DEEP LEARNING INTERPRETABILITY: MEASURING THE RELEVANCE OF CLINICAL CONCEPTS IN CNN FEATURES

1. Introduction

A large variety of tasks is being solved by algorithmic systems implementing classical Machine Learning (ML) and Deep Learning (DL) techniques [1]. DL, in particular, emerged as a better-performing substitute for hand-crafted feature extraction [2], which is more traditional for ML in health applications. Not provided with any insights about the decision-making, end-users seem to report wobbly confidence in the DL decision process [3]. Some of the inherent risks that even a perfectly well performing DL model may hide are the codification of biases and the weak accountability of decision-making. The flawed system for pneumonia risk detection analyzed by Caruana et al. in [4] is an example. Despite its high performance, the model learned to assign a lower risk of death to cases of pneumonia with concurring asthma because of misleading correlations in the data. A correct diagnosis would have taken the opposite decision given the high risk of death with this pre-existing condition. The misleading correlation (i.e. presence of asthma thus low risk of death from pneumonia) was rather a consequence of the effective care given to these patients by healthcare specialists that were promptly reacting to reduce the risk of death, consequently lowering the recorded risk for these patients. The misleading feature "presence of asthma" was captured thanks to model interpretability.

The perception of DL as a black-box that gives little insights about the final output is a limiting factor for the acceptance and consequent use of DL models by physicians [5]. The near-perfect accuracy of DL models may only be apparent for a few very specific tasks, dropping significantly in real-world practice [6]. This shows, as argued in [7], that the evaluation of DL models only on

the basis of task performance is fundamentally incomplete. The need for interpretability in the development of AI for health emerges as impellent for two main reasons. On the one hand, the interaction between physicians and AI is improved by interpretability methods. On the other hand, interpretability can be used as an alternative to the test performance to validate the model decision-making process. Interpretability can be used, for example, for pointing humans those subtle visual features in the image that make the diagnosis at the borderline between two choices, causing low inter-rater disagreement. In retinopathy [8] DL and physicians can interact to decipher borderline cases such as the detection of the plus disease in the retina of preborn babies. High performance in this task can make a considerable difference in saving babies from blindness. The performance of the combination of humans and DL were shown to be the highest in terms of the inter-rater agreement also for other application domains, e.g. cancer diagnosis [9]. Interpretability in the sense of explaining the rationale for AI decisions to its final users is therefore an important prerequisite for the application of AI to health-care, which will be further discussed in this chapter.

Most of the interpretable AI methods can be categorized according to a few factors, namely **global** vs. **local** interpretability, **built-in** vs. **post-hoc** methods and **feature** vs. **concept** attribution, described later in this chapter (in Sec. 2.2). The popular activation maps in [10], for example, generate explanations of the decision for a single input (i.e. local), without requiring to retrain the model parameters (i.e. post-hoc), highlighting the most salient input pixels (i.e. feature attribution). Concept-based methods such as the Concept Activation Vectors (CAVs) proposed by Kim et al. [11] generate explanations in terms of arbitrary high-level concepts. This method is also post-hoc and shows that clinically relevant features, that we refer to as **clinical concepts**, can be directly used to explain complex DL models. These explanations help the users of interpretable AI to think

more systematically about the relevance of specific features within the AI model [12]. By including experiments on diabetic retinopathy, Kim et. al. showed the applicability of CAVs to health-care, explaining DL decisions in terms of the presence or absence of a clinical feature. One limitation of CAVs is that they express a clinical feature only in terms of either its presence or absence, whereas continuous or categorical measures are more frequently used to describe clinical factors, e.g. the size of a lesion. Regression Concept Vectors (RCVs) were proposed to extend CAVs to continuous and categorical clinical features [13]. Research on the applicability of RCVs shows that these explanations can fit the requirements of various medical tasks ranging from histopathology [13-15], to radiomics [16] and retinopathy [14, 17]. By directly matching the semantics of the end-users, RCVs explain DL decisions in relation to well-known prognostic factors and clinical guidelines. This approach to "subject-centric" explanations (SCEs), as referred to in [18], shows promise for interactive explanations, learning about the model behavior from the outside.

The main focus of this chapter is the application of concept-based interpretability to measure the relevance of clinical concepts in DL decisions. Within the chapter, we clarify the terminology around AI interpretability, presenting an in-depth analysis of the existing tools for health applications. The concept-attribution approach of RCVs is discussed as a way of obtaining SCEs that relate to clinical concepts, fostering the interaction between physicians and DL models. The detection of plus disease in Retinopathy of Prematurity (ROP) cases is presented as the main application domain. The application to the ROP is relevant because the detection of plus disease is at the edge between two fundamentally different treatment planning strategies and causes large disagreement rates in the diagnoses. In Section 2, we review the literature concerning AI interpretability. Section 3 and 4 present the methods and the experimental results, respectively. In

Section 3.1, in particular, we introduce ROP and the task of plus disease detection. Section 5 presents insights and in-depth discussions on the analyses. The conclusions in Section 6 summarize the key points in this chapter and present a higher-level discussion on XAI research for computer-assisted diagnosis systems.

2. Related Work on Interpretable AI

2.1 Motivations

The research field in AI interpretability has grown very quickly in the last four years (see Figure 1 on the left). In the medical imaging domain, the number of publications per year concerning interpretable AI development also presents a marked increase (Figure 1 on the right). The rising interest in interpreting DL models can be traced back to the evidence that the classic metrics of model performance (e.g. classification accuracy, loss) are not sufficient to describe the model's principles of inner functioning.

[ADD Figure 1 HERE] Figure 1. Trends of the research fields in interpretable AI (on the left) and interpretable AI for medical imaging (on the right).

As Doshi Velez and Kim argue in [7], the need for interpretability stands out in problems that suffer from incompleteness in their formalization. Particularly in medical imaging, model performance in terms of the specificity and sensitivity of the predictions is evaluated for a pre-collected testing set for which experts have agreed regarding a ground-truth diagnosis. The generalization to unseen data may not hold, causing significant drops in performance in real-world applications [6]. Depending on the application, the sole measurement of task performance may lead to incomplete model evaluations on various fronts. Some models may require fairness, for example, not encoding biases that would induce gender or racial discrimination in their decisions.

Other models may require the robustness to adversarial attacks, e.g. biometrics and person identification. Model accountability (in the sense of taking responsibility for the decisions) may be another desideratum, for example in credit allowance or automated driving applications [19]. The motivation for interpretable AI development, therefore, directly stems from the application requirements. Interpretability in health-care applications aims at avoiding erroneous diagnosis since automatic predictions can be analyzed and interpreted before a final decision by an expert. In autonomous driving, interpretability mostly aims at demonstrating the causes for an accident (for insurance liability reasons, among others) once the mistake has already happened [19]. The EU's General Data Protection Regulation (GDPR), in effect since May 2018, officialized the need for safety, fairness and explainability of AI deployment in the real world. The so-called "right to obtain an explanation" provides individuals with the right to inquire about the transparency, accountability and explainability of how their data were handled by the automated decisions. For example, if an automatic system was to deny a loan application, the denied person has the right to ask for an explanation regarding the decision in the form of "meaningful information about the logic of processing".

2.2 Related Terminology

Differences in the specification of the interpretability objectives (e.g. for debugging, for explaining wrong decisions or as a way of proving model safety, fairness and accountability) inevitably lead to inconsistencies in the terminology related to interpretable AI. The words "interpretable", "explainable", "intelligible", "understandable", "transparent" and "comprehensible" have often been used interchangeably in the literature, causing confusion and different taxonomies [20-25]. A formal definition of **interpretability** was given by Doshi-Velez and Kim in [7] as that of

"explaining or presenting in understandable terms to a human" the decisions of a ML system. The concept of "interpreting" is therefore inherently linked to that of "explaining" in this definition that we adopt. In the analysis of interpretability from the perspective of social sciences [20], Miller agrees with assigning the same meaning to explainable and interpretable, in the sense of "providing explanations" to humans. Interpretability as intended in [7], however, seems to correspond to what is meant as intelligible in the taxonomy presented in [21], namely the large set of possible actions and developments to obtain a system that is "clear enough to be understood" by humans. According to the scheme in [21], being interpretable or explainable also means being intelligible, but the opposite is not necessarily true. **Intelligible** AI, in practice, does not imply the generation of explanations. This could be obtained, for example, by adding interpretability constraints on the model objective function being optimized [26], or by visualizing the network internal features [26]. The set of **explainable** models is thus a rather smaller subset than that of intelligible models, although interpretable and explainable are still being used to refer to the same purpose, which is "explaining" decisions to humans [7, 20, 21]. Transparency and comprehensibility also appear as terms related to interpretable AI [23]. The former generally refers to a set of descriptions that are most relevant to AI developers. **Transparency** is defined in [23] as the description of the structure, equations, parameter values and assumptions necessary to understand the inner model mechanisms. Lipton further divides this definition into model simulatability, decomposability of the parameters and algorithmic transparency [28]. Finally, the notion of model **comprehensibility** is described in [23] as the ability of the learning algorithm to generate meta-descriptions about its inner working mechanism that can be interpreted in natural language. The taxonomy relative to interpretable AI is nevertheless subject to continuous change and updates. Among these definitions proposed in the literature, two should be retained for this chapter, namely that of explainable AI

(XAI) as a means of generating explanations for AI decisions and that of intelligible (or interpretable) AI as a wider set of tools also including methodologies that do not necessarily aim at generating explanations. In the context of XAI, Miller further clarifies the elusivity about the concept of providing explanations [20]. The sole association between model output and possible causes is not sufficient to provide a "good explanation". These should be rather contextualized with the user's needs, promoting the interaction and answering contrastive questions, i.e. "why did the model output was class P instead of another class Q?". From these approaches to interpretability, the definitions of human-centric or Subject-Centric Explanations (SCEs) arose as a way of identifying XAI methods that are tailored to the user's needs and the requirements of the application domain. SCEs aim at fostering the interaction between end-users and ML or DL models and are expressed in the ontology of the application domain.

2.3 Related work on XAI

2.3.1 XAI for Medical Applications

Proving the safety and reliability of the model decision-making is an emerging challenge in the deployment of AI to the medical domain. Several techniques for interpretable AI development find a relevant application in the medical domain, providing interesting insights into various tasks [29-36]. A large part of the interpretable AI methods for medical applications generate visual explanations to provide justifications for the model predictions. Among these, saliency maps are the most frequently used to generate explanations in pathology [32-35], retinopathy [30], and radiology [29, 36-38].Numerous interpretable AI approaches can be categorized by the schema in Figure 2 of which the elements are described in detail.

[INSERT Figure 2 HERE] Figure 2 Categorization of interpretable AI approaches

This section reviews the main approaches and introduces the relevant technical terminology to categorize XAI methods. By focusing on explainability techniques, this section does not include inherently intelligible models (e.g. linear regression), models with built-in interpretability (e.g. a decision tree, also part of intelligible AI) and dataset exploration methods (e.g. dimensionality reduction techniques or the retrieval of influential instances [39]). We also exclude geometrical approaches such as Singular Vector Canonical Correlation Analysis (SVCCA) [40].

Some of the technical terms used to distinguish most of the current approaches to obtaining interpretable AI were introduced by Lipton in [28]. In particular, Lipton distinguishes **local** vs. **global** explanations and **built-in** vs. **post-hoc** methods. Local explanations refer to explanations that are only true for a single input. Global explanations, on the contrary, explain the model behavior for an entire set of inputs, e.g. all images of a single class in the dataset. Built-in methods, as explained by Lipton, introduce interpretability as one of the objectives of the model optimization function. These methods are included in the more general notion of intelligible AI. An example is that of inherently interpretable models, e.g. linear regression, where the linear increase of a feature value corresponds to a proportional increase in the model output. Post-hoc methods are on the other hand methods that generate explanations without requiring the retraining of the model parameters with interpretability constraints. Finally, attribution methods generate explanations by identifying either the most relevant features, in feature attribution, or the most relevant concepts, in concept attribution, to the network decisions. Referring back to Miller's formalism of explainability [20], feature attribution methods answer to the question "What would the model output be if the *value of this input feature* was different?". Concept attribution provides explanations in terms of high-level concepts that can match the semantics of the end-users. In the

medical scenario, these can be directly clinically relevant concepts. Therefore, concept attribution answers the questions of the type “Why did the model output class P, and would it answer class Q if this *clinical attribute* was different?”. The clinical attribute may be a visual feature, for example, describing the size of a lesion. The shift between feature and concept attribution is mainly at the interpretation level. In some cases, for example in imaging applications, the values of individual features (e.g. the raw input pixels) appear rather incomprehensible to humans [11]. The aim of concept attribution, as further explained in Sec. 2.3 and 3.2, is to generate explanations that can directly relate to the ontology of the receivers of the explanation.

The next sections present a review of several XAI methods. Despite being rather long, this review is not exhaustive of all the methods existing in the literature and does not include several interpretability approaches that do not generate explanations (such as intelligible models, transparent and comprehensible models). The review is organized as follows. In Sec. 2.3.2, we review the most common visualization and feature attribution methods. In Sec. 2.3.3 we introduce the related work to concept attribution. Finally, in Sec. 2.4 we review the evaluation of XAI methods in the literature.

2.3.2 Visualization Methods and Feature Attribution

Visualization methods were proposed, at first, for interpreting the remarkable increase in performance given by the application of deep Convolutional Neural Networks (CNNs) to computer vision tasks. Visualizations were proposed to either visualize the learned features or to highlight the most salient input features. These methods evolved into the generation of explanations by feature attribution. The network output for a single input (local method) is explained by a subset of the input features. These input features identify, in the case of images, the most important pixels

that are then displayed as a heatmap. In this section, we further clarify the literature related to visualization approaches and we report the formalization of feature attribution as proposed in [44].

A cornerstone of early interpretability development is the deconvolution paper by Zeiler and Fergus. Their approach is twofold. On the one hand, they visualize the filters learned by various CNNs (an idea already formulated in [41] by the Activation Maximization approach) through the inversion of the convolution operations, that is where the name "deconvolution" comes from. On the other hand, they generate saliency heatmaps by systematically occluding portions of the input image with a grey square [42]. Their occlusion sensitivity method monitors the output of the classifier to these input perturbations. Simonyan's saliency maps further develop this idea by computing the relevance of individual pixels rather than entire input regions [43]. Each value at a location of Simonyan's saliency heatmap represents the derivative of the decision function with respect to the input pixel in that same location. Guided backpropagation upgrades this formulation, using the signals from the high layers as additional guidance to avoid the flow of negative gradients [45]. The Layerwise Relevance Propagation (LRP) in [46] further expands the idea of saliency in [43]. LRP decomposes the contribution of each pixel (also obtained by a derivative operation) at each layer in the CNN computation. The propagation of this relevance through the layers is then evaluated to obtain the LRP values that are visualized as a heatmap. An entirely different approach that also generates visual explanations, often called activation maps, is that of Class Activation Mapping (CAM) [47]. Individual CNN feature maps are used to obtain a heatmap of the network's attention before being spatially averaged and linearly combined to produce the network prediction. One limitation of CAM is that it can only be applied to CNNs with a global average pooling layer, rarely used in recent state-of-the-art architectures. Grad-CAM is proposed in [48] as a generalization of CAM that directly takes into account the cascade of gradients at each CNN layer.

In this way, the activation maps can be obtained from a wider variety of CNN architectures, including those used for image captioning and query answering. When applied to classification tasks, CAM and Grad-CAM are equivalent up to a normalization constant that is proportional to the number of pixels in the feature maps [48]. The Grad-CAM framework does not generalize to multiple occurrences of same-class instances in the input image considering the gradients with respect to entire feature maps. This limitation is partially addressed by Grad-CAM++, which considers the gradients directly at the pixel level [49]. The prediction difference analysis method in [50], based on the framework in [51], proposes a probabilistic approach to generating heatmaps. The basic idea is that of estimating the relevance of a feature by measuring the change in the prediction when that specific feature is unknown. This change is obtained by evaluating the difference between the probability of the prediction when conditioning on the complete feature set and when conditioning on the feature set where that specific feature is removed [50]. The framework Local-Interpretable Model Agnostic Explanations (LIME) proposed by Ribeiro in [52], is a local post-hoc XAI method that uses linear surrogate models to generate explanations for a single input image. To clarify eventual confusions, the adjective “Local” in LIME refers to the approximation of the DL decision function in the locality of an input sample. For instance, the linear surrogate model approximates the decisions of the DL model in a neighborhood of the input sample. Using LIME to explain CNNs is similar to using a sparse linear model to approximate the complex decision function of the CNN. The first step of the application of LIME to images consists of clustering pixels into superpixels (that are used as features) using color, texture and other types of local similarities. Randomly hiding some of the superpixels generates perturbations (called samples) of the original images which can be used to compute the relevance of each superpixel to the decision-making. Some common algorithms to extract superpixels are Simple Linear Iterative

Clustering [53] and Felzenszwalb's graph-based image segmentation [54]. DeepLIFT is proposed in [55] to generate attribution scores on the basis of the difference between the neuron activations and a “reference activation” that is computed, for example, by using a blurred version of the original input image. Finally, SHapley Additive exPlanations (SHAP) [56] are proposed as a framework that unifies some of the method formulations for explaining predictions, including LRP, LIME and DeepLIFT.

All of the methods presented in this section generate post-hoc local explanations that attribute the CNN decisions to a set of input features. This approach is summarized by the formal definition by Sundararajan et al. of the framework of attribution to features [44], reported in the following.

Given an input image $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and a CNN with a decision function f mapping the input image to a class probability, the attribution vector is defined as:

$$A_f(\mathbf{x}, \mathbf{x}') = (a_1, \ldots, a_n) \in \mathbb{R}^n$$

where each a_i explains the contribution of each pixel x_i to $f(\mathbf{x}), i = 1, \ldots, n$.

2.3.3 Concept Attribution

Concept Attribution aims at addressing a key difficulty in the generation of pixel-based explanations for CNNs, namely that humans understand high-level concepts more easily than the raw input pixel values or the internal CNN activation values [11]. This section reviews the related work on interpreting CNNs by using human-friendly concepts and presents the framework of attribution to concepts, as opposed to that of attribution to features.

The reference paper for generating explanations in terms of high-level concepts is Testing with Concept Activation Vectors [11]. The interpretation of a ML or DL model, in the sense of

generating post-hoc explanations, is seen as a translation problem. The state of the model is defined by Kim et al. as a vector space E_m (e.g. the space of the CNN activations) [11]. The basis vectors in this space correspond to the input features and neural activations. Another vector space E_h is used to describe the space of high-level concepts and interactions understandable to humans, with bases vectors corresponding to the high-level concepts. Generating explanations means finding a function $g : E_m \rightarrow E_h$. The method in [11] proposes a way of obtaining the translation g. Given a concept of interest, they collect a set of example images representative of the concept. The Concept Activation Vector (CAV) is then learned in the space of the activations of a CNN layer as a linear classification task that separates the set of examples with the concept from a set of random images (not containing the concept). The CAV for that concept, for instance, is the unit vector representing the linear classifier. In other words, the CAV models the presence or absence of a human-friendly concept and it is computed as the unit weight vector representing the linear classifier that separates images with the concept from those without the concept in the space of activations of a CNN layer. The performance of the linear classifier is indicative of how well the concept is learned in the network representation. The use of linear classifiers, that are inherently intelligible, is also addressed as linear probing in [57]. The work on CAVs presents numerous extensions in the literature. The automatic extraction of visual concepts is proposed in [58] to obtain insights on the concepts learned automatically by CNNs despite not having explicit knowledge of all of them. Causal Concept Effect (CaCE) aims at establishing the causal effect of the presence of a concept in the input image [59]. The latest concept bottleneck models propose the training of DL models on images with annotations for both the ground-truth labels and the presence of concepts [60].

The Regression Concept Vectors (RCVs) in [13, 14] extend CAVs to model not only the presence or absence of a concept, but also continuous-valued measures. These measures do not need necessarily additional annotations, as they can be directly computed on the images. The development of RCVs is particularly relevant to the medical domain since clinical concepts are expressed as observed measurements that do not fit in the binary formulation of CAVs, e.g. radiomic features [16], nuclei pleomorphism [13,14], vessel features of the retina [14,17].

Finally, concept attribution is defined in [14] for a set of Q concepts $\{c_i\}_{i=1}^{Q}$ as in the following: A vector $\mathbf{v}_{c_i}$, being either the CAV or the RCV, represents a concept c_i in the activation space of a CNN layer l. The concept attribution vector $A_f(\Phi^l(\mathbf{x}), \{\mathbf{v}_{c_i}\}_{i=1}^{Q}) = (a_1, ..., a_Q)$ represents with each a_i the relevance of the concept c_i to the CNN decision function $f(\mathbf{x})$ for an input $\mathbf{X}$.

Explanations obtained with concept attribution are defined locally around the input image $\mathbf{X}$. Being independent of the pixel locations the attribution values a_i can be agglomerated in multiple ways to obtain global explanations (valid for an entire class or an entire set of inputs). Some methods such as the TCAV and the Br scores are proposed within the works on CAVs and RCVs themselves. For this reason, the box of attribution to concepts in Figure 2 leads to both local and global explanations.

2.4 Evaluation of XAI methods

In the previous sections, we presented XAI methods that generate explanations in terms of visualizations or high-level concepts. Both approaches provide immediate feedback on the network internal state and can give insights on the criteria for decision-making. We discuss in this

section the need for quantitative evaluation methods for XAI, motivated by the risk of confirmation bias if only a qualitative assessment is performed to evaluate XAI plausibility [61]. We then present a review of the studies proposing quantitative evaluations, arguing that some of the evaluation approaches do not generalize to all medical tasks. These considerations are relevant for both evaluating the existing XAI methods and developing new ones that can better suit the clinical needs.

XAI methods should highlight the relevant information behind the model's decision-making, while showing properties of robustness, implementation invariance, consistency, appropriateness and reliability [44]. Without these properties XAI methods would lose the user's trust as a meaningful way of assessing DL decisions. Testing the reliability and trustworthiness of visual explanations only by visual inspection, however, is subject to the risk of confirmation bias. Confirmation bias is defined in cognitive psychology as the human tendency to attribute greater confidence to a hypothesis, even if false, when explanations are generated for it [lomb]. From a technical perspective, the quantitative evaluation of XAI methods is complicated because of the lack of ground-truth. We do not know, in fact, what input features are important to a model. Remarkable work in the literature focuses on developing evaluation methods, particularly on evaluating the consistency of saliency maps. By the term consistency, we refer to a series of desired invariances and dependencies that XAI outcomes should present. This includes, for example, implementation and input invariance and the dependency on the model parameters. Implementation and input invariance are addressed by the works in [44, 63]. The same explanations, according to Sundararajan [44], should be generated for functionally equivalent networks, namely models with different architectures but reporting the same outputs for the same inputs. XAI outcomes should be invariant to constant shifts in the input data, although some of them show sensitivity are easily

fooled by simple changes in the background color [63]. The dependency on the model parameters is evaluated in [64] by a series of randomization tests. The similarity of the explanations is compared when the learned model parameters are re-initialized to random values layer by layer in a cascading way, and completely, namely by resetting all the parameters to random values. The outcomes of the randomization tests show that XAI methods are inconsistent and perhaps assign the wrong attribution values to the wrong features [bim]. These results further stress the need for a solid evaluation of XAI outcomes that goes beyond simple visual inspection.

Concerning XAI development for medical applications, the deployment to the clinical setting further expands the desiderata for these methods. In the first place, explanations should be targeted at helping physicians with decision-making, without requiring extra expertise in the theoretical aspects of AI systems [63]. This goes in favor of the human-centric or subject-centric (SCE) approaches that consider the user's need in the development of interpretable AI. In addition to this, Tokenaboni expands the list of desirable properties for XAI methods for clinical application. The explanations should be evaluated according to a series of factors. The first of these factors is the appropriateness of the explanation to the clinical domain. Clinically irrelevant, inconsistent and unnecessary explanations do not support physicians and should be given a lower priority. Explanations that cannot be translated into action (may this mean asking for additional analyses, for the confirmation or the modification of pre-existent choices) should also be avoided, as they do not help with the clinical workflow. Finally, Tokenaboni adds the invariance to shifts in the XAI implementation parameters as a further evaluation of the consistency. Within the medical context, a few works assess the trustworthiness and reliability of XAI visualizations [64, 35]. By using lesion contours annotations Arun et al. assess four points of saliency methods, namely their utility for localization tasks, their sensitivity to the randomization of the parameter weights, their

repeatability and reproducibility [64]. The instability of XAI visualization methods applied to emerges from this study on chest X-rays. It is important to point out that evaluating XAI methods on the basis of their localization performance as in [64], however, may not generalize to all clinical tasks. The work in [35], for example, discusses how this approach would easily fail in the context of histopathology images. This is mainly due to the fact that in these images there is not 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. It could be sufficient for the CNN to focus on one or a few instances, thus causing low evaluations of the localization capability of the CNN.

The existing studies on the evaluation of XAI methods show, finally, that the rigorous evaluation of these studies still necessitates sustained research that keeps into account the application domain [63].

3. Methods

3.1 Retinopathy of Prematurity

3.1.1 Relevant Background

We present in this section the applicative domain of the works presented in this chapter, namely the classification of plus disease in ROP. ROP affects premature babies born before 31 weeks of gestation and weighing less than 1.3 kilos. This disease of the eye causes the abnormal growth of the blood vessels in the retina to more than 14,000 premature infants per year only in the U.S.. If

prompt action is not taken, the aggressivity of ROP may remain stable or advance further. Medical treatment is required by around 10 % of the babies to avoid the degeneration of ROP. To analyze ROP an indirect ophthalmoscope is used to visually inspect the retina. With the digitalization of medical images, special cameras are used to take high-resolution pictures of the retina. These pictures are analyzed by multiple experts and can be used to track disease evolution over time, as shown in Figure 3. The ROP diagnosis consists of identifying the affected zones of the retina, staging the disease on a scale from 0 to 5 (in Figure 4). The risk if ROP advances is that of blindness due to the total detachment of the retina (as in the picture on the right in Figure 4).

[ADD Figure 3 HERE] Figure 3. ROP progression from grade 0 to 2 in the right eye of the same patient. No presence of plus is noticeable.

[ADD Figure 4 HERE] Figure 4. Example of ROP grades from 0 to 5 in multiple patients. plus is not noticeable in any of the images, except from the presence of pre-plus in the image for grade 4.

The plus disease is a condition that may co-occur to ROP, illustrated in Figure 4. In retinas affected by plus, the blood vessels appear enlarged and twisted and pre-announce the worsening of the disease. The prompt detection of this condition is necessary for preventing the exacerbation of ROP and retinal detachment. Pre-plus indicates an intermediate stage where the severity of the eventual vascular abnormalities is not yet sufficient to define the presence of plus, but it is remarkable enough to plan earlier intervention. The presence of pre-plus or plus is assessed on the basis of the coexistence of clinical factors such as increased venous dilation and arterial tortuosity. The distinction between the two diseases is very subtle, and it is often a reason for strong disagreement among experts.

[ADD Figure 5 HERE] Figure 5. ROP progression on the same patient eye from the absence of plus disease, to pre-plus and presence of plus. The ROP grade is 2 in the first 2 pictures on the left and shifts to 3 in the picture on the right.

3.1.2 Dataset for the Experiments

The dataset for the experiments consists of 4,800 de-identified posterior retinal images from a private dataset. The images were obtained by a commercially available camera, namely RetCam by Natus Medical Incorporated (in Pleasanton, CA). A total of 3,024 images was used for training the network, consisting of 1,084 images without plus, 1,074 images with signs of Pre-plus and 1,080 images containing plus. The testing set comprises 100 images, including 54 normal, 31 pre–plus disease, and 15 plus disease images. The assignment of the images to each category is obtained by the majority voting from three expert assessments. The high class imbalance between plus and normal cases is a consequence of the low prevalence of the ROP disease (only 3%).

3.1.3 Task and Classification Model

The task is the automatic ternary classification of images into normal, pre-plus and plus classes.

We reproduce the preprocessing pipeline from Brown et al. [66]. The retinal vasculature is segmented by a U-Net [67]. The network assigns, for instance, a probability to each pixel for being part of a vessel. This segmentation step bypasses the domain shifts due to variations in terms of pigmentation, illumination, and nonvascular pathology. A resizing operation is performed to uniformize the image size to 224 x 224 pixels, the input size of the CNN used for the classification. An Inception-V1 network [68], pre-trained on ImageNet, is then finetuned on the ROP dataset to classify the images as normal, pre-plus or plus. We train the CNN with stochastic gradient descent and a categorical cross-entropy loss for 100 epochs. The learning rate is maintained constant to 1e-4. Data augmentation is applied with right-angle rotations and horizontal and vertical flipping. The hyperparameters are tuned by 5-fold cross-validation as in Brown et al. [66].

3.2. Concept Attribution with Regression Concept Vectors

3.2.1 Identification of the Concepts

The starting point of attribution to concepts is the identification of clinical concepts that should be used for generating explanations. In this section we will describe the workflow for the selection of clinical concepts represented in Figure 6, starting from the collection of information to the final definition of a list of concepts and how to measure them. In the framework defined in [14], we identify mainly two sources of information that can drive the selection of clinical concepts, namely the prior knowledge on the domain and the consultation with domain experts. Figure 6 shows these two as the starting point of the workflow for the selection of clinical concepts. Although for the experiments in this chapter we directly interacted with ophthalmologists to define the clinical concepts, we clarify in this section both approaches for completeness. We present the approaches from a high-level perspective so that they could be used also in other applications. The details on the interaction process and the type of questions that led to the identification of the concepts are reported in Sec. 4.2.1.

[ADD Figure 6 HERE] Figure 6. Workflow for the selection of clinical concepts. Reproduced from [14].

We describe in the following some of the sources of prior knowledge (represented as the starting point of the workflow in the box of the top left of Figure 6) that can lead to the identification of clinical concepts. Prior knowledge can be represented in multiple ways, e.g. collections of existing guidelines, as reports of previous studies or as annotated data. We discuss each of these in detail. Existing guidelines that are followed for human decision-making constitute an important resource, being based on several years of studies and joint efforts towards identifying decisive factors. Some examples of these are the well-established Nottingham grading (NGH) for breast cancer or the Gleason score in prostate cancer grading. The guidelines specify a list of factors that should be

assessed by the pathologists to determine the tumor grade. For example, abnormalities in the appearance of nuclei and cells is one of the criteria in the NGH. In addition to the guidelines, the combination of handcrafted visual features and ML has been studied for several years before transitioning to deep learning. The handcrafted feature extraction driven by expert knowledge in the domain drives, in some cases, the extraction of powerful features with prognostic relevance [69-71]. Written reports, besides, justify the decision-making by describing the image content and the main causes that led to the diagnosis. Information to identify clinical concepts can be collected from all of these sources, namely the grading guidelines, handcrafted features and written reports. The selection of the clinical concepts performed in this way is particularly useful to verify that domain-knowledge is reflected in the layer activations of the network.

As the box on the bottom left of Figure 6 suggests, the end users of the DL algorithm, in our case the physicians (ophthalmologists), can contribute to the selection of clinical concepts. As [14] suggests, this is rather an interactive process where the list of concepts is refined over multiple iterations until the explanations satisfy the users' inquiries on the model's decision-making. The direct interaction is useful to understand the expectations of the physicians on the DL decision-making. physicians may be interested in validating that the model decisions are in line with the guidelines of clinical practice as supposed in [13]. A question of interest, in this case, could be "Is the nuclei shape relevant to the automatic classification as tumor?". Confounding factors can be specified to make sure that irrelevant features are not used to make the classification, for example, by stating "changes in color appearance do not influence the classification".

The next step in the workflow is to understand whether the factors identified by prior knowledge and end-users can translate to questions about the model decision-making. The question "is nuclei size relevant to the classification?" for example could be posed to translate the attention to the

nuclei size and shape in the NGH into a relevant question about the automated decision-making. In this case, “nuclei size” is identifiable as a concept for the analysis. For an additional example, let us suppose that from the interaction with experts it emerged that they suspect that the watermarks at the bottom of the images may influence the model's attention. The sentence “the influence of the presence of watermarks is to investigate” therefore translate into a relevant question that should be answered by concept attribution: “is the presence of watermarks a relevant factor to the decision?”. The concept “presence of watermarks” is therefore added to the list of potential concepts. Note that this is a confounding factor with no clinical relevance and the outcome of the concept attribution analysis should show that this is not a relevant concept. If otherwise, this may highlight a bias in the decision-making requiring further analyses of the model and, if needed, the retraining of the parameters.

It is important to notice at this point that the concepts do not necessarily need to be specified in terms of the input features or the training data. Additional concepts can be defined using new data with annotations or from the metadata. Some concepts can be specific to the type of data being analyzed, as undefined for some data types. RGB color measures, for instance, are undefined for single-channel image modalities, e.g. computed tomography (CT) scans. Besides, to generate the explanations the list of concepts does not need to be perfect, and it will not likely be exhaustive of all possible concepts.

3.2.2 Computing the Regression Concept Vector

In this section, we formalize the computation of RCVs as described in [14].

The output of the CNN internal layer is used to find the RCV for that layer. This procedure is post-hoc and does not require the training of the parameters. The space of the activations of layer l,

$\Phi^l(\mathbf{x})$ is considered. We extract $\Phi^l(\mathbf{x})$ for $\mathbf{x} \in X$ where X is the training dataset, a subset of it, or an additional dataset describing the concepts. For each image $\mathbf{x} \in X$ we have access to, or we can compute, a value of the clinical concept for which we are seeking the RCV. We represent this operation of accessing or evaluating the value of the clinical concept by $c(\mathbf{x})$. Given one image representing tumor cells in a tissue slide, the average number of pixels in the segmentation of the nuclei regions can represent a value for the concept "nuclei size". We seek the linear regression that can model the value of the concept $c(\mathbf{x})$ for each $\mathbf{x} \in X$ as in the following:

$$c(\mathbf{x}) = \mathbf{v}_c \cdot \Phi^l(\mathbf{x}) + error$$

The RCV for the concept C is $\mathbf{v}_C$. The RCV components can be found by applying linear least squares (LLS) estimation to $X_{concepts}$. Figure 7 illustrates the approach for a 2-dimensional space.

[ADD Figure 7 HERE] Figure 7. In this two-dimensional example, the RCV is the direction represented by the regression plane. In higher dimensions, unwanted pixel dependencies are removed by an aggregating operation of the internal layer representation.

If l is a dense layer of width p, $\mathbf{v}_C$ is a p-dimensional vector in the space of its activations. If l is a convolutional layer the output of $\Phi^l(\mathbf{x})$ has spatial and channel dimensions (height, width, channels) represented as $w \times h \times p$. The simplest way of solving LLS in this space is to flatten $\Phi^l(\mathbf{x})$ to a one-dimensional array of whp elements as in [13,57]. This operation is widely discussed in [14], where better approaches are also proposed. Unrolling the convolutional maps may cause the explosion of the dimensionality of whp. The flattening operation, besides, breaks the natural 2D structure of the representation of convolutional feature maps, assigning neighboring features to independent dimensions. A spatial aggregation, i.e. global pooling, along

the (height, width) of each feature map is a solution to this shortcoming, generating a representation of $\Phi^l(\mathbf{x})$ as a one-dimensional array of p elements. This solution, only briefly mentioned in [57] and tested in [14] improves the quality of the regression fit. A further solution proposed in [14] is adding a regularization term to the optimization:

$$\mathbf{v}_c^{ridge} = argmin_{\mathbf{v}_c}(||c(\mathbf{x}) - \mathbf{v}_c\Phi^l(\mathbf{x})||_2^2 + \lambda||\mathbf{v}_c||_2^2)$$

As opposed to CAVs, RCVs allow expressing the influence of the concept in terms of increasing values rather than its sole presence. For this reason, they are more suited to medical applications, which often consider continuous measures. 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 $\mathbf{v}_C$

3.2.3 Generating Local Explanations by Conceptual Sensitivity

In this section, we summarize how local explanations can be generated for a single input by a derivative operation [14].

The conceptual sensitivity proposed in [11] constitutes the way of generating explanations for a single input image x in terms of a concept C, and it represents the impact of changes in the concept value $c(x)$ to the network output. It is defined for CAVs, hence for binary concepts, in [11]. The same formula can be applied to RCVs for categorical and continuous concepts. In the following paragraphs, we report the definition of conceptual sensitivity for binary and multitask classification.

For a binary classification task, the conceptual sensitivity $S_c^l(\mathbf{x}) \in \mathbb{R}$ is defined as the directional derivative of the network output $f(\mathbf{x})$ over the CAV or the RCV direction $\mathbf{v}_C$, computed as a scalar product:

$$S_c^l(\mathbf{x}) = \mathbf{v}_c \cdot \frac{\partial f(\mathbf{x})}{\partial \Phi^l(\mathbf{x})}$$

$S_c^l(\mathbf{x})$ represents the network responsiveness to changes in the input along the direction of the increasing values of the concept measures. The sign of $S_c^l(\mathbf{x})$ represents the direction of change, while its magnitude represents the rate of change. When moving along the RCV direction, the output $f(\mathbf{x})$ 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 $p(y = 1|\mathbf{x})$ when the representation $\Phi^l(\mathbf{x})$ is moved towards the direction of the increasing values of the concept. Negative conceptual sensitivity can be interpreted as an increase in $p(y = 0|\mathbf{x})$ when the same shift in the representation is applied. Conceptual sensitivities scores are informative about the concept influence on the decision for the single input image.

The derivation of the scores for multiclass classification tasks is straightforward.

Given the class label k, we consider the corresponding k-th neuron in layer L. The neuron activation before softmax, $\Phi^{L,k}(\mathbf{x})$, is a vector of real numbers representing the raw prediction values. These values are then squashed by the softmax into a probability distribution, namely the probability of the label k to be assigned to the input data point $\mathbf{x}$. The conceptual sensitivity score for class k is computed as:

$$S_c^{l,k}(\mathbf{x}) = \mathbf{v}_c \cdot \frac{\partial \Phi^{L,k}(\mathbf{x})}{\partial \Phi^l(\mathbf{x})}$$

The sensitivity scores can be computed for each class k, thus obtaining a vector of K elements. Large absolute values of the conceptual sensitivity for a single class correspond to a strong impact in the decision function when the activations are shifted along the direction of the RCV. The derivative of the decision function can be obtained by stopping gradient backpropagation at the l-th layer of the network.

3.2.4 Agglomerating Scores for Global Explanations

In this section, we report two ways in the literature of agglomerating the concept sensitivity scores to obtain global explanations of model behavior for an entire set of input data, e.g. for a full class. These two ways are, for instance, the TCAV score proposed in [11] and the Br score in [14]. These ways of agglomerating score take into consideration different aspects and can be seen as complementary. Alternative scores can explore additional characteristics of the conceptual sensitivity, e.g. the ratio between positive and negative sensitivities or the largest variation. The UBS score in [16] proposes a layer-agnostic score that allows the comparison of the concept sensitivities across all layers in a CNN. Note that one score is computed for each concept analyzed. If we were to consider three concepts in our analysis, for example, three TCAV scores would be computed, namely one agglomerating all the conceptual sensitivity values obtained for the first concept, one agglomerating all of those for the second and finally one agglomerating the values for the third.

The TCAV score is defined in [11] as the fraction of k-class inputs for which the activation vector of layer l is positively influenced by the concept c

$$\mathrm{TCAV} = \frac{|\{\mathbf{x} \in X_k : S_c^{l,k}(\mathbf{x}) > 0\}|}{|X_k|}$$

where $X_k \subset X_{task}$ is the set of inputs with label k. The TCAV score is bounded between zero and one. If no images are influencing the decision with a positive gradient, TCAV is zero. In the original paper, TCAV is only defined for CAVs [11], but its application is the same for RCVs. Bidirectional relevance (Br) scores are proposed for medical tasks with two class labels symbolizing either the presence or the absence of a condition, for example tumor in [14]:

$$Br = R^2 \times \left(\frac{\hat{\mu}}{\hat{\sigma}} \right)$$

The coefficient of determination $R^2 \leq 1$ measures if the concept vector is representative of the concept. The coefficient of variation $\hat{\sigma}/\hat{\mu}$ is the standard deviation of the conceptual sensitivities over their average. This score is large when the RCV models correctly the concept values, i.e. R^2 is 1, and when the conceptual sensitivity values are consistent for all input samples, lying closely around their sample mean. Br explodes to infinite if $\hat{\sigma} = 0$. A normalization per layer is applied to scores for multiple concepts such that the highest magnitude is equal to 1. This scaling permits the comparison of the scores among concepts.

4. Experiments and Results

4.1 Network performance on the ROP task

We report in this section the results of the deep learning classification experiment on the ROP dataset.

The first step before classification is the vessel segmentation by a U-Net model, for which we report the output of some segmentations, compared to the raw images in Figure 8.

[PLACE Figure8 HERE] Figure 8. Raw input images and outputs of the vessel segmentation from the U-Net model.

The mean area under the ROC curve was computed on the validation sets, across five cross-validation splits. We obtain 0.94 (standard deviation 0.01) for the diagnosis of normal (i.e. binary classification normal vs. pre-plus/plus) and 0.98 (0.01) for the diagnosis of plus disease (i.e. binary classification plus vs. normal/pre-plus). The classification on the test set of the best model (based on cross-validation) achieves 91% accuracy on the 100 test images, sensitivity of 93% and specificity of 94%. The model loss and accuracy over training of one split of the cross-validation are reported in Figure 9.

[PLACE Figure9 HERE] Figure 9. Model loss (on the left) and accuracy (on the right) at each training epoch on the training and validation sets. Image credits: James M. Brown

4.2 Results of Concept Attribution

4.2.1 Identification of the Concepts

In the following sections, we describe the explainability results obtained on the ROP classification task using the methods described in Sec. 3.2. In this section, in particular, we report in detail the process that led to the identification of the clinical concepts used for the research described in [17]. As mentioned in Sec. 3.2.1, the definition of the concepts was mostly driven by the interaction with the physicians (opthamologists).

This interaction focused on clarifying the visual factors that are taken into consideration to diagnose plus disease in ROP. This was made by asking them to sort the images in terms of the degree of aggressiveness of ROP and in particular of the plus disease. In Figure 10, we show some images for the three classes, namely normal, pre-plus and plus. Insights about these images were discussed with the physicians. To explain the visual differences in the images, they sketched out one of the patterns showing the exacerbation of the disease in Figure 11, namely the tortuosity of the retinal vessels.

[ADD Figure10 HERE] Figure 10. Image examples for each class, namely normal, pre-plus and plus.

[ADD Figure11 HERE] Figure 11. Sketch from the interaction with ophthalmologists during the identification of clinical concepts for the explainability analysis.

Vessel tortuosity is also present in the literature of ROP as an important pattern for the detection of plus. Mathematical models, for example, describe the vessel appearance in terms of the Cumulative Tortuosity Index (CTI) [72, 73]. The aim of the feature modeling in the literature is to describe the appearance of the vessels in the whole retinal sample by computing standard statistics of the mathematical model representing the per-vessel features. These features can then be used with standard ML algorithms for automatic ROP diagnosis. The features are divided into three groups according to their computation: point-based, segment-based or tree-based. An example of a point-based feature is the average point diameter. It describes the width of the vessels, in the direction normal to the blood flow, at each location in the image. Segment-based features are computed on segments of the vessel trace. For example, average segment diameter is obtained by dividing the number of pixels in the vessel by the vessel curve length. Finally, the distance to the center of the optic disc is an example of a tree-based feature. This feature represents, for each vessel segment, the distance between the ending point of the vessel and the disc center. More details and exhaustive comparison of the feature types can be found in [43].

Following this interaction with the physicians, we considered the handcrafted feature design used for ML approaches in ROP applications [72, 73]. We computed 11 feature types following the well-established and validated approach in [73]. The pool of the extracted features was separated into two clusters to differentiate the signal originating from normal and abnormal vessels. The normal and abnormal clusters were then fit into a Gaussian Mixture Model (GMM) and the means, variances and the mixing component were used as GMM statistics. For each of the 11 types of features, we extracted eight standard statistics (minimum, second minimum, maximum, second

maximum, mean, median and second and third moments) and the five GMM statistics, obtaining a total of $11 \times (8 + 5) = 143$ handcrafted features. These features were extracted from the automated vessel segmentations obtained by the U-Net model described in 4.1. We trained 100 random forest classifiers on random train-test splits with replacement to rank the features in terms of importance according to their Gini coefficient. The median of the cumulative tortuosity index appeared in the top 5 for all 100 models, confirming the selection of tortuosity as a relevant concept. The Kernel Density Estimation (KDE) of the top 10 features is shown in Figure 12. We sorted the image samples in the training set by increasing values of the 10 retained feature types and we sampled some of the images to create the visualizations in Figure 13, Figure 14 and Figure 15. We presented these images to the physicians to collect feedback on which extracted features aligned the most with clinically relevant aspects according to them.

[ADD Figure12 HERE] Figure 12. Kernel Density Estimation of the handcrafted feature values for the three classes (normal in orange, pre-plus in green, plus in blue).

[ADD Figure13 HERE] Figure 13. Images in the training dataset sorted for increasing values of the feature “mean curvature” as defined in [73].

[ADD Figure14 HERE] Figure 14. Images in the training dataset sorted for increasing values of the feature “median curvature” as defined in [73].

[ADD Figure15 HERE] Figure 15. Images in the training dataset sorted for increasing values of the feature “Average point diameter mean” as defined in [73].

From the inspection of the sortings in Figures 13-15, the features of curvature “curvature mean” and “curvature median” appeared informative about the class differences, while the utility of the “Average point diameter mean” was not clear to the experts. By this interaction with the physicians, we refined the list of concepts to six measures covering a wide set of clinically interpretable features, including the notion of tortuosity discussed in Figure 11. Features with a

frequency of appearance lower than 10 % in the ranking were discarded. The retained measures are described in Table 1.

Table 1. Handcrafted feature description and clinical interpretation. $\kappa(s)$ describes the rate of changing velocity between points with respect to the rate of changing the curve length between points. L_c and L_x denote respectively curve and chord length. $W_n$denotes the width of the vessel in the normal direction.

Feature	Name abbreviation	Description	Clinical interpretation
curvature	CURV	$\kappa(s)$	rate of direction change
Avg Segment Diameter	ASD	$\#pixels/L_c(x)$	global dilation
Avg Point Diameter	APD	$W_n(x)$	absolute dilation
Cumulative Tortuosity Index	CTI	$\mathrm{cti}(x) = L_c(x)/L_x(x)$	curving, curling, twisting rate

[ADD Figure16 HERE] Figure 16. Examples of the vessel segmentations according to their values of the handcrafted features. The top row shows the masks retrieved from the training data having the lowest value of the feature. The bottom row shows the masks with the largest value of the feature. The mn- and md- prefixes stand respectively for mean and median.

Figure 16 shows examples of the vessel segmentations retrieved from the training data according to their minimum and maximum values of the mean and median statistics computed for the features in Table 1. The RCVs in the next section will find a direction in the activation space of the CNN layers that represents the change from the minimum to the maximum values of these features.

The analysis in this section led to the central research question in [17], namely whether the concept-based explanations of concept attributions can be used to establish a link between the handcrafted features and the deep features.

4.2.2 Computation of the RCVs

Since the pre-plus disease represents a natural progression from normal to plus disease, we compute the RCVs on the set of training images for normal and plus. This was not done in [17], where the RCVs were computed separately on the two input classes. The R^2 for multiple layers of the network are reported in Table 2, evaluating the presence of the concepts at multiple layers in the CNN, as explained in Sec. 3.2.2. Two pooling strategies for aggregating the feature maps before computing the RCVs are compared, for which we illustrate the differences in Figure 17. The results for the regularized regression are compared against multiple values of the regularization penalty in Figure 18.

Table 2. Coefficient of determination R^2 for the ROP concepts. The pooling strategy is indicated on the top left of each block. The labels of the other columns refer to the layers of Inception-V1. Higher values of R^2 reflect the stronger presence of the concept. Results partially replicated from our study in [14].

max pool	**conv1**	**Mixed3b**	**Mixed4b**	**Mixed4c**	**Mixed5c**
medianCTI R^2	0.59	0.66	0.64	0.63	0.67

meanCTI R^2	0.49	0.56	0.50	0.47	0.56
medianCURV R^2	0.65	0.72	0.69	0.67	0.71
meanCURV R^2	0.65	0.70	0.61	0.57	0.72
medianASD R^2	0.55	0.66	0.58	0.56	0.64
medianAPD R^2	0.69	0.76	0.69	0.66	**0.76**
avg pool	**conv1**	**Mixed3b**	**Mixed4b**	**Mixed4c**	**Mixed5c**
medianCTI R^2	0.68	0.75	0.70	0.72	0.72
meanCTI R^2	0.56	0.63	0.54	0.55	0.56
medianCURV R^2	0.62	0.73	0.75	0.76	0.71
meanCURV R^2	0.65	0.74	0.68	0.69	0.71
medianASD R^2	0.69	0.74	0.67	0.67	0.64
meanAPD R^2	**0.72**	**0.80**	**0.76**	**0.77**	**0.76**

[ADD Figure17 HERE] Figure 17. Comparison of the regression of concepts of curvature (mdCURV and mnCURV), dilation (mdASD, mnAPD) and tortuosity (mdCTI and mnCTI) in ROP images of class normal and plus. Note that the "md" and "mn" prefixes stand respectively for median and mean. Results replicated from our study in [14].

[ADD Figure18 HERE] Figure 18. Impact of the parameter λ (strength of the regularization) on the ridge regression with (on the left) and without (on the right) global average pooling for the ROP concepts. The pooling operation reduces the need for regularization and leads to higher values of R^2. A subset of ROP concepts is shown, representing dilation (blue and orange) and tortuosity (green). Results replicated from our study in [14].

4.2.3 Evaluation of the Conceptual Sensitivities

In this section, we present two examples of the conceptual sensitivities as local explanations, as introduced in Sec. 3.2.3.

In Figure 19, we show the sensitivities for a misclassified image. The original values of the handcrafted features (which were used as concept measures) are reported on the left of the image. The network probability of each class is shown on top of the segmentation as p_n, p_{pre} and p_{plus}. The analysis highlights the fact that higher values of curvature and tortuosity would increase the prediction probability of the plus class. Similarly, Figure 20 presents the conceptual sensitivities for a correctly classified image.

[ADD Figure19 HERE] Figure 19. Conceptual sensitivities for a misclassification of a plus image as a pre-plus. The original values of six concept measures are displayed on top of the raw input image on the left. The network probabilities for the three classes, normal, pre-plus and plus are reported as p_n, p_{pre} and p_{plus}. Image reproduced from our work in [17].

[ADD Figure20 HERE] Figure 20. Conceptual sensitivities for a correct classification of an image of the normal class. The original values of six concept measures are displayed on top of the raw input image on the left. The network probabilities for the three classes, normal, pre-plus and plus are reported as p_n, p_{pre} and p_{plus}. Image reproduced from our work in [17].

4.2.4 Global Explanations with Br

The global explanations, as explained in Sec. 3.2.4, are summarized for inputs of the normal and plus classes in Figure 21.

[ADD Figure21 HERE] Figure 21. Global Br scores on the testing set for normal and plus images. Positive scores represent a shift towards the prediction of the normal class (left) or plus class (right) when the concept measure increases. Negative scores represent a shift towards these same classes when the concept measure decreases. Figure reproduced from our work in [17].

From the global explanations, curvature median appears as the most relevant concept to detect plus images with $Br = 1.0$. Avg point diameter mean is, on the other side, the most important concept for the detection of normal cases with $Br = -0.99$. The negative score shows that an increase of the value for this clinical concept would correspond to a decrease in the network output, hence a shift towards the prediction of the normal class. Avg point diameter mean and cti median appear as equally important with $Br = 0.56$ for the detection of plus.

5. Discussion of the Results

The presence of plus disease has a relevant impact on the treatment planning for ROP. Its diagnosis is, however, highly subjective, being mostly based on the identification of vessel dilation and tortuosity. The performance for the diagnosis of plus disease of the fully automated system in Sec. 4.1 compares, if not exceeds, that of ROP experts [65]. This result aims at showing that the use of DL models can introduce objectivity in the assessment of ROP severity, supporting physicians with difficult decisions such as establishing the presence of plus disease. This result shows a high potential of improving the clinical outcomes from the integration of DL and experts, similar to the results in other medical applications [8,9].

Interpreting the model predictions is a necessary step to validate the model's decision-making. The proposed approach to ROP classification is particular since the images classified by the Inception-V1 are not continuous multi-channel inputs, like natural images, they are binary masks of vessel segmentations. Feature attribution methods to obtain visualizations may therefore not provide sufficient insights on the decision-making. Since the inputs are binary masks, this application is also challenging for concept attribution. Despite their versatility in many other applications [13-15, 62], basic visual features such as image intensity and texture cannot be extracted from the binary masks of the vessels. The concept selection had to be defined on purpose for this task. The interaction with the ophthalmologists was essential to the formulation of the clinical concepts. The tortuosity measures of CTI emerge as relevant from the exchange with the physicians in Figure 11. This result is also in agreement with the analysis of the Gini coefficients, with CTI appearing in the top 5 for all the training repetitions of the random forests model classification. The vessel curvature is another interesting feature, according to the Gini coefficients. The visual differences between images with increasing values of vessel curvature features (i.e. curvature mean and median) suggest the relation between increased curvature and the presence of plus. The selection of clinical concepts has both upsides and downsides. On the upside, arbitrary concepts can be used to formulate explanations that directly address specific answers about the application. They do not need large annotated datasets, since the concepts can be computed automatically on the images, e.g. the handcrafted vessel features. The downside of the arbitrary choice is that the selection of the concepts itself is a delicate process, as seen in Sec. 4.2.1. The selection needs multiple iterations and it requires the participation of experts, to find the clinically relevant visual patterns, and developers, to implement the modeling of such patterns.

The computation of the RCVs (in Sec. 4.2.2) is straightforward after selecting clinical concepts. The results compare the RCVs obtained by applying the improvements proposed in [14] to obtain more stable vectors, including appropriate feature map pooling and regression regularization. The average pooling of the features leads to the most stable vectors, with the regularization only leading to small improvements. It is important to note that we compute R^2 and relevance scores on unseen test data, and not on the training data as generally done in statistics. The rationale behind this is to check whether the correlation learned in the input features can generalize and is robust enough to be predictive of unseen data. This is informative on the robustness and the reliability of the explanations. The risk of CAVs of capturing spurious correlation is therefore reduced in this application of RCVs by this evaluation on test data. Yet, more research is still needed to clarify the causal link between the presence of the concept and the decision [59].

The insights given by the local (in Sec. 4.2.3) and global explanations (in Sec. 4.2.4) represent a first attempt in bridging the gap between handcrafted visual features used for plus disease detection in classic ML approaches and the data-driven learning of features that is automated in CNNs. The scores reflect the clinical expectation that emerged from the interaction with the physicians, reporting high relevance for curvature and tortuosity in the diagnosis of plus disease. The relevance of average diameter mean as a discriminant factor for normal images is yet to be investigated. The visualizations in Figure 19 and Figure 20 propose a possible way of integrating the local explanations as a tool to assist the diagnosis, showing the conceptual sensitivities, the original and segmented images and the raw values of the hand-crafted features.

6. Conclusions

This chapter covered important topics in the quest for interpretable AI in the medical domain, presenting an explainability approach with an application on ROP. The often unclear terminology has led to confusion and multiple taxonomies for interpretability [20-25]. By reviewing these works we identified in Sec. 2.2.2 the terms for which most of the taxonomies agree in the definition. Importantly, we clarified the use of interpretable, explainable and intelligible. Interpretability and explainability can be used interchangeably for referring to the generation of explanations for the model decisions. Intelligibility refers to a wider group of methods that includes inherently interpretable models [4] and the introduction of interpretability as an additional model objective [26].

We presented the framework of concept attribution as opposed to the visualization techniques that are wide-spreading in the medical community. As discussed in Sec. 2.4, visualization methods may lead to unstable explanations that do not inspire reliability [35,64]. Concept attribution comes as an alternative approach to visualizations, that can provide further insights on the network decision-making, both at the global and the local level. Being post hoc, it does not need the retraining of the parameters and it can thus be applied to any network. If a more performant and accurate model was to be developed for ROP, concept attribution could be applied to the updated model.

The use of clinical concepts to explain the decisions may foster the comparison between the explanations of the models used in multiple institutions. This is in line with future developments of AI for healthcare, with federated learning approaches also promoting the exchange of information [74]. Clinical concepts, moreover, generate explanations that are at a higher level of abstraction than heatmaps. This makes the comparison of network behavior independent from the input images used to generate the explanations. By selecting concepts that match pre-existing

guidelines, explanations can help the physicians with verifying if the same set of values of principles is followed by the model decision-making. New hypotheses on the learned clinical concepts can be tested, also to verify that the network does not contain biases. For example, CAVs and RCVs could be used to inspect if the watermarks and text annotations, often present in medical images, affect the classifications.

From a more global perspective, explaining the automated decision-making of AI is a task at the frontier of two research worlds: the clinical and the developmental. Explanations should be generated with a human-centric approach, considering the requirements of the receivers of the explanations. For this reason, domain experts and DL developers should join forces to develop methods that can make the automated choices less intimidating and more understandable for physicians, while at the same time more stable and reliable from the development perspective.

Acknowledgements

This work was possible thanks to the following projects, part of the European Union's Horizon 2020 research and innovation program: PROCESS (grant agreement No 777533), AI4MEDIA (grant agreement No 825619) and EXAMODE (grant agreement No 825292). We thank Jayashree Kalpathy-Cramer, and Michael F. Chiang for the important directives on the research and the explanations on ROP. James Brown for providing the deepROP code, the trained model weights and some of the images in this chapter. We thank Veysi Yildiz for providing the feature vectors representing the clinical concepts.

References

[1] https://www.aiforsdgs.org/

[2] Nanni, Loris, Stefano Ghidoni, and Sheryl Brahnam. "Handcrafted vs. non-handcrafted features for computer vision classification." *Pattern Recognition* 71 (2017): 158-172.

[3] Zhou, Jianlong, and Fang Chen. "DecisionMind: revealing human cognition states in data analytics-driven decision making with a multimodal interface." *Journal on Multimodal User Interfaces* 12.2 (2018): 67-76.

[4] Caruana, Rich, et al. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015.

[5] London, Alex John. "Artificial intelligence and black-box medical decisions: accuracy versus explainability." *Hastings Center Report* 49.1 (2019): 15-21.

[6] Yune, S., et al. "Real-world performance of deep-learning-based automated detection system for intracranial hemorrhage." *2018 SIIM Conference on Machine Intelligence in Medical Imaging: San Francisco*. 2018.

[7]Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." *arXiv preprint arXiv:1702.08608* (2017).

[8] Brown JM, Campbell JP, Beers A, Chang K, Ostmo S, Chan RVP, Dy J, Erdogmus D, Ioannidis S, Kalpathy-Cramer J, Chiang MF; Imaging and Informatics in Retinopathy of Prematurity (i-ROP) Research Consortium. Automated Diagnosis of Plus Disease in Retinopathy of Prematurity Using Deep Convolutional Neural Networks. JAMA Ophthalmol. 2018 Jul 1;136(7):803-810. doi: 10.1001/jamaophthalmol.2018.1934. PMID: 29801159; PMCID: PMC6136045.

[9] Wang, Dayong, et al. "Deep learning for identifying metastatic breast cancer." *arXiv preprint arXiv:1606.05718* (2016).

[10] Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[11] Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." *International conference on machine learning*. PMLR, 2018.

[12] Cai, Carrie J., et al. "Human-centered tools for coping with imperfect algorithms during medical decision-making." *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019.

[13] Graziani, Mara, Vincent Andrearczyk, and Henning Müller. "Regression concept vectors for bidirectional explanations in histopathology." *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*. Springer, Cham, 2018. 124-132.

[14] Graziani, M., et al. "Concept attribution: Explaining CNN decisions to physicians." *Computers in Biology and Medicine* 123 (2020): 103865.

[15] Graziani, Mara, et al. "Interpretable CNN Pruning for Preserving Scale-Covariant Features in Medical Imaging." *Interpretable and Annotation-Efficient Learning for Medical Image Computing*. Springer, Cham, 2020. 23-32.

[16]Yeche, Hugo, Justin Harrison, and Tess Berthier. "UBS: A Dimension-Agnostic Metric for Concept Vector Interpretability Applied to Radiomics." *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Springer, Cham, 2019. 12-20.

[17] Graziani, Mara, et al. "Improved interpretability for computer-aided severity assessment of retinopathy of prematurity." *Medical Imaging 2019: Computer-Aided Diagnosis*. Vol. 10950. International Society for Optics and Photonics, 2019.

[18] Edwards, Lilian, and Michael Veale. "Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for." *Duke L. & Tech. Rev.* 16 (2017): 18.

[19] L. M. Cysneiros, M. Raffi and J. C. Sampaio do Prado Leite, "Software Transparency as a Key Requirement for Self-Driving Cars," *2018 IEEE 26th International Requirements Engineering Conference (RE)*, Banff, AB, 2018, pp. 382-387, doi: 10.1109/RE.2018.00-21.

[20] Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial Intelligence* 267 (2019): 1-38.

[21] Clinciu, Miruna-Adriana, and Helen Hastie. "A Survey of Explainable AI Terminology." *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*. 2019.

[22] Chromik, Michael, and Martin Schuessler. "A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI." *ExSS-ATEC@ IUI*. 2020.

[23] Arrieta, Alejandro Barredo et al. “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI.” *Inf. Fusion* 58 (2020): 82-115.

[24] Adadi, Amina & Berrada, Mohammed. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2870052.

[25] Tjoa, Erico and Cuntai Guan. “A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI.” *ArXiv* abs/1907.07374 (2019): n. pag.

[26] Bertsimas, Dimitris, Angela King, and Rahul Mazumder. "Best subset selection via a modern optimization lens." *The annals of statistics* (2016): 813-852.

[27] Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert. "Feature visualization." *Distill* 2.11 (2017): e7.

[28] Lipton, Zachary C. "The mythos of model interpretability." *Queue* 16.3 (2018): 31-57.

[29] Wang, Xiaosong, et al. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[30] Gargeya, Rishab, and Theodore Leng. "Automated identification of diabetic retinopathy using deep learning." *Ophthalmology* 124.7 (2017): 962-969.

[31] González-Gonzalo, Cristina, et al. "Improving weakly-supervised lesion localization with iterative saliency map refinement." (2018).

[32] Huang, Yongxiang, and Albert CS Chung. "Evidence localization for pathology images using weakly supervised learning." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2019.

[33] Korbar, Bruno, et al. "Looking under the hood: Deep neural network visualization to interpret whole-slide image analysis outcomes for colorectal polyps." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2017.

[34] Xu, Yan, et al. "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features." *BMC bioinformatics* 18.1 (2017): 1-17.

[35] Graziani, Mara, et al. “Evaluation and Comparison of CNN Visual Explanations for Histopathology”. (Under review) (2020)

[36] Reyes, Mauricio, et al. "On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities." *Radiology: Artificial Intelligence* 2.3 (2020): e190043.

[37] Pereira, Sérgio, et al. "Automatic brain tumor grading from MRI data using convolutional neural networks and quality assessment." *Understanding and interpreting machine learning in medical image computing applications*. Springer, Cham, 2018. 106-114.

[38] Hosny, Ahmed, et al. "Deep learning for lung cancer prognostication: A retrospective multi-cohort radiomics study." *PLoS medicine* 15.11 (2018): e1002711.

[39] Koh, Pang Wei, and Percy Liang. "Understanding black-box predictions via influence functions." *arXiv preprint arXiv:1703.04730* (2017).

[40] Raghu, Maithra, et al. "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability." *Advances in Neural Information Processing Systems*. 2017.

[41] Erhan, Dumitru, et al. "Visualizing higher-layer features of a deep network." *University of Montreal* 1341.3 (2009): 1.

[42] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." *European conference on computer vision*. Springer, Cham, 2014.

[43] Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." (2014).

[44] Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." *arXiv preprint arXiv:1703.01365* (2017).

[45] Springenberg, Jost Tobias, et al. "Striving for simplicity: The all convolutional net." *arXiv preprint arXiv:1412.6806* (2014).

[46] Bach, Sebastian, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." *PloS one* 10.7 (2015): e0130140.

[47] Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[48] Selvaraju, Ramprasaath R., et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization." *Proceedings of the IEEE international conference on computer vision*. 2017.

[49] Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018.

[50] Zintgraf, Luisa M., et al. "Visualizing deep neural network decisions: Prediction difference analysis." ICLR (2017).

[51] Marko Robnik-Šikonja and Igor Kononenko. Explaining classifications for individual instances. Knowledge and Data Engineering, IEEE Transactions on, 20(5):589–600, 2008.

[52] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should I trust you?" Explaining the predictions of any classifier." *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016.

[53] Achanta, Radhakrishna, et al. "SLIC superpixels compared to state-of-the-art superpixel methods." *IEEE transactions on pattern analysis and machine intelligence* 34.11 (2012): 2274-2282.

[54] Felzenszwalb, Pedro F., and Daniel P. Huttenlocher. "Efficient graph-based image segmentation." *International journal of computer vision* 59.2 (2004): 167-181.

[55] Shrikumar, Avanti, Peyton Greenside, and Anshul Kundaje. "Learning important features through propagating activation differences." *arXiv preprint arXiv:1704.02685* (2017).

[56] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems*. 2017.

[57] Alain, Guillaume, and Yoshua Bengio. "Understanding intermediate layers using linear classifier probes." *arXiv preprint arXiv:1610.01644* (2016).

[58] Ghorbani, Amirata, et al. "Towards automatic concept-based explanations." *Advances in Neural Information Processing Systems*. 2019.

[59] Goyal, Yash, et al. "Explaining classifiers with Causal Concept Effect (CaCE)." *arXiv preprint arXiv:1907.07165* (2019).

[60] Koh, Pang Wei, et al. "Concept bottleneck models." *arXiv preprint arXiv:2007.04612* (2020).

[61] Yang, Mengjiao, and Been Kim. "Benchmarking Attribution Methods with Relative Feature Importance." *arXiv* (2019): arXiv-1907.

[62] Graziani, Mara, Henning Muller, and Vincent Andrearczyk. "Interpreting intentionally flawed models with linear probes." *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2019.

[63] Tonekaboni, Sana, et al. "What clinicians want: contextualizing explainable machine learning for clinical end use." *arXiv preprint arXiv:1905.05134* (2019).

[64] Arun, Nishanth, et al. "Assessing the (Un) Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging." *arXiv preprint arXiv:2008.02766* (2020).

[65] Brown, James M., et al. "Automated diagnosis of plus disease in retinopathy of prematurity using deep convolutional neural networks." *JAMA ophthalmology* 136.7 (2018): 803-810.

[66] Brown, James M., et al. "Fully automated disease severity assessment and treatment monitoring in retinopathy of prematurity using deep learning." *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications*. Vol. 10579. International Society for Optics and Photonics, 2018.

[67] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *International Conference on Medical image computing and computer-assisted intervention*. Springer, Cham, 2015.

[68] Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[69] Whitney, Jon, et al. "Quantitative nuclear histomorphometry predicts oncotype DX risk categories for early stage ER+ breast cancer." *BMC cancer* 18.1 (2018): 610.

[70] Wang, Xiangxue, et al. "Computer extracted features of cancer nuclei from H&E stained tissues of tumor predicts response to nivolumab in non-small cell lung cancer." (2018): 12061-12061.

[71] Lee, George, et al. "Cell orientation entropy (COrE): predicting biochemical recurrence from prostate cancer tissue microarrays." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Berlin, Heidelberg, 2013.

[72] Hart, William E., et al. "Measurement and classification of retinal vascular tortuosity." *International journal of medical informatics* 53.2-3 (1999): 239-252.

[73] Ataer-Cansizoglu, Esra, et al. "Computer-based image analysis for plus disease diagnosis in retinopathy of prematurity: performance of the “i-ROP” system and image features associated with expert diagnosis." *Translational vision science & technology* 4.6 (2015): 5-5.

[74] Chang, Ken, et al. "Distributed deep learning networks among institutions for medical imaging." *Journal of the American Medical Informatics Association* 25.8 (2018): 945-954.