialized feed-forward networks used to process data with a grid-like topology, e.g. images (2-D grid of pixels). The spatial structure of this data motivated the design of specific DNN architectures to reduce their complexity, i.e. number of parameters, and exploit the translation symmetries of the data.

The convolutional layer is the building block of CNNs, consisting of a set of small trainable filters. Each neuron is locally connected to a small region of the output of the preceding layer, removing the dense connections of DNNs. The spatial extent of this local connectivity of a neuron is a hyper-parameter called the *receptive field* or *filter size* of the neuron. The input to the layer is passed to a mathematical operation called convolution, which replaces the general matrix multiplication in Eq. 1.2. For a bidimensional input, i.e. an image $\boldsymbol{x}$, the input is convolved with a two-dimensional *filter* ($\boldsymbol{w}$) to generate a *feature map* ($(\boldsymbol{x} \circledast \boldsymbol{w})(i, j)$), as in the following:

$$(\boldsymbol{x} \circledast \boldsymbol{w})(i,j) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} \boldsymbol{x}(i-m, i-n)\boldsymbol{w}(m,n), \tag{1.7}$$

where $\boldsymbol{w}(i-m, j-n)$ is the kernel at the pixel location (m, n) and $\boldsymbol{x}(i-m, i-n)$ is the input image at the location $(i-m, i-n)$[1]. The purpose of the filter is to help with detecting features in the input by activating for a certain pattern. A single filter may be useful to detect the same pattern at multiple locations. Based on this assumption, CNN parameters are tied to store a single filter that can be used at any location rather than one filter per location. This particular form of parameter sharing gives the convolutional layer the property of *translation equivariance*[2], meaning that if the input is translated, the output is translated accordingly. The activations at all locations in the feature map are passed to an elementwise non-linear activation function such as a Rectified Linear Activation Unit (RELU) that thresholds the activations above zero: $max(0, \boldsymbol{z})$. An aggregation operation such as max pooling or local average pooling is then used to make the representation approximatively invariant to small translations, hence locally invariant[3].

1.3 Interpretability of Machine Learning

Interpretability is a concept rather complex to define in a unique way. The next section starts from the analysis of the etymology and of existing definitions that I published in Graziani, Dutkiewiczk, Calvaresi, Pereira Amorima, Yordanova, Vered, Nair, Abreu, Blanke, Pulignano, O. Prior, Lauwaert, Reijers, Depeursinge, Andrearczyk & Müller (2021), clarifying the definitions used in this thesis. Sections 1.3.2 discusses the connotation of interpretability as a social relationship of trust. This is used in Section 1.3.3 to further specify the requirements for ML interpretability development in the clinical context.

1.3.1 Etymology and Definitions

A clear and unique definition of terms such as interpretable, explainable, transparent and fair does not yet exist in the context of ML, nor in the broader context of Artificial Intelligence (AI) (Graziani, Dutkiewiczk, Calvaresi, Pereira Amorima, Yordanova, Vered, Nair, Abreu, Blanke, Pulignano, O. Prior, Lauwaert, Reijers, Depeursinge, Andrearczyk & Müller 2021). The terminology used by multiple research groups presents several discordances, particularly when referring to the terms (i) interpretable and explainable, (ii) transparent and decomposable, and (iii) intelligible and interpretable, about which I comment in the following.

The meaning assigned to the words interpretable and explainable (i) emerges as one of the main dividing points in the literature. Several researchers equate these two terms (Miller 2019, Adadi & Berrada 2018, Arya et al. 2019, Clinciu & Hastie 2019, Murdoch et al. 2019). An even larger number of works suggests, however, that most of the academics differentiate interpretability from explainability (Rudin 2019, Lipton 2018, Biran & Cotton 2017, Montavon et al. 2018, Mittelstadt et al. 2019, Chromik & Schuessler 2020, Arrieta et al. 2020, Palacio et al. 2021).

[1] For further explanations, we redirect the reader to Chapter 9 of Goodfellow et al. (2016).
[2] Formally, a function $f(\boldsymbol{x})$ is equivariant to a transformation $g(\cdot)$ if $f(g(\boldsymbol{x})) = g(f(\boldsymbol{x}))$.
[3] A function $f(\boldsymbol{x})$ is invariant to a transformation $g(\cdot)$ if $f(g(\cdot)) = f(\cdot)$

Similarly, transparency (ii) is lightly intended as a synonym of interpretability in some publications (Murdoch et al. 2019, Arrieta et al. 2020), while it is used with the meaning of model decomposability (as defined by Lipton (2018)) in other papers (Clinciu & Hastie 2019, Chromik & Schuessler 2020). As Mittelstadt et al. (2019) explain, transparency can also be seen as understanding the functioning of the model, for example, by acknowledging particular properties such as monotonicity (Rudin 2019, Nguyen & Martínez 2019).

The concept of intelligibility (iii) is equated to inherent interpretability in Arya et al. (2019), while it is used meaning the introduction of interpretability constraints in the model design by Clinciu & Hastie (2019) and Montavon et al. (2018). Acknowledging these main differences is important to understand the points of view of the multiple research groups in this field.

The inconsistencies in the taxonomy caused confusion that led to several unifying papers with the intent of clarifying the approaches Lipton (2018), Arrieta et al. (2020), Montavon et al. (2018), Adadi & Berrada (2018), Arya et al. (2019). The technicalities and implementation details of the interpretability methods have been used to define most of the taxonomy papers. Most works do not consider the perspective of other experts that are also involved in the use of ML: lawyers, sociologists and ethicists. This should be a concern, since using terminology that is understandable and usable solely by the people in ML design may cause having the helpless being led by the clueless (Miller et al. 2017)[4]. In other words, if interpretability is not developed in collaboration with social scientists, there is a high risk of creating AI systems only for other researchers in AI. Section 1.3.2 further dives into this aspect, clarifying the need for a unified perspective from the social and technical sciences.

These considerations drove my preliminary research on the historical formation and the original meaning of the words used in ML interpretability (Graziani, Dutkiewiczk, Calvaresi, Pereira Amorima, Yordanova, Vered, Nair, Abreu, Blanke, Pulignano, O. Prior, Lauwaert, Reijers, Depeursinge, Andrearczyk & Müller 2021). Table 1.1 reports my research on the etymology of the terms, shedding light on their roots, history and intrinsic meaning. From this and the other review works, I derived the definition of interpretability reported in the following:

Key Term 1

A multidisciplinary definition of ML **interpretability** is:
Given a ML system, interpretability constitutes a set of techniques or model properties that make the output generation process of the system explainable and understandable to humans. This can be achieved by introducing interpretability *by design*, i.e. before training the model parameters or by generating post-hoc explanations that do not affect the training of the model parameters. Achieving interpretability is an iterative process that should be adapted to the receiver's requirements. Interpretability analyses should foster the accountability of the system, empowering the user with the information needed to accept or deny the automated outcome.

As underlined by the definition, ML interpretability is strictly connected to the human ability of understanding information. Cognitive psychology describes the process of **understanding** as the ability of the human brain to infer or make predictions within the area of the semantic memory. The semantic memory is wired by connections of neurons

[4]In the original paper, this problem is formulated as that of "the inmates running the asylum".

that are created and consolidated by positive enforcement. A high-level model of such neural connections identifies areas that are specialized for reacting to specific stimuli (e.g. numbers, words, shapes, colors, actions, sounds). Depending on what kind of information is being understood, these areas may be used individually or share functions (Ward 2019).

Key Term 2

The **understandability** of something is here used to identify the property of an object, may this be a model or the outcome of interpretability methods, to be understood by a human. Because the wiring of the neurons constituting the areas in the semantic memory is a result of individual experiences, understandability incorporates some degree of subjectivity and variability, e.g. what is understandable to someone may not be understandable to someone else. The addressees of the interpretability results in this work are mostly physicians without prior knowledge of ML. Understandability here does not require any prior training concerning the feature extraction, hyper-parameter selection and training of ML models. The criteria used to establish whether some information is understandable for clinical use are further discussed in Section 1.3.3. Multiple clinicians are asked to evaluate the understandability of the proposed analyses (see Section 3.5).

The last column of Table 1.1 (ML definition) summarizes the definitions of other terms such as explainability and transparency that I will refer to in the thesis.

1.3.2 Perspective from the Social Sciences

From a sociological perspective, interpretability is a natural requirement that has a parallelism with human decision-making (Coeckelbergh 2020). We expect bankers to explain why they reject a loan, doctors to explain why they discontinue treatment, and politicians to explain why they want to implement a certain policy. Similarly to these human relationships, any AI system should establish a social interaction with its user (Hilton 1990). One of the goals of the interaction should be to help the user improving his *mental model* of the tool, namely his understanding of the system (Hoffman et al. 2018). This interaction has a social connotation since it can be seen as the negotiation of a "social contract of trust" between the human and the system (Graziani, Dutkiewiczk, Calvaresi, Pereira Amorima, Yordanova, Vered, Nair, Abreu, Blanke, Pulignano, O. Prior, Lauwaert, Reijers, Depeursinge, Andrearczyk & Müller 2021). Depending on the mental model, the user defines how much he can rely on the system, deciding when to accept (and refuse) the automated outcomes. In the long term, this reliance transforms into trust and sustained uptake of the system.

The interaction itself is, however, difficult to obtain. Humans and ML systems represent the information in very distinct ways, speaking two different languages. A large part of human reasoning is mostly based on high-level concepts that interact with each other to form a semantic representation. On the contrary, semantic meaning is not directly represented by most ML models. DL, in particular, operates on complex numeric features such as input pixel values, internal activations and weights of intermediate layers (Kim et al. 2018). Conventional metrics of model accuracy, specificity and sensitivity do not suffice to meet the human requirement of gaining understanding and transparency about the automated data processing (Doshi-Velez & Kim 2017). The interpretability analysis shall thus clarify the features considered by the model, helping the user to understand the model priorities when making a prediction. This can only be achieved if the user is

Table 1.1: Etymology of the terms related to interpretability and corresponding definition in the ML domain as in Graziani, Dutkiewiczk, Calvaresi, Pereira Amorima, Yordanova, Vered, Nair, Abreu, Blanke, Pulignano, O. Prior, Lauwaert, Reijers, Depeursinge, Andrearczyk & Müller (2021).

ID	Word	Etymology		ML Definition
1	Interpretability, Interpretable	From late Latin interpretabilitis from Latin interprĕtor, interprĕtāri (to interpret).	To interpret, comment, explain, expose, illustrate, to translate.	To translate, expose, and comment the generation process of one or multiple ML system's outcomes, making the overall process understandable by a human.
2	Explainability, Explainable	From 1600 use of explain + -able adapted from Latin explāno, explānāre (to explain).	To explain, clarify, expose, illustrate, state clearly.	To indicate with precision, to illustrate what features or high-level concepts were used by the ML system to generate predictions for one or multiple inputs.
3	Transparency, Transparent	Medieval Latin adaptation of the words trans (on the other side) and pārĕo, pārēre (to appear, to show).	To see through.	A *transparent* ML system is non-opaque: the roles of the individual components, the learned paradigms, and the overall behavior of the model are known and can be simulated by a human user.
4	Intelligibility, Intelligible	From Latin intelligibilis, intellegibilis (undestandable).	To understand, comprehend, decipher.	An intelligible ML system is an understandable system with inherent interpretability.
5	Accountability, Accountable	From 1770 use of accountable + -ity, adapted from Old French acont derived from Latin compŭto, compŭtāre, which has multiple meanings including to count, to estimate, to judge and to believe.	Used from the 1610s with the sense of "rendering an account", meaning providing a statement answering for conduct.	An accountable ML system is expected to justify its outcomes and behavior.
6	Reliability, Reliable	From Scottish of the 1560s "raliabill", derived from Old French relier a derivation of the latin rĕlĭgo, rĕlĭgāre (to tie, to bind).	From the 1570s used with the sense of to depend, to trust, typically used in the expression "to rely on something/-someone".	To be consistently good and be worthy of trust.
7	Auditability, Auditable	From Latin noun auditŭs, auditūs.	The sense of hearing, the act of hearing, audition. Used in the sense of official audience, judicial hearing or examination.	An "auditable" ML system provides instructions to perform an official audience of the model by extra documentation and functionalities.
8	Liability, liable	From liable, derived from Latin lĭgo, lĭgāre (to tie, to bind).	Legal responsibility for acts.	Legal liability of a product implementing ML, particularly in the case where something goes wrong.
9	Robustness, Robust	From French robuste, derived from Latin robustus, robustum (strong, resistant).	The literal meaning is oaken, made of oak. Used in the figurative sense of strong, vigorous and resistant.	Robust ML systems provide secure and reliable results also in case of adversarial attacks, domain shifts and outliers.

actively considered in the development of interpretability. The following section illustrates how this can be done in the clinical context.

1.3.3 Clinical Requirements for Model Interpretability

Tonekaboni et al. (2019) argue that the application of ML to clinical settings represents a relevant use case for interpretability, motivated by the high stakes, the complexity of the modeling task and the need for reliability. Physicians are the sole people legally accountable for any diagnosis and decision-making, hence accepting ML suggestions is seen as taking an acknowledged risk that may affect the survival and life quality of the patient. Interpretability is also seen the ethical requirement to provide "a factual, direct, and clear explanation of the decision-making process, especially in the event of unwanted consequences" (Floridi et al. 2018, Robbins 2019). Making a mistake may impact strongly the life of the patient, hence the ML application cannot be allowed to take decisions independently, differently from other contexts, e.g. recommendation systems. This sets a major requirement, namely that ML tools for clinical use should aid the diagnosis by interacting with the experts.

This work mainly focuses on the requirements of ML experts and physicians, but there may be other addressees for the explanations. Patients, in the first place, are entitled to an explanation if the physician decides to rely on the automated outcome. Developers may use interpretability for debugging before deployment. Software houses may also be interested in the interpretability analysis to ensure the reliability of their tools before deployment.

The indications of the prospective study ran by Tonekaboni et al. (2019) further confirm the physicians' need of interpretability to justify the clinical decision-making to patients and colleagues. In this study, they identify three requirements: (i) explanations should be appropriate to the clinical task and they should not obfuscate the model behavior by providing redundant information; (ii) explanations should be actionable, namely, they should identify timely and with parsimony the most relevant information that would help physicians making decisions; and (iii) explanations should be consistent to data or parameter shifts that do not modify the model outcome.

In this thesis, I envision a circular life-cycle of automated tools for supporting the diagnosis similar to that depicted in Figure 1.2, where the collaboration between ML developers and physicians is exploited at multiple stages of the development. I argue that physicians should be part, not only of the data collection and annotation stages as in the current practices but also of the model evaluation process. This may be possible thanks to interpretability toolboxes that are understandable to physicians and that can be used to evaluate their reliability on the model. The collected feedback can be used to improve at the same time the degree of understandability of the explanations given by the interpretability toolboxes and the performance of the models.

1.4 Research Questions and Objectives

The previous chapters summarize the basic notions and definitions in the area concerning the interpretability of DL predictions for Medical Image Analysis (MIA) tasks. Based on my research interests, my analysis of the requirements in Sec. 1.3.3 and the literature review in Chapter 2, I identified a research question that is not yet sufficiently covered by the academic literature:

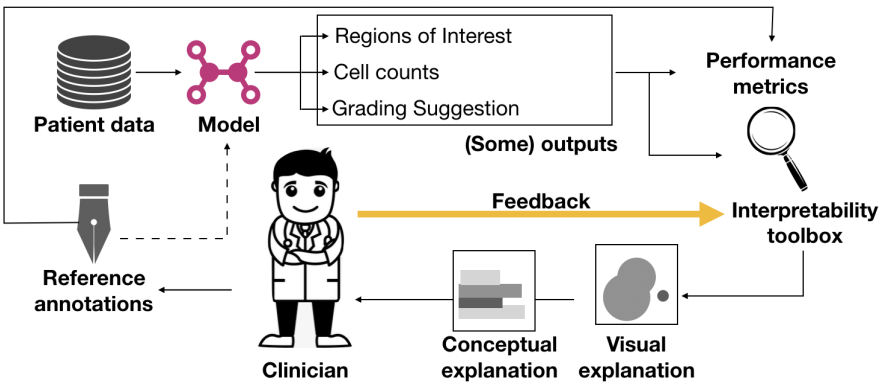


Figure 1.2: A physician-centered software development scheme to achieve improved interpretability and performance of decision support toolboxes for cancer diagnosis in digital pathology.

Main Research Question

Can we make DL models for medical image classification more understandable to physicians? How can this analysis be used to improve model generalization?

This research question has a general focus on medical image classification. For simplicity, this thesis work mostly concerns applications to digital pathology, although it can be extended to other imaging modalities. Particularly, this work applies to the detection of tumor metastases in Whole Slide Images (WSIs) of breast lymph node sections. The wider impact of this work is demonstrated for texture analysis in Graziani, Andrearczyk & Müller (2019), radiology in Yeche et al. (2019) and eye fundus in Graziani, Brown, Andrearczyk, Yildiz, Campbell, Erdogmus, Ioannidis, Chiang, Kalpathy-Cramer & Müller (2019).

The objectives of the thesis are the following:

1. developing new interpretability methods and explanations for DL-based medical image classification that show improved understandability by physicians;

2. developing new DL approaches that use interpretability as a means to improve the quality of the DL models in terms of their performance and generalization.

1.5 Thesis structure

The thesis is structured as follows.

The purpose of Chapter 2 is to introduce the main concepts about digital pathology and to present a literature review of interpretability methods in this context. At the beginning of Chapter 3 (in Section 3.2), I further contribute to the literature review by proposing a quantitative evaluation of the reliability and consistency of the two most frequently used interpretability methodologies in the field, namely Gradient-weighted Class Activation Mapping (Grad-CAM) by Selvaraju et al. (2017) and Locally Interpretable Model-agnostic Explanations (LIME) (Ribeiro et al. 2016). This evaluation contributes to the thesis focus since they show the important limitations of the existing methods in terms of consistency,

understandability and reliability, as also remarked by other studies (Adebayo et al. 2018, Rudin 2019, Babic et al. 2021).

Chapter 3 targets the first objective of the thesis (in Section 1.4), which is generating post-hoc explanations that are more understandable to the physicians. The approaches in these chapters overcome the limitations in Chapter 2 and provide a way to generate more user-centric and user-friendly explanations than traditional methods. The new methods are evaluated by an interactive interface that was developed to collect feedback from the physicians, showing how my vision in Section 1.3.3 (in Figure 1.2) can be translated into concrete practice.

Chapter 4 focuses on the second objective of the thesis, namely improving the model performance and generalization. The method in Section 4.2 uses the interpretability approach developed in Chapter 3 to change the representations learned by the model and preserve scale covariance, improving the performance over the existing baseline. Section 4.3 presents a general framework that allows physicians to guide CNN training by identifying which clinical features should be considered by the model during training and which should be discarded.

The aim of Chapter 5 is to discuss the assets and limitations of the methods presented in this work, together with the potential impact that some of the methodologies may have on the future research scenario, and the software market in digital pathology.

Chapter 7 summarizes the conclusions that should be derived from this work.

1.6 Contributions

The contributions in this thesis are based on some existing approaches in the literature such as Concept Activation Vectors (CAV) (Kim et al. 2018), LIME Ribeiro et al. (2016), Multi-task Learning (MTL) Caruana (1997) and domain adversarial training (Ganin et al. 2016). Some of the works presented in the manuscript are reported from the peer-reviewed and published works where I contributed the most as the first author Graziani et al. (2018), Graziani, Brown, Andrearczyk, Yildiz, Campbell, Erdogmus, Ioannidis, Chiang, Kalpathy-Cramer & Müller (2019), Graziani, Muller & Andrearczyk (2019), Graziani, Andrearczyk & Müller (2019), Graziani, Andrearczyk, Marchand-Maillet & Müller (2020), Graziani, Lompech, Müller & Andrearczyk (2021), Graziani, Palatnik de Sousa, B. R. Vellasco, Costa da Silva, Müller & Andrearczyk (2021). This thesis, however, also contains new pieces of work that I developed that are still undergoing the reviewing process, such as the architecture merging multi-task learning and domain adversarial training presented in Section 4.3.

The most notable contributions of this thesis are:

1. The quantification of the reliability and consistency of existing interpretability tools for digital pathology reported in Section 3.2. The proposed evaluation shows that the existing methods have important limitations, some of which are overcome in Chapters 3 and 4 as the second contribution of this work

2. The development of new post-hoc explainability methods that can be applied to multiple imaging modalities, for which the methods are described in Sections 3.3 and 3.4. The proposed techniques are easier to understand by physicians than the existing ones, as shown by the evaluation with user tests in Section 3.5. They also show improved consistency and reliability.

3. The development of methodologies that build on top of the work in Chapters 2 and 3 to improve the model performance and generalization. In particular, the methodology developed in Section 3.4 is used as a building block to: (i) introduce an interpretable change that preserves scale-covariance in the features learned by a pre-existing CNN architecture for the magnification regression of digital pathology images, as shown in Section 4.2 and (ii) develop a novel CNN architecture for tumor detection in WSI that shows improved generalization to new acquisition centers, as reported in Section 4.3. In both cases, the experimental evidence shows performance improvements over traditional approaches.

The code for the experiments presented in this manuscript is available at https://github.com/maragraziani.

1.7 List of publications

Some of the contributions in this thesis appear in the following publications and preprints.

Chapters 1 and 2 include and adapt work from:

- Graziani, M., Dutkiewicz, L., Calvaresi, D., Pereira Amorim, J., Yordanova, K., Vered, M., Nair, R., Abreu, P. H., Blanke, T., Pulignano, V., O. Prior, J., Lauwaert, L., Reijers, W., Depeursinge, A., Müller, H., Andrearczyk, V. A *Global Taxonomy of Interpretable AI: Unifying the Terminology for the Technical and Social Sciences.* Submitted to Artificial Intelligence Reports.

Chapter 2 also contains work from:

- Graziani, M., Deligand , F., Eggel, I., Bobak, M., Andrearczyk, V., Müller, H. (2020). *Breast Histopathology with High-Performance Computing and Deep Learning.* In Computer and Informatics.

- Graziani, M., Marini, N., Otálora, S., Ciompi, F., Aztori, M., Fragetta, F., Müller, H. *How should AI for Decision Support be Integrated in Fully Digital Pathology Workflows?* In preparation.

Chapter 3 extends the work in:

- Graziani, M., Palatnik de Sousa, I., Vellasco B.R., Marley M., Costa Da Silva, E., Müller, H., Andrearczyk, V. (2021) *Sharpening Local Interpretable Model-agnostic Explanations for Histopathology: Improved Understanding and Reliability.* In Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention, 2021.

- Graziani, M., Lompech, T., Müller, H., Andrearczyk, V. (2021) *Evaluation and Comparison of CNN Visual Explanations for Histopathology.* In Proceedings of the Explainable Agency for AI Workshop at the Association for the Advancement of Artificial Intelligence Conference, 2021.

- Graziani, M., Andrearczyk, V., Marchand-Maillet, S., Müller, H. (2020). *Concept attribution: Explaining CNN decisions to physicians.* In Computers in Biology and Medicine, 103865.

- Graziani, M., Brown, J.M., Andrearczyk, V., Yildiz, V., Campbell, J.P., Erdogmus, D., Ioannidis, S., Chang, M.F., Kalpathy-Cramer, J. Müller, H. (2019) *Improved interpretability for computer-aided severity assessment of retinopathy of prematurity.* In: SPIE Medical Imaging, San Diego, CA, USA, 2019

- Graziani, M., Andrearczyk, V., Müller, H. (2018). *Regression concept vectors for bidirectional explanations in histopathology.* In Understanding and Interpreting Machine Learning in Medical Image Computing Applications (pp. 124-132). Springer, Cham.

Chapter 4 reports the work in:

- Graziani, M., Lompech, T., Müller, H., Depeursinge, A., Andrearczyk, V. (2021). *On the Scale Invariance in State of the Art CNNs Trained on ImageNet.* In Machine Learning and Knowledge Extraction 3 no. 2 (pp. 374-391). Multidisciplinary Digital Publishing Institute

- Graziani, M., Otálora, S., Marchand-Maillet, S., Müller, H., Andrearczyk, V. (2021) *Learning Interpretable Pathology Features by Multi-task and Adversarial Training Improves CNN Generalization.* In review at Nature Machine Intelligence (submitted in July, 2021).

- Graziani, M., Lompech, T., Müller, H., Depeursinge, A., Andrearczyk, V. (2020). *Interpretable CNN Pruning for Preserving Scale-Covariant Features in Medical Imaging.* In Interpretable and Annotation-Efficient Learning for Medical Image Computing (pp. 23-32). Springer, Cham.

Some considerations in the thesis have been taken from the co-authored papers:

- Andrearczyk, V., Graziani, M., Müller, H., Depeursinge, A., (2020) *Consistency of Scale Equivariance in Internal Representations of CNNs.* In Proceedings of the Irish Machine Vision and Image Processing Conference, 2020.

Other publications that were not included in this manuscript are:

- Graziani, M., Andrearczyk, V., Müller, H. (2020). *Visualizing and interpreting feature reuse of pretrained CNNs for histopathology.* In Proceedings of the Irish Machine Vision and Image Processing Conference, 2020.

- Bobak, M., Habala O., Tran, V., Cushing, R., Valkering, O., Belloum, A. S. Z., Graziani, M. and Müller, H., (2020). *Process Data Infrastructure and Data Services.* In Computers And Informatics.

- Bobák, M., Hluchý, L., Habala, O., Tran, V. Cushing, R., Valkering, O., Belloum, A. S. Z., Graziani, M., Müller, H., Madougou, S. and Maassen, J., (2020). *Reference Exascale Architecture - Extended Version.* In Computer And Informatics.

- Graziani, M., Müller, H., Andrearczyk, V. (2019). *Interpreting intentionally flawed models with linear probes.* In Proceedings of the IEEE International Conference on Computer Vision Workshops.

- Ladislav Hluchy, Martin Bobák, Henning Müller, Mara Graziani, Jason Maassen, Hanno Spreeuw, Matti Heikkurinen, Jörg Pancake-Steeg, Stefan Spahr, Nils Otto vor dem Gentschen Felde, Maximilian Höb, Jan Schmidt, Adam S. Z. Belloum, Reginald Cushing, Piotr Nowakowski, Jan Meizner, Katarzyna Rycerz, Bartosz Wilk, Marian Bubak, Ondrej Habala, Martin Seleng, Stefan Dlugolinsky, Viet Tran and Giang Nguyen, *Heterogeneous exascale computing*, in: INES 2018 conference, Springer, 2019.

- Martin Bobák, Ladislav Hluchý, Mara Graziani and Henning Müller, *Machine Learning in Medical Imaging*, Radenci, Slovenia, 2019.

Finally, during this time of my Ph.D. work I had the pleasure to co-supervise the following works by M.Sc. and B.Eng. students:

- *Développement d'une interface web pour l'évaluation du diagnostic assisté par ordinateur (CADeval).* Thesis winning the Best Bachelor Thesis Award in 2021, written by Nicolas Costantin for the degree of Bachelor in Management Information Technology at the Haute Ecole de Gestion of the University of Applied Sciences of Western Switzerland (HES-SO Valais), Sierre, Switzerland.

- *Evaluating concept-based guidance of CNN training for breast histopathology.* Report of the internship work written by Priscille-Anne Tavernier for the degree of M.Sc. in Electronics, Signal Processing and Artificial Intelligence at Ecole National Supérieure de l'Electronique et de ses Applications (ENSEA), Cergy, France

- *Analysis of scale covariance in CNNs pre-trained on ImageNet* and *Evaluation of visual explainability methods for histopathology.* Internship works done respectively in 2019 and 2020 by Thomas Lompech for the degree of M.Sc. in Sciences du numérique et Image et Multimédia at the Ecole Nationale Supérieure d'Electrotechnique, d'Electronique, d'Informatique, d'Hydraulique et des Télécommunications (ENSEEIHT), Toulouse, France.

- *Dense-coverage of WSI patch extraction: limitations and challenges.* Internship work done in 2020 by Francois Deligand for his M.Sc. in Sciences du numérique et Image et Multimédia at the Ecole Nationale Supérieure d'Electrotechnique, d'Electronique, d'Informatique, d'Hydraulique et des Télécommunications (ENSEEIHT), Toulouse, France.

Chapter 2

Interpretable Deep Learning for Digital Pathology

2.1 Convolutional Neural Networks for Digital Pathology

The methods in this thesis are presented for the specific application to digital pathology tasks. Where not clearly stated otherwise, the task considered is the detection of Breast Cancer Metastases in Lymph Nodes (BCMLN). With an estimated number of affected women worldwide of 271,270 in 2018 and an increasing number of women dying from this disease (2.8 % increase from 2017 with 41,760 estimated deaths), breast cancer is the second leading cause of cancer death among women (Siegel et al. 2019, Ehteshami Bejnordi 2017). Being the most likely target for initial metastases, axillary lymph nodes are analyzed to determine the spreading stage to neighboring areas. The traditional workflow to diagnose BCMLN is the same as that for the diagnosis of breast cancer, where a tissue sample is carefully inspected at the microscope by a pathologist. Importantly, traditional microscopes are increasingly being replaced by digitalized approaches (Fraggetta et al. 2017, Stathonikos et al. 2013, Griffin & Treanor 2017), with institutions transitioning to fully digital workflows as that in Figure 2.1.

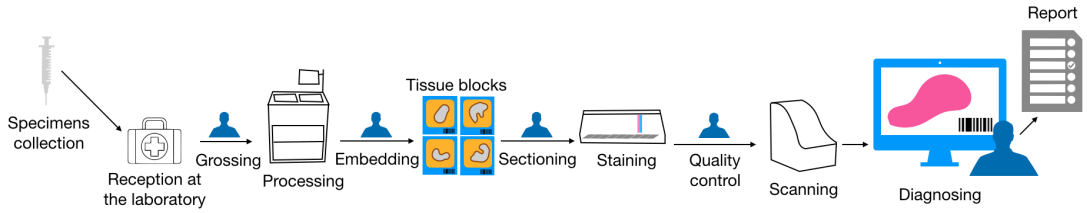


Figure 2.1: Digital workflow. Illustration adapted from Graziani, Marini, Otálora, Ciompi, Aztori, Fragetta & Müller (2021).

In this workflow, the collection of the specimens and shipment to the laboratory are the initial steps. After grossing and processing, fixation is performed to preserve the tissue, which is embedded into paraffin, and cut into thin slices (i.e. sectioning) mounted onto glass slides. Most of these operations are now automatized with minimal user interaction needs[5]. The slices are then stained with different reactors to identify tissue structures and cellular features. Hematoxylin and Eosin (H&E) is the most common staining procedure.

[5] An example can be found in the Leica Automated Tissue Processor `leica-microsystems.com`

Hematoxylin highlights the nucleus cytoplasm, membrane, and chromatin patterns. Eosin produces a tri-tonal staining effect shading epithelial cell cytoplasm with deep magenta, collagen with light pink and nucleoli with purple. Other staining techniques that are not analyzed in this work are Immunohistochemistry (IHC) and In-Situ Hybridization (ISH).

The stained slides are passed through high-resolution slide scanners, which capture digital images of the slide at the micron level (up to 160nm per pixel). WSIs can show cellular details as those shown by a microscope and are stored in a pyramidal structure with intermediate layers being down-sampled versions of the original image. The use of digital slides opens a broad range of new possibilities and a wide set of functionalities that may assist clinicians in their daily routines (Griffin & Treanor 2017). Most importantly, it sets a solid base for developing CNNs that learn patterns from the image archives (Ilse et al. 2020, Zhang et al. 2019, Gurcan et al. 2009, Janowczyk & Madabhushi 2016, Litjens et al. 2017, Ehteshami Bejnordi et al. 2017).

Before introducing automated approaches, it is important to understand the features that pathologists consider to determine the severity, the type of cancer (i.e. ductal, lobular and in-situ or invasive) and the prognosis. Elston & Ellis (1991) showed that tumor grade is an important prognostic indicator, representing the aggressive power of the tumor. This is assessed by looking at the three factors illustrated in Figure 2.2, namely (i) the formation of glands, (ii) the degree of nuclear pleomorphism, and (iii) the mitotic rate. The diagnosis process is time-consuming and error-prone, with the rate of over-looking small metastases higher than 60% (Van Diest et al. 2010). The grading of the tumor severity is also very subjective, reporting high inter-observer variability (Ehteshami Bejnordi et al. 2017).

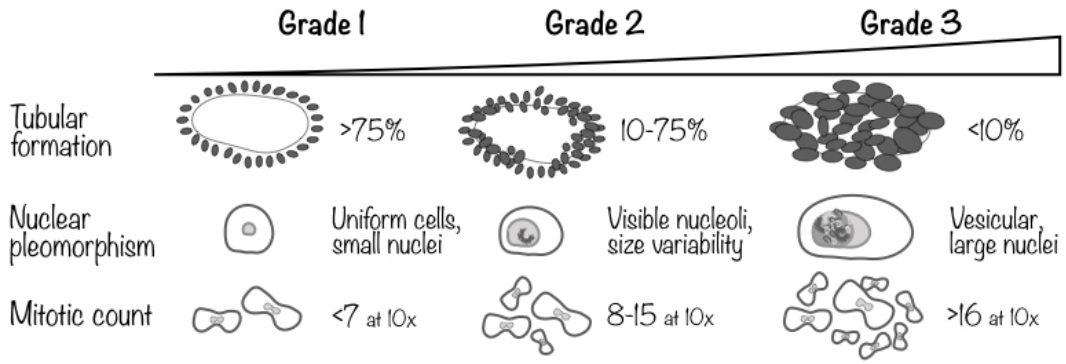
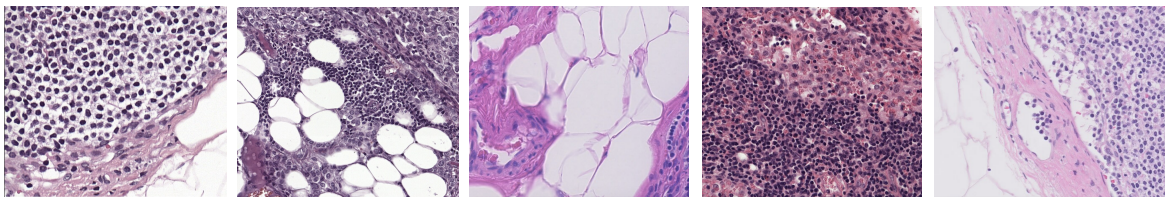


Figure 2.2: Prognostic indicators used for tumor grading. Adapted from `pathology.jhu.edu/breast/staging-grade/`, last access July 2020.

AI-based approaches may help to overcome the shortcomings of subjective evaluation, for example by identifying malignant areas, by providing objective measures that characterize the tissue structure and by giving diagnosis and prognosis suggestions. The detection and segmentation of tumor regions are two very common tasks, for which DL models and in particular CNNs are the most frequently chosen approaches (Litjens et al. 2017, Bera et al. 2019, Campanella et al. 2019). Training CNNs on pathology images presents multiple challenges, which are described by several reviews on the matter (Janowczyk & Madabhushi 2016, Campanella et al. 2019, Litjens et al. 2017, Gurcan et al. 2009). WSIs contain hundreds of thousands of pixels, and CNN training requires hundreds of them with local annotations of the tumor. Generating annotations of tumor contours is, besides, a tedious and expensive process for pathologists, that rarely reaches pixel-level precision (Janowczyk



(a) Center 0 (b) Center 1 (c) Center 2 (d) Center 3 (e) Center 4

Figure 2.3: Examples of staining variability within the Camelyon dataset.

& Madabhushi 2016). The existing datasets have instance- or patient-level annotations, while pixel-level annotations are provided for only a few collections. For the case of breast cancer, Camelyon[6] is a well-curated collection providing more than 300 WSIs with pixel-level annotations. In addition, the learning process is hindered by the heterogeneity of the data, coming from differing staining, fixation and slicing procedures, multiple scanner resolutions, and the possible presence of artifacts in the images. As a result, decreases in performance are observed when testing models on data collected from a different institutions (Tellez et al. 2019). An example of data heterogeneity in Camelyon is shown in Figure 2.3, with WSIs taken from multiple acquisition centers.

Increasingly sophisticated CNNs-based approaches have been developed over the past years, outperforming traditional ML (Bera et al. 2019, Janowczyk & Madabhushi 2016). Transfer learning uses CNNs as a means of feature extraction. In this case, the network weights learned on datasets of natural images such as ImageNet (Deng et al. 2009) are reused to extract the activations obtained with histopathology inputs. The deep features are then used to train a linear classifier as in Xu et al. (2017). Alternatively, CNN weights pre-trained on ImageNet are fine-tuned on pathology images to refine the feature extraction process. The fine-tuning of a standard CNN architecture on pathology images was proven by Mormont et al. (2018) to lead to better results than transfer learning and it thus constitutes the backbone of recent approaches. Architectures such as Residual Neural Network (ResNet) (He et al. 2016*a*) and Inception (Szegedy et al. 2016) are employed in multiple papers (Liu et al. 2017, Ehteshami Bejnordi et al. 2017, Gamper, Kooohbanani & Rajpoot 2020, Kandel & Castelli 2020). The evaluation study by Ehteshami Bejnordi (2017), in particular, compares the performance of multiple CNN-based pipelines for BCMLN detection to pathologist performance. Multiple approaches deal with the scarcity of pixel-level annotations. Ilse et al. propose multiple-instance learning to handle digital slides as a bag of words, exploiting weak instance-level annotations to train the detection of tumorous regions also on slides without pixel-level annotations (Ilse et al. 2018, 2020). In teacher-student designs, a teacher model is trained on the pixel-level annotations and a student model learns to generalize on the weakly annotated data (Cheng et al. 2020). Some methods directly address the staining variability causing the data heterogeneity. Tellez et al. (2018) show that H&E staining augmentation can improve generalization to unseen acquisition centers. Invariant designs are also proposed in Lafarge et al. (2017), Otálora et al. (2019) as a means to obtaining robustness to staining variability, and in Veeling et al. (2018) to induce rotation invariance. In addition to CNNs, Generative Adversarial Networks (GANs) are increasingly used as digital pathology applications. The work in (Xu et al. 2019) proposes a GAN-based approach to virtually re-stain slides, for example, converting H&E into IHC.

[6] https://camelyon17.grand-challenge.org/, accessed August 2021

The pipeline used in this work presents a standard approach to train CNNs to detect tumorous regions, also called Region Of Interest (ROI), for BCMLN. The performance of multiple approaches for this task was assessed and compared to pathologist performance by Ehteshami Bejnordi et al. (2017). In the context of the PROCESS project, I compared multiple network architectures and their performance depending on the resource availability. From my analyses on CNN fine-tuning in Graziani, Eggel, Andrearczyk et al. (2020), the Inception V3 (Szegedy et al. 2016) backbone led to the best performances and I thus chose this model as a starting point for interpretability analyses. The next section reviews the main interpretability approaches in the field, justifying my attention to post-hoc explanations in Chapters 3 and 4.

2.2 Interpreting Deep Learning Models

This section reviews the literature on interpretability techniques, that counts until 2020 more than 70,000 papers containing either "XAI", "explainability", or "interpretability"[7]. An exhaustive and complete overview of these works is not the main purpose of this section, which rather aims at analyzing the main differences among the techniques. Section 2.2.1 reviews interpretability methods for CNN models and visual inputs. Some of the methods mentioned in this section are, however, not restricted to explaining only CNNs. Section 2.2.2 presents the applications in the context of digital pathology, reporting the main results and insights in the literature.

2.2.1 Interpretability of CNNs for Visual Inputs

An intuitive categorization of interpretability methodologies is the one given by Montavon et al. (2018). The authors identify three "dimensions" of interpretability as represented in Figure 2.4, along which the existing interpretability methods can be clustered together. The three clusters differ from each other depending on the object that is being inter-

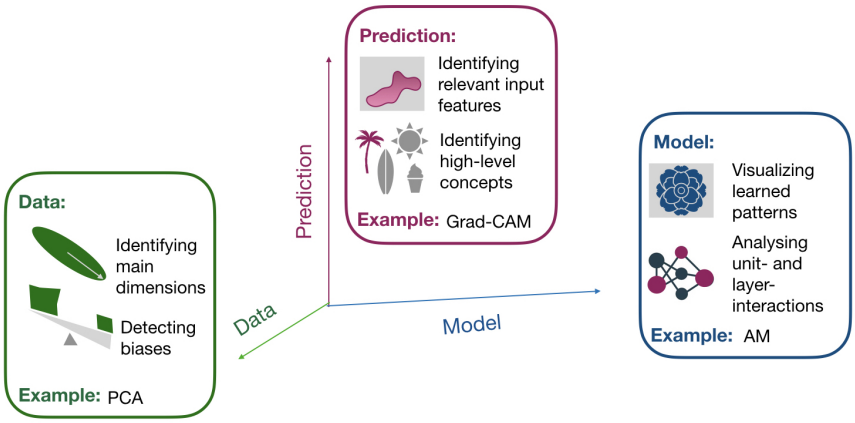


Figure 2.4: The three dimensions of interpretability. Inspired by the review in Montavon et al. (2018). In the examples, PCA refers to Principal Component Analysis, Grad-CAM to the work on Gradient-weighted Class Activation Mapping by Selvaraju et al. (2017) and AM stands for Activation Maximization by Erhan et al. (2009).

[7] According to `app.dimensions.ai` as accessed in August 2021.

preted, thus on the scope of the interpretability analysis. The interpretability methods gathering along the first dimension (shown in green in Figure 2.4), aim at interpreting the data. They focus on identifying the most informative dimensions for the task being solved and highlighting possible biases. Dimensionality reduction techniques such as the classic Principal Component Analysis (PCA) (Pearson 1901) and the more recent Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018) are examples of techniques that gather on this axis. The second dimension identifies another object of the interpretability analysis, which is the prediction. Approaches on this line attempt at explaining why a certain input led to a given output. Later in this section, I will discuss how this objective can be obtained in multiple ways leading to further groupings, for example of methods that identify relevant input areas (Lapuschkin et al. 2015, Selvaraju et al. 2017) or that show what high-level concepts are used to reach the prediction (Kim et al. 2018). The third dimension concerns the interpretation of the model, including approaches that aim to explain how the model behaves and how the layers interact with each other. Examples of techniques in this category are the Activation Maximization (AM) in the technical report written by Erhan et al. (2009), feature visualization approach developed by Olah et al. (2017) and the network dissection method in Bau et al. (2017).

The grouping in Figure 2.4 gives an intuitive map of how the multiple approaches can be organized on a three-dimensional space. Notwithstanding, this classification does not consider important differences in the implementation of the techniques that have been used to separate the methods in sub-categories. As Camburu (2020) explains in her Ph.D. dissertation, multiple groupings of the methods exist, and displaying all of them would only be overwhelming for the reader. For this reason, we only focus here on the key distinctions between the methods and we point the reader to the review by Arrieta et al. (2020) for more information.

Apart from the object of the analysis (i.e. the data, the model or the outcome), methods can be grouped into built-in and post-hoc interpretability. Post-hoc methods can be further separated based on two factors, namely the level of opacity that they can deal with (i.e. model-agnostic against model-dependent methods) and the form of the outcome (e.g. feature-based, concept-based explanations, among others). Finally, the granularity of the interpretability analysis is also an important parameter, distinguishing local against global interpretability. I explain these points in detail in the next paragraphs.

Built-in interpretability Built-in interpretability aims at building models that are interpretable by construction. This can be obtained by following two paths: (i) developing models with a transparent design and inherent interpretability (ii) adding a self-explanatory module that generates explanations for the model predictions. In (i), transparency can be introduced in model design in multiple ways. Introducing parameter sparsity constraints is one method to identify relevant features (Li et al. 2019). Transparency may also be obtained by adopting functions that have intelligible properties, e.g. monotonicity (Nguyen & Martínez 2019). Alternatively, interpretability constraints can be added to the optimization objectives. The interpretable decision sets by Lakkaraju et al. (2016) illustrate how interpretability constraints can be defined in terms of rule parsimony, non-redundancy and class coverage. As for built-in interpretability for DL, Cyntia Rudin advocates the importance of learning interpretable intermediate features (Rudin 2019). Her co-authored work in Chen et al. (2020), for instance, proposes the concept-whitening transformation to align the axes of intermediate layers of a CNN with predefined concepts. In (ii), an explanation generator is added to the DL architecture to obtain a

self-explanatory design (Camburu 2020). Self-explanatory models include as an additional objective the generation of an explanation for the predictions. Lei et al. (2016) introduce a module that selects subsets of the input features before these are passed to the network to compute the prediction. The network outcome is based only on the input features subset, which hence corresponds to an explanation for the outcome. Other self-explanatory models generate multi-modal explanations, e.g. textual description for visual inputs. Differently from the work in Lei et al. (2016), these models require additional supervision for the explanations (Park et al. 2018).

Post-hoc explanations Introduced by Ribeiro et al. (2016), the term *post-hoc* refers to methods that aim at explaining already trained models. The main advantage of post-hoc methods is that they do not alter the network training, hence they do not compromise the predictive performance for interpretability. Besides, these methods can be applied to most existing models that have been already used over the past years, such as Inception (Szegedy et al. 2016) and ResNet (He et al. 2016*a*). An important distinction to make here is between model-agnostic and model-dependent models. Model-agnostic methods do not need any access to the internal model's logic and/or state (e.g. model parameters), and only rely on the input and output pairs. They consider the model to be interpreted as a black box where only the output for a given input is observable. As a result, model-agnostic methods can be applied to all models. Perturbation methods such as occlusion (Zeiler & Fergus 2014), LIME (Ribeiro et al. 2016) and SHapley Additive exPlanations (SHAP) (Lundberg & Lee 2017) are model-agnostic. Post-hoc approaches can be further grouped depending on the form of the generated explanations into (i) feature attribution, (ii) feature visualization, (iii) concept attribution and (iv) surrogate explanations. Two additional strategies that are not related to the context of this thesis are case-based and textual explanations, for which we refer the reader to Arrieta et al. (2020) for further information. Figure 2.5 illustrates examples of the categories that are described in the following.

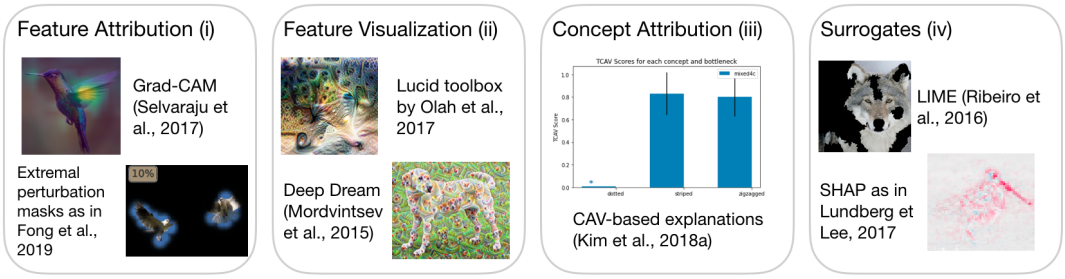


Figure 2.5: Classification of post-hoc interpretability methods.

Feature attribution methods (i) aim at identifying the input features that are the most relevant to the prediction. Referring to Miller's *why questions* Miller (2019) about the model outcome, feature attribution methods mainly answer to questions of the type "*What would the model output be if the value of this input feature was changed?*". Perturbation methods, first introduced by Zeiler & Fergus (2014) with occlusion sensitivity, look at the impact of input feature perturbations on the model output. This approach is further developed by Fong into finding extremal perturbations: minimal regions that most modify a layer's activation (Fong et al. 2019). The majority of feature attribution methods evaluate the impact of perturbations by accessing the model gradients. Perturbation

sensitivity with saliency maps is now a reference approach, computing the gradient of the class of interest w.r.t. the input pixels Simonyan et al. (2014). Saliency maps have some important limitations due to the instability and saturation of the gradients, which were addressed by the gradient backpropagation method by Springenberg et al. (2015) and Shrikumar's Deep Learning Important FeaTures (Deep-LIFT) Shrikumar et al. (2017). The contribution scores of each input pixel are assigned by Deep-LIFT by looking at the difference of the activations w.r.t. a reference activation. Integrated gradients developed by Sundararajan et al. (2017) is another method that compares the activations to those obtained with an input reference. In this method, the gradients are accumulated for all perturbations, computing the integral of the gradients along the straight-line path that goes from the reference input to the modified input. The work on Layer-wise Relevance Propagation (LRP) shows that accurate explanations are obtained by replacing the gradients with their Taylor approximation and by evaluating pixel relevance in a layer-wise manner (Lapuschkin et al. 2015, Montavon et al. 2017). In his work on Class Activation Mapping (CAM), Zhou shows that a GAP operation before the decision layer can be used to project the weighted relevance of individual feature maps back onto the input image (Zhou et al. 2016). This method is extended into Grad-CAM by using the gradients as the weighting function of the feature maps (Selvaraju et al. 2017). In the same work, the authors prove that Grad-CAM is a generalization of CAM that gives the same result up to a normalization constant. Model-agnostic feature attribution methods such as LIME (Ribeiro et al. 2016) and SHAP values (Lundberg & Lee 2017) can also be used to identify relevant input features. Since these methods use surrogate models, more details will be provided in the appropriate section. All these feature attribution methods generate, for image inputs, visual explanations in the form of heatmaps (Lapuschkin et al. 2015, Montavon et al. 2017, Zhou et al. 2016, Selvaraju et al. 2017, Sundararajan et al. 2017, Springenberg et al. 2015, Fong & Vedaldi 2017, Simonyan et al. 2014, Zeiler & Fergus 2014). Multiple works evaluated the reliability of these visualizations, pointing to important limitations that urge for improvement. A simple operation such as adding a constant shift to the input data leads to diverse, and misleading, pixel saliency values (Kindermans et al. 2019). The sanity checks proposed by Adebayo et al. (2018) provide additional reasons to question the reliability of gradient-based methods to generate explanations. In the experiments, the authors randomize progressively the network parameters to evaluate the impact on the explanations of breaking the patterns learned during training. Eight gradient-based approaches are considered, including integrated gradients, saliency and Grad-CAM. The results show that, except for Grad-CAM, the explanations are insensitive to the randomization, suggesting that the explanations are not representative enough of the model's behavior. The Grad-CAM method, despite passing the sanity checks, presents other limitations. It is argued by Chattopadhay et al. (2018) that Grad-CAM cannot well explain occurrences of multiple object instances in a single image, the reason for which the authors propose an improved version called Generalized Grad-CAM++ (Grad-CAM++). As I further explain in Section 2.2.2, this is particularly important in digital pathology since the images contain multiple repetitions of cell instances.

Feature visualization (ii) aims at uncovering the patterns that are learned by intermediate layers and units. Proposed already in a technical report of 2009 written by Erhan et al. (2009), AM starts from an input of random noise and optimizes each pixel value to obtain a pattern that maximally activates a single unit. This method is further developed by Olah et al. (2017), who evaluated the impact of solving the optimization problem in the Fourier domain and of adding regularization in terms of transformation robustness

and frequency penalization. Approaches on the same line are the inverted representations in Mahendran & Vedaldi (2015) and DeepDream (Mordvintsev et al. 2015). Deconvolutions also constitute an early approach to visualize internal activations, obtained by inverting the convolution operation (Zeiler & Fergus 2014, Xu et al. 2014).

The idea of concept attribution (iii) was firstly proposed by Kim et al. (2018) with CAVs. The key idea in this approach is the learning of concepts in the internal activations of a layer. This is done by solving a binary classification problem that aims at distinguishing images that show examples of a concept from random images without the concept. A linear classification model was used for this work, following the idea of linear classifier probes introduced by Alain & Bengio (2016). The performance of the linear classifier is indicative of how well the concept is learned in the network representation. In principle, concepts are defined arbitrarily in the form of queries to interpret the model (Kim et al. 2018). Ghorbani et al. (2019) remove the need for concept queries by performing an unsupervised search of concepts in the latent space of a network. Goyal et al. (2019) further extend the analysis with CAVs to the evaluation of concept-based causal relationships. This work is, however, at an early stage with synthetic data and it is only published as a pre-print work. Last in the list, surrogate explanations (iv) group all the techniques that require the creation of a surrogate model to generate explanations. A proxy model, e.g. a rule-list or a linear classifier, is trained to approximate the DL decision function. In (Chakraborty et al. 2020), for example, rule-lists are used as a proxy. In LIME, a linear model is used as a proxy, which is trained in a neighborhood of the input perturbations around the decision boundary.

Granularity of the analysis The granularity at which the analysis is made differentiates local from global explanations (Lipton 2018). Instance-wise, *local* explanations are given by methods that analyze the prediction of a single input at a time. Several feature attribution methods provide local explanations (Ribeiro et al. 2016, Lundberg & Lee 2017, Simonyan et al. 2014, Montavon et al. 2017, Zhou et al. 2016, Selvaraju et al. 2017, Sundararajan et al. 2017, Fong & Vedaldi 2017, Lapuschkin et al. 2015). On the contrary, when the inner working principles of the entire model or the behavior for a class or multiple classes are explained, the outcome is a *global* analysis. For example, the global behavior of a DL model is explained by its distillation into a soft decision tree in Frosst & Hinton (2017) and by the extraction of rule-lists in Chakraborty et al. (2020).

2.2.2 Applications to Digital Pathology

This section reviews the applications of interpretability methods to digital pathology, highlighting the main features and findings of the existing works. While only a few works focus on directly learning interpretable features (Diao et al. 2021), multiple papers show in the form of visual heatmaps the results of feature attribution explanations (Korbar et al. 2017, Huang & Chung 2019, Palatnik de Sousa et al. 2019, 2020). Korbar et al. (2017) apply Grad-CAM to visualize the activation maps of a ResNet model trained on pathology images of colorectal cancer. An important limitation for the application of class activation maps to digital pathology inputs is discussed in Paschali et al. (2019). CAM and Grad-CAM strongly depend on the effective receptive field of the network being interpreted. As explained in Section 2.1, CNNs used in digital pathology are often obtained by the fine-tuning of deep CNNs such as ResNet and Inception on the histopathology inputs. The effective receptive fields in these networks increase their size with depth, leading to

very diffused maps with little focus at the cellular level (Paschali et al. 2019). CAM is also used by Huang & Chung (2019) in their CELNet architecture, that combines CAM, saliency maps and attention weights for the localization of tumor regions in WSI patches. Pirovano et al. (2020) propose a multi-scale analysis of feature importance by applying CAM at multiple points in their multiple-instance learning model. Another interesting analysis is the one in Palatnik de Sousa et al. (2019), where the application of the model-agnostic method for feature attribution LIME is studied for breast WSIs inputs. The authors identify in the super-pixel creation algorithm of LIME one of the main limitations for its application to pathology inputs. Their follow-up work in Palatnik de Sousa et al. (2020) shows that the region proposal for the super-pixel choice in LIME can be optimized by genetic strategies.

Feature visualization is proposed in Xu et al. (2017) and Pirovano et al. (2020). In the former, Xu et al. (2017) analyze an Support Vector Machine (SVM) classifier of tumor regions trained on handcrafted and CNN features. The authors show the patches that maximally activate individual neurons, following the line of work of AM. They also overlay on the SVM's confidence scores onto the WSI. The analysis in Pirovano et al. (2020) shows, by applying AM, that CNNs learn to recognize spindle-shaped cells and clustered lymphocytes to detect cancerous areas in breast tissue.

A few works focus on using attention weights as a means of improving model interpretability (Katharopoulos & Fleuret 2019, Ilse et al. 2020). The work by Ilse et al. (2020), in particular, shows a reconstruction of the WSI where the patches are weighted by the model attention weights and patches with small weights are shown in small size and with little opacity. This visualization summarizes the importance of each patch during the learning process.

2.3 Summary

This chapter introduced the background on digital pathology for breast cancer and the existing interpretability works in this context. Several DL-based approaches exist in the literature to analyze pathology images of breast tissue (Janowczyk & Madabhushi 2016, Ehteshami Bejnordi et al. 2017). At the core of these methods, there is a CNNs architecture such as ResNet or Inception. Fine-tuning is used to refine the weight values from those learned during pre-training on ImageNet. The staining variability and the lack of precise pixel-level annotations are two of the main challenges in this field, which are addressed by staining augmentation techniques, domain adversarial training and weak supervision. Interpretability analyses of CNN models in this context are mostly focused on the application of existing methods to the backbone architectures. Most of the works show feature attribution heatmaps such as those obtained by applying Grad-CAM (Paschali et al. 2019, Pirovano et al. 2020) and LIME (Palatnik de Sousa et al. 2019). Few AM visualizations of the features learned by the network are in Pirovano et al. (2020). The next chapter will evaluate feature attribution approaches, showing some important limitations and introducing a new methodology that addresses the existing shortcomings and improves the understandability and reliability of the explanations.

Chapter 3

Improving the Understandability of Post-hoc Explanations

3.1 Motivation

The previous chapter presented a review of existing interpretability methods. It is now important to examine their limitations for the application to digital pathology, to clarify where improvement is needed. The work in Adebayo et al. (2018) already proved that several gradient-based approaches show little reliability. New explanations should pass the sanity checks by Adebayo et al., and show their sensitivity to shifts in the model parameters. Explanations that are too complicated to understand by the physicians, besides, would be of little use in the clinical context (Tonekaboni et al. 2019). It is thus important to assess the explanations' appropriateness, understandability and alignment with clinical factors. Part of this evaluation can be done *a priori* by software developers, to establish whether the explainability technique is good enough to be tested with real users, in this case, with pathologists (Hoffman et al. 2018). This chapter starts exactly from this point. Section 3.2 proposes an *a-priori* evaluation framework of explanations for digital pathology.

Sections 3.3 and 3.4 deal with the thesis objective number 1. (see Chapter 1, Section 1.4), that is improving the understandability (and reliability) of the explanations. The proposed contributions are based on the main hypothesis that prior expert knowledge is a valuable source of information to overcome the limitations of existing techniques. Section 3.3 addresses the fact that visual heatmaps often have a blurred appearance with little sharpness on the nuclei instances in the image (Paschali et al. 2019). I propose a method that uses the existing information on the nuclei contours in the images to sharpen the visualizations and improve their clarity.

Section 3.4 addresses a further limitation of existing techniques, namely the fact that visual heatmaps such as CAM and LIME do not give explanations in terms of clinical features such as those in Figure 2.2 (in Section 2.1). The pixels highlighted by the heatmaps are those that would mostly affect the model output if changed. No explanation is given as to what pattern causes this strong correlation with the output. Section 3.4 proposes to focus on concept-attribution techniques to address this issue. The proposed method evaluates the relevance that the model attributes to clinical features such as nuclei size, shape and appearance. These features are known and understood by physicians since they are used in clinical practice, and they can thus be used to evaluate the alignment of

automated outcomes with clinical factors.

A user-centric evaluation is key, at this point, to verify the impact of the proposed methods on pathologists, collecting their impressions and opinions about the reliability, safety and accountability of the model. For this reason, Section 3.5 introduces an online interface that is used to interact with pathologists. The results in this section are still preliminary, and more work is currently being done on this topic.

The content in this chapter is adapted from my previous works: Graziani, Lompech, Müller & Andrearczyk (2021), Graziani, Palatnik de Sousa, B. R. Vellasco, Costa da Silva, Müller & Andrearczyk (2021), Graziani et al. (2018), Graziani, Andrearczyk, Marchand-Maillet & Müller (2020). The code implemented to run the experiments is available at `https://github.com/maragraziani`.

3.2 Evaluation of Interpretability for Digital Pathology

This section proposes quantitative metrics to evaluate a priori visual explanations for histopathology.

3.2.1 Related work

Few works in the literature evaluate post-hoc explanations for digital pathology. The multiple audiences of the explanations make the evaluation extremely challenging (Weller 2019). The work in (Tonekaboni et al. 2019) conducted in-person interviews with physicians to determine the specific requirements of explainability for clinical use. Some evaluation criteria are derived from the results in this work. Most of the interviewed participants strengthened the importance of obtaining domain-appropriate information. This implies the fact that explanations should provide new, concise and precise information. Most importantly, the explanations should be useful for the clinicians, hence they should be easy enough to understand and helpful to make decisions about the course of action for the patient. These aspects are discussed later in this chapter (see Section 3.5 since they are rather subjective and require a user-based evaluation. Arun et al. (2021) is a related work that evaluates the localization capability of multiple (including Grad-CAM) for chest x-ray images of pneumothorax and pneumonia. In this application, the lesion contours are available and the appropriateness of the explanations can be evaluated by localization metrics. The method proposed in their work, however, cannot be directly applied to pathology images because of the structural difference between chest x-rays and WSIs. WSIs do not have a clear central subject on the foreground, but rather a structural disposition of many instances (e.g. connective, adipose, or epithelium cells) at several scales. This work thus proposes a methodology that is tailored to digital pathology.

3.2.2 Methods

Datasets Most of the experiments in this manuscript concern BCMLN detection in publicly available data collections. The experiments in this section use Camelyon by Litjens et al. (2018) and the breast subset of PanNuke by Gamper, Koohbanani, Graham, Jahanifar, Khurram, Azam, Hewitt & Rajpoot (2020). Camelyon includes data collected for the challenges in BCMLN patient stage grading run in the years 2016 and 2017. The collected images sum up to a total of 1169 WSIs. An example of the images is given

in Figure 2.3 in Section 2.2.2. Annotations of metastasis type (i.e. negative, macro-metastases, micro-metastases, isolated tumor cells) are available for all slides, whereas manual segmentations of tumor regions are provided for only 320 slides. The patches assigned to the tumor category are sampled from inside the annotated tumor regions, whereas non-tumor patches are extracted from outside the annotated regions and from the negative instance-level labeled WSIs.

The PanNuke dataset is a collection of tissue images from multiple organs with semi-automatic annotations of the nuclei contours and types. The semi-automatic instance segmentation tool developed by the authors of the data collection was used to assign labels of neoplastic, inflammatory, connective, epithelial, and dead nuclei (Gamper, Koohbanani, Graham, Jahanifar, Khurram, Azam, Hewitt & Rajpoot 2020). In the breast subset, the multiple nuclei types are present at multiple ratios, with neoplastic nuclei being the most frequent and no dead nuclei. Where not stated otherwise, the data from these collections are pre-processed following the approaches used by the participants in the Camelyon challenge, as described by Ehteshami Bejnordi et al. (2017). Patches of 224×224 pixels are extracted at the highest magnification level from the tissue areas. Since the Camelyon WSIs contain large portions of background, the Otsu's thresholding method is performed to isolate the tissue areas on the lowest resolution images (Otsu 1979). The staining variability across the multiple acquisition centers and datasets is reduced by the normalization in Reinhard et al. (2001). Oversampling is applied to the PanNuke images to balance their under-representation. For each input, we extract smaller image patches located in the center, upper left, upper right, bottom left and bottom right corners of the image. Table 3.1 reports the training, validation and test splits used for the experiments in this section and in Sections 3.3 and 3.5. The three pre-existing PanNuke folds were used to separate the patches in the splits by using two folds in the training set and the third fold in the internal testing set. No PanNuke images were used for the external validation since all the three folds contain images for the multiple centers. The code for the extraction of the patches was released in the context of the EU project process and is available online[8].

Table 3.1: Summary of the train, validation, internal and external test splits used for the experiments in Sections 3.2, 3.3, 4.3

		Cam16	Cam17 (5 Centers)					PanNuke (3 Folds)		
	Label		C. 0	C. 1	C. 2	C. 3	C. 4	F. 1	F. 2	F. 3
Train	Neg.	12954	31108	25137	38962	25698	0	1425	1490	0
	Pos.	6036	8036	5998	2982	1496	0	2710	2255	0
Val.	Neg.	0	325	0	495	0	0	0	0	0
	Pos.	0	500	0	500	0	0	0	0	0
Int. Test	Neg.	0	0	274	483	458	0	0	0	1475
	Pos.	0	500	999	0	0	0	0	0	2400
Ext. Test	Neg.	0	0	0	0	0	500	0	0	0
	Pos.	0	0	0	0	0	500	0	0	0

Network Architecture and Training The CNN architecture is Inception V3 (Szegedy et al. 2016) with ImageNet pre-trained weights. The model is finetuned on the histopathol-

[8] `https://github.com/medgift/PROCESS_L1`

ogy training images to solve the binary classification task of distinguishing positive samples against negative ones. This solution outperforms other architectures in Graziani, Eggel, Andrearczyk et al. (2020) and is thus used for the analyses. Three fully-connected layers (2048, 512 and 256 neurons respectively) with a dropout probability of 0.80 and a prediction layer are added on top of the pre-trained features. The weighted binary cross-entropy loss is used to address the strong class imbalance in the training data. L2 regularization is used with a coefficient of 0.01 on the fully-connected layers. The optimization is solved with SGD and standard parameters, i.e. 0.90 Nesterov momentum (Nesterov 1983). Early stopping is performed on the validation loss to stop the training process, with 5 epochs of patience (convergence is reached after 60 epochs on average). The model performance is measured by the average Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) over ten runs with multiple initialization seeds, reaching 0.82 ± 0.0011 and 0.87 ± 0.005 for the internal and external test sets respectively. The model training on a single GPU NVIDIA V100 takes 20 hours on average. The same GPU is used to run the evaluation experiments.

Explainability Techniques The experiments focus on Grad-CAM and LIME, which are methods that generate visual explanations largely applied in medical imaging. In the following, we clarify the implementation details of these two techniques. As already explained in Section 2.2.1, CAM produces a localization map by visualizing the contribution of each feature map before these are spatially averaged by the GAP and linearly combined to produce the network prediction. Grad-CAM, which is illustrated in Figure 3.1, directly takes into account the cascade of gradients to determine the weights of each feature map. The importance weights $\alpha_k^c \in \Re$ for a class c and the k-th feature map are obtained by computing the following:

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i \sum_j}^{\text{GAP}} \underbrace{\frac{\delta \hat{y}^c}{\delta A_{ij}^k}}_{\text{gradients}}, \tag{3.1}$$

where $A_{i,j}^k \in \Re$ is the activation of the k-th feature map at location (i, j), and $\hat{y}^c$ is the model output for class c. CAM and Grad-CAM were shown to be equivalent up to a normalization constant that is proportional to the number of pixels in the feature maps (Selvaraju et al. 2017). Grad-CAM++ is a further development that considers the gradients at the pixel level rather than those of the entire feature maps. Grad-CAM++ explanations partially address the shortcomings of considering the entire feature maps, like the difficulty to operate when multiple occurrences of instances of the same class occur in a single image (Chattopadhay et al. 2018).

The second technique in this evaluation is LIME, which is formulated as the optimization problem:

$$\xi(\boldsymbol{x}) = \underset{g \in G}{\operatorname{argmin}} \quad \mathcal{L}(f, g, \pi_{\boldsymbol{x}}) + \Omega g. \tag{3.2}$$

This formulation minimizes the explanatory infidelity $\mathcal{L}(f, g, \pi_{\boldsymbol{x}})$ for the CNN's decision function of a potential explanation g, given by a surrogate model G. The minimization is solved in a neighborhood $\pi_{\boldsymbol{x}}$ defined around a given sample $\boldsymbol{x}$. Ωg is a measure of the opaqueness of the explanation g. The explanation is often obtained by a ridge regression model trained on the perturbed instances, which are weighted by the pairwise cosine similarity with the original instance. For this model, Ωg is the number of non-zero weights

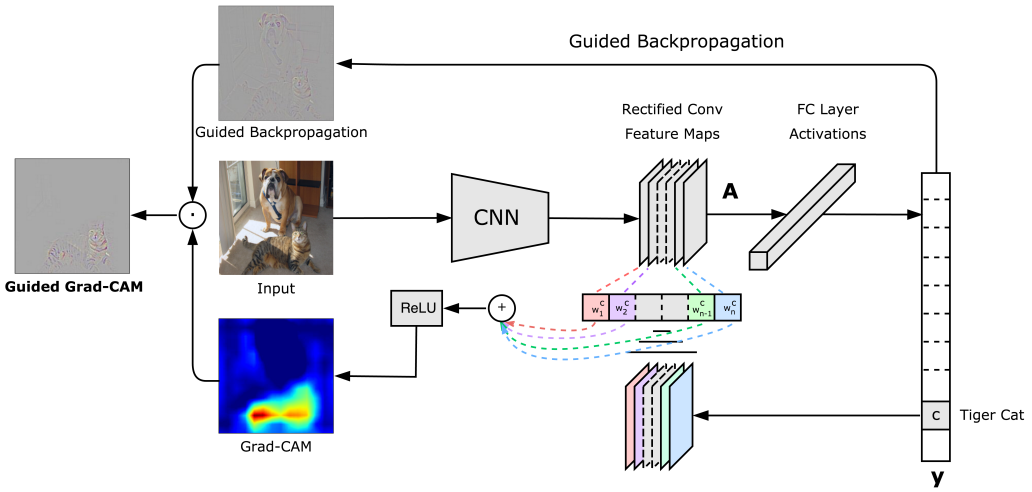


Figure 3.1: Illustration of Gradient-weighted Class Activation Mapping (Grad-CAM). Replicated from http://gradcam.cloudcv.org/ as accessed in August, 2021.

of the linear regression surrogate. This value should be left low enough to be intelligible by humans with some experience in machine learning. The coefficients of this linear model, i.e. the *explanation weights*, explain the importance of each super-pixel to the model outcome.

The neighborhood $\pi_{\boldsymbol{x}}$ is obtained by perturbations of the input image x. The image is first divided into representative image sub-regions called super-pixels. Quickshift (Vedaldi & Soatto 2008) is the default algorithm to generate super-pixels in LIME, and two diffused alternatives are Simple Linear Iterative Clustering (SLIC) (Achanta et al. 2012) and Felzenszwalb's graph-based image segmentation (FHA) (Felzenszwalb & Huttenlocher 2004). These methods cluster the image pixels using color, texture and other types of local similarities. The perturbations are then obtained by filling random super-pixels with black pixels. LIME explanations can be found in the medical imaging literature, with applications in radiology (Reyes et al. 2020) and histopathology (Palatnik de Sousa et al. 2019, 2020).

Evaluation of Visual Similarity and Alignment with Clinical Factors The first evaluation concerns the following two points: (i) the accordance of methods in terms of their visual similarity and (ii) their alignment with clinical factors. Point (i) is evaluated by the Structural Similarity Index Measure (SSIM). This measure is used in the literature to remove duplicate images and quantify image similarity. It is obtained as a weighted sum of three measures that compare image luminance, contrast and structure. For more details on the implementation, the reader may consider Wang et al. (2004). The SSIM ranges from 0 (no structural similarity) to 1 (identical structural similarity) and it is computed on heatmap pairs obtained for the same input image from two differing methods. The second measure (ii) is obtained by following the line of work in Zhou et al. (2018). The Intersection over Union (IoU) is used to establish the overlap of the explanations with specific image regions, e.g. background, neoplastic nuclei, epithelial nuclei. Also known

as the Jaccard's index, the IoU is defined as follows:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FP}}, \tag{3.3}$$

where TP is the number of True Positives, FN is the number of False Negatives and FP is the number of False Positives. The results of this evaluation are in Section 3.2.3 under points (i) and (ii).

Evaluation of Consistency and Repeatability The consistency of LIME visualizations is evaluated over variations in the method hyper-parameters (for the results, see Section 3.2.3, point (iii)). The similarity of the visualizations obtained for slightly increasing values of the hyper-parameters is evaluated by the SSIM. The method is evaluated based on its dependency on the number of samples used to solve the local linear classification task, the number of super-pixels used and the starting seed.

Randomization Tests Explanations obtained from a network with trained parameters are compared to those from a network with randomly initialized parameters. The randomization is performed in a cascading manner, for instance, the CNN weights are randomized in progression from the top layer to the bottom one (Adebayo et al. 2018). If no clear change is present between the explanation of the trained CNN and that with randomly initialized weights, then no clear link can be established between the network weights and the explanation. The similarity between the original explanation and the one obtained with random weights (up to a given layer) is computed by the SSIM.

The fully randomized network is also used to compute the IoUs with each nuclei type. The IoUs are compared to those of a trained network, to verify whether the explanations become more aligned with clinically relevant factors after training.

3.2.3 Results

Visual Inspection Figure 3.2 illustrates the explanations obtained for four of the 200 inputs used for the experiments. The inputs are selected to showcase multiple patch-based classification outcomes, namely True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). Note that the TP, TN, FP and FN used here differ from the pixel-wise ones used in Eq. 3.3, being at the patch-level rather than at the pixel-level. The semantic segmentation of the nuclei is overlayed on the original images. To enable a fair comparison across CAM heatmaps, the results are normalized between zero and one according to the maximum and minimum values of the heatmaps for all testing inputs. As expected, heatmaps of negative predictions (both TNs and FNs) have lower values than those for TPs, with the mean values being 1.53 for TNs, 1.83 for FNs and 4.03 for TPs. Large absolute values gather for all heatmaps on the areas containing neoplastic nuclei.

Figure 3.3 visually compares the explanations obtained with LIME. The maximum number of features is used for the explanations, corresponding to using all the superpixels in the images. The neighborhood size is set to the default value of 1,000 samples. The results obtained with FHA and SLIC seem harder to interpret than those obtained with the Quickshift segmentation method. The last column in this image shows the proposed improvement of this explanation method that is introduced in Section 3.3.

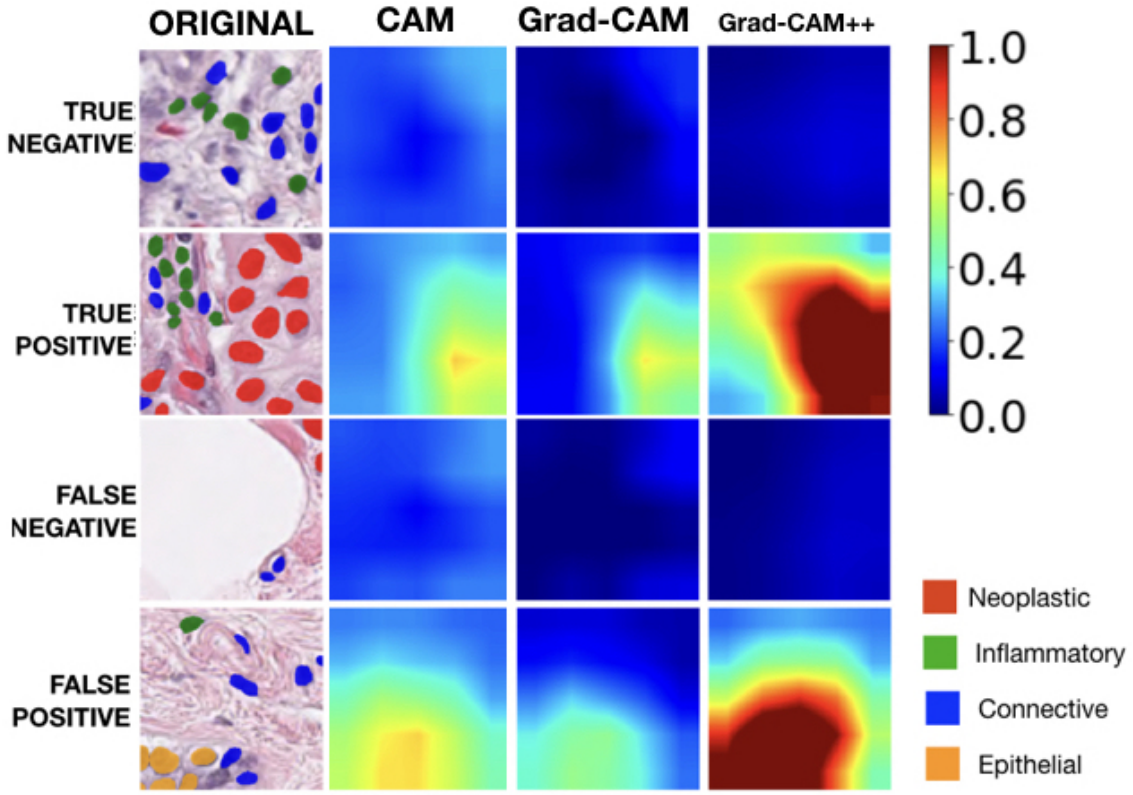


Figure 3.2: Qualitative comparison of Class Activation Mapping (CAM), Gradient-weighted CAM (Grad-CAM) and its improved version Grad-CAM++. Reproduced from the original work in Graziani, Lompech, Müller & Andrearczyk (2021).

Quantitative Evaluation In the following, we report the results of these quantitative evaluations: (i) the visual similarity among the methods, namely their agreement on the most salient regions in the input; (ii) the alignment of the explanations with the clinically relevant factor of nuclei neoplasticity, which indicates the presence of tumor; (iii) the consistency and repeatability of LIME; (iv) the sensitivity of LIME and CAM to randomization of the model parameters.

Point (i), namely the agreement of the heatmaps, is shown by high SSIM when the network is confident about the predicted class. This is given by comparing the average SSIM values for pairs of XAI methods, for which the results are shown in Fig. 3.4. The XAI methods agree more on negative predictions than on positive ones, with SSIM values above 0.50 for all couples.

The alignment of the explanations with clinical factors (ii) is quantified as the IoU of the heatmaps with the segmentation masks of functionally different nuclei, for which the results are in Fig. 3.5. The IoU is computed for 100 testing images of the PanNuke dataset containing at least one neoplastic nucleus (indicative of the presence of tumor). The heatmaps are thresholded, as in Zhou et al. (2018), so that they activate on average for 60% of the pixels of the positive class images. We obtain one IoU score per image and annotation type. Because some nuclei types are not present on some subsets of images, the IoU for a given annotation type is computed only on the subset of images that contains at least one instance of this type. The IoU of the heatmaps generated for a

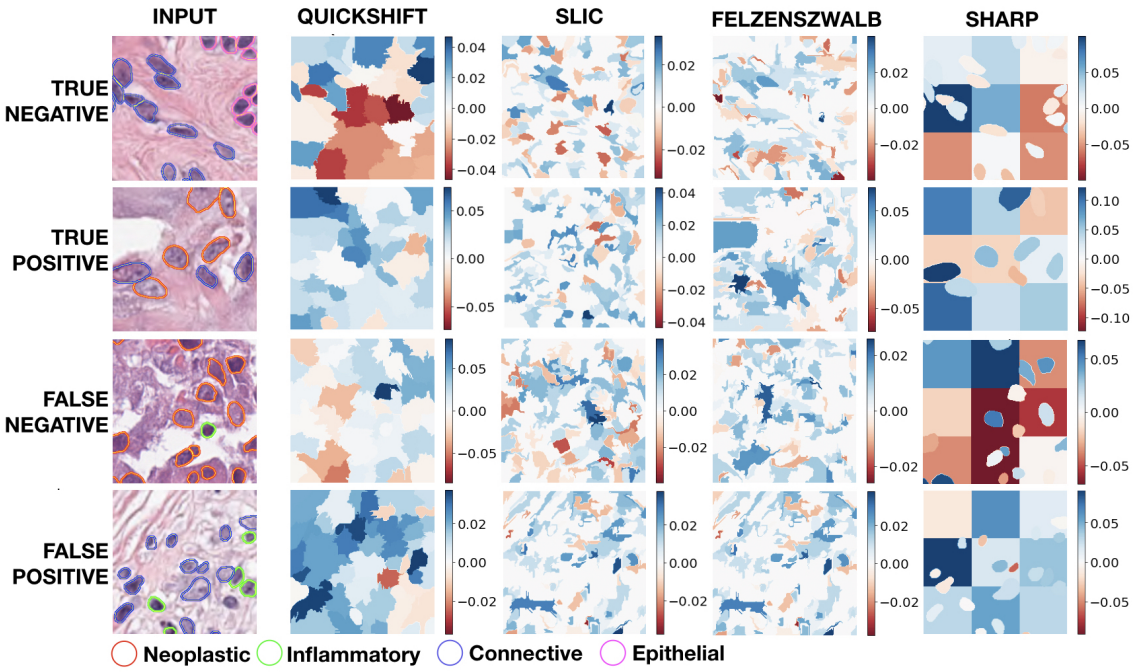


Figure 3.3: Qualitative comparison of Local Interpretable Model-agnostic Explanations (LIME) for multiple segmentation methods: Quickshift, Simple Linear Iterative Clustering (SLIC), Felzenszwalb and the new method proposed in Section 3.3, called Sharp-LIME.

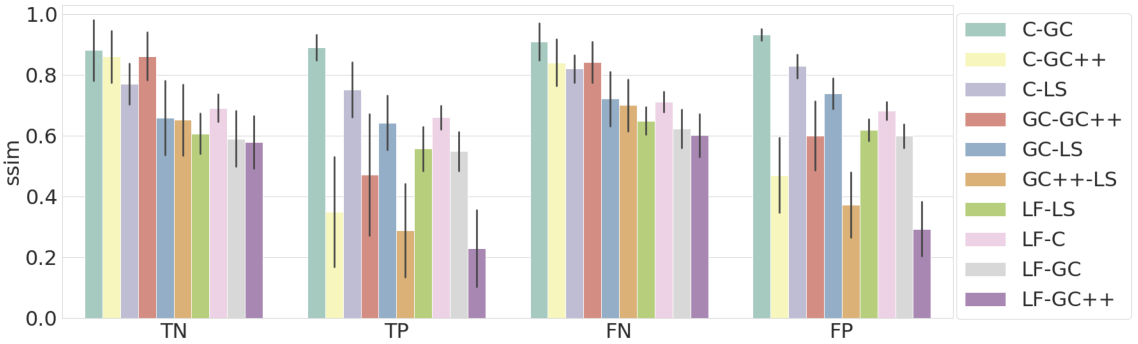


Figure 3.4: Agreement of the heatmaps, measured as the average SSIM between pairs of methods for the network outcomes TN: True Negative, TP: True Positive, FN: False Negative, FP: False Positive. The error bars represent the standard deviation of the SSIM values. Results from Graziani, Lompech, Müller & Andrearczyk (2021).

CNN with fully randomized weights is added as a baseline for comparison[9]. The results show that the heatmaps have higher IoU values for neoplastic nuclei, although there is no significant difference between the explanations generated from a trained network and one with random weights.

The next experiments assess the consistency of the explanations by quantifying their sensitivity to parameter changes and the re-initialization with multiple seeds. Since activation maps do not require the tuning of hyper-parameters nor an initial seed, these

[9]This is not the same as the cascaded randomization test proposed in Adebayo et al. (2018), that is reported in Fig. 3.8

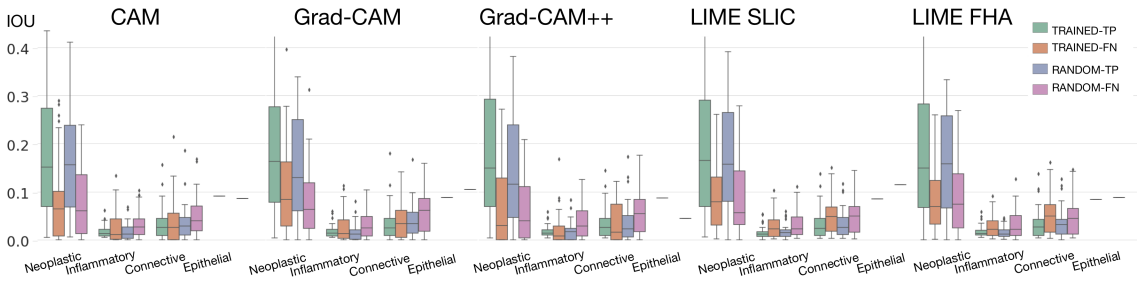
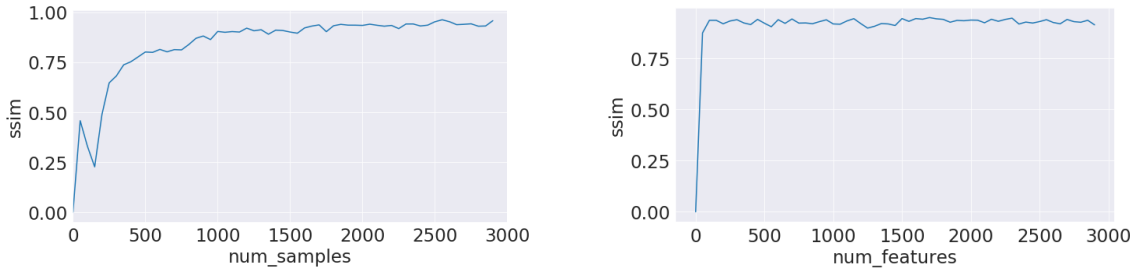


Figure 3.5: Alignment of the explanations with clinical factors, obtained by quantifying the IoU between the heatmaps and the nuclei types in PanNuke testing data. The IoU of a network with randomly initialized weights (RANDOM-TP and RANDOM-FN) is added as a baseline for comparison. Replicated from Graziani, Lompech, Müller & Andrearczyk (2021).



(a) Neighborhood size.

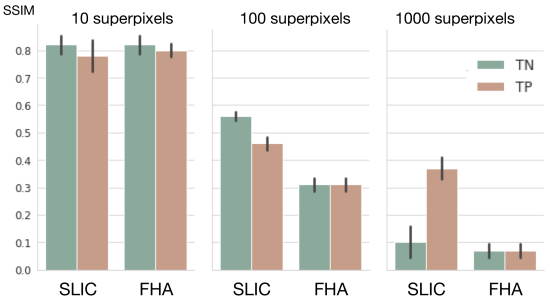
(b) Number of superpixels.

Figure 3.6: SSIM between heatmaps obtained from LIME when a parameter differs by a shift of 50. The studied parameters are the number of samples in (a) and the number of superpixels in (b). For a given value N on the x-axis, the plot represents the SSIM between the heatmap obtained with N $N - 50$ image perturbations. E.g. at point 1000 on the x-axis the graph shows the SSIM between heatmaps obtained with 1000 and 950 perturbations. The number of superpixels is set to 100 in (a) and the neighborhood size to 1000 in (b). Replicated from Graziani, Lompech, Müller & Andrearczyk (2021).

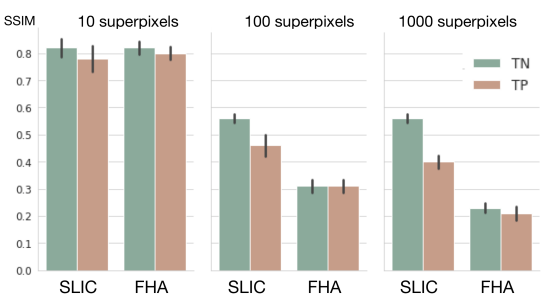
analyses are only reported for LIME. Figures 3.6a and 3.6b show the SSIM against small shifts in the values of, respectively, the neighborhood size and the number of super-pixels. The shifts are performed in the range of zero to 3000 with a step of 50.

The repeatability of LIME visualizations is shown in Figure 3.7. The SSIM of the heatmaps obtained with 25 initialization seeds is evaluated depending on the hyper-parameter values for the number of superpixels and the neighborhood size. The figure compares the repeatability of the visualizations for 10, 100 and 1000 superpixels with neighborhoods of 100 and 1000 samples. High repeatability (SSIM around 0.80) is obtained only with 10 superpixels.

Figure 3.8 shows the results obtained from the cascading randomization test (iv). The plot shows the SSIM between the original heatmap (from the trained CNN) and the one after the randomization at each layer. The test is only passed CAM-based methods.



(a) Neighborhood size of 100 samples.



(b) Neighborhood size of 1000 samples.

Figure 3.7: SSIM evaluating LIME repeatability over 25 repetitions for LIME with multiple random seeds. Error bars report the standard deviation. Replicated from Graziani, Lompech, Müller & Andrearczyk (2021).

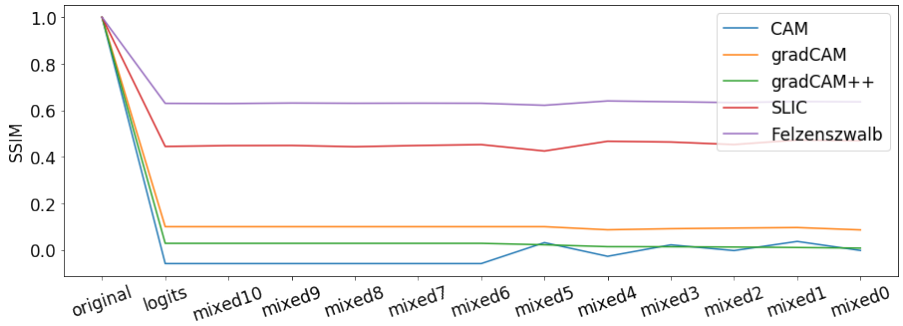


Figure 3.8: Cascading randomization results, showing the SSIM between the heatmaps of a trained CNN and those generated as the CNN weights are randomized in the cascading way.

3.3 Sharpening the Visualizations of Local Interpretable Model-agnostic Explanations: Sharp-LIME

This section reports the work in Graziani, Palatnik de Sousa, B. R. Vellasco, Costa da Silva, Müller & Andrearczyk (2021). The main assumption used to conduct this work is that prior knowledge can be used to improve the quality of the existing explanations. Most of the variability of LIME explanations seen in the results from Section 3.2 seems due to the strong dependency of the method on the generated set of super-pixels, a limitation also mentioned in Palatnik de Sousa et al. (2019). In this section, I further analyze this dependency between the super-pixels and the explanations, proposing a method that uses nuclei contours to obtain sharper and more understandable explanations than the ones obtained with default segmentation methods. Existing datasets with manual annotations of the nuclei contours such as the Pannuke (Gamper, Koohbanani, Graham, Jahanifar, Khurram, Azam, Hewitt & Rajpoot 2020) and Kumar image collections are used as prior information about the super-pixels. Sharper visualizations are obtained by directly using the existing nuclei contours as super-pixels. When the contours are not available, these are predicted by the segmentation output of the Mask Region-based Convolutional Neural Network (Mask-RCNN) in Kumar et al. (2017). The following sections clarify the details on the architecture and methods.

3.3.1 Related work

The main objective of LIME is introduced in Eq. 3.2 in Section 3.2.1. Previous works in the literature have studied the dependency between the super-pixel and the explanations. The work in Palatnik de Sousa et al. (2019), in particular, proposes a systematic, manual search for parameter heuristics that would generate super-pixels that visually correspond to expert annotations on breast pathology images. The follow-up analysis in Palatnik de Sousa et al. (2020) shows that the quality and consistency of the super-pixel can be further improved if the search is performed by genetic algorithms rather than manually. Both the manual and automatic solutions proposed by these papers, however, appear impractical for clinical use. Both algorithms require a considerable amount of time to search the hyper-parameter space and generate a single explanation, which may be undesired if the explanations were to use in everyday clinical practice. Manual search, besides, may lack objectivity in the ranges used for the parameter search.

Central to this work is the identification of nuclei contours. When these are not available, they are predicted by the Mask-RCNN in Kumar et al. (2017). The Mask-RCNN model belongs to the family of Region-based Convolutional Neural Networks (RCNNs) for object detection and segmentation. These models use selective search to identify regions in the image that may be likely to contain an object. Proposed by He et al. (2017), Mask-RCNN adds the pixel level position of the target instance to the model objectives, improving the target detection accuracy and performing an additional instance segmentation task. This architecture uses a backbone ResNet to generate multiple feature maps at multiple scales. The feature maps are then used to generate the region proposals and the segmentations. Mask-RCNN has been applied to detect and segment nuclei in multiple tissue types and applications (Graham et al. 2019, Jung et al. 2019). Graham et al. (2019) compared this method to other segmentation approaches such as fully convolutional networks (Long, Shelhamer & Darrell 2015), SegNet (Badrinarayanan et al. 2017), and to some methods specifically developed for nuclear segmentation such as the ones in Raza et al. (2019) and in Vu et al. (2019). The performance of Mask-RCNN is further improved by adding color normalization and post-processing in Jung et al. (2019). Mask-RCNN is chosen for this work since it can easily separate clustered nuclei through its region proposal module. This is a desirable and important feature to generate explanations that focus on individual nuclei since overlapping nuclei or nuclei that are too close to each other may be merged into a single nucleus by other methods.

3.3.2 Methods

Datasets and CNN Architecture The dataset in Section 3.2 is used to train the patch-based classification model that distinguishes tumor from non-tumor patches and that is the object of the interpretability analysis. The breast-cancer images from the Kumar data collection (Kumar et al. 2017) are used for the nuclei segmentation model since these data contain manual annotations of the nuclei contours. The dataset collects WSIs from multiple organs with annotated nuclei boundaries.

The classification model is the same as in Section 3.2. The nuclei segmentation model is the Mask-RCNN model in Kumar et al. (2017), for which already trained weights are made available for download by the authors of the paper[10]. The model uses a ResNet 50 (He et al. 2017) as the convolutional backbone, which is fine-tuned from ImageNet pre-

[10] `shorturl.at/fiuFN` (accessed on September 2021).

training on the Kumar dataset. The R-CNN model detects the nuclei entities and generates bounding-boxes as region proposals for the segmentation. From these, pixel-level masks of each nuclei instance are produced by optimizing the Dice Similarity Coefficient (DSC). The DSC is formally defined as:

$$\text{DSC} = \frac{2 \times \text{TP}}{2 \times \text{TP} + FP + FN}. \tag{3.4}$$

Sharp-LIME The overview of the Sharp-LIME method is illustrated in Figure 3.9. Patches at high magnification are extracted at first from the Camelyon dataset. For the PanNuke inputs, the images are oversampled by extracting smaller patches at five locations, as explained in Section 3.2.2. Not strictly requiring manual annotations, this approach can generalize to inputs coming from other datasets that do not have nuclei contours. It is the case of the Camelyon inputs, for which Sharp-LIME uses the automatic segmentation of nuclei contours obtained by Mask R-CNN (which is trained on the Kumar dataset). Manual annotations of regions of interest may also be drawn directly by end-users to probe the network behavior for specific input areas. Once the segmentation is

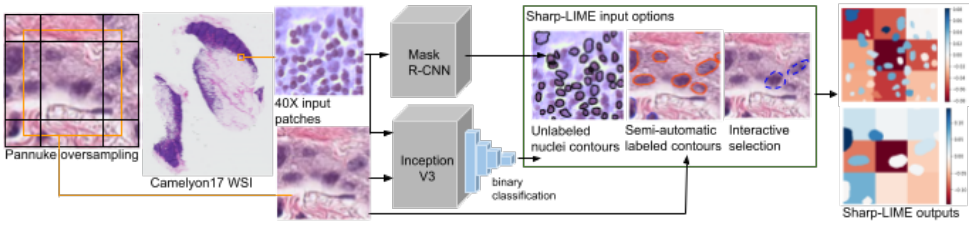


Figure 3.9: Overview of Sharp-LIME. Inception V3 classifies tumor from non-tumor patches at high magnification sampled from the input WSIs. Manual or automatically suggested nuclei contours (by Mask R-CNN) are used as input to generate the Sharp-LIME explanations on the right. Replicated from Graziani, Palatnik de Sousa, B. R. Vellasco, Costa da Silva, Müller & Andrearczyk (2021).

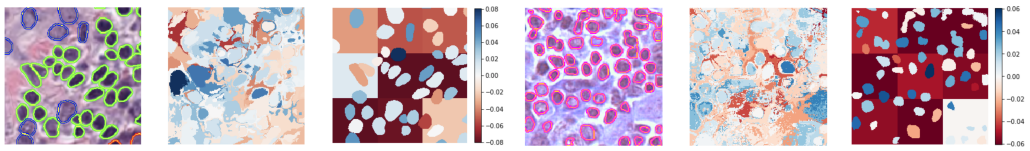
obtained, the input image is split into nuclei contours and background. The background is further split into 9 squares of fixed size. This splitting reduces the difference between nuclei and background areas since overly large super-pixels may achieve large explanation weights by sheer virtue of their size. This division, besides, might help with improving the heatmap precision on possible parts of the background that may be relevant to the network outcome. The code to replicate the experiments (developed with Tensorflow $>$ 2.0 and Keras 2.4.0) is available at `github.com/maragraziani/sharp-LIME`, alongside the trained CNN weights. Experiments were run using a GPU NVIDIA V100. A single Sharp-LIME explanation requires roughly 10 additional seconds to the traditional inference and nuclei segmentation times of the models (i.e. around 5 and 10 seconds, respectively).

Evaluation The proposed method is evaluated against the state-of-the-art LIME.The evaluation focuses on the PanNuke data, for which ground-truth annotations are available. The evaluation replicates some of the tests already presented in Section 3.2, for instance, the sanity checks concerning consistency and repeatability and the quantification of the alignment of the explanations with clinical factors. For the latter, the importance of a neoplastic nucleus, which is an indicator of a tumor, is measured by the sign and magnitude of the explanation weight. The cascading randomization test in Adebayo et al. (2018) is

also performed by assigning random values to the model weights starting from the top layer and progressively descending to the bottom layer. We expect the cascading randomization test to show near-zero SRCC for both techniques since by randomizing the network weights, the network output is randomized much as the explanations. Finally, the robustness to constant input shifts is evaluated as in Kindermans et al. (2019). The results of the sanity checks are validated by statistical testing. The pairwise non-parametric Kruskal test for independent samples is used for the comparisons between descriptive statistics of the weights assigned to multiple nuclei types. The paired t-test is used to compare LIME weights obtained from a randomly initialized and a trained network. Spearman's Rank Correlation Coefficient (SRCC) is used to evaluate the similarity of the ranking of the most important super-pixels. The repeatability and consistency for multiple seed initializations are evaluated by the SRCC, the Intraclass Correlation Coefficient (ICC) (two-way model), and the coefficient of variation (CV) of the explanation weights.

3.3.3 Results

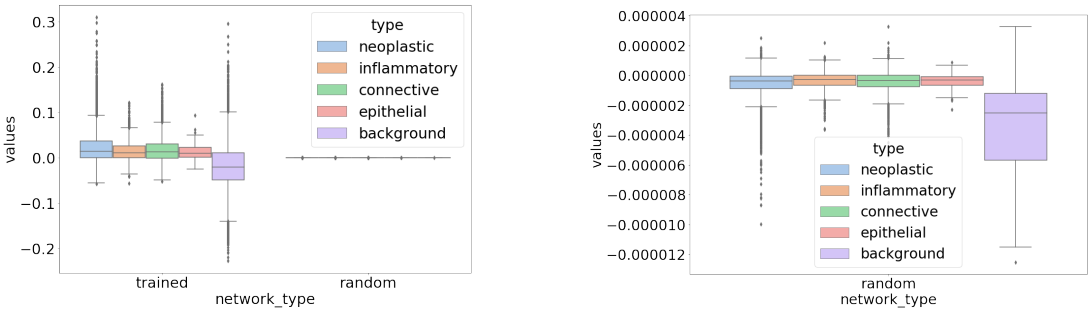
Visual Inspection Already in the previous chapter, Figure 3.3 showed Sharp-LIME explanations against LIME explanations obtained with the Quickshift, SLIC, and Felzenszwalb segmentation algorithms on PanNuke inputs. More examples of Sharp-LIME against LIME (obtained with the default segmentation algorithm Quickshift) are shown in Figure 3.10. All the results on PanNuke can be inspected in the GitHub repository at `github.com/maragraziani/sharp-LIME` (as last accessed in August 2021). The results on Camelyon can be inspected by using the interactive tool at `https://cadeval.p645.hevs.ch/` (as accessed in August 2021) as explained in Section 3.5.



(a) (b)

Figure 3.10: From left to right, original image with annotated nuclei contours, standard LIME and sharp LIME for an input from a) PanNuke and b) Camelyon.

Attention to neoplasticity The Sharp-LIME explanation weights are distributed over neoplastic, inflammatory, connective, epithelial nuclei and the background. The explanation weights assigned by Sharp-LIME to each of these instance categories are quantified in the box plots in Figure 3.11a. The weights of the neoplastic nuclei, with average value 0.022 ± 0.03, are significantly larger than those of the background squared super-pixels, with average value -0.018 ± 0.05. Explanation weights of the neoplastic nuclei are also significantly larger than those of inflammatory, neoplastic and connective nuclei (Kruskal test, p-value < 0.001 for all pairings). Sharp-LIME weights are compared to those obtained by explaining a random CNN, that is the model with randomly initialized parameters. The comparison is shown by the boxplot in Figure 3.11b. The Sharp-LIME explanation weights for the trained and random CNN present significant differences (paired t-test, p-value< 0.001) and the explanations for the randomized network are all at almost-zero values.



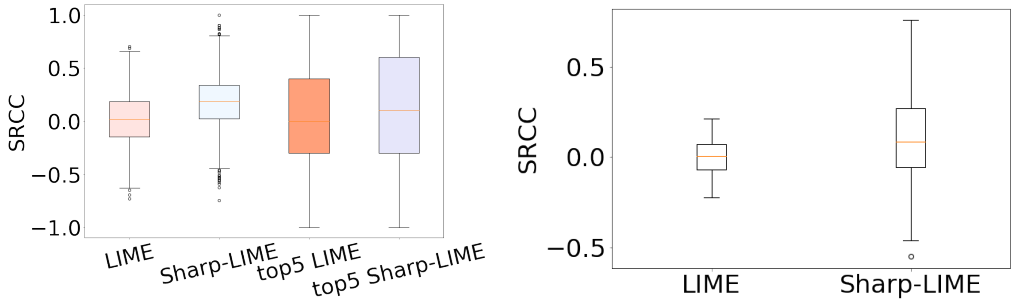
(a) (b)

Figure 3.11: a) Comparison between Sharp-LIME explanation weights for a trained and a randomly initialized CNN; b) Zoom on the random CNN in a). These results can be compared to those obtained for standard LIME in Section 3.2.3, Figure 3.5. Replicated from Graziani, Palatnik de Sousa, B. R. Vellasco, Costa da Silva, Müller & Andrearczyk (2021)

Consistency Figures 3.12a and 3.13a show the results from the evaluation of Sharp-LIME consistency. Sharp-LIME rankings appear in Figure 3.12a more consistent than the standard implementation of LIME, showing lower sensitivity to the seed initialization. The mean of LIME SRCC is, in fact, significantly lower than that of Sharp-LIME, 0.015 against 0.18 (p-value$<$ 0.0001). In addition, the SRCC of the five super-pixels with the highest ranking is compared in the same figure, with average LIME explanation weights 0.029 and 0.11 for Sharp-LIME.

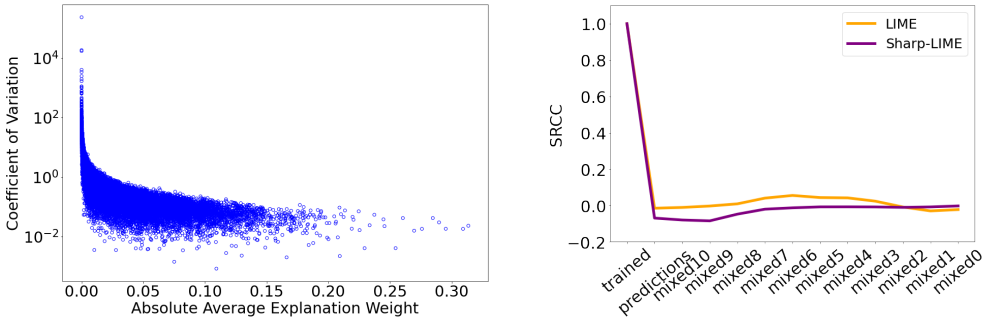
Super-pixels with a large average absolute value of the explanation weight are more consistent across re-runs of Sharp-LIME, as shown by Figure 3.13a and by their lower value of the CV. The ICC of the most salient super-pixel in the image, i.e. first in the rankings, further confirms the largest agreement of Sharp-LIME, with ICC 0.62 against the 0.38 of LIME.

The cascading randomization of network weights shows nearly zero SRCC in Fig-



(a) (b)

Figure 3.12: Evaluation of consistency and robustness by the SRCC. a) Consistency to three re-runs with changed initialization seeds. SRCC of the entire and top-5 super-pixel rankings. The means of the distributions are significantly different (paired t-test, p-value$<$ 0.001); b) Robustness to constant input shifts, quantified by the SRCC of the super-pixel rankings for all inputs. Results from Graziani, Palatnik de Sousa, B. R. Vellasco, Costa da Silva, Müller & Andrearczyk (2021)



(a) (b)

Figure 3.13: a) Consistency over multiple re-initializations. CV against average explanation weight for three re-runs with multiple seeds; b) Cascading Randomization test. The SRCC of the super-pixel rankings is monitored at each layer. Replicated from Graziani, Palatnik de Sousa, B. R. Vellasco, Costa da Silva, Müller & Andrearczyk (2021).

ure 3.13b. This result was expected, according to the considerations in 3.3.2. A visual example of LIME robustness to constant input shifts is given in Figure 3.14. The SRCC of LIME and Sharp-LIME is compared for original and shifted inputs with unchanged model prediction in Figure 3.12b. Sharp-LIME is significantly more robust than LIME (t-test, p-value$<$ 0.001).

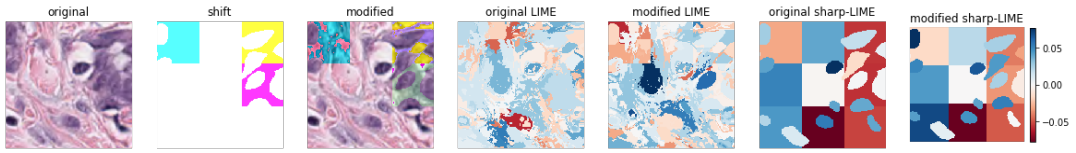


Figure 3.14: Robustness to constant input shifts. From the left to the right, the original input image, the applied shift, the modified image, the LIME and Sharp-LIME explanations for the original and the shifted inputs. Reproduced from Graziani, Palatnik de Sousa, B. R. Vellasco, Costa da Silva, Müller & Andrearczyk (2021).

The results in this section demonstrate the improvements brought by Sharp-LIME in terms of consistency, robustness and repeatability of the explanations. The improvements in the understandability of the explanations are demonstrated by user-evaluation tests, which are reported in a dedicated section of this chapter, i.e. Section 3.5.

3.4 Explainability with Clinical Features: Regression Concept Vectors

The visualizations obtained in the previous sections with feature attributions reveal an important limitation. A region in the heatmap with highly activated values, does not yet clarify the characteristics in this region causing such high activations. It is not clear if the nuclei appearance, the background texture or other factors are causing the high activation values in the heatmaps for histopathology images. In other words, feature attribution does not provide a clear interpretation of the clinical features that are used by the model to make decisions. This can be addressed by concept-based attribution, which can complement the visualizations going beyond pixel-level relevance. The method of Regression Concept Vectors (RCVs), introduced in this section, aims at improving the understandability of the explanations by analyzing concept-based explanations in terms of clinical features of nuclei morphology and appearance. The work reported in this section is adapted from the conference work in Graziani et al. (2018)[11] and the peer-reviewed extension in Graziani, Andrearczyk, Marchand-Maillet & Müller (2020). The work proposed in this chapter is also applied to handwritten digits and retinopathy images in Graziani, Andrearczyk, Marchand-Maillet & Müller (2020). The reader may refer to this work for the application of this approach not only to digital pathology but to a wider range of applications including computer vision and ophthalmology.

3.4.1 Related work

The method in this work is based on the idea, proposed in 2018 by Kim et al. (2018), of learning concepts in the representations learned by a CNN. Concept learning finds its origin in the late Sixties, being defined in cognitive psychology as the identification of attributes that can distinguish exemplars from non-exemplars of multiple categories (Bruner et al. 1967). This definition is adapted into a traditional machine learning task where examples of a category are classified based on a list of concepts. A concept, in this context, represents a feature that is distinctive for the examples of a category, may this be concrete (e.g. people, places, objects, shapes) or abstract (e.g. actions, emotions). Instances of a zebra, for example, can be described through the concepts of animal, four-legged, striped, horse-shaped, black and white, and so on. These concepts are either present or not present and are thus encoded as binary labels. In the example above, all the listed concepts assume a value of one since they can all be found in a zebra.

In the work by Kim et al. (2018), the concepts are learned in the intermediate activations of a CNN. Their work is mostly developed for the computer vision task of object detection. For a given object category, the concepts are boolean-valued functions representing the presence or absence of a given attribute, e.g. striped texture. The concepts are learned by training a linear classification model on the activations of an arbitrary intermediate layer, using ground-truth binary labels for each attribute. The idea of using linear classifiers at intermediate layers also follows the related work on linear probing of deep networks, proposed by Alain & Bengio (2016). In this work, the authors evaluate the linear separability of the object categories at each layer, showing that the separability of the classes increases with network depth. The TCAV score, introduced in Kim et al. (2018), is computed by counting the fraction of how many images in the input instances of one object category respond with a positive increase of the predicted probability (for a

[11] Springer license number 5131960559402 for reproduction

given class) if the concept was present in the intermediate layer. The score represents to what extent a concept c is used for making the prediction, and it is defined as follows:

$$\text{TCAV} = \frac{|\{\boldsymbol{x} \in X_k : S_c^{l,k}(\boldsymbol{x}) > 0\}|}{|X_k|}, \tag{3.5}$$

where X_k is the set of inputs with label k and $S_c^{l,k}(\boldsymbol{x})$ is the sensitivity of the model prediction to the concept, computed for class k at layer l and for the input $\boldsymbol{x}$. The sensitivity to a concept is computed for a multi-class classification model in Eq. 3.6.

$$S_c^{l,k}(\boldsymbol{x}) = \boldsymbol{u}_c \cdot \frac{\partial \phi^{L,k}(\boldsymbol{x})}{\partial \phi^{l}(\boldsymbol{x})}. \tag{3.6}$$

$\phi^{L,k}(\boldsymbol{x})$ is a vector of real numbers representing the raw prediction values for the k-th class for the input image $\boldsymbol{x}$ and $\boldsymbol{u}_c$ is the CAV for concept c. The derivative of the decision function is obtained by stopping gradient backpropagation at the l-th layer of the network. Note that the TCAV score is bounded between zero and one. If no images are influencing the decision with a positive gradient, TCAV is zero. If all images influence the decision, then the TCAV score is one.

Multiple works exist that followed the approach of CAV. They are applied to interpret the classification of ophthalmology images in Fang et al. (2020) and the retrieval of pathology images in Cai et al. (2019). The interactive framework in Cai et al. (2019) collected feedback by pathologists on an interactive image retrieval system. To understand the system, the participants formulated questions in terms of concepts that are often used during cancer screening, such as *How would the decision change if there was less stroma in the tissue? What if the nuclei appeared larger and with less regular texture?* These questions refer to the criteria in Figure 2.2. It must be noted that these features are considered by pathologists in their full range of expression and not only as binary characteristics that are either present or absent. The variability of nuclei size, for example, is expressed on a continuous scale, increasing gradually from small to large when moving from a low tumor grade to a high tumor grade. Encoding the concepts as binary variables is an important limitation of the CAVs by Kim et al. (2018) for the application to digital pathology. The work proposed in this section mainly aims at addressing this limitation, modeling the concepts as continuous variables rather than binary ones.

3.4.2 Methods

Datasets and models The datasets used for this study are the Camelyon (Litjens et al. 2018) and the Kumar (Kumar et al. 2017) collections[12], for which the details are given in Section 3.3.2.

From the Camelyon dataset, 41,039 patches are extracted from random locations at the highest magnification (i.e. 40x) from WSIs acquired at centers 0, 1, 2 and 3. The validation set is built by sampling 2726 additional patches from WSIs coming from these centers. The test set is built by sampling 3996 patches from the last center, i.e. center 4, in a stratified way (same number of tumor and non-tumor images). Staining normalization and online data augmentation (random flipping, brightness, saturation and hue perturbation) are used to reduce the domain shift between the centers. The WSIs of breast tissue from

[12]Since PanNuke was released in late 2020, this dataset was not available yet at the time when this work was initially developed, hence we included only two datasets.

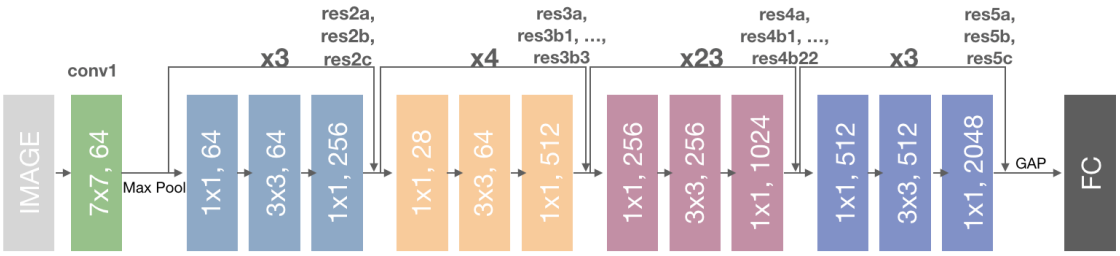


Figure 3.15: Illustration of the ResNet 101 (He et al. 2016*a*) architecture. The size of the filter and the number of channels is written inside the layer, e.g. the first layer, *conv1*, has filter size 7x7 and 64 channels. The residual blocks in multiple colors are repeated the number of times that is indicated on top of the skip connection (e.g. x3, x4, etc.). The name of the layers at the end of each residual block is reported on top. The last layer is a fully-connected layer with a node for each class.

the Kumar dataset (Kumar et al. 2017) are used to extract 300 patches with annotated nuclei contours, from which it is possible to extract measures that are representative of pathology features such as nuclei area and texture.

The analyses are performed on the ResNet 101 (He et al. 2016*a*) model shown in Figure 3.15, which is trained on the training split from the Camelyon data. The last layer is replaced by a single node with a sigmoid activation function to solve the binary classification of tumor against non-tumor patches. The first convolutional layer is referred to as *conv1*, while the layers at the end of each residual block are *res2a*, *res2b*, etc. The model weights are fine-tuned from the pre-trained weights on ImageNet. The BCE loss is optimized with SGD and Nesterov momentum (Nesterov 1983) and standard hyper-parameters (learning rate 10×-4, momentum 0.90). The AUC of the model on the testing set is 0.70. The model is trained using a GPU NVIDIA K80 (11.5 hours for 15 epochs).

Extraction of clinical measures A key component of the RCV approach is the extraction of clinical measures, which are hereafter called concept measures. The concept measures can be chosen arbitrarily, depending on the main scope of the analysis. Developers may interact with experts, for example, to understand the type of questions that the interpretability analysis should address and design the concept measures accordingly. Pre-existent handcrafted features extracted from the images can be used as concept measures, being developed by the joints efforts of multiple experts to represent the geometry, shape and appearance of the tissue, the cells and the stroma (Khan et al. 2015, Bhargava et al. 2020). The exhaustive evaluation of all possible concepts is unfeasible, hence collecting this information before the analysis is important to delimit the analysis to a finite number of measurable concepts.

For the application to BCMLN, the main objective is the validation of the alignment between the learned features and the guidelines of clinical practice. The grading system in Figure 2.2 (Section 2.1) is taken as the starting point to define the concepts, being a well-established reference for breast cancer. The concept of nuclear pleomorphism, namely of variations in the nuclear size, shape or chromatin appearance of the cells, is suitable to a representation by visual features. The questions that this method tries to answer to are of the type: *"Are large nuclei more relevant to the prediction of tumor than small*

ones?" Changes in size are represented by evaluating the area of each nucleus in the image, which is expressed as the sum of pixels in the nuclei areas. Hyper-chromatic nuclei result in a changed appearance from normal ones, a variation that can be modeled by texture descriptors (Khan et al. 2015). Already used in Khan et al. (2015), Haralick's texture descriptors include measures of the texture contrast, correlation, homogeneity, dissimilarity and Angular Second Moment (ASM) (Haralick et al. 1973). The concept measures are computed on a small set of visual examples (i.e. 300 samples in these experiments) that is called $X_{concepts}$. The segmentation of the nuclei instances in the Kumar data are used to evaluate the measures of nuclei area and texture. If no nuclei contours were available, automatic segmentation methods such as the ones in Otálora et al. (2020), Jung et al. (2019), Badrinarayanan et al. (2017), Graham et al. (2019) would have been used.

Computing the Regression Concept Vector The RCV of a concept c is computed by seeking the linear regression that, for an input image $\boldsymbol{x}$, predicts the value of the associated concept measure $c(\boldsymbol{x}) \in \Re$, on the basis of the features $\phi^l(\boldsymbol{x})$ (where $\phi^l(\boldsymbol{x}) \in \Re^{w \times h \times p}$ for convolutional layers of width w, height h and p channels) learned by the intermediate convolutional layer l in the CNN:

$$c(\boldsymbol{x}) = \boldsymbol{v}_c \cdot \overbrace{\frac{1}{Z} \sum_i \sum_j}^{\text{GAP}} \phi^l(\boldsymbol{x}) + error. \tag{3.7}$$

In this equation, the RCV for concept c is the vector $\boldsymbol{v}_c \in \Re^p$. The RCV represents the direction of the strongest increase of the concept measures for the concept c and it is normalized to obtain a unit vector. The components of the RCV are found by solving the linear regression problem by linear least squares (LLS) estimation. In addition to the formulation as a regression, a spatial aggregation of GAP along the (height, width) of each feature map is introduced to address the shortcomings of flattening the features to a one-dimensional array of whp elements, which is used in Kim et al. (2018) to solve the linear classification task and find the CAV. When flattening the representations of a convolutional layer l, the number of dimensions of the unrolled convolutional maps easily grows to millions. This may affect the regression since there could be the risk of finding spurious correlations. There is, besides, a loss of information since the 2D structure of the space is lost and it is not possible to discern anymore the relative position of the pixels on the grid. A feed-forward network with only dense layers cannot detect a circle or a square because the structure of the pixels is broken and no information is kept about the original relationships between pixels, e.g. about the neighboring pixels in the vertex of the square). The GAP operation generates a representation of $\phi^l(\boldsymbol{x})$ as a one-dimensional array of p elements. This solution improves the quality of the regression fit by aggregating the information in the intermediate representations. Note that if l is a dense layer with p units, then the GAP operation is not needed and $\boldsymbol{v}_c$ is a p-dimensional vector in the space of its activations.

Sensitivity to a concept The sensitivity of the CNN outcome to a concept S_c, called conceptual sensitivity[13], represents how much the concept measure affects the network's

[13]Note that the term conceptual sensitivity was used by Kim et al. (2018) and it does not refer to the output classification sensitivity commonly known as recall.

prediction for a given input. Being defined for a single input, the conceptual sensitivity is a local explanation. For a binary classification task, $S_c \in \Re$ is computed as follows:

$$S_c^l(\boldsymbol{x}) = \boldsymbol{v}_c \cdot \frac{\partial f(\boldsymbol{x})}{\partial \phi^l(\boldsymbol{x})}, \tag{3.8}$$

where $\frac{\partial f(\boldsymbol{x})}{\partial \phi^l(\boldsymbol{x})}$ is the directional derivative of the network output $f(\boldsymbol{x})$ w.r.t. the the RCV direction $\boldsymbol{v}_c$. The directional derivative is obtained by projecting the partial derivative along the RCV vector by computing the scalar product between the two. $S_c^l(\boldsymbol{x})$ represents the network responsiveness to changes in the input that result in a translation along the direction of the increasing values of the concept measures. The sign of $S_c^l(\boldsymbol{x})$ represents the direction of change, while its magnitude represents the rate of change. When moving along the RCV direction, the output $f(\boldsymbol{x})$ may either increase (positive conceptual sensitivity), decrease (negative conceptual sensitivity) or remain unchanged (conceptual sensitivity equals zero). In a binary classification network with a single neuron in the decision layer, the decision function is a logistic regression over the activations of the penultimate layer. A positive value of the sensitivity to a concept can be interpreted as an increase of $p(y = 1|\boldsymbol{x})$ when the representation $\phi^l(\boldsymbol{x})$ is moved towards the direction of the increasing values of the concept (that is the RCV $\boldsymbol{v}_c$). A negative conceptual sensitivity can be interpreted as an increase in $p(y = 0|\boldsymbol{x})$ when the same shift in the representation is applied.

Global scores Global scores of concept relevance such as TCAV (Kim et al. 2018) are obtained as the fraction of k-class inputs for which the activation vector of layer l was positively influenced by a concept c. TCAV, however, does not consider the magnitude nor the sign of the single conceptual sensitivities. To address this limitation, we propose the bidirectional relevance score Br:

$$Br = R^2 \frac{\hat{\mu}}{\hat{\sigma}}. \tag{3.9}$$

In this equation, the coefficient of determination $R^2 \leq 1$ indicates how well the RCV represents the concept in the internal CNN activations. It measures whether the concept vector is actually representative of the concept by evaluating its predictive performance on unseen data. This value is divided by the coefficient of variation $\hat{\sigma}/\hat{\mu}$, which describes the relative variation of the scores around their mean, being the standard deviation of the scores over their average. Br is large when two conditions are met, namely R^2 is close to 1 and the coefficient of variation is small (the values of the sensitivity scores lie closely concentrated near their sample mean). Br goes to infinite if $\hat{\sigma} = 0$. After computing Br for multiple concepts, the scores are scaled to the range [-1, 1] by dividing by the maximum absolute value. Such scaling permits a fair comparison among concepts since these are represented by multiple RCVs. With the set of analyzed concepts being reasonably large, a score close to the absolute value of one can be considered as large. This means that the concept has a considerable impact on the increase (in case of positive sign) of the outcome probability. Bidirectional scores provide an intuitive explanation of the impact of a concept on a binary classification outcome. Their extension to multi-class classification tasks is out of the scope for the experiments in this chapter, for which we redirect the reader for more information to Graziani, Andrearczyk, Marchand-Maillet & Müller (2020).

3.4.3 Experiments

The experiments in this section aim to demonstrate that RCVs constitute a valuable alternative to feature attribution explanations that can be applied to histopathology to obtain further insights on the features learned by a CNN and a clearer understanding of the image characteristics influencing the prediction.

Correlation of the concepts with tumor areas The first experiment aims at evaluating the correlation between the concept measures and tumorous tissue. Table 3.2 reports the Pearson correlation between the concept measures and the ResNet101 predictions for the images in the Kumar dataset. Note that for these images, the ground truth on tumorous areas is not available so it is not possible to evaluate the correlation of these concepts with the ground truth annotations.

Table 3.2: Pearson correlation between the concept measurements and the network prediction.

	correlation	ASM	eccentricity	Euler	area	contrast
ρ	$\mathbf{-0.30}$	-0.20	-0.10	0.10	0.30	$\mathbf{0.40}$
p-value	≤ 0.001	≤ 0.001	≤ 0.01	≤ 0.001	≤ 0.001	≤ 0.001

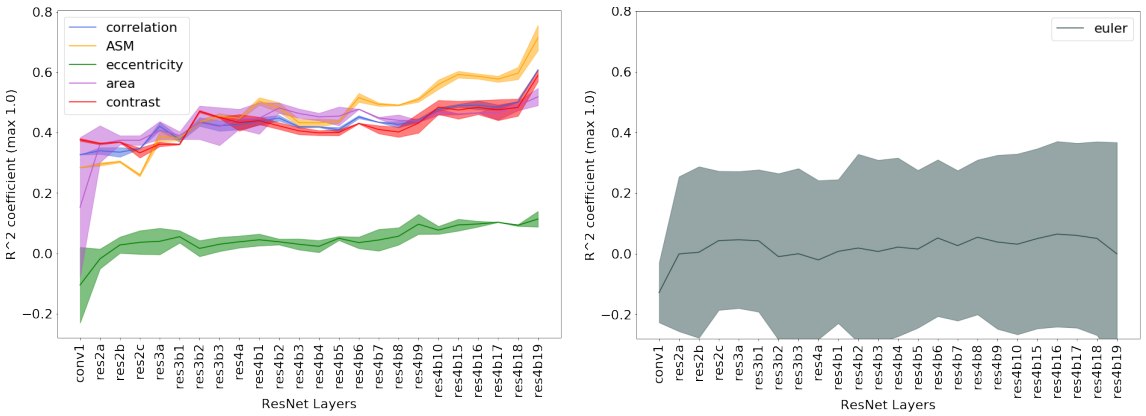
Layer-wise Performance of the Regression Multiple intermediate layers in the network can be used to compute the RCVs. The performance of the linear regression, measured by it determination coefficient R^2, expresses the percentage of variation that is captured by the regression and it is indicative of how well a certain concept is (linearly) learned at each layer. Almost all the concepts are learned already from the early network layers, as shown in Figure 3.16. *eccentricity* and *Euler* are the only two concepts that cannot be regressed at any layer, reporting almost zero mean of the R^2, suggesting that the learned RCVs might be simply random directions. The concepts *eccentricity* and *Euler* are thus excluded from the remaining analysis because they are not learned sufficiently well in the activation space.

As mentioned in Section 3.4.2, the features extracted from the intermediate layer may be aggregated by a GAP operation. The improvements given by this observation are shown in Table 3.3.

Table 3.3: Impact of GAP on the R^2 of the RCVs for breast histopathology. The labels in the other columns refer to the CNN layers, as in the Keras implementation of ResNet101.

	no pooling			GAP		
	res3a	res4a	res5a	res3a	res4a	res5a
area	0.43	0.47	0.46	0.03	0.32	**0.52**
contrast	0.37	0.45	0.43	0.02	0.42	**0.57**
ASM	0.38	0.44	0.50	0.28	0.52	**0.62**
correlation	0.41	0.42	0.48	0.18	0.54	**0.62**

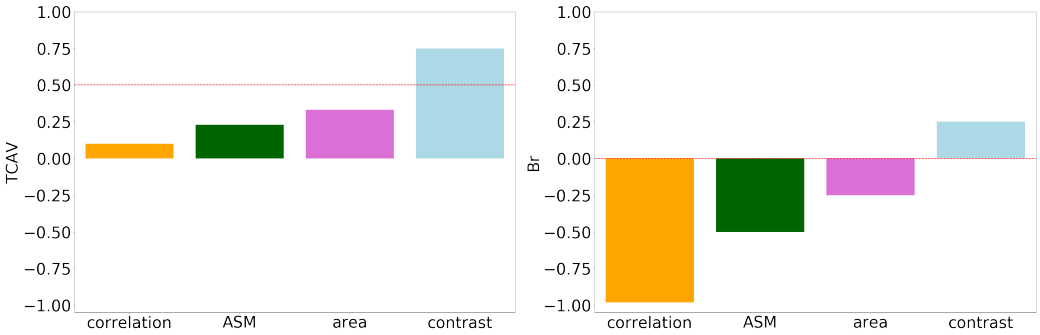
Global scores The comparison between Br scores and the TCAV baseline is shown in Figure 3.17. The scores are computed from the activations of the layer where the RCVs



(a) (b)

Figure 3.16: (a)R^2 at multiple layers in the network. Results were averaged over three reruns. 95% confidence intervals are reported. (b) The RCVs for the concept *Euler* show high instability of the determination coefficient. Replicated from Graziani et al. (2018).

obtains the highest R^2, hence from res5a. *Contrast* has high TCAV$= 0.75$ and $Br = 0.25$. The impact of *correlation* appears even stronger with $Br = -1$ and TCAV$= 0.1$.



(a) TCAV baseline. (b) RCVs.

Figure 3.17: Comparison of TCAV ($\in [0, 1]$) and Br ($\in [-1, 1]$) scores. *Contrast* is relevant according to both measurements. Br scores show that higher *correlation* drives the decision towards the non-tumor class. Scores for the unstable *Euler* are approximately flattened to zero by Br. Replicated from Graziani et al. (2018).

Local explanations The conceptual sensitivity scores in Eq. 3.8 already explain why the CNN assigns the input image to a certain class. These scores may be used to facilitate the interaction between the pathologists and the CNN, for example by providing local explanations for each input patch. Figure 3.18 gives an example of the application of RCV to evaluate the relevance of Haralick's texture descriptors. As expected, the explanations for multiple patches have a consistent sign and only varying magnitudes. This suggests that the behavior of the network is consistent with the tested inputs.

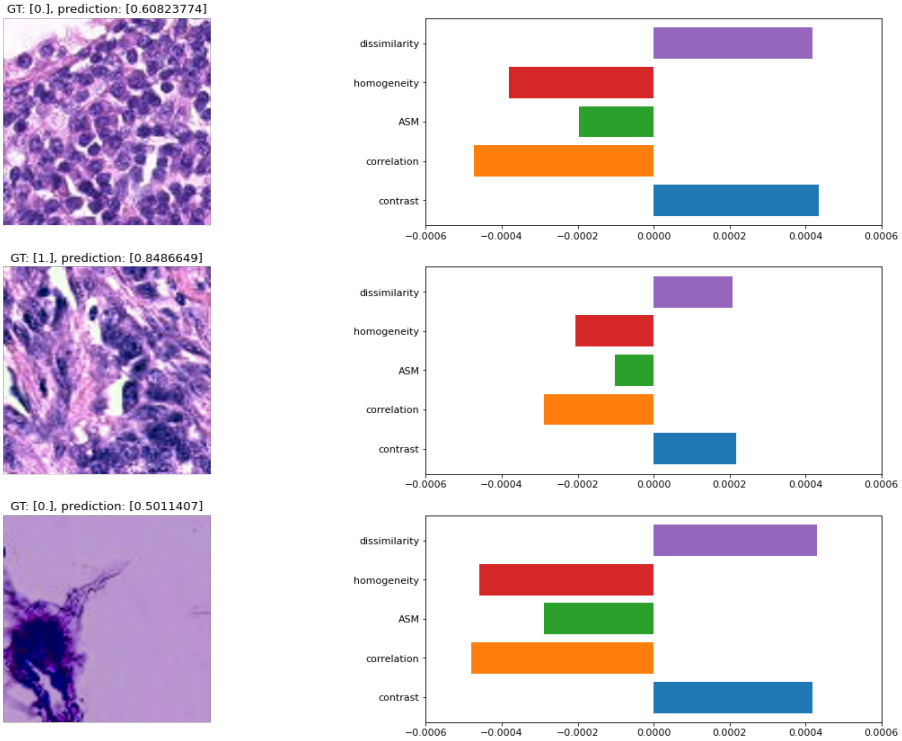


Figure 3.18: Visualization of the local explanations for a single instance determined by the values of the concept sensitivity.

3.5 User-Centric Evaluation with Domain Experts

The previous sections propose Sharp-LIME and RCVs as improvements to existing interpretability techniques.

The objective of this section is to verify that the proposed explanations are more understandable to domain experts than the state-of-the-art methods in the literature, i.e. Grad-CAM (Selvaraju et al. 2017) and LIME (Ribeiro et al. 2016). An interface that allows pathologists to interact with the proposed explanation methods is introduced in this section as a way to perform user tests and collect feedback. This part is essential to implement the vision introduced in Section 1.3.3, for instance, to collect domain-expert feedback during development. Being still at the development phase, the evaluation is performed on retrospective and well-annotated data, and it is far from the real demands of everyday clinical practice. With this simplified task, it is possible to isolate the quality of the explanation methods from the complexity of the task. The work in this section is only preliminary. Additional experiments are being developed in this direction to try to include additional experts in the near future.

3.5.1 Related work

According to Hoffman et al. (2018), there are multiple stages in the evaluation of the interpretability methods. The first stage is the *a-priori* evaluation of the explanation goodness by developers and ML experts. This evaluation aims at measuring the quality and reliability of the explanations and it follows the lines of what has been proposed in Section 3.2. The following stages are the evaluation of user satisfaction, comprehension and perfor-

mance when using the explanations. The goal of these stages is to evaluate whether the users are satisfied with the explanations and whether these improve their understanding of the system and their ability to make diagnoses. Inevitably, the feedback needs to be collected from expert pathologists, and this evaluation is thus more expensive than the a-priori evaluation. Notwithstanding, there is a strong rationale for evaluating the interaction between the user and the system that justifies the additional evaluation. Whether a highly accurate system will be used in clinical practice, in fact, strongly depends on the context and the capabilities of the users to understand and trust the system (Chromik & Schuessler 2020). A non-clear and difficult-to-understand explanation may cost additional time while not being useful to the physicians. This aspect has been slightly overseen in the literature, with only 5% of the surveyed papers in Adadi & Berrada (2018) containing an evaluation of the interpretability methods.

Performing a user-centric evaluation of interpretability is challenging for multiple reasons. As pointed out by Weller (2019), multiple types of people with diverse backgrounds and scopes are involved in the software development stage, and this may bias the way the evaluation is performed. For example, software providers may favor the creation of comforting explanations, regardless of their usefulness and reliability, to induce trust and sustained use by the users. Biasing the evaluation is a risk that should be avoided, as people tend to accept explanations as rightful even when these are empty in terms of informative content (Lombrozo 2006). In addition, the lack of ground truth about interpretability outcomes further complicates the evaluation. Multiple explanations can be equally valid and yet be understood differently by people with different backgrounds. The acts of understanding and interpreting are strongly subjective and the collection of feedback by users is a necessary component already within the software development phase.

Most importantly, the tests of user satisfaction and comprehension aim at clarifying whether the explanations are understandable, useful and informative to the user. These evaluation criteria are depicted as relevant by multiple researchers (Yang et al. 2019, Doshi-Velez & Kim 2017). In Doshi-Velez & Kim (2017), in particular, two of the three evaluation stages discussed in the paper, namely the human-grounded and application-grounded evaluations, include the collection of human feedback. The authors proposed the selection of the best explanation between two options as a human-grounded metric that can be used to evaluate which explanation is the most useful to the users. Our work builds upon this idea and proposes a human-grounded evaluation of the developed methods for the context of digital pathology.

3.5.2 Methods

Interactive web-system An interactive web-based visual interface is used for handling the interaction between the users and the DL software. The interface uses the opensource javascript framework React[14] and the OpenSeadragon visualizer to display the images under an interactive lens that can be used to zoom in and out the images in real time[15]. The visual interface, shown in Figure 3.19, provides the following functionalities: (i) select one image from the five centers in Camelyon dataset (Litjens et al. 2018) (ii) display the image with the possibility to zoom in at multiple magnification levels (iii) annotate one ROI from which patches of 224×224 pixels are extracted with a stride of 10 pixels (both vertical and horizontal) (iv) predict the tumor probability of the extracted patches by using

[14] https://it.reactjs.org/ (last accessed September 2021)
[15] https://openseadragon.github.io (last accessed in September 2021)

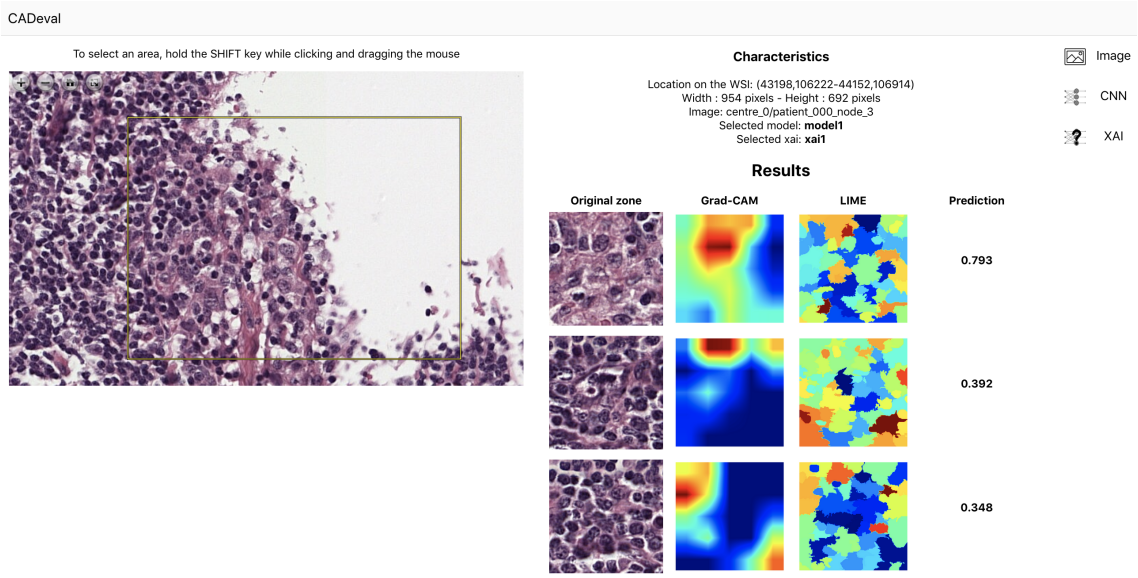


Figure 3.19: Prototype of the interactive web-based interface for the evaluation of explainability outcomes for Whole Slide Images (WSIs).

the Inception V3 model described in Section 3.2 (v) compare multiple explanation methods on the first ten patches with the highest prediction. A video-recorded demonstration of the tool functionalities is available at `shorturl.at/uCP04` (as accessed in October 2021). The methods LIME (Ribeiro et al. 2016), Grad-CAM (Selvaraju et al. 2017), and the proposed Sharp-LIME and RCV are used for the comparison.

The system functionalities of patch extraction, inference and interpretability are run on a GPU NVIDIA V100. The inference time together with the extraction of patches is below 5 seconds. Depending on the method selected by the user and on the dimension of the selected ROI, the generation of the explanations takes between 30 and 90 seconds, hence 7.5 seconds per patch on average.

Evaluation of understandability, confidence, limitations and impacts Six pathologists were involved in the evaluation. The pathologists were first requested to reply to a series of questions regarding their concerns on the integration of DL methods within the clinical workflow and the relevance of interpretability in this context. They were asked to state their expectations about the explanations and whether concept-based explanations in terms of clinical features may be beneficial to their understanding[16]. They were presented with three pixel-level feature attribution visualizations, namely LIME, Sharp-LIME and Grad-CAM, being asked to choose the most understandable method among these three. The experts were then asked whether the explanations increase their confidence in the model's decision-making and to point out any limitations of the explanations. Finally, the experts were asked to describe the impact of the predictions on their diagnostic, provided that they could verify the model's correct functioning.

[16]Note that, at this point of the analysis, the experts were not presented the explanations obtained with RCVs since this feature of the interactive interface was yet under development.

3.5.3 Results

Prior expectations All the feedback provided by the pathologists can be accessed online at https://bit.ly/2WFomX7 (as last accessed in October 2021). The users reported high expectations on visualization methods in the form of heatmaps. The visualization of the areas used by the model to make the decision emerged from this analysis as a minimum requirement for the integration of DL in clinical practice. This requirement was reinforced not only in the context of assisted diagnosis, but also of predictions of cell proliferation values, mitotic counts and grading suggestions. Quantitatively, 66.7% of the pathologists (four out of six) stated that visualizing the areas used to make the decision would improve their confidence in the model.

Understandability If asked to choose between three visualization methods, 60% of the pathologists (three out of five, one missed this question) chose Sharp-LIME over Grad-CAM and LIME as the most understandable method. This further confirms the improvements in the understandability and clarity given by the Sharp-LIME explanations that were claimed in Section 3.3.

Confidence in the model The clinicians stated that both visual and concept-based explanations may be useful to increase their confidence in the model. Only 33% (two out of six) of the experts confirmed that the visualized Sharp-LIME explanations increased their confidence in the model. This percentage raised to 50% for the explanations in terms of measures of nuclei pleomorphism such as those provided by the RCVs[17].

Limitations The participants were also asked about the limitations of the proposed explainability methods. Two out of six pathologists highlighted that the strict magnification requirement and image sizes of the heatmaps were perceived as a main obstacle to understanding the visual explanations. The analysis of small input patches of 224×224 pixels differs from the way they generally operate, which consists in looking at cellular details as much as at the contextual information. The diagnosis is made at a multi-scale level, where multiple zoom-in and zoom-out operations are done before reaching a decision. This method, however, is not followed by the Inception V3 model used for the analyses and, as a consequence, the pixel-level explanations can only show the relevance of the areas on the small input patches.

Impact on the diagnosis Pathologists were asked to select one or multiple follow-up actions that they would take under certain circumstances from a list of five options (no change in the diagnosis, check the areas highlighted by the model, double check of the entire slide, ask for help to a colleague, ask for additional analyses). The analyzed circumstances are (i) the model predicts tumor on a slide that was already diagnosed as negative by the pathologist (ii) the model and the pathologist disagree on the diagnosis (iii) the model points to an area that was not inspected by the pathologist (iv) the model disagree on the diagnosis for a hard case. The results of this analysis are reported in Figure 3.20. In case of discrepancy in the diagnoses pathologists stated that they would "always want an explanation" for the model outcomes. The model may be used to validate some cases that were already diagnosed as negative. If some areas are suggested as tumorous by the

[17]Note that this value is based on the a-priori analysis and needs further validation by additional experiments with RCV explanations. This is currently undergoing work.

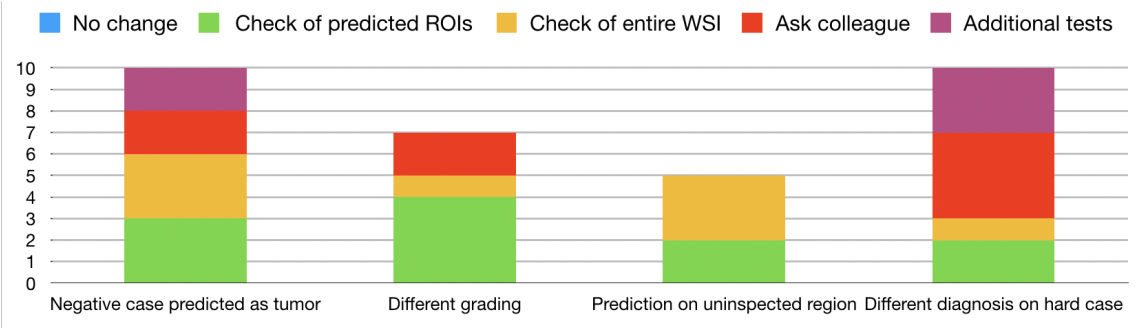


Figure 3.20: Results from surveying 6 experts on the follow-up actions that may be taken in four possible scenarios of model outcomes, namely in case of (i) tumor found by the model in a case predicted as negative (ii) discordance between the pathologists and the model diagnosis (iii) model predicts tumor on a region that was not inspected (iv) discordance between the diagnoses on a case already labeled as difficult to diagnose.

model, pathologists would perform a further check to evaluate whether any region was overseen, and, in some cases, they would analyze again the entire slide. For negative, discordant, or difficult cases, the model decisions may motivate clinicians to seek support from colleagues and from further analyses with other staining methods such as IHC.

While it is difficult to obtain quantitative comparisons and these analyses are only preliminary, I believe that this expert feedback, although subjective, is an essential evaluation that points to the strengths and limitations of the proposed approaches.

3.6 Strengths and Limitations

Clarity and reliability As the evaluation of the existing methods in Section 3.2 pointed out, an important challenge of pixel-level feature attribution methods is how to provide a sufficient level of detail and resolution in the explanations. CAM and LIME explanations do not agree on which areas generate a positive prediction, as demonstrated by Figure 3.4. The visualizations do not differentiate between the nuclei in the foreground and the background. The high activated regions in the heatmaps generated by CAM in Figure 3.2, for example, almost cover the entire patch, appearing out-of-focus. Even LIME explanations with a large number of super-pixels such as those obtained with SLIC and FHA (in Figure 3.3) fail at representing the relevance of meaningful instances such as the nuclei in the image. The proposed Sharp-LIME visualizations are sharper and more understandable than LIME and Grad-CAM, and this is confirmed by the user-tests with pathologists in Section 3.5. Pathologists preferred Sharp-LIME explanations over CAM and LIME. The explanation weights of Sharp-LIME, besides, are larger in magnitude than LIME ones and they show lower variability in Figure 3.13a. These enhancements are given by choosing super-pixels that are meaningful to the model and that are kept unchanged between the multiple re-runs. The super-pixels with large weights in one Sharp-LIME explanation are assigned again a high weight by re-runs of the method with a new seed. The instability to multiple seed re-initialization is reduced by the super-pixels in Sharp-LIME as demonstrated in Figure 3.12a.

Nuclei relevance against background The results in Figure 3.11a demonstrate that, according to Sharp-LIME explanations, the attention to neoplastic nuclei is higher than that given to the background and other nuclei types, i.e. inflammatory, epithelial and connective. The context in the background explains the negative class, with large and negative explanation weights on average. This result is important since it clarifies that the model is paying attention to neoplasticity, which is a correct indicator of the tumor, and that it is not being confounded by other information in the background. Despite a similar result is also obtained by standard LIME in Figure 3.5, in this case, the higher attention to neoplasticity is misleading since it is due to data bias. The sanity-check performed in the same figure shows that LIME and Grad-CAM explanations do not differ for a randomly initialized CNN, and hence are independent of the trained parameters. On the contrary, the proposed Sharp-LIME passes this same test, demonstrating invariance to data bias in Figure 3.11b. Sharp-LIME weights for a randomly initialized model, in fact, are very close to zero for all nuclei types.

Explanations with clinical features such as nuclear morphology Once clarified that the CNN focuses on neoplastic nuclei to predict tumorous tissue from WSIs, it is yet unclear *what* features of these regions are used by the CNN to generate the prediction. The proposed RCV method allows us to analyze the network behavior in terms of concepts at a higher abstraction level than that of input pixel importance. This is also an advantage since the indications of clinical practice for tumor grading in Figure 2.2 find an intuitive translation into measurable features that can be used as concept measures. RCVs explanations extend the state-of-the-art method of CAV to concepts that do not necessarily have a binary expression and that are commonly used in clinical practice. The analysis with RCV in Figure 3.17 compares multiple descriptors of nuclear morphology such as the area and texture at a global level, i.e. for all tumor class testing images. The explanations illustrate that variations of the texture contrast and correlation are used by the model to predict tumor. The Br scores in Figure 3.17 mirror the expectations given by the Pearsons' correlation analysis in Table 3.2 and are in line with the criteria for tumor grading depicted in Figure 2.2. From the preliminary user tests, it emerges that this type of information may be useful to physicians to increase their confidence in the model.

Versatility of RCVs The RCV approach can be applied with high versatility to a variety of image classification tasks. As a result, this method already impacted other research work and imaging modalities. The work in Yeche et al. (2019) applies RCVs to Computer Tomography (CT) images, proposing global scores that evaluate the relevance of a concept in all layers at once. Concept-based explanations are applied to interpret skin lesions in Lucieri et al. (2020) and retinopathy of prematurity in Graziani, Brown, Andrearczyk, Yildiz, Campbell, Erdogmus, Ioannidis, Chiang, Kalpathy-Cramer & Müller (2019). RCVs are applied in Graziani, Muller & Andrearczyk (2019) for explaining the classification of object textures and for investigating the learning dynamics of state-of-the-art CNNs.

Additional complexity and annotations Both Sharp-LIME and RCVs require additional complexity and annotations compared to the baseline visualization methods of LIME and Grad-CAM. The increased complexity is given by the segmentation of nuclei instances by an additional CNN in Sharp-LIME and by the computation of the regression

model and the gradients in RCVs. As mentioned in the relative sections, the time complexity required to compute these steps is neglectable as compared to the CNN training time, being a matter of seconds against several hours. The nuclear segmentation step, besides, is a pre-processing phase that is already present in various histopathology pipelines that require nuclei identification and counting, sometimes even for the entire WSI (Janowczyk & Madabhushi 2016, Janowczyk et al. 2019). Compared to generating segmentations for the entire WSI, a forward pass for a single patch requires only a few seconds, and it is an affordable cost to obtain more insights into the model's inner functioning.

Non-orthogonality of concept directions One limitation of RCVs and CAVs pointed out by Chen et al. (2020) is that the concept vectors learned in both methods are not decorrelated. This may lead to two pitfalls: (i) if the latent space is not mean-centered, then most of the learned directions would point towards where the data lies (ii) if the latent space is strongly stretched in one direction, two differing concepts may lead to very similar vectors with cosine similarity close to one. This limitation is mostly linked to the fact that standard models do not achieve the orthogonality of the concepts without explicit regularization.A safety check may be used to evaluate the angle and the cosine similarity between pairs of concepts as proposed in (Andrearczyk et al. 2020). The authors in the paper, besides, propose a concept-whitening module that can be used during training to obtain orthogonal concept directions.

Limits of non-causal analyses One important limitation of the proposed explanation methods is their sole reliance on the correlation between the model prediction and the concept measures. The causal link between the model prediction and the concept measures is not considered when generating the explanations. The pixel- and concept-level explanations proposed in this chapter only illustrate which features and concepts most correlate with the model's prediction. The explanations may be confounded by correlations that are present in the data but not causally relevant to the model Goyal et al. (2019). Confounding variables may be a concept, for example, a watermark in the image, that is highly correlated with one class. The presence of a watermark with a high correlation for a class may be a confounding variable for the explanations, even if the classifier is powerful enough to not consider the watermark to predict the correct class. In such a case, the explanations may still point out the watermark as a relevant concept. This is an important limitation to keep in mind, particularly for the experiments on the RCVs. The relevance scores in Figure 3.17 only describe correlations and do not aim at describing cause-effect relationships in the model decision-making. Future work may look into how to address this point.

Limits of single-scale explanations Pixel-level explanations are strongly dependent on the scale at which the model is working. In the experiments in this chapter, the models learn from inputs at the highest WSI magnification level available in the datasets, i.e. 40x, and the explanations are also generated for images at 40x. The analysis of the images only at such a high magnification strongly differs from the way expert pathologists operate. The results from the user evaluation demonstrate that pathologists expect heatmaps that are informative on the context and illustrate whether multi-scale information was considered by the model. This is a limitation that not only applies to the proposed contributions in this work, but to multiple explanation methods. The visualizations obtained for individual high-resolution patches may be recomposed as in a puzzle to form a higher-level

visualization map of the entire WSI, as in Hägele et al. (2020). Even in this case, however, the limitation persists as the analysis of the network behavior is focused on the only scale analyzed by the model. Multi-scale explanations may thus be obtained only for models that directly analyze the input images at multiple scales.

3.7 Open Questions

The results in this chapter show that applying off-the-shelf techniques is not sufficient to obtain sufficiently understandable and reliable explanations for some domains. Particularly for digital pathology, the methods should be adapted to the images and requirements of the field. Future developments should try to include the multi-scale context and causal dependencies in the explanations.

How to introduce causal reasoning in the generation of explanations is still an open question. Research on the Causal Concept Effect (CaCE) in Goyal et al. (2019) designed a perturbation operation to estimate the average causal effect between an arbitrary concept and the CNN prediction. This work is, however, only preliminary and performed in the controlled setting of synthetic data. Explanations with the "Anchors" proposed by Ribeiro et al. (2018) may be an interesting investigation for future work on Sharp-LIME. These explanations also move beyond the sole description of correlations, searching for sufficient local conditions for obtaining an invariant prediction to perturbations.

The integration of multi-scale resolution in the explanations cannot be achieved if only a single-scale model is analyzed by the interpretability analysis. Integrating contextual information into the explanations is another open topic for further research. If a multi-scale architecture was used, more challenges may arise in the generation of explanations that can easily adapt to the diverse types of information contained at multiple scales. Future work may aim at generating multi-scale explanations for ad-hoc multi-scale architectures such as those in Hashimoto et al. (2020).

3.8 Summary

This chapter aimed at addressing one part of the main research question of this thesis, which is whether interpretability can be used to explain DL mechanics to physicians. I started by evaluating the existing methods in the literature to interpret CNNs outcomes in digital pathology. From this analysis, the lack of clarity, detail and reliability of the explanations emerged as key factors that required further development. I thus aimed at addressing these limitations as much as improving the understandability of the explanations. Considering the requirements for clinical use in Section 2.2.2 of domain appropriateness, usefulness and understandability of the explanations, I proposed Sharp-LIME and RCVs. The results obtained by these techniques demonstrate that interpretability analyses can benefit from the introduction of external knowledge. The simplicity of Sharp-LIME is also its strength. Much clearer explanations than those generated by standard LIME are obtained by implementing a simple change. The proposed super-pixels clearly distinguish semantically diverse entities in the images, e.g. nuclei types and background, leading to sharp visualizations. Neoplastic nuclei obtained the highest explanation weights, proving their relevance against the background and other nuclei types.

The RCV analysis demonstrates that the network learns features that contain information about the size and texture of nuclei, which influence the gradients to make the

prediction and are thus relevant to the classification. The generation of explanations in terms of clinical features is an asset of the RCV method, which can be easily understood by physicians as a complementary explanation method that describes both the global and the local behavior of the CNN.

The next chapter will demonstrate how the work proposed in this chapter can be used to drive the development of CNNs for digital pathology.

Chapter 4

Improving Model Performance with Interpretability

4.1 Motivation

The post-hoc interpretability methods in the previous chapter can provide explanations, but cannot act on the training process nor modify the learned features. If we find an undesired behavior, e.g. attention to watermarks, we naturally wonder how to correct such behavior. Invariance to specific features such as scale, rotation or domain can be induced in the deep representations to remove undesired patterns. Random rotations and flipping of the inputs, for example, are a standard data augmentation procedure that induces the CNN to learn features that are robust to affine transformations (Shorten & Khoshgoftaar 2019). Specific architectural changes can also promote invariances. Group convolutions, for example, are used to guarantee rotation equivariant features in Cohen & Welling (2016). Changes of the optimization process such as adding training objectives or amplifying the gradient backpropagation are also used to act on the feature learning. The adversarial training proposed in Ganin et al. (2016) can generate domain invariant features by reverting the gradients coming from a domain classifier trained on top of the features. These works all start from the hypothesis that certain properties of invariance or equivariance can improve the model and lead to increased performance. The methods differ depending on where they introduce ad-hoc changes, i.e. in the input (Shorten & Khoshgoftaar 2019), the architecture (Cohen & Welling 2016, Kanazawa et al. 2014, Marcos et al. 2018, Worrall & Welling 2019, Ghosh & Gupta 2019), or the optimization process (Ganin et al. 2016). The group convolutions in Cohen & Welling (2016) act on the convolution operation (by repeating it multiple times for rotations of the same kernel), whereas adversarial training in Ganin et al. (2016) acts on the gradients (by introducing a gradient reversal operation). Despite Lafarge et al. (2017) adapted the gradient reversal to a different type of feature, i.e. staining, there is not yet a clear approach for changing what the network is learning into what it should learn. The aim of this chapter is thus to analyze whether interpretability can be used as a starting point to introduce "must have" patterns in the learned features.

The main objective here is to demonstrate that interpretability can be used not only as a passive analysis of what the network has learned but also to actively modify the feature extraction process, introducing desired behaviors. The main hypothesis is that interpretability analyses can be used, together with the feedback from experts in the

domain, to identify changes in the model that could improve the performance. Concept-based attribution with RCVs, identifies concepts that are learned by intermediate layers of the network. If confounding concepts are learned by the model, we may change the architecture in such a way that information about these concepts is discarded. Similarly, we may encourage the learning of important concepts representing discriminant features.

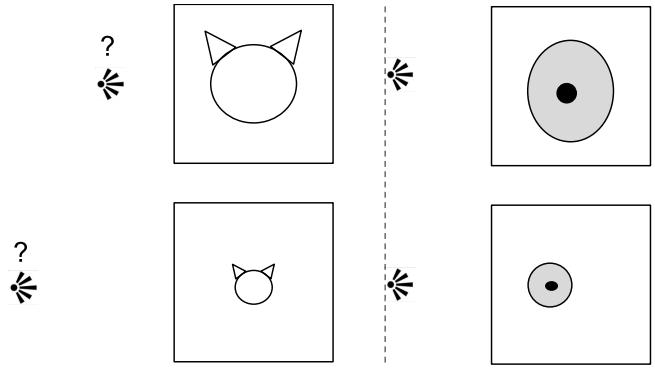
The developments in the following sections propose two methods that implement this perspective. Section 4.2 proposes a pruning module that is based on RCVs. The module analyzes the presence of information about scale at intermediate layer features and prunes off the layers that introduce invariance to this information. Preserving information about scale is, in fact, important in medical applications where the observation viewpoint is known and the size of an instance in the image may have an associated meaning, e.g. tumor extension. Section 4.3 describes an architecture that can be used to encourage (or discourage) the learning of arbitrary concepts. This architecture combines two successful techniques, namely multi-task learning (Caruana 1997) and adversarial training (Ganin et al. 2016) to accentuate the learning of specific concepts in the internal model representations. The evaluation is performed by quantifying with ablation studies the observed improvement in performances given by the proposed methods.

The content of Section 4.2 is adapted from the conference work in Graziani, Lompech, Müller, Depeursinge & Andrearczyk (2020) and the peer-reviewed work Graziani, Lompech, Müller, Depeursinge & Andrearczyk (2021). Section 4.3 reports my work in Graziani, Otálora, Marchand-Maillet, Müller & Andrearczyk (2021), which is currently under review.

4.2 Preserving Scale-covariant Features with Interpretable Pruning

The method proposed in this section improves the performance of a magnification regression model by developing an interpretable pruning module. The pruning uses the interpretability technique of RCVs in Section 3.4 to identify which network layers should be pruned to preserve scale-covariant features.

With this work, I aim at evaluating whether interpretability can be used to improve the performance of transfer learning from pre-training on the ImageNet dataset of natural images (Deng et al. 2009). Transfer learning is widely applied in medical imaging, leading to improvements in terms of model accuracy and speed of convergence (Litjens et al. 2017). Despite the considerable domain shift given by the reduced number of classes and the limited color, texture and object variability (Raghu et al. 2019), basic features learned during pre-training such as color, edges and textures are re-used by medical imaging applications (Huh et al. 2016, Graziani, Andrearczyk & Müller 2019). The invariance to multiple object scales is also one of the features implicitly learned by the layers, since objects appear naturally at multiple distances from the observation point, hence at multiple scales. As the illustration in Figure 4.1a shows, the observation viewpoint is unknown in natural images, and instances of a category covering the input area at different ratios belong to the same class because they represent the same object. In medical images, however, the viewpoint is controlled (as in Figure 4.1b) and voxel spacing has a known corresponding physical dimension. Scale is informative, if not decisive, in some tasks such as estimating the lesion size. A specific design that retains the helpful features learned during pre-training and discards scale invariance may thus perform better than both standard



(a) unknown viewpoint (b) fixed viewpoint

Figure 4.1: Illustration of (a) an unknown and varying viewpoint typical in natural images that requires scale-invariant analysis and (b) a controlled viewpoint in which a difference in size carries crucial information that is discarded by a scale invariant analysis. Replicated from Graziani, Lompech, Müller, Depeursinge & Andrearczyk (2021).

transfer and training from scratch.

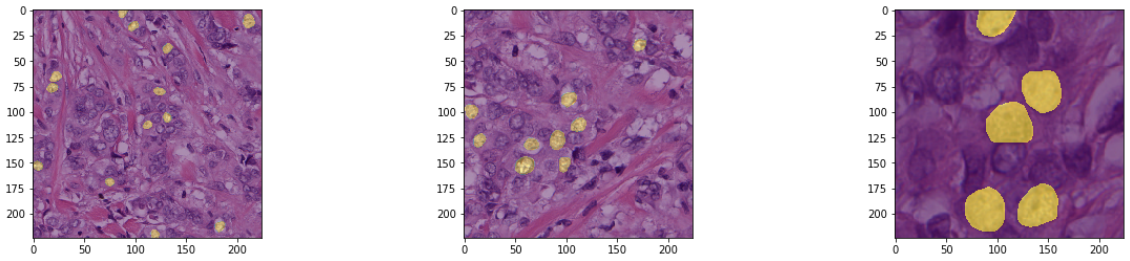
This study may help to build models that predict the magnification range of images for which the physical dimension of voxels is unknown, e.g. magnification level not reported. This may have a positive impact on the use of large and growing open-access biomedical data repositories such as PubMed Central[18] to extend existing medical datasets (Müller et al. 2020).

4.2.1 Related work

Features that are covariant to a scale transformation[19] are obtained with ad-hoc designs in the literature (Kanazawa et al. 2014, Marcos et al. 2018, Worrall & Welling 2019, Ghosh & Gupta 2019). Built-in covariance is proposed in Worrall & Welling (2019), for example, by enforcing the disentanglement of the features for transformations including rotation and scale variations. These methods, however, require large datasets for training the model parameters and transfer learning remains the most common practice to apply deep learning to medical imaging (Raghu et al. 2019). The behavior of pre-trained CNNs on ImageNet is analyzed in multiple studies, some of which pay particular attention to the encoding of scale-related information (Raghu et al. 2019, Yosinski et al. 2015, Aubry & Russell 2015, Lenc & Vedaldi 2015). The work in Yosinski et al. (2015) proposes an analysis of manually selected deep activations that respond to faces viewed at different scales. Aubry & Russell (2015) use computer-generated images to control attributes (concept measures, including scale) of a single object and visualized the effect on the internal representations. In Lenc & Vedaldi (2015), the regression of geometric image transformations (e.g. image flips and half-rescaling) is studied to learn the homomorphic transformations in the feature space that account for the transformations of the input. Importantly, the conclusion from this study is that scale invariance is implicitly learned on ImageNet since the model accuracy is not improved by reversing the scaling transformations in the feature space.

Related work concerning pruning approaches also exists in the literature (Molchanov

[18] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/
[19] also referred to as continuous features with this transformation.



(a) 10X (b) 15X (c) 40X

Figure 4.2: Examples of histopathology images at 10, 15 and 40X with nuclei segmentations. Replicated from Graziani, Lompech, Müller, Depeursinge & Andrearczyk (2021).

et al. 2017, Liu et al. 2016), with medical applications for PAP smear imaging in Wang et al. (2020) and Chest X-rays in Fernandes & Yen (2021). These methods mostly focus on identifying the importance of individual elements in the network, such as individual neurons in the work proposed by Molchanov et al. (2017), or individual filters and feature maps in the approaches developed by Fernandes & Yen (2021) and Wang et al. (2020). The pruned networks achieve similar performance, if not better, than the original networks. The asset of network pruning is that even if not providing massive increases in network performance it improves training convergence and it reduces the number of parameters to be trained and thus the computational complexity of the models (Fernandes & Yen 2021). This allows the training and fine-tuning of the models on smaller datasets, as shown by the study on PAP smears in Wang et al. (2020). Liu et al. (2016) dealt with multiple object scales by specific-design observations that can make their pruning responsive to multiple object scales. Differently from the existing works (Molchanov et al. 2017, Wang et al. 2020), the method proposed in this work prunes off entire network layers based on the quantification of scale information at each layer. An explicit design as that used in Liu et al. (2016) is not needed for this method, nor the expensive computations of evolutionary strategies used in Fernandes & Yen (2021). Any architecture pre-trained on ImageNet can be analyzed and pruned by the method here proposed.

4.2.2 Methods

Datasets The experiments in this paper involve two datasets since the scale analysis is performed on natural images and the proposed final architecture is evaluated on a medical image analysis task. For the scale quantification part, images with manual annotations of bounding boxes are selected from the publicly available PASCAL-VOC dataset (Everingham et al. 2010). The analysis is restricted to three object categories and images containing a single bounding box, chosen among the available annotated classes. These are *albatross* (ID: n02058221, 441 images), *kite* (ID: n01608432, 406 images) and *racing car* (ID: n04037443, 365 images).

For the histopathology application, the data consist of 141 Whole Slide Images (WSI) of Estrogen Receptor-positive Breast Cancer (ERBCa+) taken from the collection in Janowczyk & Madabhushi (2016). For these images of $2,000 \times 2,000$ pixels, manual annotations of $12,000$ nuclei are available. Image regions of 224×224 pixels are extracted as image patches from the WSIs. A total of $69,019$ patches with nuclei segmentation masks are split into training, validation and test partitions (approximately 60%, 20%, 20% respec-

Split/# patches	5X	8X	10X	15X	20X	30X	40X	*Total*
Train	94	2,174	4,141	7,293	9,002	10,736	11,638	45,078
Validation	8	588	1,197	2,132	2,604	3,504	3,150	12,733
Test	36	428	900	1,728	2,198	2,802	3,166	11,208
Total	138	3,190	6,238	11,153	13,804	16,592	17,904	69,019

Table 4.1: Number of ERBCa+ patches extracted per magnification and partition. Adapted from Graziani, Lompech, Müller, Depeursinge & Andrearczyk (2021).

tively) as shown in Table 4.1. To not introduce bias, all the patches from a single image are assigned to the same data partition. The imbalance in the magnification categories is due to the area covered by each magnification level. The average nuclei area is extracted for each input image by computing the average number of pixels in the relative nuclei segmentation mask. Example images with overlaid segmentation masks are displayed in Figure 4.2.

Architectures Inception V3 (Szegedy et al. 2016) and ResNet 50 (He et al. 2016*b*) are used for the analysis with pre-trained ImageNet weights. The networks produce a vector of probabilities $f(\mathrm{X}) \in [0,1]^{1000}$, where $\sum_{i=1}^{1000} f(\mathrm{X})[i] = 1$. The histopathology task is the prediction of the magnification of the images of the histopathology images in the data. This is done following the approach in Otálora et al. (2018), which proposed to first predict the average nuclei area in the patches and then mapping these to the magnification category that has the closest mean average value of the nuclei areas in the training set. This mapping approach was used since it outperforms the direct classification of the magnification as shown in Otálora et al. (2018). Transfer to the histopathology data is performed from both the original and pruned architectures. The average area of the nuclei is predicted by a single-unit dense layer. The model is trained to minimize the Mean Squared Error (MSE) loss between the true areas and the predicted ones. The nuclei area is expressed for each image as the average number of pixels within the segmentation of the nuclei present in the image. The networks are implemented in Keras and trained for five epochs with an Adam optimizer and standard hyper-parameters (learning rate 1e-4, batch size 32, and default values of the exponential decay rates). The full pipeline is shown in Figure 4.3 and the source code is available on GitHub for reproducibility[20]. Training times on a single NVIDIA GPU V100 are below one hour for each model.

Bounding-Box Size vs. Image Size Indications of scale are commonly used to relate the dimensions of two objects. In design modeling and cartography, the scale is the ratio comparing the length of the represented segment to the one in the real world (i.e. 1 cm:1000 Km). Computer vision and image processing mostly refer to the act of scaling, namely the transformation that generates a new image with a larger or smaller number of pixels. One may intuitively think of the scaling transformation $g_\sigma(\cdot)$ as a reshaping operation that can be performed on the inputs. Input size and object scale are, however, represented differently by the CNN. A "train-test" resolution discrepancy was already observed in Touvron et al. (2019) during network inference.

[20] `shorturl.at/gAQZ2`

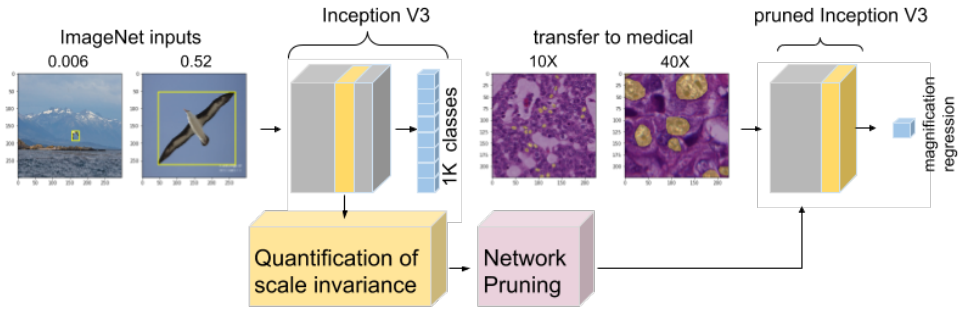


Figure 4.3: Pipeline of scale quantification and consequent network pruning for better transfer to medical tasks. The bounding boxes for inputs of the ImageNet class *albatross* and the segmentation masks for the ERBCa+ inputs (at 10 x and 40 x magnifications) are overlaid in yellow on the images. The bounding box ratios r are on top of the ImageNet inputs. The layer in yellow is the most informative about scale according to our quantification. The pruned network drops the layers after this point. Replicated from Graziani, Lompech, Müller, Depeursinge & Andrearczyk (2021).

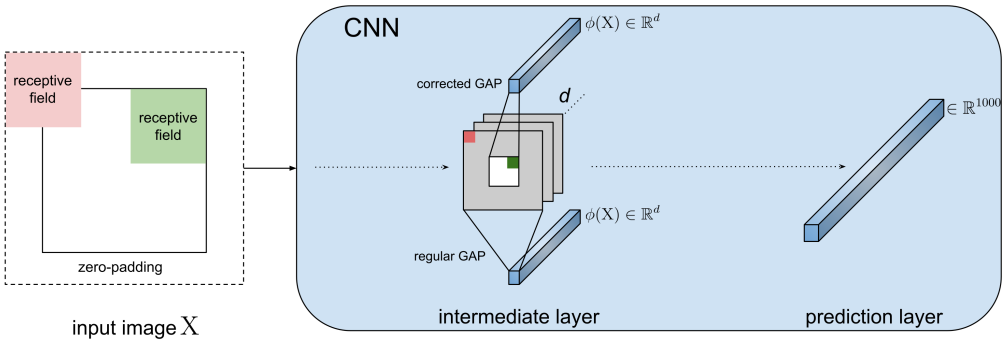


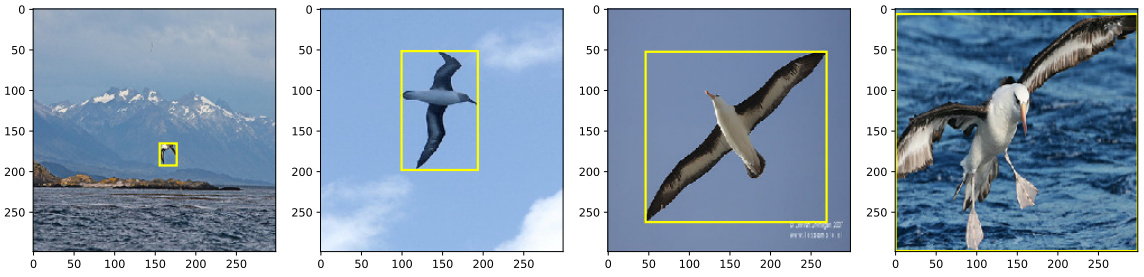
Figure 4.4: Illustration of the working principle of the corrected GAP. The colored receptive fields in the input image (**left**) are associated with the colored neurons in the feature maps (**center**). In the Convolutional Neural Network (CNN), activations used for the corrected GAP (**top**) are displayed in white that is, activations of the neurons with a receptive field contained in the input image. All activations are used for the regular GAP (**bottom**). Replicated from Graziani, Lompech, Müller, Depeursinge & Andrearczyk (2021).

The first experiment in Section 4.2.3 wants to demonstrate the hypothesis that input size and scale are two different types of information. Information about image size, for instance, is encoded in the features from the padding effect of early convolutional layers. This is verified by introducing the corrected GAP operation illustrated in Figure 4.4, which discards the activations at the border of the feature maps since these are affected by padding operations. The corrected GAP averages only the activations of the neurons with a receptive field contained entirely in the input image. The experiment evaluates whether image size can be regressed from noise inputs in the intermediate layers of the CNN. To only analyze image size, images of white noise of varying sizes are used as inputs, since they do not contain any object nor related scale. If the network encodes information about the image size differently from the object scale, then the input size should be possible to obtain from the noise inputs. If this information is encoded from the padding at early layers, then the regression with the corrected GAP should fail as this

operation discards the edges of the feature maps. Therefore, the regression of the image scale with and without the corrected GAP is compared to show that current state-of-the-art CNN architectures encode information about the image size. The regression vector $\mathbf{v}$ in Equation (4.2) is sought to regress the image width s_i.

For the other experiments, the input size is fixed to the default of Inception V3, i.e. $S_i = 299 \times 299$. Images are chosen so that they contain a single object. In this context, image scale is pragmatically defined as the solid angle of the object in the image, namely the proportion of the field of view occupied by an object (Yan & Huang 2021). A small bounding box corresponds to a smaller space in the field of view of the camera, and thus a smaller solid angle. Scale measures are thus defined as the ratio $r = \frac{S_b}{S_o} = \frac{h_b \times w_b}{h_o \times w_o}$, where h_b and w_b are the bounding box height and width. The image area is $S_o = h_o \times w_o$, where h_o and w_o are respectively the original image width and height.

Examples of multiple scale measures for the same object category are shown in Figure 4.5.



(a) $r = 0.006$ (b) $r = 0.154$ (c) $r = 0.524$ (d) $r = 0.975$

Figure 4.5: Examples of albatross images and their respective scale measures used for learning the regression. Replicated from (Graziani, Lompech, Müller, Depeursinge & Andrearczyk 2021).

Quantification of scale information The act of scaling is defined in image processing as a transformation $g_\sigma(\cdot)$ that generates a new image with a larger or smaller number of pixels, depending on the scaling factor σ. The aim of the scale quantification module proposed in this work is quantifying the covariance (4.1) of a mapping $\phi(\cdot)$ to a transformation $g(\cdot)$, where covariance is defined as follows[21]:

$$\phi(g(\cdot)) = g'(\phi(\cdot)). \tag{4.1}$$

This corresponds to seeking for a linear transformation $g'_\sigma(\cdot)$ that is a predictable transformation of $g_\sigma(\cdot)$ in the input space. The covariance being measured is that of $\phi(\boldsymbol{x})$, namely of the averaged feature maps of intermediate layers (or the activations of fully-connected layers) for the input image $\boldsymbol{x}$. Measuring the covariance of the function $\phi(\boldsymbol{x})$ means, therefore, finding a transformation $g' : \mathbb{R}^d \to \mathbb{R}^d$ in the feature space that predicts a transformation $g : \mathbb{R}^{h \times w} \to \mathbb{R}^{h \times w}$ of the input image. This is done by searching the regression vector $\boldsymbol{v}$ (i.e. the RCV) in the feature space to predict the scaling factor σ as[22]:

$$\sigma = \sum_i v_i \phi_i(g_\sigma(\boldsymbol{x})) = \boldsymbol{v} \cdot \phi(g_\sigma(\boldsymbol{x})). \tag{4.2}$$

[21] Note that invariance is defined as $\phi(g(\cdot)) = \phi(\cdot)$. Equivariance is a particular case of covariance, when $g'(\cdot) = g(\cdot)$.

[22] For simplicity, we omit the intercept. In Equation (4.2), the intercept is v_0 with $\phi_0(g_\sigma(\mathrm{X})) = 1$.

The transformation $g'_\sigma(\cdot)$ can be represented as a translation matrix (in $\mathbb{R}^d$) by σ along $\boldsymbol{v}$, so that $g'_\sigma(\phi(\boldsymbol{x})) = \phi(\boldsymbol{x}) + \boldsymbol{v} \cdot \sigma$. The RCV in Eq. 4.2 is $\boldsymbol{v}$, and it corresponds to the representation of the concept "scale" at an intermediate layer.

The regression is sought at several layers in the network to compare different depths. Aggregation is performed on the feature maps in the form of GAP to obtain the feature vector $\phi(\boldsymbol{x})$ (except for the prediction layer which is already pooled). The determination coefficient R^2 is used to evaluate the prediction of the scale ratio r on unseen test data of the same class[23]. This evaluation is informative about the scale-covariance of the features. The R^2 is a measure between zero and one when the prediction of the regression on the test samples is better than predicting their mean. When the prediction of the model is worse than the mean, the R^2 is negative. To maintain the score in a [0,1] range the test R^2 is normalized by evaluating $\frac{e^{R^2}}{e}$, with values below $\frac{1}{e}$ evidencing bad performance.

Pruning strategy Network pruning is performed by comparing the test R^2 to identify the layer where the scale covariance is the highest. This evaluation is averaged across different object categories to remove the dependence on the class of the inputs. The layer with the highest test R^2 (the yellow layer in Fig. 4.3) is where the scale covariance is the highest. Layers deeper than this one are pruned off the architecture and GAP is added to obtain a vector of the aggregated features.

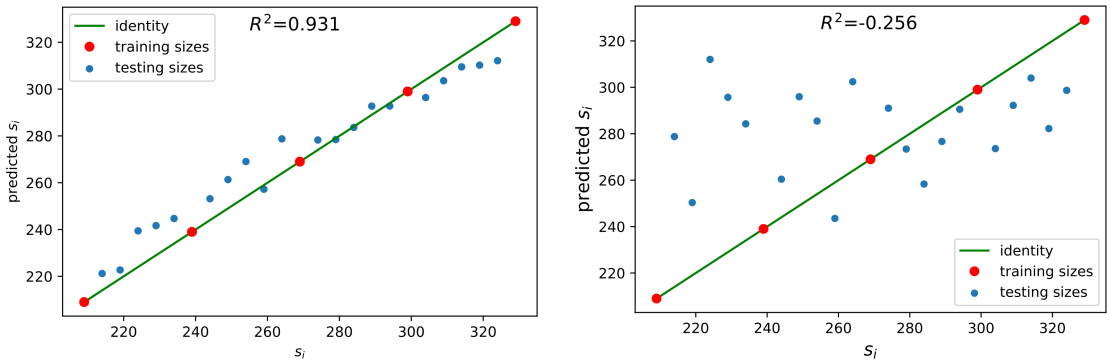
Evaluation The transfer learning experiments are evaluated by the Mean Average Error (MAE) and Cohen's kappa coefficient. MAE is used to evaluate the regression of the average nuclei areas, while Cohen's kappa coefficient is used to measure the inter-rater reliability of the prediction of the magnification classes.

4.2.3 Experiments and Results

Input size against scale The experiments start by demonstrating that a scaling operation $g_\sigma(\cdot)$ of a factor σ cannot be performed as a simple input reshaping operation, since the CNN features encode information about image size differently from the object scale. The regression of s_i is learned from five noise images and evaluated on 20 held-out images. A small number of images is intentionally used to illustrate the simple linear correlation. Similar results are obtained when using more images. The results show that we can regress the size for the model with the regular GAP in deep layers, with the R^2 close to one in Figure 4.6a. On the contrary, Figure 4.6b shows that we cannot regress the size information when aggregating the feature maps using the corrected GAP ($R^2 < 0$). In light of these results, we do not associate the input size to the measure of object scale in the subsequent analyses.

Note that since the receptive fields grow throughout the network, the region of activations unimpacted by the paddings reduces up to a point where no activation remains for the corrected GAP. Because of this limitation, this method is only used to show the impact of zero-padding but it cannot be used for the analysis of scale invariance throughout the entire network.

[23]We compute $R^2 = \frac{\sum_{i=1}^{N}(\hat{r}_i - \bar{r})}{\sum_{i=1}^{N} r_i - \bar{r}}$, were N is the number of test data samples, $\hat{r}$ is the ratio predicted by the regression model, $\bar{r}$ is the mean of the true ratios $\{r_i\}_{i=1}^{N}$.



(a) regression with regular GAP

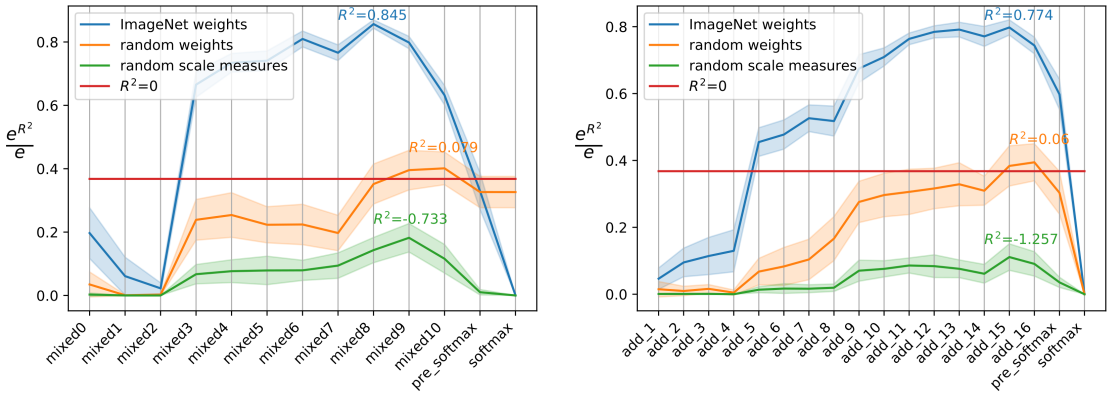
(b) regression with corrected GAP

Figure 4.6: Regression of size s_i at layer *mixed 0* with noise inputs. The R^2 is shown for the prediction of scale measures on held-out noise images. Results obtained for (**a**) Regular GAP; (**b**) Corrected GAP. Replicated from Graziani, Lompech, Müller, Depeursinge & Andrearczyk (2021).

Scale quantification The next experiments focus on the regression of scale measures in ImageNet pre-trained models for the object categories albatross (ID: n02058221), race car (ID: n04037443) and kite (ID: n01608432). 70% of the input class images are used to learn the regression, while the remaining ones are held out for evaluating the determination coefficient. Figure 4.7a compares the scale regression at multiple depths in a randomly initialized Inception V3 (orange line) and one trained on ImageNet (blue line). A baseline in which the regression is trained with random concept measures obtained from shuffling the scale concept measures before regression is also shown in the figure (green line). Similar results are obtained for the other classes[24].

Improvement of transfer to pathology Here are reported the experiments on the transfer to the histopathology task. The original Inception V3 and ResNet 50 networks are compared to their pruned counterparts in terms of performance in the nuclei area and magnification prediction in Table 4.2. The MAE is computed over ten repetitions for multiple seed initializations of the dense connections of the last prediction layer. The standard deviation is reported in brackets. Cohen's kappa coefficient is used to evaluate the prediction of the magnification category. The results show significant improvements when the networks are pruned at the layer suggested by the pruning strategy for both tasks. This validates the utility of the proposed scale invariance analysis. The non-parametric Wilcoxon signed-rank test is used to evaluate the statistical significance (p-value ¡ 0.001 for the MAE and kappa with both networks). The average MAE (standard deviations reported in brackets) between the true nuclei areas and those predicted by the pruned Inception V3 are respectively 55.33 (31.16) for 5X images, 42.15 (11.39) for 8X, 34.65 (0.15) for 10X, 33.28 (0.69) for 15X, 48.38 (5.26) for 20X and 81.05 (15.67) for 40X images.

[24]For the precise values of the R^2 the reader may refer to the extensive results in the paper Graziani, Lompech, Müller, Depeursinge & Andrearczyk (2021).



(a) Inception V3 (b) ResNet 50

Figure 4.7: Comparison of regression (RCV) of scale measures at different layers on the albatross ImageNet class (ID: n02058221). The regression is evaluated as the R^2 of the prediction of scale measures on held-out images and $\frac{e^{R^2}}{e}$ is plotted for better visualization. Values above the red line $R^2 = 0$ show a predictive regression better than the average of ratios r. Average and standard deviations are reported for 25 runs.

Table 4.2: Mean Average Error (MAE) of the nuclei area regression (in pixels) and Cohen's kappa coefficient between the true and predicted magnification categories. Results are averaged across ten repetitions; the standard deviation is reported in brackets.

model	layer	MAE (std)	kappa (std)
pre-trained IV3	mixed10	81.85 (11.08)	0.435 (0.02)
from scratch IV3	mixed10	82.30 (17.92)	0.560 (0.09)
pruned IV3	mixed8	**54.93** (4.32)	**0.571** (0.05)
pre-trained ResNet 50	add16	70.08 (12.49)	0.610 (0.03)
from scratch ResNet 50	add16	95.66 (21.39)	0.461 (0.09)
pruned ResNet 50	add15	**54.76** (3.10)	**0.623** (0.04)

4.3 Learning Diagnostic Features with Multi-task Adversarial CNNs

This section introduces a methodology that guides the training of CNNs towards learning arbitrary concepts. The goal is to exploit the prior knowledge of physicians to guide the feature design of the model. The learned representations are encouraged to contain information about arbitrary diagnostic factors such as nuclei morphology and density. This may be used to ensure pathologists that the features used by the network align with their clinical requirements. Confounding factors such as staining variations can be discarded from the learned features.

The proposed architecture is obtained by building on top of successful techniques such as multi-task learning (Caruana et al. 2015) and domain adversarial training (Ganin et al. 2016). Learning diagnostic factors is introduced as an additional training objective, whereas discarding confounding features is modeled as an adversarial task. The CNN designed as such is encouraged to learn representations containing information about di-

agnostically relevant concepts representing nuclei morphology and density, called hereafter *desired targets*. A gradient reversal operation (Ganin et al. 2016, Xie et al. 2017) is used to obtain invariance to *undesired targets*, namely to the domain differences of the multiple acquisition centers, which are due to the tissue staining, fixation, processing and digitalization.

While multi-task learning (Caruana 1997) and adversarial learning (Ganin et al. 2016) are widely used techniques, fundamental in these contributions is their combination for steering the learning process. Balancing multiple tasks such as the regression of the nuclei contours and density is a technique that only recently arose interest in the digital pathology landscape (Gamper, Kooohbanani & Rajpoot 2020). The joint optimization is non-trivial and here I propose a novel exploration for the histopathology field. We analyze the benefits of an uncertainty-based approach to weight the multiple losses, showing that it best handles the convergence and stability of the joint optimization.

4.3.1 Related works

Similarly to how learning happens in humans, multi-task architectures aim at simultaneously learning multiple tasks that are related to each other (Ruder 2017). Multi-task learning has been successful in various applications, such as natural language processing (Subramanian et al. 2018), computer vision (Kokkinos 2017), autonomous driving (Leang et al. 2020), radiology (Andrearczyk et al. n.d.) and histology (Gamper, Kooohbanani & Rajpoot 2020). The preliminary work by Gamper, Kooohbanani & Rajpoot (2020), in particular, shows a decrease in the loss variance as an effect of multi-task for oral cancer, suggesting that this work may have a high potential for histology applications.

Multi-task architectures divide into two families depending on the hard or soft sharing of the parameters, both illustrated in Figure 4.8. In architectures with hard parameter sharing such as the one proposed in this paper, multiple supervised tasks share the same input and some intermediate representation (Caruana 1997). The parameters learned up to this intermediate point are called *generic parameters* since they are shared across all tasks. In soft parameter sharing, the weight updates are not shared among the tasks and the parameters are task-specific, introducing only a soft constraint on the training process (Duong et al. 2015).

As explained by Caruana (1997), multi-task learning leads to various benefits if the tasks are linked by a valid relationship, namely if what is learned for each task can help the other tasks to be learned better. The variations in the observed data must be explained by factors that are shared by two or more tasks (Goodfellow et al. 2016). Figure 4.9 helps understanding this concept by illustrating the explanation in Caruana (1997). The scenarios described in Figures 4.9 (a) and (c) suppose that the learning of two related tasks generates signals that contain extra information from which both can benefit. The additional task introduces an inductive bias in the model optimization that leads to more general and robust representations than traditional or multimodal learning. Let us suppose that a complex model, e.g. a CNN, is trained on the main task M. In the optimization objective of M has two local minima, represented as the set $\{a, b\}$. The auxiliary task A is related to the main task, with which it shares the local minimum in a in Figure 4.9 (a). The joint optimization of M and A is likely to identify the shared local minima a as the optimal solution (Caruana 1997). The search is biased by the extra information given by task A towards representations that lay at the intersection of what could be learned individually for each task. Under these conditions, the multi-task configuration improves

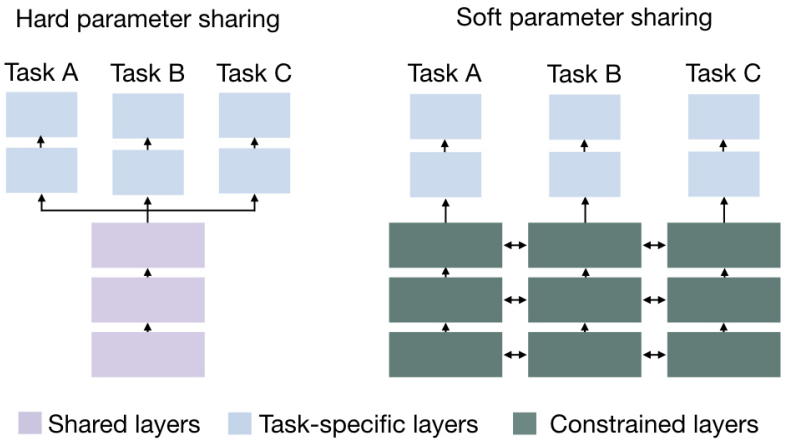


Figure 4.8: Hard and soft parameter sharing for multi-task learning. Adapted from Ruder (2017).

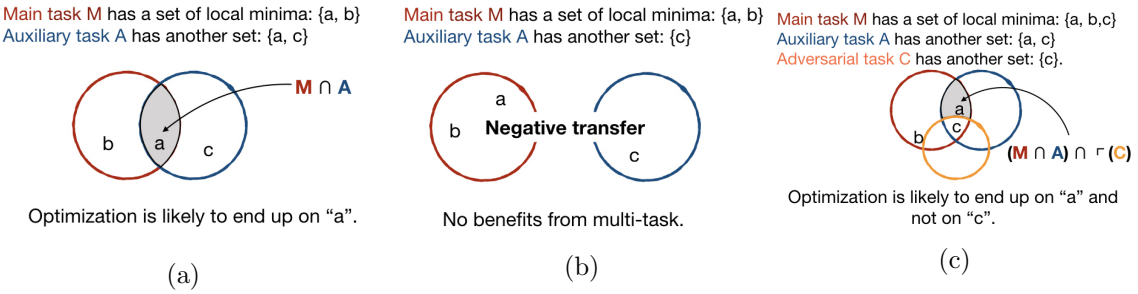


Figure 4.9: Intuitive illustration about multi-task learning in (a): given two related tasks M and A, the optimization process is driven to choose solutions that satisfy both tasks. In (b) no connection exists between the tasks, hence the multi-task approach may result in a negative transfer, providing only sub-optimal models for all the tasks. In (c), an adversarial task is added and the optimization is pushed to representations that satisfy both main and auxiliary tasks, but that avoid the minimum of the adversarial task.

the generalization error bounds and reduces the risk of overfitting (Baxter 1995). The speed of convergence is also increased since fewer training samples are required per task (Baxter 2000). If there is no valid relationship between the multiple tasks as in Figure 4.9 (b), then there are no local minima being shared and a negative transfer may happen without positive improvements to the performance. No relevant benefits are remarked, in this case, and there may be an eventual loss in performance. Finally, Figure 4.9 (c) shows an extension of the concepts in (Caruana 1997) with the addition of an extra adversarial task C. In this case, the main task M has local minima in $\{a, b, c\}$, but the minimum in c is also a solution of the adversarial task C. By being adversarial to C, the optimization is likely to prefer solutions that satisfy M and A, while avoiding solutions that satisfy the adversarial task C. Hence, the solution a should be favored by the concurrent action of both tasks A and C.

Note that losses from the multiple tasks contribute to the same objective function that is optimized during training. Depending on the tasks and on the losses used, multiple strategies for weighting the contributions can be adopted. A review of the multiple

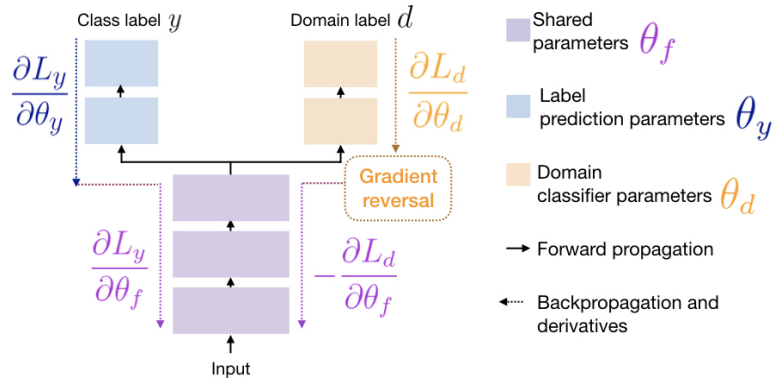


Figure 4.10: Illustration of domain adversarial training, where the label prediction loss is L_y and the domain prediction loss is L_d. Adapted from Ganin et al. (2016).

weighting strategies is given by the benchmarking paper of Gong et al. (2019). The authors claim no clear winner among the approaches, with often a uniform weighting strategy being sufficient. Alternatively to uniform weighting, dynamical task re-weighting during training is proposed by Leang et al. (2020) and uncertainty estimates are used in Kendall et al. (2018) to directly learn the best weights for each task.

The approach of adversarial training is illustrated in Figure 4.10. Proposed in Ganin et al. (2016), it represents a way to address the so-called problem of domain adaptation, namely the minimization of the domain shift in the distributions of the training (also called source distribution) and testing data (i.e. target). Typically treated as either an instance re-weighting operation (Gong et al. 2013) or as an alignment problem (Long, Cao, Wang & Jordan 2015), domain adaptation is handled by adversarial learning as the optimization of a domain confusion loss. A domain classifier discriminates between the source and the target domains during training and its parameters are optimized to minimize the error when discriminating the domain labels. This can be extended to more than two domains by a multi-class domain classifier. The adversarial learning of domain-related features is obtained by a gradient reversal operation on the branch learning to discriminate the domains. Because of this operation, the network parameters are optimized to maximize the loss of the domain classifier, thus making multiple domains impossible to distinguish one from another in the internal network representation. This causes a competition between the main task and the domain branch during training that is referred to as a min-max optimization framework. As a downside, the optimization of adversarial losses may be complicated, with the min-max operation affecting the stability of the training (Ganin et al. 2016). Convergence can be promoted, however, by following the training schedule in Lafarge et al. (2017), which activates and cyclically de-activates the gradient reversal branch.

4.3.2 Methods

Datasets The datasets used for the experiments are the Camelyon (Litjens et al. 2018) and PanNuke (Gamper, Koohbanani, Graham, Jahanifar, Khurram, Azam, Hewitt & Rajpoot 2020) as described in Section 3.2.2. The same training, validation and testing splits in Table 3.1 are used for the experiments.

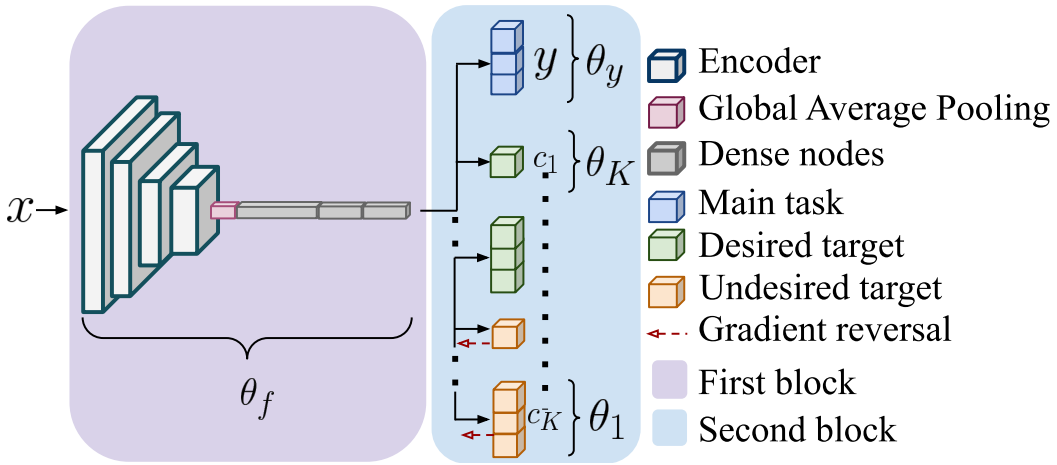


Figure 4.11: Multi-task adversarial architecture for guiding model training with arbitrary desired and undesired target features to learn.

Proposed Architecture The proposed architecture is described for a general application with pre-defined features since it is conceived to be applied to the classification of multiple image types and not only for digital pathology. The diagnosis of cancerous tissue in breast microscopy images is proposed as an application for which the implementation details are described later in this Section.

We assume that a set of N observations, i.e. the input images, is drawn from an unknown underlying distribution and split into a training subset $\{\boldsymbol{x}_i\}_{i=1}^{n}$ and a testing subset $\{\boldsymbol{x}_i\}_{i=n+1}^{N}$. The main task, namely the one for which we aim at improving the generalization, is the prediction of the image labels $\boldsymbol{y} = \{y_i\}_{i=1}^{n}$, for which ground truth annotations are available. A CNN of arbitrary structure is used as a feature encoder, of which the features are then passed through a stack of dense layers. The model parameters up to this point are defined as $\boldsymbol{\theta}_f$. The parameters of the label prediction output layers are identified by $\boldsymbol{\theta}_y$. The structure described up to this point replicates a standard CNN with a single main task branch that is addressing the classification. The remaining parameters of the architecture implement (i) the learning of auxiliary tasks by multi-task learning (Caruana 1997) and (ii) the adversarial learning of detrimental features to induce invariance in the representations, as in the domain adversarial approach by Ganin et al. (2016). We combine these two approaches by introducing K extra targets representing desired and undesired tasks that must be introduced to the learning of the representations. The targets are modeled as the prediction of the feature values $\{c_{k,i}\}_{i=1}^{N}$, where $k \in 1, \ldots, K$ is an index representing the extra task being considered. The feature values may be either continuous or categorical. Additional parameters $\boldsymbol{\theta}_k$ are trained in parallel to $\boldsymbol{\theta}_y$ for the K extra targets. We refer to all model outputs for all inputs $\boldsymbol{x}$ as $f(\boldsymbol{x}) \in \mathcal{R}^{K+1}$.

The architecture is illustrated in Figure 4.11 and consists of two blocks. The first block is used to extract features from the input images. A state-of-the-art CNN of arbitrary choice without the decision layer is used as a feature encoder generating a set of feature maps. The feature maps are passed through a Global Average Pooling (GAP) operation that is performed to spatially aggregate the responses and connect them to a stack of dense layers. For this specific architecture, we use a stack of three dense layers of 1024, 512 and 256 nodes respectively. The second block comprises one branch per task, taking as input the output of the first block. The main task branch consists of the prediction

of the labels $\mathbf{y}$ and has as many dense nodes as there are of unique classes in $\mathbf{y}$. For binary classification tasks, e.g. discrimination of tumorous against non-tumorous inputs, the main task branch has a single node with a sigmoid activation function. K branches are added to model the extra targets. We refer to *extra* tasks for all the additional targets to the main task whether desired or undesired. *Auxiliary* tasks refer to the modeling of the desired targets, while *adversarial* tasks refer to that of undesired targets. The extra tasks are modeled by linear models as in Graziani et al. (2018). For continuous-valued targets, the extra branch consists of a single node with a linear activation function. For categorical targets, the extra branch has multiple nodes followed by a softmax activation function. A gradient reversal operation (Ganin et al. 2016) is performed on the branches of the undesired targets to discourage the learning of these features.

Objective Function The objective function of the proposed architecture balances the losses of the main task and the extra tasks for the desired and undesired targets. This is obtained by a combination of multi-task and adversarial learning. The main task loss is $\mathcal{L}_y^i(\boldsymbol{\theta_f}, \boldsymbol{\theta_y}) = \mathcal{L}_y(\boldsymbol{x}_i, y_i; \boldsymbol{\theta_f}, \boldsymbol{\theta_y})$, where $\boldsymbol{\theta}_f$ are the parameters of the first block (namely of the CNN encoder and the dense layers) in Figure 4.11 and $\boldsymbol{\theta}_y$ those of the main task branch in the second block of the same figure. The extra parameters $\boldsymbol{\theta}_k$ $(k \in 1, \ldots, K)$ are trained for the branches of the desired and undesired target predictions, with the loss being $\mathcal{L}_k^i(\boldsymbol{\theta}_f, \boldsymbol{\theta}_k) = \mathcal{L}_k(\boldsymbol{x}_i, c_{k,i}; \boldsymbol{\theta}_f, \boldsymbol{\theta}_k)$.

Training the model on n training an $(N - n)$ testing samples consists of optimizing the function:

$$E(\boldsymbol{\theta}_y, \boldsymbol{\theta}_f, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K) = \lambda_m \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_y^i(\boldsymbol{\theta}_f, \boldsymbol{\theta}_y) + \sum_{k=1}^{K} \lambda_k \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_k^i(\boldsymbol{\theta}_f, \boldsymbol{\theta}_k). \tag{4.3}$$

The gradient update is:

$$\boldsymbol{\theta}_f \leftarrow \boldsymbol{\theta}_f - \left(\lambda_m \frac{\partial \mathcal{L}_y^i}{\partial \boldsymbol{\theta}_f} + \sum_{k=1}^{K} \lambda_k \alpha_k \frac{\partial \mathcal{L}_k^i}{\partial \boldsymbol{\theta}_f} \right), \tag{4.4}$$

$$\boldsymbol{\theta}_y \leftarrow \boldsymbol{\theta}_y - \lambda_m \frac{\partial \mathcal{L}_y^i}{\partial \boldsymbol{\theta}_y}, \tag{4.5}$$

$$\boldsymbol{\theta}_k \leftarrow \boldsymbol{\theta}_k - \lambda_k \frac{\partial \mathcal{L}_k^i}{\partial \boldsymbol{\theta}_k}, \tag{4.6}$$

where λ_m and λ_k are positive scalar hyper-parameters to tune the trade-off between the losses. For each extra branch, the hyper-parameter $\alpha_k \in \{-1, 1\}$ is used to specify whether the update is adversarial or not. A value of $\alpha_k = -1$ activates the gradient reversal operation and starts an adversarial competition between the feature extraction and the corresponding k^{th} extra branch. The main task is only trained on the training data, since $\mathcal{L}_y^i = 0$ for $i > n$ in Eq. (4.4) and (4.5) as in Ganin et al. (2016). The extra tasks are learned on both training and test data. The training on test data can also be removed since it is not always possible to fully retrain a network for new data.

Loss weighting strategy The proposed architecture requires the combination of multiple objectives in the same loss function. The vanilla formulation in Eq. 4.3 simply performs a weighted linear sum of the losses for each task. This is the predominant approach used in

prior work with multi-objective losses (Gong et al. 2019) and adversarial updates (Ganin et al. 2016, Lafarge et al. 2017). The appropriate choice of a weighting strategy for the multiple task losses is a major challenge of this setting. The tuning of the hyper-parameters may reveal tedious and non-trivial due to the combination of classification and regression tasks with different ranges of the loss function values (e.g. combining the bounded binary cross-entropy loss in [0,1] with the unbounded MSE loss).

An optimal weighting approach may be learned simultaneously with the other tasks by adding network parameters for the loss weights λ_m and λ_k. The direct learning of λ_m and λ_k, however, would just result in weight values quickly converging to zero. Kendall et al. (2018) proposed a Bayesian approach that makes use of the homoscedastic uncertainty of each task to learn the optimal weighting combination. In loose words, homoscedastic uncertainty reflects task-dependent confidence in the prediction. The main assumption to obtain an uncertainty-based loss weighting strategy is that the likelihood of the task output can be modeled as a Gaussian distribution with the mean given by the model output and a scalar observation noise σ:

$$p(\boldsymbol{y}|f(\boldsymbol{x})) = \mathcal{N}(f(\boldsymbol{x}), \sigma^2) \tag{4.7}$$

This assumption is also applied to the outputs of the extra tasks. The loss weights λ_m and λ_k are then learned by optimizing the minimization objective given by the negative log-likelihood of the joint probability of the task outputs given the model predictions. To clarify this concept, let us focus on a simplified architecture with the main task being the logistic regression of binary labels (e.g. tumor v.s. non-tumor) with noise σ_1 and one auxiliary task consisting of the linear regression of feature values $\boldsymbol{c} = \{c_i\}_{i=1}^N$, with noise σ_2. The minimization objective for this multi-task model is:

$$-\log p(\boldsymbol{y}, \boldsymbol{c}|\boldsymbol{f}(\boldsymbol{x})) \propto \frac{1}{2\sigma_1^2}\mathcal{L}_y(\boldsymbol{\theta}_f, \boldsymbol{\theta}_y) + \frac{1}{2\sigma_1^2}\mathcal{L}_k(\boldsymbol{\theta}_f, \boldsymbol{\theta}_k) + \log\sigma_1 + \log\sigma_2 \tag{4.8}$$

By minimizing Eq. 4.8 w.r.t. σ_1 and σ_2, the optimal weighting combination is learned adaptively based on the data (Kendall et al. 2018). As σ_1 increases, the weight for its corresponding loss decreases, and vice-versa. The last term $\log\sigma_1 + \log\sigma_2$, besides, acts as a regularizer discouraging each noise to increase unreasonably. This construction can be extended easily to multiple regression outputs and the derivation for classification outputs is given in Kendall et al. (2018).

Configuration for the histopathology task The classification of breast histopathology images containing tumor from those of normal tissue is the main task used for the experiments. Inception V3 pre-trained on the ImageNet (Szegedy et al. 2016) is used as the backbone CNN for feature encoding. The parameters up to the last convolutional layer are kept frozen to avoid overfitting to the pathology images. The output of the CNN is passed through the GAP and the three fully-connected layers as illustrated in Figure 4.11. The fully-connected layers have respectively 2048, 512 and 256 units. A dropout probability of 0.80 and L2 regularization are added to these three fully-connected layers to avoid overfitting. The main task is the detection of patches containing tumor as a binary classification task. The branch consists of a single node with a sigmoid activation function connected to the output of the third dense layer. The architecture as described up to here, hence without extra branches, is used as the baseline for the experiments. The extra tasks consist of either the linear regression or the linear classification of continuous or categorical labels respectively. For linear regression, the extra branch is a single node with a

linear activation function. The MSE between the predicted value and the label is added to the optimization function in Eq. 4.3. For the linear classification, the extra branch has as many dense nodes as the number of classes, and a softmax activation function, also connected to the third dense layer. The Categorical Cross-Entropy (CCE) loss is added to the optimization in Eq. 4.3.

The architecture is trained end-to-end with mini-batch Stochastic Gradient Descent (SGD) with standard parameters (learning rate of 10^{-4} and Nesterov momentum of 0.90). The main task loss function is the class-Weighted Binary Cross-Entropy (WBCE). The class weights are set to weigh more heavily every instance of the positive class, for instance, they are set to the ratio of negative samples $136774/29513+136774 = 0.82$ for the positive class and the ratio of positive samples 0.18 for the negative class.

The convergence of the network is evaluated by early stopping on the total validation loss with patience of 5 epochs. The AUC is used to evaluate model performance. For each experiment, the performance variation due to initialization is evaluated over five runs with varying starting seeds and unchanged data splits. The performance on multiple test splits is evaluated by bootstrapping of the test sets. A number of 50 test sets of 7589 images (the total number of test images in the two sets) are obtained by sampling with repetition from the total pool of testing images. This method evaluates the variance of the test set without prior assumption on the data distribution and it shows the performance difference due to variation of the sampling of the population.

Breast Cancer Targets The experiments focus on the integration of four desired and one undesired target with multiple combinations. Learning the desired features is expected to improve the solution robustness and generalization of the model over the baseline. Discarding the undesired targets may introduce invariance to confounding factors in the deep features. The grading of breast tissue introduced in Section 2.1 is used to identify the key diagnostic features for breast cancer. The desired and undesired features that can be derived from this grading are illustrated in Figure 4.12. Note that only the cancer indicators at the nuclear level are used for the experiments since the input images are at the highest magnification. Variations of the nuclei size, appearance (e.g. irregular, heterogeneous texture) and density shown in Figure 4.12 are modeled as real-valued variables. Because of the heterogeneity of the data, we also guide the network training to discard information about staining and tissue representation differences in the images. The processing center of the slides is modeled as an undesired target, encouraging feature invariance to staining and acquisition differences. Hand-crafted features representing the variations in the nuclei size and appearance are automatically extracted either from the images or from the nuclear contours. The nuclear contours are available in the form of manual annotations only for the PanNuke data. Automated contours of the nuclei in the Camelyon images are obtained by a multi-instance deep segmentation model. This model is the Mask-RCNN model (He et al. 2017) described in Section 3.2, which was developed by Kumar et al. (2017) for the nuclei segmentation challenge and for which fine-tuned weights are available for re-use. The R-CNN identifies nuclei entities and then generates pixel-level masks by optimizing the Dice score. ResNet 50 (He et al. 2017) is used for the convolutional backbone.

The number of pixels inside nuclear contours is averaged for each input image to represent variations of the nuclei area, referred to as *area* in the experiments. Nuclei *density* is estimated by counting the nuclei in the image. Haralick descriptors of texture contrast and correlation (Haralick 1979) are also extracted from the entire input images

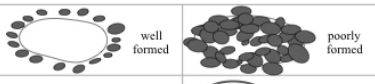
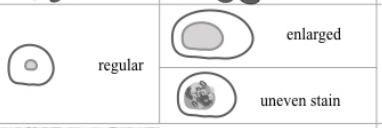
Concept	Clinical reference	Description	Visual examples	Magnification	Source	Type	Task
Count of cavities	NGH tubular formation	tumour cells in gland structure		low	annotation or automated	D	auxiliary regression
Nuclei area	NGH nuclear pleomorphism	abnormality in size		high	annotation or automated	C	auxiliary regression
Nuclei Texture		vesicular appearance					
Mitotic count	NGH mitotic count	number of mitosis		high	annotation or automated	D	auxiliary regression
Nuclei density	Proliferation index	Reproduction rate		any	annotation or automated	D	auxiliary regression
Staining	Staining procedure	dye applied on the tissues		any	metadata	D	adversarial classification

Figure 4.12: Control targets for breast cancer. C and D stand for continuous and discrete respectively.

as in Graziani et al. (2018). Being continuous and unbounded measures, the values for these features are normalized to have zero mean and unitary standard deviation before training the model. In the paper, we refer to these features as *area*, *density*, *contrast* and *correlation*. The values of these features are used as prediction labels for the auxiliary target branches, which are also named as the feature that they should predict. These auxiliary branches perform a linear regression task, trying to minimize the MSE between the predicted value of the feature and the extracted values used as labels.

Information about the center that performed the data acquisition is present in the dataset as metadata. This is modeled as a categorical variable that may take values from 0 to 7, namely one for each known center in the training data. Since there is no specific information on acquisition centers in Camelyon16 and PanNuke, these have been modeled as two distinct acquisition centers in addition to the five known centers of Camelyon17. This information is partly inaccurate since we know that in both datasets more than a single acquisition center was involved Litjens et al. (2018), Gamper, Koohbanani, Graham, Jahanifar, Khurram, Azam, Hewitt & Rajpoot (2020). The noise introduced by this information may limit the benefits introduced by the adversarial branch but it should not affect negatively the performance. In the future, unsupervised domain alignment methods may also be explored. The prediction of this variable is added to the architecture as an undesired target branch, referred to as *center* in the experiments.

4.3.3 Experiments and Results

Desired and undesired targets are added as extra branches in the second block of the architecture following multiple configurations. The experiments initially focus on adding one extra branch at a time to identify the benefits of encouraging each task individually. Subsequently, the most promising branches are combined to evaluate whether their combination may further improve performance from the one obtained in the single-branch experiments. The undesired target branch is finally added to the most performing combinations to induce staining invariance in the learned features. The following combinations of extra tasks are tested in the experiments: *density*, *area*, *contrast*, *correlation*, *center*, *center* + *density*, *center* + *area*, *center* + *density* + *area*. The gradient reversal operation

is only active for the *center* branch.

The experiments compare the vanilla and the uncertainty-based functions for weighting the optimization targets. Where not stated otherwise, the average AUC (avg. AUC) over ten repetitions with multiple initialization seeds is used for the evaluation. In the vanilla configuration, the loss weight values are set to 1 for all branches. The standard deviation is computed over ten repetitions of the network training with multiple seed initializations.

Baseline results The results in Table 4.3 are reported using unique IDs to identify the configurations tested in the experiments with numbers ranging from 1 to 8. Two columns are used to report the results on the internal (int.) and external (ext.) test sets. The results of the baseline model, i.e. of model-ID 1, are shown in the first row of the table. In this model, only the main task branch is trained and no extra tasks are used. The baseline model leads to internal (int. hereafter) avg AUC 0.819 ± 0.001, and external (ext. hereafter) avg. AUC $0,868 \pm 0.005$. Training the baseline on a GPU NVIDIA V100 takes approximately 19 hours.

Single-branch results The models with IDs from 2 to 5 represent a combination of the main task with a single extra branch. Model-ID 2, for example, is given by the combination of the main task branch with the additional task *area*, namely of predicting the area of the nuclei in the images. For these models, Table 4.3 reports the results of both the vanilla and the uncertainty-based weighting strategies of the multiple losses. A single auxiliary branch already outperforms the baseline. Model-ID 3, for example, encourages the learning of nuclei *count* and obtains int. avg AUC 0.836 ± 0.005, and ext. avg. AUC $0,890 \pm 0.009$. Model-ID 4 encourages the learning of image *contrast* and leads to int. avg. AUC 0.835 ± 0.008, ext. avg. AUC 0.876 ± 0.007. The models with a single additional branch in these experiments require between 6 and 17 hours of training before reaching convergence with the uncertainty estimation weighting strategy. Longer times than these are required by the vanilla configuration, which may take between 24 and 34 hours before reaching convergence.

Multi-branch results The combination of all the branches in model-ID 8 leads to the best performance on the int. test (int. avg. AUC 0.874 ± 0.009), with an increase of 0.05 AUC points compared to the baseline. On the external test set, the best generalization is achieved by adding *count* as a desired target, leading to ext. avg. AUC 0.890 ± 0.009. The models with the uncertainty-based weighting of the losses take between 8 and 15 hours to reach convergence on the GPU used for the experiments. The vanilla configuration may require up to 40 hours to converge.

Results with the adversarial branch The addition of the *center* adversarial branch in model-ID 6 leads to the best model overall with an overall avg. AUC (on both internal and external sets) at 0.824 ± 0.006 for the uncertainty trained model. This represents a significant improvement compared to the overall avg. AUC 0.79 ± 0.001 of the baseline model, with $p - value < 0.001$. The statistical significance of the results is evaluated by the non-parametric Wilcoxon test (two-sided) applied on the bootstrapping of the test set.

Sanity checks To confirm the benefit of the added related tasks, the results are compared with those obtained with random noise as additional targets. This experiment is

performed as a sanity check, where an auxiliary task is trained to predict random values. As expected, the overall, internal and external avg. AUCs are lower for this experiment and have larger standard deviations (overall avg. AUC 0.819 ± 0.04, int. test AUC 0.834 ± 0.001 and ext. avg. AUC 0.879 ± 0.03). This shows that the selected tasks are more relevant to the main task than the regression of random values.

Table 4.3: Average AUC on the main task and standard deviations from different starting points of the network parameter initialization. Results for the vanilla and uncertainty based weighting strategies. The adversarial task, i.e. *center*, is marked by an overline.

ID	main	area	count	contrast	$\overline{center}$	int. test		ext. test	
1	x					0.819±0.001		0.868±0.005	
						vanilla	unc.	vanilla	unc.
2	x	x				0.718±0.11	0.834±0.01	0.560±0.06	0.871±0.01
3	x		x			0.853±0.03	0.836±0.005	0.874±0.02	**0.890±0.009**
4	x			x		0.854±0.07	0.835±0.008	0.883±0.02	0.876±0.007
5	x				x	0.845±0.10	0.822±0.005	0.884±0.04	0.871±0.005
6	x		x		x	0.863±0.06	0.841±0.004	0.623±0.10	**0.890±0.01**
7	x	x	x		x	0.838±0.05	0.848±0.003	0.490±0.03	0.864±0.01
8	x	x	x	x	x	0.858±0.02	**0.874± 0.009**	0.686±0.20	0.825±0, 01

At this point, one may ask if the additional tasks were learned by the guided architectures. For model-ID3 (trained with the uncertainty-based weighting strategy), the prediction of the nuclei *count* values has an average determination coefficient $R^2 = 0.81 \pm 0.05$, showing that the concept was learned during training, passing from an initial MSE of the prediction of 0.46 to 0.17 at the end of training. Similar results apply to the other model-IDs 2 to 4 when only a single branch is added. Table 4.4 compares the performance on the extra-tasks to learning the concepts directly on the baseline model activations, where the network parameters are not optimized to learn the extra tasks. The classification of the *center* in model-ID 5 has low accuracy since the gradient reversal is used during training. The centers of the validation sets are predicted with accuracy 0.29 ± 0.01 at the end of the training (starting from an initial accuracy of 0.53 ± 0.01). When more extra tasks are optimized together the performance on the side tasks is affected, with Model-IDs 6, 7 and 8 not reporting high R^2 values. The average R^2 of nuclei *count* for model-ID 6, for example, decreases from -2.25 ± 0.05 and plateaus at around -0.63 ± 0.05.

Table 4.4: Performance on the extra-tasks for the baseline and guided models with the uncertainty-based strategy. The average and standard deviation of the determination coefficient are reported (the closer to 1 the better).

ID	area	count	contrast
baseline	0.66 ± 0.003	0.85 ± 0.007	0.56 ± 0.01
2	**0.70** ± 0.005	-	-
3	-	**0.88** ± 0.004	-
4	-	-	**0,64** ± 0.003

Visualization of the embeddings Figure 4.13 shows the dimensionality reduction of the internal representations learned by the baseline and model-ID 3. The visualization is obtained by applying the UMAP method by McInnes et al. (2018) (the hyper-parameters

Figure 4.13: Uniform Manifold Approximation and Projection (UMAP) representation of the internal activations of the baseline and guided model-ID3 (obtained with the UMAP default hyper-parameter set up). The top row shows the activations at the last convolutional layer of both models, known as mixed10 in the standard implementation of Inception V3 (Szegedy et al. (2016)). The bottom row shows the activations of the first fully-connected layer after the GAP operation.

for the visualization were kept to the default values of 15 neighbors, 0.1 minimum distance and local connectivity of 1). Note that the model-ID 3 selected for visualization was trained with the uncertainty-based weighting strategy. In the representation, the two classes are represented with different colors, whereas the size of the points in the plot is indicative of the values of nuclei counts in the images. The top row shows the projection of the internal representation of the last convolutional layer (known as mixed10 in the standard Inception V3 implementation) of the two models. The bottom row shows the projection of the first fully-connected layer after the GAP operation. Since the nuclei count values were normalized to zero mean and unit variance, these are represented in the plot as ranging between a minimum of -2 and a maximum of 2. For clarity of the representation, the image shows the UMAP of a random sampling of 4000 input images.

4.4 Strenghts and Limitations

From understanding pre-training to reducing the prediction error The scale quantification method in Section 4.2 increased our understanding of the features that are learned during pre-training on natural images. The incorporation of scale-related information is an interesting point for the medical application since the scale has an associated measure in the physical world (e.g. at a 40x magnification the field of view corresponds to 5 millimeters on the specimen), and it thus carries meaningful information. The invariance to scale is implicitly learned on the ImageNet inputs, but it is reached only towards the last layers before softmax as illustrated by the blue line in Figure 4.7. Note that the sanity checks that are shown by the green and the orange lines in the same figure ensure the validity of this result. Scale ratios are regressed better than random values (as shown by the green line) and the architecture with random weights does not contain information of scale (as shown by the orange line).

The understanding gained by this analysis turns into a valuable advantage to change the network architecture by pruning. The prediction error measured by the MAE significantly reduces from the baseline for the pruned models in Table 4.2. A better prediction of the magnification range and a higher kappa coefficient than the baseline are also observed in the same table, showing that the pruning of the layers learning scale-invariant features is beneficial to the task.

Enhancing generalization by introducing training guidance The strength of the multi-task adversarial architecture in Section 4.3 is that external guidance can be introduced to learn representations that generalize to new inputs. Already when a single extra task is added to the training, for example in model-ID3, the representations of the positive class organize in a more compact cluster than in the baseline model (as shown by the UMAP visualization in Figure 4.13). The representations on the right side of the figure (for model-ID3) also appear more structured than those on the left, being organized as following a direction for increasing values of the nuclei count (suggested as a gray line). The AUC of the baseline is already outperformed by adding a single auxiliary branch, with int. avg AUC 0.836 ± 0.005 against 0.819 ± 0.001 and ext. avg. AUC $0,890 \pm 0.009$ against $0,868 \pm 0.005$.

Introducing transparent changes The proposed methods introduce, in both cases, transparent changes to the baseline architectures. The pruning of the baseline layers, for instance, introduces a directly interpretable change to the architecture, since it is done by removing the layers that are shown to learn invariant features to scale. Similarly, the extra tasks in the multi-task adversarial architectures are modeled as inherently interpretable regression tasks. This approach, therefore, also favors model transparency, ensuring that specific features of the data are learned during network training. The features of *area* and *contrast* modeled in the previous chapter (in Section 3.4 as linear regression tasks to interpret the baseline model are here used to encourage the learning of discriminant factors and drive the classification.

Handling model complexity The two methods in this chapter do not require a marked increase in the model complexity. Improving the performance by increasing the model capacity is, in fact, not the objective of the developments in this thesis. The interpretable pruning reduces the model complexity, with the pruned models requiring the training of

51% and 19% less of parameters than the original Inception V3 and ResNet 50, respectively. The multi-task adversarial architecture only requires a neglectable increase of the number of parameters compared to Inception V3. Each extra task requires the training of only 2049 additional parameters, namely the 0.008% of Inception V3.

Self-tuning of hyper-parameters Another benefit of the proposed multi-task adversarial network is that the loss weights are balanced during training without any additional tuning nor a hyper-parameter search. The auxiliary and adversarial tasks introduced in the multi-task adversarial architecture in Section 4.3 are balanced by the uncertainty-weighting approach in the same end-to-end training. Task-dependent uncertainty is used to balance structurally different losses such as MSE and BCE (Kendall et al. 2018). With the uncertainty-based weighting strategy, the architecture did not require any specific tuning of the loss weights, whereas a fine-tuning of the weighting parameters appears highly necessary in the vanilla approach, particularly for the combinations with more than one extra task (model-IDs 6, 7, 8). The manual fine-tuning of the loss weights in the vanilla approach may lead to the over-specification of the model to the specific requirements of the test data considered in this study. These results not only extend the preliminary work by Gamper, Kooohbanani & Rajpoot (2020) to a different histology tissue and model architecture but also give more insights on how to handle multiple auxiliary losses and adversarial losses without requiring tedious tuning of hyper-parameters.

Versatility of the approaches The pruning and the multi-task adversarial approaches proposed in this chapter are built on top of the concept-based analysis obtained with RCVs.These methods inherit the versatility of the RCVs since concept measures can be chosen arbitrarily depending on the application. An additional task for the multi-task adversarial architecture can be the learning of mitosis detection since this is used to detection of tumor in breast cancer lymph nodes.

Need for annotations The methods proposed in this chapter require additional annotations of the clinical features and the scale measures. In both cases, this is only a minor limitation since a few annotations are already sufficient to train the models. The experiments in Section 4.2, for instance, were purposely performed on a reduced set of images with bounding box annotations (i.e. 25 images in total).

Limitations of numerous additional branches Our experiments show that the auxiliary tasks become harder to learn when they are scaled up in number, with model-ID 8 having a lower R^2 for the regression of the individual features than those reported for model-IDs 2 to 5 in Table 4.4. As explained also by Caruana (1997), the poor performance on the extra tasks is not necessarily an issue as long as these help with improving the model performance and generalization on unseen data. Further research is necessary, however, to verify if the AUC can be improved by the combination of all the extra tasks.

4.5 Impact and Open Questions

The work in this chapter shows how interpretability and expert knowledge can be used pro-actively during the training of CNNs to drive the representation learning process. The scale quantification with RCVs represents an intuitive and easy-to-apply method to

investigate the representation of scale at intermediate layers. Deep features (up to the penultimate layer) are linearly scale-covariant. This is an important observation since multiple models are used as feature extractors for medical imaging applications where scale has a physical meaning.

Clinically relevant and easy-to-interpret concept measures are introduced as extra tasks in the multi-task adversarial architecture. The resulting model is more robust and it improves its generalization compared to the baseline. This method may influence the development of human-computer interfaces to create datasets and annotations that may be used by pathologists to introduce their feedback during training.

Open questions concern the use of alternative approaches to the uncertainty-based weighting strategy used in this work, such as those analyzed in (Leang et al. 2020). The results on *center* in Table 4.3 do not show a marked improvement by the adversarial branch. This could be due to the lack of annotations about the acquisition centers in the PanNuke dataset. An unsupervised domain adaptation approach such as the domain alignment layers proposed by Carlucci et al. (2017) may be used to discover this latent information. Depending on the application, a different loss weighting approach may be used for the adversarial task and other undesired control targets can also be included, such as rotation, scale and image compression methods.

Extracting extra features solely from unlabeled data may be a direction for future work, where the idea of unsupervised concept discovery in Ghorbani et al. (2019) is exploited to build the multi-task adversarial architecture without requiring any annotation of the concepts. The combination of this architecture with weakly supervised learning could be interesting for future work since the labels of the additional tasks could be used during training as weak supervision.

4.6 Summary

This chapter aimed at addressing the second part of the main research question of this thesis, which is whether interpretability can be used to improve the performance and generalization of existing models. The first approach that I proposed is an intuitive application of RCVs. Given a concept of interest, i.e. image scale, I evaluated whether it is possible to modify the model in such a way that this concept is retained by the deep features. Since scale covariance already existed in the features extracted at some layers, the proposed change consists of pruning off those layers where this information is lost. The pruning method does not allow, however, to introduce information about some concepts of interest within the deep features. I thus conceived a multi-task adversarial architecture that aims at promoting the learning of features that are relevant to the main task. The extra tasks may be used as a weak-supervision to extend the training data with unlabeled datasets at a marginal cost of some extra automatic processing such as the extraction of nuclei contours or texture features.

Chapter 5

Discussion

5.1 Main Findings

The results of this thesis generated insights on interpretability techniques for deep learning methods within the context of medical imaging tasks.

User-centric development The experiments in this work (in particular those in Sections 3.2 and 3.5) underline the importance of integrating human users already at the early stages of the development of interpretability methods. As discussed in Section 1.3.3, the addressee of the explanations can have expectations about what features may be predictive for the model. The contributions in this work provide a methodology to verify whether these features are used by the model or not. Domain-expert knowledge should steer the interpretability analysis to address the doubts and expectations that experts may have about the models. The user tests confirm this point, showing that the explanations generated with the input from the experts are clearer and more understandable than those obtained by off-the-shelf methods. This demonstrates that user feedback is relevant to develop improved versions of the existing methods. Therefore, the experiments in Chapter 3 demonstrate some of the benefits of a user-centric approach to the development of interpretability. This vision is reflected by other works in the literature. Doshi-Velez & Kim (2017), for example, describes two types of tests with users that may be used to evaluate interpretability. Similarly, in Hoffman et al. (2018), it is discussed that the user's satisfaction can be used as a metric to evaluate if the interpretability method needs further development.

A further benefit given by the user-centric approach adopted in this work is the reliability of the new methods. The quantitative evaluation proposed in Section 3.2 highlights that the explanations generated for standard computer vision tasks have important pitfalls in terms of consistency and repeatability. Being quantitative, this evaluation guarantees that we did not only consider qualitative aspects of the explanations and reduced the risk of confirmation bias. It is known that, as humans, we tend to accept explanations even when these are empty of real informative content (Lombrozo 2006). The Sharp-LIME method is more consistent than standard LIME, assigning in Figure 3.12a high importance to the super-pixels that have a semantic meaning, i.e. the nuclei. High explanation values are assigned to neoplastic nuclei, in particular, in Figure 3.11a. This result is not obtained if the network parameters are randomly initialized, as illustrated in Figure 3.11b. This result is important because it suggests that Sharp-LIME explanations are not only

more consistent but also more robust to data bias. The same behavior cannot be observed for the standard LIME explanations in Figure 3.5.

Single-target post-hoc explanations The post-hoc explanations developed in Chapter 3 are focused on answering simple and targeted questions. Sharp-LIME, for example, has the main objective of clarifying whether the CNN that classifies tumorous tissues pays more attention to neoplastic nuclei than to the image background. The RCV method in Section 3.4 is used with the specific target of evaluating what descriptive features are used by the model. The narrow focus of these analyses is a strength, since it leads to explanations that are targeted to the physicians than individual pixel relevances (Ribeiro et al. 2016, Selvaraju et al. 2017, Chattopadhay et al. 2018, Zhou et al. 2016).

New insights on BCMLN with CNNs New insights about the model behavior for the BCMLN prediction are obtained by applying Sharp-LIME and RCVs to this task. The CNN attention is more focused on the nuclei instances than on the background in Figure 3.11a. As already discussed, higher attention is paid to neoplastic nuclei than to other types. The follow-up question derived from the analysis with Sharp-LIME concerns the type of features in these regions that are used to make the predictions. The results obtained with RCVs in Figure 3.17 address this question by evaluating the relevance of morphometric features such as nuclei size and appearance. Variations in the texture of the images are explained by RCVs as more relevant to the model than variations in the nuclei sizes. These results show the relevance of the nuclei appearance to make predictions. This is in line with the clinical criteria of grading tumor by observing the degree of nuclear pleomorphism described in Figure 2.2, according to which large nuclei with hyperchromatic appearance are assigned a high tumor grade. Nuclear shape, orientation and morphology, besides, are demonstrated to be predictors of the prognosis for breast tumor in Lu et al. (2018) and in Whitney et al. (2018). Note that the larger importance given by the CNN to texture features in Figure 3.17 is not surprising, since it is discussed in the literature that CNNs pre-trained on ImageNet are strongly biased towards recognizing textures (Hermann et al. 2020).

Beyond post-hoc explanations The methods in Chapter 4 aim at going beyond the generation of explanations to inspect model behavior and target the modification and correction of the training procedure. Guidance is introduced on the CNN training by two different approaches. In the first method in Section 4.2, some of the layers in the architecture are discarded to remove an undesired behavior of the original model, i.e. the introduction of scale invariance in the features. The scale invariance that is implicitly learned from pre-training on ImageNet is, in fact, detrimental to the transfer to medical tasks where scale has an associated physical meaning. As a result, the proposed pruning strategy significantly reduces the prediction error in Table 4.2.

Model training is guided in Section 4.3 by defining additional tasks that can improve the generalizability of the features learned by the model. Concept measures such as nuclei area, texture and density are introduced as additional outputs that the model should learn to predict. These tasks are modeled as the learning of a RCV. The additional tasks are joined to an adversarial task trained with gradient reversal to obtain invariance to domain-specific features. The results in Table 4.3 show that even the simplest combination with only a single additional branch can improve the model performance and generalization

over the baseline. I find these observations in line with the preliminary results obtained by Gamper, Kooohbanani & Rajpoot (2020).

5.2 Discussion of the experimental setting

Most of the results presented in this work were performed in a fixed setting with clearly defined tasks, architectures and data types. In the following, I discuss how the results may change under different conditions.

Choice of the architectures The experiments analyzed mainly two types of architectures, namely Inception V3 (Szegedy et al. 2016) and ResNet 101 (He et al. 2016*a*). It is worth discussing at this point, how well the methods developed in this thesis would apply to different architectures from the ones that were considered for the experiments. The Sharp-LIME method is a model-agnostic technique, that can be applied to any image classifier.The RCVs can be computed on the output of convolutional or fully connected layers, as long as the internal values of their activations can be accessed at inference time. I think that an interesting application of the interpretable pruning strategy may be found on graph-based convolutional models. In this type of network, finding the optimal pruning that preserves the most informative content is an interesting problem, and the combination of RCVs and the pruning strategy may lead to interesting solutions.

Scale-dependency The histopathology application poses particular challenges because of the multi-scale content of the images, which can be analyzed at increasing magnification levels to detect multiple structures. Evidence on the BCMLN task has shown that the analysis at the highest magnification level leads to the best predictive performance (Ehteshami Bejnordi et al. 2017), although this analysis strongly differs from the multi-scale approach at which pathologists operate. Recent work has looked at the benefits of combining information at multiple scales (Hashimoto et al. 2020) and this may be an interesting starting point for future developments of the works proposed in this thesis. Sharp-LIME, RCVs, and the multi-task adversarial architecture are valid for a fixed magnification at 40 X. The choice of the super-pixels and concepts should be changed for different magnifications. The analysis of the nuclei morphology performed at 40 X becomes the analysis of the tissue organization and tubular formation at a magnification of 10x. Similarly, considering the nuclei contours as Sharp-LIME super-pixels may not be the best approach for multi-scale analyses. The best strategy of scaling Sharp-LIME to multiple input magnifications is probably to develop an interactive segmentation tool that pathologists may directly use to evaluate the relevance of a given region.

Extension to new imaging modalities and tasks The methods in this thesis are presented for the task of BCMLN detection. They can adapt, however, relatively easy to other imaging modalities and tasks. The extension of RCVs for multi-class classification problems and different image types is discussed in Graziani, Andrearczyk, Marchand-Maillet & Müller (2020). The method is applied to the classification of handwritten digits and retinopathy inputs. This method has found further developments for other imaging modalities such as skin cancer Lucieri et al. (2020) and CT images Yeche et al. (2019). Sharp-LIME may be applied to chest X-ray images, although more work should be done to define super-pixels that can segment semantically meaningful regions in the

lungs (Palatnik de Sousa et al. 2021). As for the multi-task adversarial architecture, since the additional tasks can be arbitrarily chosen, this may be easily extended to new modalities and tasks.

Definition of the concept measures The definition and annotation of concepts is a requirement for concept-based attribution methods such as RCVs and CAVs. Identifying concepts that can be used for the analysis and translating these into measurable attributes is, in fact, a crucial step in the analysis with RCVs. Designing the concepts may require interaction with experts. Pre-existing hand-crafted features can be used as concept measures as in Graziani, Brown, Andrearczyk, Yildiz, Campbell, Erdogmus, Ioannidis, Chiang, Kalpathy-Cramer & Müller (2019).The work in Ghorbani et al. (2019) demonstrates that the need for concept annotations can be removed by implementing an unsupervised clustering approach that identifies shared features among the examples of a concept. A similar approach could be used to discover clinically relevant concepts. One may also think of using concepts that are not directly visible to the human eye. The work in Munk et al. (2021) shows that CNNs can predict biomarkers of age and gender from eye fundus images, although these are scarcely visible, if not at all, to the human eye. This may suggest a new line of research where relevant patient information is included in the analysis to verify, for example, whether the network can learn these types of concepts.

Chapter 6

Conclusions and Future Directions

In this thesis, I developed new post-hoc explainability techniques and new architectures that improve the understandability and generalization of DL models for medical image classification. The challenge that I addressed is that the internal mechanisms of DL models are difficult to interpret and show poor generalization performances when applied to inputs from new domains. Without interpretability, the opaqueness and limited generalization hinder the reliability of CNNs, affecting their applicability to everyday clinical practice. The contributions in this work are thus relevant to improving our understanding of what features are used by the models to make predictions.

Focusing on the digital pathology task of BCMLN detection, I showed by addressing two main objectives that new methods can be developed to provide more understandable explanations than the existing ones and that new architectures can be built to improve model generalization through interpretability. In Chapter 3, I developed and validated two new post-hoc explainability techniques that can explain the prediction of tumorous tissue in breast tissue slides by pointing to neoplastic nuclei instances and determining which factor, if the morphology, the size, or the appearance of the nuclei is relevant to the automated prediction. In Chapter 4, I proposed an interpretable pruning module that is beneficial to the transfer of pre-trained parameters from natural images to digital pathology inputs, and I designed a CNN architecture that incorporates human directives as additional tasks and that can focus on desired features and forget undesired ones. The proposed methods take inspiration from existing developments in computer vision and tailor the approaches to the requirements of the field of medical imaging. The developed models provide higher transparency and reliability at the cost of limited additional complexity.

The results that I obtained lead me to answer the research question of whether the understandability and generalization of DL models can be improved by new interpretability approaches affirmatively. The methods in this work provide ML developers and physicians with additional tools that could be used to understand whether an opaque model such as a CNN is good enough for clinical use and how its generalization could be improved to fit the variability of real-world data. Other tasks than BCMLN can be considered for these methods, as demonstrated by their application on eye imaging (Graziani, Brown, Andrearczyk, Yildiz, Campbell, Erdogmus, Ioannidis, Chiang, Kalpathy-Cramer & Müller 2019), skin lesion analysis (Lucieri et al. 2020), computer vision (Graziani, Andrearczyk, Marchand-Maillet & Müller 2020, Andrearczyk et al. 2020) and radiology (Yeche et al. 2019). These publications underline the impact of my developments within the research community and they highlight the importance of developing techniques that can be customized depending on the application.

Ultimately, this work fits in the broader research on the partnership between humans and machines, with the goal of understanding what type of DL-based support may best improve human decision-making. Future works should be performed in collaboration with clinicians to define data acquisition and annotation procedures that can establish a ground truth for training and evaluating explainability techniques. No indications exist, currently, about what kind of information is globally well-understood by pathologists as a valuable explanation for DL decisions. In some cases, the introduction in the study of the patients' feedback should also be considered, as patients are the receivers of the clinical decisions. The cognitive processes used to analyze the DL outcomes and the explanations are likely to be diverse for the patient, the technical, and the clinical staff. The collaboration between physicians, ML engineers, and social scientists should thus be fostered, for example, by designing interactive systems for testing, annotating and explaining diagnoses. Interpretability could be used, moreover, as a tool to discover biomarkers in the representations learned by deep models, mining new knowledge from the automated learning processes. ML experts and physicians could collaborate towards using interpretability to detect features that are yet unknown, but that show a causal association with the data generation process.

From a technical standpoint, further research on the interpretability methods here proposed should address the limitation that they have in common with most of the explanation methods in the current literature, namely that the vast majority of current explanation methods rely on the sole input-output correlations and cannot describe causal relationships with the data generation process. Approaches such as the causal concept effect in Goyal et al. (2019) should be further developed and tested on medical tasks. Generative models such as generative adversarial networks may be used to synthesize versions of the same pathology image with slightly modified features (e.g. increased chromatin or size) and evaluate the causal effect of such modifications on the network output.

Main terms

a-priori Latin expression with the literal meaning of "from the former". This term describes the inductive process of thinking from the causes to effects. Generally it is contrasted to a posteriori knowledge, which is based on experimental evidence.. 33

concept Something that is conceived in the human mind such as an idea, an image, a notion or a thought. 14

concept measure Feature representative of a concept that can be measured on a set of visual samples or annotated by experts. 50, 51, 53

diagnosis Judgement about the exact character of a disease or other problem, made after examination. 16, 23, 24, 58, 78

explainability To indicate with what features or high-level concepts are used by the ML model to generate predictions. 12

explanation weights Coefficients of the linear surrogate model explaining the weight of each super-pixel in LIME (Ribeiro et al. 2016). 37, 44–46, 60

feature A distinctive part that gives characterization to something by its prominence. 8

generalization Expected value of the model's error on new inputs. Definition from Goodfellow et al. (2016). 7, 8, 17–19, 25, 76, 78, 81, 83, 87, 88, 90, 93

intelligible Synonym of inherently understandable, that does not require further explanations. 15

interpretability To translate, expose abd comment about the generation proces of one or multiple outcomes by a ML system. 8, 12–14

model decomposability Each part of the model (e.g. input, parameters, calculations) admits an intuitive explanation. Defined by Lipton (2018). 13

performance How well a human or a ML system can perform a task. It can be measured in multiple ways, e.g. accuracy, precision, recall, specificity, ROC curve. 7, 8, 16, 25, 26, 30, 52, 53, 65, 66, 68, 72, 76, 81, 82, 84, 86–88, 90, 91

prognosis Doctor's judgement about the expected development of a disease, a statement of what the likely future situation is. 24, 90

specimen A small amount of blood, urine or tissue used for testing. 23

staining Technique used to enhance contrast in histology samples. 23–25, 31, 35, 59, 65, 74, 75, 81, 82

transparency Providing a non-opaque output generation process. See pg. 13

understandability The degree of comprehensibility of the provided information by a human with little to no experience in ML. In the clinical context, understandability requirements include explanation conciseness, usefulness, and consistency.. 14

understanding To know the meaning of something, to believe something is true because you have been told something that causes you to think so. 7, 13

Acronyms

AI Artificial Intelligence. 12, 14, 24

AM Activation Maximization. 27, 29, 31

AUC Area Under the Receiver Operating Characteristic (ROC) Curve. 36, 50, 81, 83

BCE Binary Cross-Entropy. 11, 50

BCMLN Breast Cancer Metastases in Lymph Nodes. 23, 25, 26, 34, 50, 90, 91, 93

CAM Class Activation Mapping. 29–31, 36

CAV Concept Activation Vectors. 18, 30, 49, 51, 60

CNN Convolutional Neural Network. 7, 11, 12, 18, 19, 24–27, 30, 31, 36, 48, 62, 63, 74

CT Computer Tomography. 60, 91

Deep-LIFT Deep Learning Important FeaTures. 29

DL Deep Learning. 7, 8, 10, 11, 14, 16, 17, 24, 27, 30, 31, 56–58, 62, 93

DNN Deep Feed-forward Neural Network. 10, 11

DSC Dice Similarity Coefficient. 44

GAN Generative Adversarial Network. 25

GAP Global Average Pooling. 4, 29, 36, 70, 72

GD Gradient Descent. 11

Grad-CAM Gradient-weighted Class Activation Mapping. 17, 29–31, 36

Grad-CAM++ Generalized Grad-CAM++. 29

LIME Locally Interpretable Model-agnostic Explanations. 17, 18, 28–31, 36

MAE Mean Average Error. 72, 73, 86

Mask-RCNN Mask Region-based Convolutional Neural Network. 42, 43, 81

MIA Medical Image Analysis. 16

ML Machine Learning. 5, 7, 8, 10, 12–16, 25, 55

MSE Mean Squared Error. 69, 80–82, 84

MTL Multi-task Learning. 18

PCA Principal Component Analysis. 27

RCNN Region-based Convolutional Neural Network. 43

RCV Regression Concept Vector. 48, 50, 51, 53, 55, 60, 88

RELU Rectified Linear Activation Unit. 12

ResNet Residual Neural Network. 25, 28, 30, 31, 43

ROI Region Of Interest. 26, 56, 57

SGD Stochastic Gradient Descent. 11, 36, 50

SHAP SHapley Additive exPlanations. 28, 29

Sharp-LIME Sharp Local Interpretable Model-agnostic Explanations. 2, 44, 55

SSIM Structural Similarity Index Measure. 37, 38

SVM Support Vector Machine. 31

UMAP Uniform Manifold Approximation and Projection. 27, 84

WSI Whole Slide Image. 17, 19, 21, 25, 31, 34, 35

Bibliography

Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P. & Süsstrunk, S. (2012), 'Slic superpixels compared to state-of-the-art superpixel methods', *IEEE transactions on pattern analysis and machine intelligence* **34**(11), 2274–2282.

Adadi, A. & Berrada, M. (2018), 'Peeking inside the black-box: a survey on explainable artificial intelligence (xai)', *IEEE access* **6**, 52138–52160.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M. & Kim, B. (2018), Sanity checks for saliency maps, *in* 'Proceedings of the 32nd International Conference on Neural Information Processing Systems', NIPS'18, Curran Associates Inc., Red Hook, NY, USA, p. 9525–9536.

Alain, G. & Bengio, Y. (2016), 'Understanding intermediate layers using linear classifier probes', *arXiv preprint arXiv:1610.01644* .

Andrearczyk, V., Fontaine, P., Oreiller, V. & Depeursinge, A. (n.d.), Multi-task deep segmentation and radiomics for automatic prognosis in head and neck cancer, *in* 'Workshop on PRedictive Intelligence in MEdicine (PRIME) at MICCAI'.

Andrearczyk, V., Graziani, M., Müller, H. & Depeursinge, A. (2020), Consistency of scale equivariance in internal representations of cnns, *in* 'Proceedings of the Irish Machine Vision and Image Processing Conference (IMVIP) 2020', 31 August-3 September 2020.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. et al. (2020), 'Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai', *Information Fusion* **58**, 82–115.

Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M. et al. (2021), 'Assessing the (un) trustworthiness of saliency maps for localizing abnormalities in medical imaging', *Radiology: Artificial Intelligence* .

Arvidsson, I., Overgaard, N. C., Marginean, F.-E., Krzyzanowska, A., Bjartell, A., Åström, K. & Heyden, A. (2018), Generalization of prostate cancer classification for multiple sites using deep learning, *in* '2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)', IEEE, pp. 191–194.

Arya, V., Bellamy, R. K., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A. et al. (2019), 'One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques', *preprint arXiv:1909.03012* .

Aubry, M. & Russell, B. C. (2015), Understanding deep features with computer-generated imagery, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 2875–2883.

Babic, B., Gerke, S., Evgeniou, T. & Cohen, I. G. (2021), 'Beware explanations from ai in health care', *Science* **373**(6552), 284–286.

Badrinarayanan, V., Kendall, A. & Cipolla, R. (2017), 'Segnet: A deep convolutional encoder-decoder architecture for image segmentation', *IEEE transactions on pattern analysis and machine intelligence* **39**(12), 2481–2495.

Bau, D., Zhou, B., Khosla, A., Oliva, A. & Torralba, A. (2017), Network dissection: Quantifying interpretability of deep visual representations, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 6541–6549.

Baxter, J. (1995), Learning internal representations, *in* 'Proceedings of the eighth annual conference on Computational learning theory', pp. 311–320.

Baxter, J. (2000), 'A model of inductive bias learning', *Journal of artificial intelligence research* **12**, 149–198.

Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. (2019), 'Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology', *Nature reviews Clinical oncology* **16**(11), 703–715.

Bhargava, H. K., Leo, P., Elliott, R., Janowczyk, A., Whitney, J., Gupta, S., Fu, P., Yamoah, K., Khani, F., Robinson, B. D., Rebbeck, T. R., Feldman, M., Lal, P. & Madabhushi, A. (2020), 'Computationally derived image signature of stromal morphology is prognostic of prostate cancer recurrence following prostatectomy in african american patients', *Clinical Cancer Research* **26**(8), 1915–1923.
URL: *https://clincancerres.aacrjournals.org/content/26/8/1915*

Biran, O. & Cotton, C. (2017), Explanation and justification in machine learning: A survey, *in* 'IJCAI-17 workshop on explainable AI (XAI)', Vol. 8, pp. 8–13.

Box, G. E. (1976), 'Science and statistics', *Journal of the American Statistical Association* **71**(356), 791–799.

Brown, J. M., Campbell, J. P., Beers, A., Chang, K., Ostmo, S., Chan, R. P., Dy, J., Erdogmus, D., Ioannidis, S., Kalpathy-Cramer, J. et al. (2018), 'Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks', *JAMA ophthalmology* **136**(7), 803–810.

Bruner, J. S., Goodnow, J. J. & Austin, G. A. (1967), 'A study of thinking. new york: Science editions', *Tinc., l962* .

Cai, C. J., Reif, E., Hegde, N., Hipp, J., Kim, B., Smilkov, D., Wattenberg, M., Viegas, F., Corrado, G. S., Stumpe, M. C. & Terry, M. (2019), *Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making*, Association for Computing Machinery, New York, NY, USA, p. 1–14.
URL: *https://doi.org/10.1145/3290605.3300234*

Camburu, O. (2020), Explaining deep neural networks, PhD thesis, University of Oxford.

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. (2019), 'Clinical-grade computational pathology using weakly supervised deep learning on whole slide images', *Nature medicine* **25**(8), 1301–1309.

Carlucci, F. M., Porzi, L., Caputo, B., Ricci, E. & Bulo, S. R. (2017), Just dial: Domain alignment layers for unsupervised domain adaptation, *in* 'International Conference on Image Analysis and Processing', Springer, pp. 357–369.

Caruana, R. (1997), 'Multitask learning', *Machine learning* **28**(1), 41–75.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M. & Elhadad, N. (2015), Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, *in* 'International Conference on Knowledge Discovery and Data Mining'.

Chakraborty, M., Biswas, S. K. & Purkayastha, B. (2020), 'Rule extraction from neural network trained using deep belief network and back propagation', *Knowledge and Information Systems* **62**(9), 3753–3781.

Chattopadhay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. (2018), Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, *in* '2018 IEEE winter conference on applications of computer vision (WACV)', IEEE, pp. 839–847.

Chen, Z., Bei, Y. & Rudin, C. (2020), 'Concept whitening for interpretable image recognition', *Nature Machine Intelligence* **2**(12), 772–782.

Cheng, H.-T., Yeh, C.-F., Kuo, P.-C., Wei, A., Liu, K.-C., Ko, M.-C., Chao, K.-H., Peng, Y.-C. & Liu, T.-L. (2020), Self-similarity student for partial label histopathology image segmentation, *in* 'European Conference on Computer Vision', Springer, pp. 117–132.

Chromik, M. & Schuessler, M. (2020), A taxonomy for human subject evaluation of black-box explanations in xai., *in* 'ExSS-ATEC@ IUI', p. 1.

Clinciu, M.-A. & Hastie, H. (2019), A survey of explainable ai terminology, *in* 'Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)', pp. 8–13.

Coeckelbergh, M. (2020), *AI Ethics*, MIT Press.

Cohen, T. & Welling, M. (2016), Group equivariant convolutional networks, *in* 'International conference on machine learning', PMLR, pp. 2990–2999.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009), ImageNet: A large-scale hierarchical image database, *in* '2009 IEEE conference on computer vision and pattern recognition', IEEE, pp. 248–255.

Diao, J. A., Wang, J. K., Chui, W. F., Mountain, V., Gullapally, S. C., Srinivasan, R., Mitchell, R. N., Glass, B., Hoffman, S., Rao, S. K. et al. (2021), 'Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes', *Nature communications* **12**(1), 1–15.

Doshi-Velez, F. & Kim, B. (2017), 'A roadmap for a rigorous science of interpretability', *CoRR* **abs/1702.08608**.

Duong, L., Cohn, T., Bird, S. & Cook, P. (2015), Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser, *in* 'Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (volume 2: short papers)', pp. 845–850.

Ehteshami Bejnordi, B. (2017), Histopathological diagnosis of breast cancer using machine learning, PhD thesis, [Sl: sn].

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J. A. W. M., & the CAMELYON16 Consortium (2017), 'Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer', *JAMA* **318**(22), 2199–2210.
URL: *https://doi.org/10.1001/jama.2017.14585*

Elston, C. W. & Ellis, I. O. (1991), 'Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up', *Histopathology* **19**(5), 403–410.

Erhan, D., Bengio, Y., Courville, A. & Vincent, P. (2009), 'Visualizing higher-layer features of a deep network', *University of Montreal* **1341**(3), 1.

Ertosun, M. G. & Rubin, D. L. (2015), Automated grading of gliomas using deep learning in digital pathology images: a modular approach with ensemble of convolutional neural networks, *in* 'AMIA Annual Symposium Proceedings', Vol. 2015, American Medical Informatics Association, p. 1899.

Everingham, M., Gool, L., Williams, C. K., Winn, J. & Zisserman, A. (2010), 'The Pascal Visual Object Classes (VOC) Challenge', *Int. J. Comput. Vision* **88**(2), 303–338.
URL: *https://doi.org/10.1007/s11263-009-0275-4*

Fang, Z., Kuang, K., Lin, Y., Wu, F. & Yao, Y.-F. (2020), Concept-based explanation for fine-grained images and its application in infectious keratitis classification, *in* 'Proceedings of the 28th ACM international conference on Multimedia', pp. 700–708.

Felzenszwalb, P. F. & Huttenlocher, D. P. (2004), 'Efficient graph-based image segmentation', *International journal of computer vision* **59**(2), 167–181.

Fernandes, F. E. & Yen, G. G. (2021), 'Pruning of generative adversarial neural networks for medical imaging diagnostics with evolution strategy', *Information Sciences* **558**, 91–102.
URL: *https://www.sciencedirect.com/science/article/pii/S0020025521000189*

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F. et al. (2018), 'Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations', *Minds and Machines* **28**(4), 689–707.

Fong, R. C. & Vedaldi, A. (2017), Interpretable explanations of black boxes by meaningful perturbation, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 3429–3437.

Fong, R., Patrick, M. & Vedaldi, A. (2019), Understanding deep networks via extremal perturbations and smooth masks, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision', pp. 2950–2958.

Fraggetta, F., Garozzo, S., Zannoni, G., Pantanowitz, L. & Rossi, E. (2017), 'Routine digital pathology workflow: The Catania experience', *Journal of Pathology Informatics* **8**(1), 51.

Frosst, N. & Hinton, G. (2017), Distilling a neural network into a soft decision tree, *in* 'Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017, co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017)'.

Gamper, J., Koohbanani, N. A., Graham, S., Jahanifar, M., Khurram, S. A., Azam, A., Hewitt, K. & Rajpoot, N. (2020), 'Pannuke dataset extension, insights and baselines', *CoRR abs/2003.10778 (2020)* .

Gamper, J., Kooohbanani, N. A. & Rajpoot, N. (2020), Multi-task learning in histopathology for widely generalizable model, *in* 'Workshop on AI for overcoming global disparities in cancer care at the 8th International Conference on Learning Representations (ICLR. 2020)'.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M. & Lempitsky, V. (2016), 'Domain-adversarial training of neural networks', *The Journal of Machine Learning Research* **17**(1), 2096–2030.

Ghorbani, A., Wexler, J., Zou, J. Y. & Kim, B. (2019), 'Towards automatic concept-based explanations', *Advances in Neural Information Processing Systems* **32**, 9277–9286.

Ghosh, R. & Gupta, A. K. (2019), Scale steerable filters for locally scale-invariant convolutional neural networks, *in* 'Workshop on Theoretical Physics for Deep Learning at the International Conference on Machine Learning'.

Giusti, A., Caccia, C., Cireşari, D. C., Schmidhuber, J. & Gambardella, L. M. (2014), A comparison of algorithms and humans for mitosis detection, *in* '11th International Symposium on Biomedical Imaging (ISBI)', IEEE, pp. 1360–1363.

Gong, B., Grauman, K. & Sha, F. (2013), Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation, *in* 'International Conference on Machine Learning', PMLR, pp. 222–230.

Gong, T., Lee, T., Stephenson, C., Renduchintala, V., Padhy, S., Ndirango, A., Keskin, G. & Elibol, O. H. (2019), 'A comparison of loss weighting strategies for multi task learning in deep neural networks', *IEEE Access* **7**, 141627–141632.

Goodfellow, I., Bengio, Y. & Courville, A. (2016), *Deep Learning*, MIT Press. http://www.deeplearningbook.org.

Goyal, Y., Feder, A., Shalit, U. & Kim, B. (2019), 'Explaining classifiers with causal concept effect (cace)', *arXiv preprint arXiv:1907.07165* .

Graham, S., Vu, Q. D., Raza, S. E. A., Azam, A., Tsang, Y. W., Kwak, J. T. & Rajpoot, N. (2019), 'Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images', *Medical Image Analysis* **58**, 101563.

Graziani, M., Andrearczyk, V., Marchand-Maillet, S. & Müller, H. (2020), 'Concept attribution: Explaining CNN decisions to physicians', *Computers in Biology and Medicine* p. 103865.
URL: *http://www.sciencedirect.com/science/article/pii/S0010482520302225*

Graziani, M., Andrearczyk, V. & Müller, H. (2018), Regression concept vectors for bidirectional explanations in histopathology, *in* 'Understanding and Interpreting Machine Learning in Medical Image Computing Applications', Springer, pp. 124–132.

Graziani, M., Andrearczyk, V. & Müller, H. (2019), Visualizing and interpreting feature reuse of pretrained cnns for histopathology, *in* 'Irish Machine Vision and Image Processing Conference'.

Graziani, M., Brown, J. M., Andrearczyk, V., Yildiz, V., Campbell, J. P., Erdogmus, D., Ioannidis, S., Chiang, M. F., Kalpathy-Cramer, J. & Müller, H. (2019), Improved interpretability for computer-aided severity assessment of retinopathy of prematurity, *in* 'Medical Imaging 2019: Computer-Aided Diagnosis'.

Graziani, M., Dutkiewiczk, L., Calvaresi, D., Pereira Amorima, J., Yordanova, K., Vered, M., Nair, R., Abreu, P. H., Blanke, T., Pulignano, V., O. Prior, J., Lauwaert, L., Reijers, W., Depeursinge, A., Andrearczyk, V. & Müller, H. (2021), 'A global taxonomy of interpretable ai: Unifying the terminology for the technical and social sciences', *submitted to Artificial Intelligence Reviews* .

Graziani, M., Eggel, I., Andrearczyk, V. et al. (2020), 'Breast histopathology with high-performance computing and deep learning', *Computing and Informatics* **39**(4), 780–807.

Graziani, M., Lompech, T., Müller, H., Depeursinge, A. & Andrearczyk, V. (2020), Interpretable cnn pruning for preserving scale-covariant features in medical imaging, *in* 'Interpretable and Annotation-Efficient Learning for Medical Image Computing', Springer, pp. 23–32.

Graziani, M., Lompech, T., Müller, H., Depeursinge, A. & Andrearczyk, V. (2021), 'On the scale invariance in state of the art cnns trained on imagenet', *Machine Learning and Knowledge Extraction* **3**(2), 374–391.

Graziani, M., Lompech, T., Müller, H. & Andrearczyk, V. (2021), Evaluation and comparison of cnn visual explanations for histopathology, *in* 'Explainable Agency in Artificial Intelligence at AAAI21', pp. 195–201.

Graziani, M., Marini, N., Otálora, S., Ciompi, F., Aztori, M., Fragetta, F. & Müller, H. (2021), 'How should ai for decision support be integrated in fully digital pathology workflows?', *preprint* .

Graziani, M., Muller, H. & Andrearczyk, V. (2019), Interpreting intentionally flawed models with linear probes, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops', pp. 0–0.

Graziani, M., Otálora, S., Marchand-Maillet, S., Müller, H. & Andrearczyk, V. (2021), 'Learning Interpretable Pathology Features by Multi-task and Adversarial Training Improves CNN Generalization', *submitted to Nature Machine Intelligence (August 2021)* .

Graziani, M., Palatnik de Sousa, I., B. R. Vellasco, M. M., Costa da Silva, E., Müller, H. & Andrearczyk, V. (2021), Sharpening local interpretable model-agnostic explanations for histopathology: Improved understandability and reliability, *in* 'Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention 2021', pp. 0–0.

Griffin, J. & Treanor, D. (2017), 'Digital pathology in clinical use: where are we now and what is holding us back?', *Histopathology* **70**(1), 134–145.

Gulshan, V., Peng, L., Coram, M., Stumpe, M. C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J. et al. (2016), 'Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs', *JAMA* **316**(22), 2402–2410.

Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M. & Yener, B. (2009), 'Histopathological image analysis: A review', *IEEE reviews in biomedical engineering* **2**, 147–171.

Hägele, M., Seegerer, P., Lapuschkin, S., Bockmayr, M., Samek, W., Klauschen, F., Müller, K.-R. & Binder, A. (2020), 'Resolving challenges in deep learning-based analyses of histopathological images using explanation methods', *Scientific reports* **10**(1), 1–12.

Haralick, R. M. (1979), 'Statistical and structural approaches to texture', *Proceedings of the IEEE* **67**(5), 786–804.

Haralick, R. M., Dinstein, I. & Shanmugam, K. (1973), 'Textural features for image classification', *IEEE Transactions On Systems Man And Cybernetics* **3**(6), 610–621.

Hashimoto, N., Fukushima, D., Koga, R., Takagi, Y., Ko, K., Kohno, K., Nakaguro, M., Nakamura, S., Hontani, H. & Takeuchi, I. (2020), Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images, *in* 'Proceedings of the IEEE/CVF conference on computer vision and pattern recognition', pp. 3852–3861.

He, K., Gkioxari, G., Dollar, P. & Girshick, R. (2017), Mask r-cnn, *in* 'Proceedings of the IEEE International Conference on Computer Vision (ICCV)'.

He, K., Zhang, X., Ren, S. & Sun, J. (2016*a*), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 770–778.

He, K., Zhang, X., Ren, S. & Sun, J. (2016*b*), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 770–778.

Hermann, K., Chen, T. & Kornblith, S. (2020), 'The origins and prevalence of texture bias in convolutional neural networks', *Advances in Neural Information Processing Systems* **33**.

Hilton, D. J. (1990), 'Conversational processes and causal explanation.', *Psychological Bulletin* **107**(1), 65.

Hoffman, R. R., Mueller, S. T., Klein, G. & Litman, J. (2018), 'Metrics for explainable ai: Challenges and prospects', *CoRR abs/1812.04608* .

Huang, Y. & Chung, A. C. (2019), Evidence localization for pathology images using weakly supervised learning, *in* 'International conference on medical image computing and computer-assisted intervention', Springer, pp. 613–621.

Huh, M., Agrawal, P. & Efros, A. A. (2016), What makes ImageNet good for transfer learning?, *in* 'Workshop on Large Scale Computer Vision Systems at NeurIPS 2016'.

Ilse, M., Tomczak, J. M. & Welling, M. (2020), Deep multiple instance learning for digital histopathology, *in* 'Handbook of Medical Image Computing and Computer Assisted Intervention', Elsevier, pp. 521–546.

Ilse, M., Tomczak, J. & Welling, M. (2018), Attention-based deep multiple instance learning, *in* 'International conference on machine learning', PMLR, pp. 2127–2136.

Janowczyk, A. & Madabhushi, A. (2016), 'Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases', *Journal of pathology informatics* **7**.

Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M. & Madabhushi, A. (2019), 'Histoqc: an open-source quality control tool for digital pathology slides', *JCO clinical cancer informatics* **3**, 1–7.

Jung, H., Lodhi, B. & Kang, J. (2019), 'An automatic nuclei segmentation method based on deep convolutional neural networks for histopathology images', *BMC Biomedical Engineering* **1**, 1.

Kanazawa, A., Sharma, A. & Jacobs, D. W. (2014), Locally scale-invariant convolutional neural networks, *in* 'Advances in Neural Information Processing Systems'.

Kandel, I. & Castelli, M. (2020), 'How deeply to fine-tune a convolutional neural network: a case study using a histopathology dataset', *Applied Sciences* **10**(10), 3359.

Katharopoulos, A. & Fleuret, F. (2019), Processing megapixel images with deep attention-sampling models, *in* 'International Conference on Machine Learning', PMLR, pp. 3282–3291.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. (2019), 'Key challenges for delivering clinical impact with artificial intelligence', *BMC medicine* **17**(1), 1–9.

Kendall, A., Gal, Y. & Cipolla, R. (2018), Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 7482–7491.

Khan, A., Sirinukunwattana, K. & Rajpoot, N. (2015), 'A global covariance descriptor for nuclear atypia scoring in breast histopathology images', *IEEE Journal of Biomedical and Health Informatics* **19**, 1637–1647.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F. & Sayres, R. (2018), Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV), *in* J. Dy & A. Krause, eds, 'Proceedings of the 35th International Conference on Machine Learning', Vol. 80 of *Proceedings of Machine Learning Research*, PMLR, pp. 2668–2677.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D. & Kim, B. (2019), 'The (un) reliability of saliency methods', *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer International Publishing* .

Kokkinos, I. (2017), Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 6129–6138.

Korbar, B., Olofson, A. M., Miraflor, A. P., Nicka, C. M., Suriawinata, M. A., Torresani, L., Suriawinata, A. A. & Hassanpour, S. (2017), Looking under the hood: Deep neural network visualization to interpret whole-slide image analysis outcomes for colorectal polyps, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops', pp. 69–75.

Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A. & Sethi, A. (2017), 'A dataset and a technique for generalized nuclear segmentation for computational pathology', *IEEE Transactions on Medical Imaging* **36**(7), 1550–1560.

Lafarge, M. W., Pluim, J. P. W., Eppenhof, K. A. J., Moeskops, P. & Veta, M. (2017), Domain-adversarial neural networks to address the appearance variability of histopathology images, *in* M. J. Cardoso, T. Arbel, G. Carneiro, T. Syeda-Mahmood, J. M. R. Tavares, M. Moradi, A. Bradley, H. Greenspan, J. P. Papa, A. Madabhushi, J. C. Nascimento, J. S. Cardoso, V. Belagiannis & Z. Lu, eds, 'Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support', Springer International Publishing, Cham, pp. 83–91.

Lakkaraju, H., Bach, S. H. & Leskovec, J. (2016), Interpretable decision sets: A joint framework for description and prediction, *in* 'Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining', pp. 1675–1684.

Lapuschkin, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R. & Samek, W. (2015), 'On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation', *PLoS ONE* **10**.

Leang, I., Sistu, G., Bürger, F., Bursuc, A. & Yogamani, S. (2020), Dynamic task weighting methods for multi-task networks in autonomous driving systems, *in* '2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)', IEEE, pp. 1–8.

LeCun, Y., Haffner, P., Bottou, L. & Bengio, Y. (1999), Object recognition with gradient-based learning, *in* 'Shape, contour and grouping in computer vision', Springer, pp. 319–345.

Lei, T., Barzilay, R. & Jaakkola, T. (2016), Rationalizing neural predictions, *in* 'Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing', pp. 107–117.

Lenc, K. & Vedaldi, A. (2015), Understanding image representations by measuring their equivariance and equivalence, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 991–999.

Li, Y., Lin, S., Zhang, B., Liu, J., Doermann, D., Wu, Y., Huang, F. & Ji, R. (2019), Exploiting kernel sparsity and entropy for interpretable cnn compression, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 2800–2809.

Lipton, Z. C. (2018), 'The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery.', *Queue* **16**(3), 31–57.

Litjens, G., Bandi, P., Ehteshami Bejnordi, B., Geessink, O., Balkenhol, M., Bult, P., Halilovic, A., Hermsen, M., van de Loo, R., Vogels, R. et al. (2018), '1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset', *GigaScience* **7**(6), giy065.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B. & Sánchez, C. I. (2017), 'A survey on deep learning in medical image analysis', *Medical image analysis* **42**, 60–88.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. & Berg, A. C. (2016), Ssd: Single shot multibox detector, *in* 'European conference on computer vision', Springer, pp. 21–37.

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G. E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P. Q., Corrado, G. S., Hipp, J. D., Peng, L. & Stumpe, M. C. (2017), 'Detecting cancer metastases on gigapixel pathology images.', *CoRR* **abs/1703.02442**.

Lombrozo, T. (2006), 'The structure and function of explanations', *Trends in cognitive sciences* **10**(10), 464–470.

Long, J., Shelhamer, E. & Darrell, T. (2015), Fully convolutional networks for semantic segmentation, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 3431–3440.

Long, M., Cao, Y., Wang, J. & Jordan, M. (2015), Learning transferable features with deep adaptation networks, *in* 'International conference on machine learning', PMLR, pp. 97–105.

Lu, C., Romo-Bucheli, D., Wang, X., Janowczyk, A., Ganesan, S., Gilmore, H., Rimm, D. & Madabhushi, A. (2018), 'Nuclear shape and orientation features from h&e images predict survival in early-stage estrogen receptor-positive breast cancers', *Laboratory investigation* **98**(11), 1438–1448.

Lucieri, A., Bajwa, M. N., Alexander Braun, S., Malik, M. I., Dengel, A. & Ahmed, S. (2020), On interpretability of deep learning based skin lesion classifiers using concept activation vectors, *in* '2020 International Joint Conference on Neural Networks (IJCNN)', pp. 1–10.

Lundberg, S. M. & Lee, S.-I. (2017), A unified approach to interpreting model predictions, *in* 'Proceedings of the 31st International Conference on Neural Information Processing Systems', NIPS'17, Curran Associates Inc., Red Hook, NY, USA, p. 4768–4777.

Mahendran, A. & Vedaldi, A. (2015), Understanding deep image representations by inverting them, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 5188–5196.

Marcos, D., Kellenberger, B., Lobry, S. & Tuia, D. (2018), Scale equivariance in CNNs with vector fields, *in* 'FAIM workshop at the International Conference on Machine Learning'.

McInnes, L., Healy, J., Saul, N. & Großberger, L. (2018), 'Umap: Uniform manifold approximation and projection', *Journal of Open Source Software* **3**(29), 861.

Miller, T. (2019), 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial intelligence* **267**, 1–38.

Miller, T., Howe, P. & Sonenberg, L. (2017), 'Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences', *preprint arXiv:1712.00547* .

Mittelstadt, B., Russell, C. & Wachter, S. (2019), Explaining explanations in ai, *in* 'Proceedings of the conference on fairness, accountability, and transparency', pp. 279–288.

Molchanov, P., Tyree, S., Karras, T., Aila, T. & Kautz, J. (2017), Pruning convolutional neural networks for resource efficient inference, *in* 'Proceedings of the International Conference on Learning Representations, 2017'.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W. & Müller, K.-R. (2017), 'Explaining nonlinear classification decisions with deep taylor decomposition', *Pattern Recognition* **65**, 211–222.

Montavon, G., Samek, W. & Müller, K.-R. (2018), 'Methods for interpreting and understanding deep neural networks', *Digital Signal Processing* **73**, 1–15.

Mordvintsev, A., Olah, C. & Tyka, M. (2015), 'Inceptionism: Going deeper into neural networks'.

Mormont, R., Geurts, P. & Marée, R. (2018), Comparison of deep transfer learning strategies for digital pathology, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition workshops', pp. 2262–2271.

Müller, H., Andrearczyk, V., del Toro, O. J., Dhrangadhariya, A., Schaer, R. & Atzori, M. (2020), Studying public medical images from the open access literature and social networks for model training and knowledge extraction, *in* 'International Conference on Multimedia Modeling', Springer, pp. 553–564.

Munk, M. R., Kurmann, T., Márquez-Neila, P., Zinkernagel, M. S., Wolf, S. & Sznitman, R. (2021), 'Assessment of patient specific information in the wild on fundus photography and optical coherence tomography', *Scientific reports* **11**(1), 1–10.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. (2019), 'Definitions, methods, and applications in interpretable machine learning', *Proceedings of the National Academy of Sciences* **116**(44), 22071–22080.
URL: *https://www.pnas.org/content/116/44/22071*

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., Topol, E. J., Ioannidis, J. P., Collins, G. S. & Maruthappu, M. (2020), 'Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies', *The British Medical Journal (BMJ)* **368**.

Nesterov, Y. (1983), 'A method for solving the convex programming problem with convergence rate o($1/k^2$)', *Proceedings of the USSR Academy of Sciences* **269**, 543–547.

Nguyen, A.-p. & Martínez, M. R. (2019), 'Mononet: towards interpretable models by learning monotonic features', *arXiv preprint arXiv:1909.13611* .

Olah, C., Mordvintsev, A. & Schubert, L. (2017), 'Feature visualization', *Distill* . https://distill.pub/2017/feature-visualization.

Otálora, S., Atzori, M., Andrearczyk, V., Khan, A. & Müller, H. (2019), 'Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology', *Frontiers in Bioengineering and Biotechnology* **7**, 198.

Otálora, S., Atzori, M., Andrearczyk, V. & Müller, H. (2018), Image magnification regression using densenet for exploiting histopathology open access content, *in* 'MICCAI 2018 - Computational Pathology Workshop (COMPAY)'.

Otálora, S., Atzori, M., Khan, A., Jimenez-del Toro, O., Andrearczyk, V. & Müller, H. (2020), 'Systematic comparison of deep learning strategies for weakly supervised gleason grading', *Medical Imaging 2020: Digital Pathology* .

Otsu, N. (1979), 'A threshold selection method from gray-level histograms', *IEEE transactions on systems, man, and cybernetics* **9**(1), 62–66.

Palacio, S., Lucieri, A., Munir, M., Hees, J., Ahmed, S. & Dengel, A. (2021), 'Xai handbook: Towards a unified framework for explainable ai', *preprint arXiv:2105.06677* .

Palatnik de Sousa, I., Bernardes Rebuzzi Vellasco, M. M. & Costa da Silva, E. (2020), 'Evolved Explainable Classifications for Lymph Node Metastases', *preprint: arXiv* .

Palatnik de Sousa, I., Vellasco Bernardes Rebuzzi, M. M. & Costa da Silva, E. (2019), 'Local interpretable model-agnostic explanations for classification of lymph node metastases', *Sensors (Basel, Switzerland)* **19**.

Palatnik de Sousa, I., Vellasco, M. M. B. R. & Costa da Silva, E. (2021), 'Explainable artificial intelligence for bias detection in covid ct-scan classifiers', *Sensors* **21**(16).
URL: *https://www.mdpi.com/1424-8220/21/16/5657*

Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T. & Rohrbach, M. (2018), Multimodal explanations: Justifying decisions and pointing to the evidence, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 8779–8788.

Paschali, M., Naeem, M. F., Simson, W., Steiger, K., Mollenhauer, M. & Navab, N. (2019), 'Deep learning under the microscope: improving the interpretability of medical imaging neural networks', *preprint arXiv:1904.03127* .

Pearson, K. (1901), 'LIII. On lines and planes of closest fit to systems of points in space', *The London, Edinburgh, and Dublin philosophical magazine and journal of science* **2**(11), 559–572.

Pirovano, A., Heuberger, H., Berlemont, S., Ladjal, S. & Bloch, I. (2020), Improving interpretability for computer-aided diagnosis tools on whole slide imaging with multiple instance learning and gradient-based explanations, *in* 'Interpretable and Annotation-Efficient Learning for Medical Image Computing', Springer, pp. 43–53.

Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. (2019), Transfusion: Understanding transfer learning for medical imaging, *in* 'Advances in Neural Information Processing Systems', Vol. 32, Curran Associates, Inc.
URL: *https://proceedings.neurips.cc/paper/2019/file/eb1e78328c46506b46a4ac4a1e378b91-Paper.pdf*

Raza, S. E. A., Cheung, L., Shaban, M., Graham, S., Epstein, D., Pelengaris, S., Khan, M. & Rajpoot, N. M. (2019), 'Micro-net: A unified model for segmentation of various objects in microscopy images', *Medical image analysis* **52**, 160–173.

Reinhard, E., Adhikhmin, M., Gooch, B. & Shirley, P. (2001), 'Color transfer between images', *IEEE Computer graphics and applications* **21**(5), 34–41.

Reyes, M., Meier, R., Pereira, S., Silva, C., Dahlweid, M. P. M., Tengg-Kobligk, H., Summers, R. & Wiest, R. (2020), 'On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities', *Radiology: Artificial Intelligence* **2**, e190043.

Ribeiro, M. T., Singh, S. & Guestrin, C. (2016), "Why Should I Trust You?": Explaining the Predictions of Any Classifier, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', KDD '16, Association for Computing Machinery, New York, NY, USA, p. 1135–1144.
URL: *https://doi.org/10.1145/2939672.2939778*

Ribeiro, M. T., Singh, S. & Guestrin, C. (2018), Anchors: High-precision model-agnostic explanations, *in* 'Proceedings of the AAAI conference on artificial intelligence', Vol. 32.

Robbins, S. (2019), 'A misdirected principle with a catch: explicability for ai', *Minds and Machines* **29**(4), 495–514.

Ruder, S. (2017), 'An overview of multi-task learning in deep neural networks', *arXiv preprint arXiv:1706.05098* .

Rudin, C. (2019), 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence* **1**(5), 206–215.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. (2017), Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, *in* '2017 IEEE International Conference on Computer Vision (ICCV)', Vol. 128, pp. 618–626.

Shorten, C. & Khoshgoftaar, T. M. (2019), 'A survey on image data augmentation for deep learning', *Journal of Big Data* **6**(1), 1–48.

Shrikumar, A., Greenside, P. & Kundaje, A. (2017), Learning important features through propagating activation differences, *in* 'International Conference on Machine Learning', PMLR, pp. 3145–3153.

Siegel, R. L., Miller, K. D. & Jemal, A. (2019), 'Cancer statistics, 2019', *CA: A Cancer Journal for Clinicians* **69**(1), 7–34.
URL: *https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21551*

Simonyan, K., Vedaldi, A. & Zisserman, A. (2014), Deep inside convolutional networks: Visualising image classification models and saliency maps, *in* Y. Bengio & Y. LeCun, eds, '2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings'.

Springenberg, J., Dosovitskiy, A., Brox, T. & Riedmiller, M. (2015), Striving for simplicity: The all convolutional net, *in* 'ICLR (workshop track)'.

Stathonikos, N., Veta, M., Huisman, A. & van Diest, P. (2013), 'Going fully digital: Perspective of a Dutch academic pathology lab', *Journal of Pathology Informatics* **4**(1), 15.

Subramanian, S., Trischler, A., Bengio, Y. & Pal, C. J. (2018), Learning general purpose distributed sentence representations via large scale multi-task learning, *in* 'Proceedings of the International Conference on Learning Representations (ICLR 2018)'.

Sundararajan, M., Taly, A. & Yan, Q. (2017), Axiomatic attribution for deep networks, *in* 'International Conference on Machine Learning', PMLR, pp. 3319–3328.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. (2016), Rethinking the inception architecture for computer vision, *in* 'IEEE Conference on Computer Vision and Pattern Recognition'.

Tellez, D., Balkenhol, M., Karssemeijer, N., Litjens, G., van der Laak, J. & Ciompi, F. (2018), H and e stain augmentation improves generalization of convolutional networks for histopathological mitosis detection, *in* 'Medical Imaging 2018: Digital Pathology', Vol. 10581, International Society for Optics and Photonics, p. 105810Z.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F. & van der Laak, J. (2019), 'Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology', *Medical image analysis* **58**, 101544.

Tonekaboni, S., Joshi, S., McCradden, M. D. & Goldenberg, A. (2019), What clinicians want: contextualizing explainable machine learning for clinical end use, *in* 'Machine Learning for Healthcare Conference', PMLR, pp. 359–380.

Touvron, H., Vedaldi, A., Douze, M. & Jégou, H. (2019), Fixing the train-test resolution discrepancy, *in* 'Advances in Neural Information Processing Systems'.

van der Laak, J., Litjens, G. & Ciompi, F. (2021), 'Deep learning in histopathology: the path to the clinic', *Nature medicine* **27**(5), 775–784.

Van Diest, P. J., Van Deurzen, C. H. & Cserni, G. (2010), 'Pathology issues related to sn procedures and increased detection of micrometastases and isolated tumor cells', *Breast disease* **31**(2), 65–81.

Vedaldi, A. & Soatto, S. (2008), Quick shift and kernel methods for mode seeking, *in* 'European conference on computer vision', Springer, pp. 705–718.

Veeling, B. S., Linmans, J., Winkens, J., Cohen, T. & Welling, M. (2018), Rotation equivariant cnns for digital pathology, *in* 'International Conference on Medical image computing and computer-assisted intervention', Springer, pp. 210–218.

Vu, Q. D., Graham, S., Kurc, T., To, M. N. N., Shaban, M., Qaiser, T., Koohbanani, N. A., Khurram, S. A., Kalpathy-Cramer, J., Zhao, T. et al. (2019), 'Methods for segmentation and classification of digital microscopy tissue images', *Frontiers in bioengineering and biotechnology* **7**, 53.

Wang, H., Cruz-Roa, A., Basavanhally, A., Gilmore, H., Shih, N., Feldman, M., Tomaszewski, J., Gonzalez, F. & Madabhushi, A. (2014), 'Mitosis detection in breast cancer pathology images by combining handcrafted and convolutional neural network features', *Journal of Medical Imaging* **1**(3).

Wang, P., Wang, J., Li, Y., Li, L. & Zhang, H. (2020), 'Adaptive pruning of transfer learned deep convolutional neural network for classification of cervical pap smear images', *IEEE Access* **8**, 50674–50683.

Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. (2004), 'Image quality assessment: from error visibility to structural similarity', *IEEE transactions on image processing* **13**(4), 600–612.

Ward, J. (2019), *The student's guide to cognitive neuroscience*, Routledge.

Weller, A. (2019), Transparency: motivations and challenges, *in* 'Explainable AI: Interpreting, Explaining and Visualizing Deep Learning', Springer, pp. 23–40.

Whitney, J., Corredor, G., Janowczyk, A., Ganesan, S., Doyle, S., Tomaszewski, J., Feldman, M., Gilmore, H. & Madabhushi, A. (2018), 'Quantitative nuclear histomorphometry predicts oncotype dx risk categories for early stage er+ breast cancer', *BMC cancer* **18**(1), 1–15.

Worrall, D. E. & Welling, M. (2019), Deep scale-spaces: Equivariance over scale, *in* 'preprint arXiv:1905.11697'.

Xie, Q., Dai, Z., Du, Y., Hovy, E. & Neubig, G. (2017), Controllable invariance through adversarial feature learning, *in* 'Advances in Neural Information Processing Systems'.

Xu, L., Ren, J. S., Liu, C. & Jia, J. (2014), 'Deep convolutional neural network for image deconvolution', *Advances in neural information processing systems* **27**, 1790–1798.

Xu, Y., Jia, Z., Wang, L.-B., Ai, Y., Zhang, F., Lai, M., Eric, I. & Chang, C. (2017), 'Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features', *BMC bioinformatics* **18**(1), 1–17.

Xu, Z., Moro, C. F., Bozóky, B. & Zhang, Q. (2019), 'Gan-based virtual re-staining: a promising solution for whole slide image analysis', *preprint arXiv:1901.04059* .

Yan, E. & Huang, Y. (2021), Do cnns encode data augmentations?, *in* '2021 International Joint Conference on Neural Networks (IJCNN)', IEEE, pp. 1–8.

Yang, F., Du, M. & Hu, X. (2019), 'Evaluating explanation without ground truth in interpretable machine learning', *preprint arXiv:1907.06831* .

Yeche, H., Harrison, J. & Berthier, T. (2019), UBS: A Dimension-Agnostic Metric for Concept Vector Interpretability Applied to Radiomics, *in* 'Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support'.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T. & Lipson, H. (2015), Understanding neural networks through deep visualization, *in* 'Deep Learning workshop at the International Conference on Machine Learning, 2015'.

Zeiler, M. D. & Fergus, R. (2014), Visualizing and understanding convolutional networks, *in* 'European conference on computer vision', Springer, pp. 818–833.

Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J. et al. (2019), 'Pathologist-level interpretable whole-slide cancer diagnosis with deep learning', *Nature Machine Intelligence* **1**(5), 236–245.

Zhou, B., Bau, D., Oliva, A. & Torralba, A. (2018), 'Interpreting deep visual representations via network dissection', *IEEE transactions on pattern analysis and machine intelligence* **41**(9), 2131–2145.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. (2016), Learning deep features for discriminative localization, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 2921–2929.