Look at this, not that! – Improving the alignment of PIP-Net with domain knowledge

Franziska Fobbe

8th September 2023

Master Thesis
Computer Science - Data Science, University of Twente
Artificial Intelligence, Universitè Paris-Saclay

Examination Committee
Prof. Dr. Christin Seifert
Dr. med. Han Hegeman
Dr. ir. Meike Nauta
Ing. Jeroen H. Geerdink

Abstract

Interpretable computer vision models like PIP-Net can push the application of machine learning models in critical domains like radiology. Based on PIP-Nets prototypes we propose two measures to evaluate a model's adherence to domain knowledge (which we define as not relying on spurious correlations of medically irrelevant features (shortcuts) and as being focused on regions of interest). We show that the methods are sufficient to discriminate models that adhere more to domain knowledge and can provide an additional dimension to traditional evaluation metrics like accuracy and AUC. By knowing which prototype correspond to shortcuts we can improve the models adherence to domain knowledge by retraining and reinitialising the classification layer of PIP-Net. For regions of interest the same strategy does not work and we propose and discuss a loss term that could improve the results in future research.

Medical Abstract

Objective We aimed to improve an interpretable machine learning model (PIP-Net), to align its decision to medical domain knowledge for hip fractures. Domain knowledge is defined as not relying on medically irrelevant features (shortcuts), such as hospital specific "L" and "R" tokens [1], and instead on fracture areas (regions of interest).

Methods We defined two measures (Intersection over Union and Prototype Importance) which combined give a metric to quantify domain knowledge adherence. With these we finetuned the model to correct shortcut learning, and proposed measures to improve reliance on regions of interest.

Results We found that shortcut learning can be mitigated by reinitialising the classification layer of PIP-Net. The same was not true when only focusing on Regions of interest. We proposed and discussed a new loss term that could lead to better coverage of Regions of Interest, improving the results in the future.

Conclusions Practitioner's feedback can be used to create a machine learning model that has high accuracy, is interpretable and adheres to domain knowledge. The model needs to be tested on more datasets and should go through iterative testing within the hospital to assess the quality more extensively.

1 Introduction

Applications of machine learning based tools have become increasingly used in radiology practice [2, 3], though not as pervasive as prior predictions suggested [4].

One reason for the slow uptake could be the black-box nature of neural networks, which do not explain their predictions in a human-understandable way. This is a major drawback for high-stakes decision-making (such as political or medical decisions), where the reasoning for decisions might need to hold up in court or under public scrutiny [5], or might even cost lives and livelihoods. It has led to calls for the development of interpretable models [6, 7, 8], which are built to be understood, instead of explainable models, which add post-hoc explanations on top of already established deep learning models. Explainable model additions, for example saliency maps can sometimes be misleading and should undergo sanity checks to avoid false confidence [9].

One class of interpretable models, especially for image categorisation, are prototypical models, first developed by Chen et.al. [10], which extract representative features (called prototypes) from a convolutional neural network (CNN) backend which are an input to an interpretable method such as a logistic regression or a decision tree. Combining these two model classes creates inherently interpretable models that can achieve similar accuracies to the black-box models and are therefore a natural choice for high stakes decision making. PIP-Net is a recent addition to the prototypical model class, which optimises the model for semantic understanding, by prioritising semantic cohesion of prototypes. As it also outperforms other prototypical models such as ProtoPNet [10] and is on par with ProtoPool [11] and therefore we chose it as the baseline model for this thesis.

However, while an interpretable model explains its reasoning, it does not mean that the reasoning is aligned with the human perception of the problem it is solving. For example a model can pick up spurious correlations in the data, leading to so-called shortcut learning [12, 1]. If we want a model to be used in practice, it is important to build trust in the decision making, which means that a radiologist needs to agree with the reasoning of the model. The classification of whether or not a X-ray depicts a fracture, should be based on the area of the image that shows the fracture.

Taking advantage of the interpretable nature of PIP-Net, we can use the additional information that the model gives to improve PIP-Net to adhere closer to domain knowledge, by discouraging it to look at (known) shortcuts and encouraging it to look at (known) regions

of interest (ROI). By quantifying the influence of ROIs and shortcuts on the decisions of the model, we open up a new evaluation dimension as an extension to traditional performance metrics, which can be used to assess the effectiveness of the model. This thesis aims to answer the following two research questions:

RQ1: How can we measure the ability to detect a) (known) spurious correlations and b) the alignment with domain knowledge for prototypical models?

RQ2: How can we adapt prototypical models, such that they a) do not rely on spurious correlations and b) better align with domain knowledge?

We wrote the thesis in cooperation with the ZGT Almelo-Hengelo, and domain knowledge was provided by Han Hegeman, MD PhD, a trauma surgeon with a special interest in hip fractures. The context in which these questions are answered, are therefore limited to this application.

The rest of the thesis is structured as follows: Section 2 summarises related works, Section 3 explains the architecture and training process of PIP-Net. The datasets, their customisation and the annotation process is described in Section 4. Research Question 1 is answered in Section 5 and Research Question 2 is expanded upon in Section 6. Finally, Section 8 discusses the results, and Section 9 concludes the thesis.

2 Related Work

Part-Prototype Models Interpretable models (see [13] for a first overview of methods and definitions) have been researched more as an answer to calls for transparent algorithmic decision making. This also led to the inception of part-prototype models for image classification tasks [10]. Prototypes are a learned representation of a feature in an image, similar to identification keys used in entomology when identifying an insect (see [14]). These prototypes can then be used as an input to interpretable decision making functions e.g. decision trees, rules or linear models [15, 16, 17], thereby making use of both the nonlinear optimisation properties of neural networks and the interpretability of the decision-making models. In part-prototype models, prototypes are first learned from an image corpus in the convolutional neural network (CNN) backbone and then identified (for example through a distance measure) in an image to be classified. The identified prototypes are compared to similar instances from the corpus, and therefore the model does not only rely on the prototype location (like a saliency map) but through the comparisons it becomes clear what semantic features it has picked up.

While Chen et al. [10] used a predetermined number of prototypes per class, Nauta et al. [18] showed that prototypes can be redundant. Multiple works [19, 20, 21] built on this foundation and adapted the original idea to reduce the number of prototypes further to reduce the explanation size. Taking human perceptive similarity of prototypes into account, PIP-Net [22] uses self-supervised representation learning to learn the prototypes and further improve the interpretability of the prototypes. Therefore PIP-Net is uniquely suited to produce interpretable prototypes and will form the basis of this thesis. More detail about PIP-Net can be found in Section 3.

Shortcut Learning Deep Neural Networks have been known to learn spurious correlations or become so-called "Clever Hans" predictors [12]. Shortcut learning can be hard to detect [23] and can not only lead to a reduced accuracy on the test set but also to unintended consequences [24], which could be especially tragic in the medical domain. As shown in [25] an interpretable model like PIP-Net can be used to identify clinically irrelevant artefacts and they can subsequently be removed from the model. Building on these results, Section 5 will identify "Clever Hans" artefacts in hip fracture x-ray datasets (see Section 4) and Section 6 will expand on it.

Alignment with Domain Knowledge In interpretable models, it is necessary that the model's explanation can be understood by a human, even though Borowski et al. [26] showed that neural networks use different strategies in recognising objects than humans. Makino et al. [27] shows that also in the field of radiology, a deep learning model focuses on (spurious) features usually ignored by radiologists. Similarly, Nauta et al. [25] shows that the PIP-Net model can likewise produce prototypes that are relevant to medical domain knowledge, but not exclusively. This alignment with domain knowledge is also evaluated and discussed in Section 5.

3 PIP-Net

In this section, we shortly describe PIP-Net [22], which is used in this thesis. PIP-Net combines a CNN backbone, that extracts latent features corresponding to prototypes with a linear layer. PIP-Net's reasoning can be seen as an inherently interpretable function. We briefly summarise the model in this paper for the context of further exploration. The model is described in detail in [22].

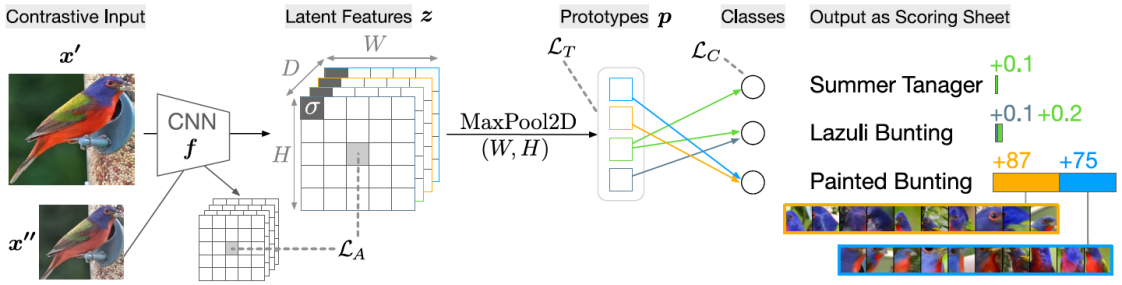


Figure 1: Schematic depiction of the PIP-Net architecture, as described in [22]. Two images, one original, one transformed are forwarded through the CNN backbone, and the prototypes are pretrained and fine-tuned by minimising the alignment and tanh loss of the latent features of both images. The identified prototypes are used as an input to the linear classification layer. The result can be shown as a scoring sheet reasoning.

Architecture Details A schematic representation of PIP-Net's architecture is shown in figure 1.

The CNN backbone learns prototypical representations which are represented as 1-dimensional prototype presence scores in the last layer of the network. The CNN backbone can consist of any context specific architecture, e.g. ResNet [28] or ConvNeXt [29], because different architectures can be appropriate for a use case. We use ConvNeXt in this thesis. Two images, one of which is a transformed version of the other, are processed by the CNN (shared weights) resulting in a convolutional output z for each image. The final layer consists of D (number of prototypes) feature maps with dimension $(H \times W)$. After applying a softmax over D, such that $\sum_d^D z_{h,w,d} = 1$, a patch $(z_{h,w,:})$ is forced to belong to exactly one prototype. This has the additional effect that the last layer can be interpreted as a saliency map of the prototype's existence over the image. By applying a max-pooling function per feature map $z_{:,:,d}$ we can identify the presence of a prototype in an image and forward the resulting image encoding p to the next stage.

The image encoding p is the input to the sparse linear classification layer, with weights $\omega_c \in \mathbb{R}_{\geq 0}^{D \times K}$. The learned weights per class represent the relevance of a prototype to the class. The output from this layer can therefore be interpreted as a scoring sheet, where the score for a class is the sum of all present prototypes multiplied by their weights.

Training process The training process consists of two distinct training stages: (i) The self-supervised pre-training stage, which pre-trains the last layer of the CNN to be a probability map of a prototype distribution. (ii) A training stage, which trains the model to achieve a high classification accuracy.

During self-supervised pretraining only the CNN backbone is trained. The aim is to learn semantic similarities and not to achieve a high classification accuracy. To attain this, the net is been given contrasting inputs, e.g. an image and a transformed version of the image. The aim of the stage is to minimize the difference in the feature representation of both images in the final layer of the CNN. This is achieved by the alignment loss:

$$L_A = -\frac{1}{HW} \sum_{(h,w) \in H \times W} log(z'_{h,w,:} \cdot z''_{h,w,:}) \quad (1)$$

To prevent a trivial solution, a second loss, the tanh-loss L_T is utilised to ensure that every prototype should be present at least once in a mini-batch.

$$L_T(\mathbf{p}) = -\frac{1}{D} \sum_d^D log(tanh(\sum_b^B \mathbf{p}_b) + \epsilon) \quad (2)$$

The two losses are combined into the overall loss of this stage: $\lambda_A * L_A + \lambda_T * L_T$

After the first training step, the linear layer will be trained along with the CNN, and a classification loss is added to the previous loss term.

The third loss term is the standard negative log-likelihood loss for classification tasks.

$$L_C(x, y) = \sum_{n=1}^{N} \frac{1}{\sum_{n=1}^{N}} - w_{y_n} x_{n,y_n} \quad (3)$$

where y is the target, x is the prediction, N is the size of the mini-batch, and w is the weight associated with the predicted class (1 if no weights are chosen).

They are combined into the total loss as

$$\lambda_A * L_A + \lambda_T * L_T + \lambda_C * L_C$$

4 Experimental Setup

To set up the experiments we introduce the dataset, model parameters and the evaluation metrics, which are used to create a baseline model that are analysed in Sections 5 and 6.

Datasets. The datasets for answering the research questions have been provided by the ZGT Almelo-Hengelo and consist of X-rays of hips, some of which depict fractures. The datasets include one high-quality dataset depicting trochanteric fractures of the hip (named HIP-TF). These fractures are are often obvious even to the untrained eye. The second dataset (HIP-CF) includes multiple different fracture types, but mostly column fractures. These fractures are harder to spot and the fracture area usually takes up a smaller percentage of the overall X-ray than in the HIP-TF dataset[1]. All datasets have binary labels ("fracture" and "no fracture"), which indicate the presence or absence of a fracture in the X-ray.

To answer the research questions the datasets are annotated with the positions of the artificially inserted patches and regions of interest.

Shortcuts. We added rectangular patches to the images in the HIP-CF dataset, which presence is positively correlated with the "no fracture" class. The rectangles are assigned random RGB values from the interval $[200, 255]$. Their size is roughly the same as an image patch from PIP-Net (image size divided by 10), so their presence in a patch can be confidently attributed to a found shortcut, and not to other information found in the patch. To make sure that no informative part of the image is covered, the centre of the image (the middle 50% of the total width) is excluded from patch placement and a gaussian blur (radius: image width divided by 20) is applied to the rest of the image to find the darkest part of the X-ray. For every image the exact location of the shortcut is recorded. To evaluate the effect of shortcut learning on the PIP-Net model, three different shortcut probabilities are applied. The dataset HIP-CF$^{50\%}$ includes shortcuts in 50% of the images, the dataset HIP-CF$^{70\%}$ includes shortcuts in 70% of the images and the dataset HIP-CF$^{100\%}$ includes shortcuts in all of the images. Figure 3 shows an example.

Regions of Interest. Regions of interests in the context of these datasets are defined as the areas of the X-ray that visibly depict fractures. The images were inspected and annotated with the *labelme* software, as a polygon with the label "fracture". The coordinates of the polygon are saved for each image. To ensure the annotation quality random samples were controlled by domain experts and subjected to inter-annotator agreement analysis (see Appendix B for more details).

A summary of the datasets can be seen in table 1.

Model Training. All of the datasets defined above are classified with a PIP-Net model with the same configurations: ConvNeXt Tiny as backbone [29], a learning rate of 0.05 for the linear classification layer and 0.0001 for the backbone. We use a batch size of 64, 16 epochs for pretraining, and 85 epochs for training (16 of those with a frozen backbone), as introduced in [25].

Metrics. Traditionally, models like the ones we present here are evaluated according to their prediction quality.

Performance metrics like accuracy

$$Accuracy = \frac{\text{Number of Correct Predictions}}{\text{Total number of predictions made}}$$

or AreaUnderCurve (AUC), which calculates the Area under the Receiver Operating Characteristics (ROC) curve, plotting the True Positive Rate (TPR)

$$TPR = \frac{TP}{TP + FP}$$

against the False Positive Rate

$$FPR = \frac{FP}{TP + FP}$$

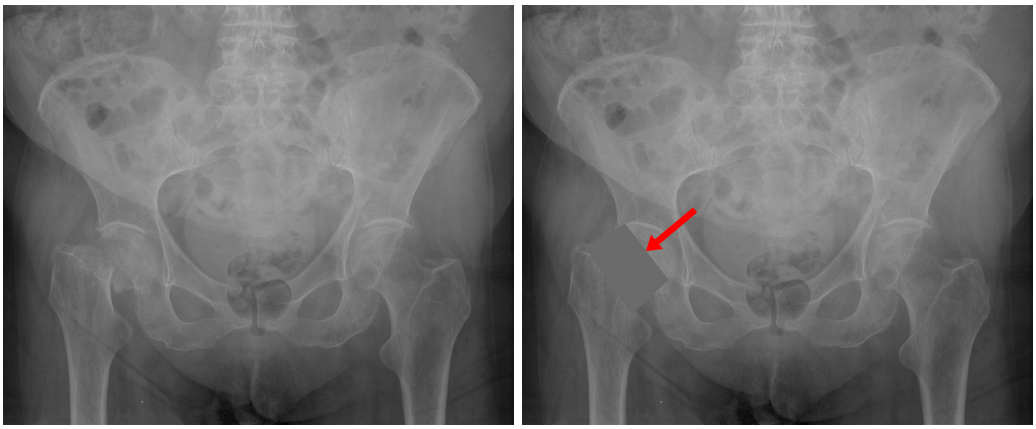
(where TP is the number of true positive predictions, and FP is the number of false positive predictions) have not been designed to evaluate the adherence to domain knowledge and can therefore not distinguish between models on this basis.

Baseline model. The baseline runs for the two raw datasets (HIP-CF and HIP-TF) as well as the datasets with artificially added patches (HIP-CF$^{50\%}$, HIP-CF$^{70\%}$, HIP-CF$^{100\%}$) evaluated with accuracy and AUC is in table 2. The results show the average performance from three runs, with the standard deviation of the metrics in parentheses. All runs have a comparatively high performance (accuracy above 0.94). While the table shows that accuracy is increasing with a higher correlation of artificially added patches, it would not be possible from these results alone to judge if a model would be determined to be useful to deploy in medical practice.

[1] For more details on the different fracture types refer to Appendix B

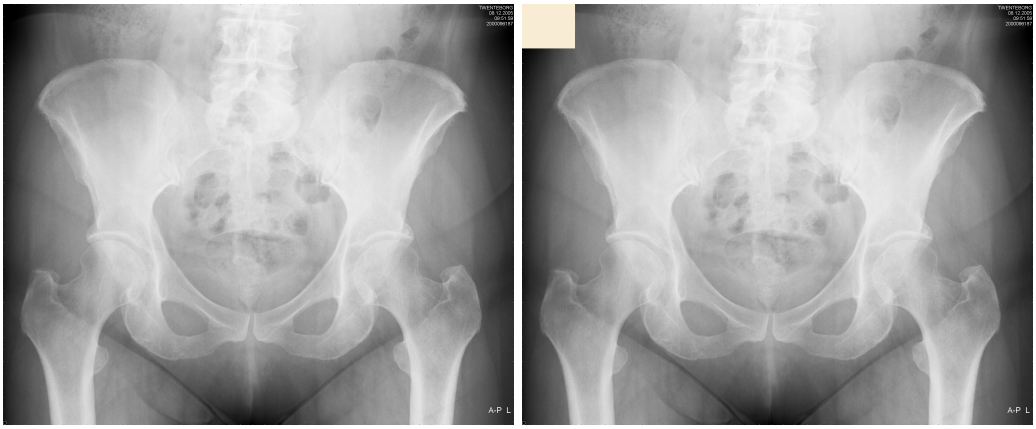
	Description	Remarks	Label	
			Train	Test
HIP-TF	Dataset from hospital Ziekenhuisgroep Twente (ZGT). ZGT's database of digital radiography was queried for hip X-rays which were taken on the suspicion of hip fracture between 2005 and 2022 (patient age $\geq$ 16). The fracture class (trochanteric fracture vs. no fracture) label was extracted from the electronic health record and crossmatched with the DBC code (financial registration code for diagnostics and treatment). All images were converted from DICOM format to png.	2115 images of fracture class have ROI annotations	# fracture 1590 # no fracture 9235	# fracture 472 # no fracture 2654
HIP-CF	Dataset from hospital Ziekenhuisgroep Twente (ZGT). ZGT's database of digital radiography was queried for hip X-rays which were taken on the suspicion of hip fracture between 2005 and 2018 (patient age $\geq$ 21). The fracture class (column fracture vs. no fracture) label was extracted from the electronic health record and crossmatched with the DBC code (financial registration code for diagnostics and treatment). All images were converted from DICOM format to png.	$\approx$200 ROI annotations in fracture class	# fracture 3468 # no fracture 4080	# fracture 859 # no fracture 1005
HIP-CF$^{50\%}$		50% of no fracture class include shortcut	# fracture 3468 # no fracture 4080	# fracture 859 # no fracture 1005
HIP-CF$^{70\%}$		70% of no fracture class include shortcut	# fracture 3468 # no fracture 4080	# fracture 859 # no fracture 1005
HIP-CF$^{100\%}$		100% of no fracture class include shortcut	# fracture 3468 # no fracture 4080	# fracture 859 # no fracture 1005

Table 1: Description of the datasets, including their number of samples. All datasets have binary labels indicating the presence or absence of a fracture.



(a) the original X-ray (b) The X-ray with ROI covered

Figure 2: An X-ray with column fracture on the left side of the image.
a) the ROI is visible, b) the image is covered by a patch of the average pixel colour.



(a) The original image (b) The image with added shortcut

Figure 3: An X-ray from the "no fracture" class. a) original image, b) with added shortcut in the upper left corner.

Class	Dataset	Accuracy	AUC	worst class accuracy
fracture	HIP-TF	0.944 (0.019)	0.981 (0.003)	0.925 (0.026)
	HIP-CF	0.946 (0.001)	0.983 (0.001)	0.942 (0.001)
no fracture	HIP-CF$^{50\%}$	0.957 (0.001)	0.991 (0.000)	0.954 (0.003)
	HIP-CF$^{70\%}$	0.962 (0.003)	0.992 (0.000)	0.958 (0.004)
	HIP-CF$^{100\%}$	0.995 (0.001)	0.999 (0.000)	0.991 (0.002)

Table 2: Baseline runs for all datasets. Values are average over three runs, the standard deviation is in parentheses. All runs are fitting a binary classification. For the first two runs the "fracture" class is associated with domain knowledge, through the regions of interest. For the last three runs the "no fracture" class is associated with domain knowledge, through the different shortcut correlations. All runs have very comparable performances and it is not possible to detect domain knowledge alignment from the table.

5 Quantification

When looking at table 2 it is not possible to pick the model which explanations are most similar to human intuition, as the evaluation metrics are roughly the same across all models. However, PIP-Net provides additional information about the classification. The locations of the activated prototypes can give us further insights into the model quality that go beyond the established methods.

For an example why it is important to evaluate adherence to domain knowledge see figure 4, where you can see a comparison between the prototype distribution after the training on HIP-CF and HIP-TF datasets. Looking at table 2 the models seem to be of equal quality, but the HIP-TF model has only very few prototypes, which have a high overlap with the ROI, whereas the HIP-CF model has a much higher prototype count for the same class, which are scattered throughout the image. A good measure would instead assess the visual alignment of the prototypes with the interested regions, as well as the mathematical importance of the regions to the classification, as we want our most important prototypes to be located in the interested areas of the X-ray.

Prototype placement A prototype that is aligned with domain knowledge should be located at a ROI and should not be located at a shortcut. Therefore we propose the Intersection over Union (IOU) to quantify how much the prototype is overlapping the interested region. We define the area of the shortcut as ROI_{sc}, the area of the fracture as ROI_{fr} and the area or the prototype patch as A_p. To facilitate the formula, both ROI_{sc} and ROI_{fr} are referred to as ROI_*. To better compare the results of this calculations for both application cases we use a correction term, where we divide the bigger area by the smaller one. [2]

$$IOU = \frac{A_p \cap ROI_*}{A_p \cup ROI_*} \times correction$$

A higher IOU with ROIs and a lower IOU with shortcuts is evidence for a higher adherence to domain knowledge. We apply the IOU measure to all activated prototypes of an image to create a candidate lists of prototypes that depict shortcuts or ROIs.

[2] As the ROI_{fr} is usually a lot bigger in this instance than A_p the IOU would become very small, therefore we correct the term by the differences in size in the two areas. We calculate the correction term as $\frac{ROI_{fr}}{A_p}$. The opposite is true for ROI_{sc}, which is smaller than A_p. In this case we calculate the correction term as $\frac{A_p}{ROI_{sc}}$

To summarise the resulting distribution over all prototypes into one value, we report $\overline{IOU}$:

$$\overline{IOU} = Mean(IOU|IOU \neq 0)$$

We only consider instances of $IOU \neq 0$ as the artificial shortcuts are only present in the "no fracture" class, and ROIs are only defined in the "fracture" class. The IOU is therefore 0 by default in these cases.

However, the current localisation of the prototypes is not very accurate (see [30] for more details), and only one instance of the prototype is considered in this method [3]. Therefore, this measure on its own is not regarded as sufficient for the evaluation of the prototype's alignment with human reasoning.

Prototype Importance A prototype should not only be located at the interested region, but also be important for the classification. To quantify how much the area of the shortcut or the ROI is influencing the classification, the area can be masked (in the case of ROI, see figure 2) or compared to the original image without shortcut (in the case of the shortcut analysis, see figure 3). A prototype that is depicting the interested area can then not be activated anymore. Comparing the activated prototypes and the resulting classification weights gives a measure of the importance of the interested area towards the whole classification. The prototype presence scores of the model are called $\mathbf{p}$ and are associated with their weights ω_C in the linear layer for each class C. These weights are non-negative ($\omega_c \in \mathbb{R}_{\geq 0}^{D \times K}$) and $\mathbf{p} \in [0, 1]$. *covered* indicates the interested areas are not visible (ROI_{fr} or ROI_{sc} are not masked or not visible). *uncovered* indicates the image includes its ROI_*.

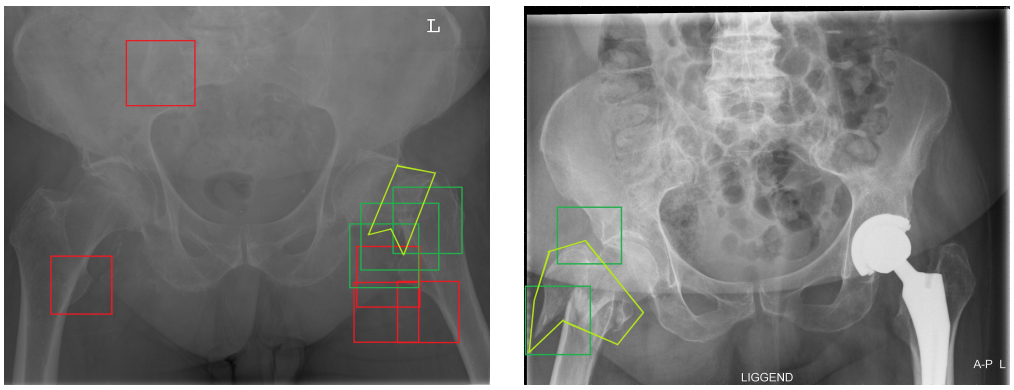
$$PI = \frac{\mathbf{p}_{covered} \cdot \omega_{C,covered}}{\mathbf{p}_{uncovered} \cdot \omega_{C,uncovered}}$$

The prototypes whose presence scores were changed the most between the covered and uncovered images are collected in a second candidate list of prototypes that depict shortcuts or ROIs.

As masking or removing the shortcut can change image properties such as the average pixel value, or trigger the activation of other shortcuts this measure is not sufficient to determine which prototypes are responding to interested areas.

Combining the Measures Both Prototype placement and Prototype importance create candidate lists for prototypes that depict ROIs

[3] a prototype can be activated at multiple patches in the image, but only the patch with the highest probability is considered for the calculation



(a) Example output from the HIP-CF model. While some prototypes are overlapping the ROI, there are even more scattered around the image.

(b) Example output form the HIP-TF model. The only two prototypes are overlapping with the ROI; The model aligns well with domain knowledge.

Figure 4: Example images showing the prototype locations in the X-ray images. The outline of the region of interest is shown in yellow, prototypes which overlap with the ROI are marked in green, prototypes outside the ROI are red

or shortcuts. Examples of these results can be seen in Figures 5a and 5b. As argued above either of these candidate lists are not pure, meaning they also can include prototypes that are not depicting shortcuts and ROIs. However, intersecting these lists leaves only candidates with high probability of depicting an interested area. By manually checking all prototype lists against the visualisations of the relevant prototypes and their positioning on the images, we found the candidate lists to be consistent with human evaluation. Therefore we conclude that the methods are able to assess adherence to domain knowledge in PIP-Net.

5.1 Results

Applying IOU and PI to the same models makes the differences between their adherence to domain knowledge apparent. While all five models have a very similar performance (see Table 2), differences become clear when looking at Table 3. Table 3 includes the number of prototypes for the relevant class. As the ROIs are only defined for the "fracture" class and the shortcuts are correlated with the "no fracture" class, the relevant class shifts between the two application cases. The column $\overline{IOU}$ depicts the average Intersection over Union for prototypes which have an IOU bigger than 0. The higher the $\overline{IOU}$, the more the prototypes are aligned with the interested area. The column $\overline{PI}$ depicts the average prototype importance for images where the prototype importance changed when comparing covered and uncovered images. The higher the $\overline{PI}$, the higher the influence of the interested area on the final classification. Finally, the column "# of identified prototypes" shows the length of the intersected candidate lists created by both measures. The number indicates the amount of prototypes that depict the interested areas.

Regions of Interest. Adherence to the region of interest can be seen both in Prototype Location and Prototype importance. As ROIs are defined for the fracture class both model HIP-TF and HIP-CF can be compared to see that HIP-TF has a higher adherence to domain knowledge than HIP-CF. HIP-TF only has 2.3 prototypes in the relevant class on average. Both IOU and average prediction percentage are quite high and most of the prototypes are identified as depicting the ROI. This means that the model has a very high alignment with domain knowledge, an impression that is confirmed by looking at prediction visualisations such as in Figure 4. Comparing this to the values from the HIP-CF model, it can be seen that this model has a lot more prototypes in the relevant class, but a much lower percentage of those are identified as being relevant to the identified class after intersecting the candidate. Based on this analysis it becomes clear that the HIP-CF adheres less to domain knowledge and leaves room for improvement.

Shortcuts. The classification's dependence on shortcuts becomes visible when looking at average IOU and average change of predictions. Shortcuts are added to the HIP-CF dataset, correlating with the "no fracture" class, therefore to assess the models dependence on shortcuts we analyse the models in the second section of Table 3. Comparing the HIP-CF run to the ones with an increasing correlation with shortcuts it can be seen that the number of prototypes in the relevant class is decreasing with higher cor-

relation, but the number of identified shortcut prototypes is increasing. Therefore a higher percentage of the overall classification is depended on the existence of shortcuts. This is confirmed through looking at average percentage of prediction, which is also increasing along with the shortcut correlation. On the other hand the average IOU stays constant. To look at this phenomenon more closely it is helpful to inspect the average IOU per prototype, as can be seen in figure 6. The graphs show that the average IOU and the average prototype importance per prototype. With a higher correlation with the shortcuts, the number of identified prototypes is increasing, which can be seen in the higher overlap between the two distributions. While the three shortcut correlation runs were almost indistinguishable in table 2, table 3 and figure 6 show that the proposed measures make it clear how much the runs are diverging from domain knowledge by their shortcut dependence.

Bringing it all together. For both ROIs and shortcuts the proposed measures give the ability to distinguish between models that are closely aligned with domain knowledge and those that are not. Therefore the measures provide another evaluation dimension and can be used to evaluate changes which we propose in section 6.

6 Model Optimisation / Adaptation

The evaluation from section 5 enables the identification of prototypes that depict shortcuts or ROIs. Using this information it is possible to retrain the classification layer create a model that adheres more closely to domain knowledge. We use two different approaches to achieve this. First, for the models that are trained on shortcuts, we exclude the prototypes that depict shortcuts. Second, for the models that are not trained on shortcuts, we restrict the prototypes to those that are depicting ROIs. By treating the problems separately we can distinguish the effectiveness of the approaches from each other.

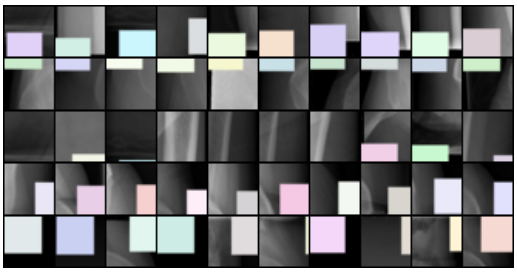
Excluding shortcut prototypes. When the shortcut prototypes are known, the most intuitive next step is to exclude them from the classification layer and continue training for a few epochs with the remaining prototypes [4]. These runs can be seen in Table 4, under the method no-SP$_{wo}$. However, this method leads to an decrease in accuracy (e.g. the accuracy for HIP-CF[7]0% changed from 0.962 to 0.939), especially when a high dependency on the shortcuts exists in the datasets. The decrease in accuracy comes entirely from the decrease in accuracy of the "no fracture" class, which is correlated with the shortcut. As can be seen the average IOU has been decreasing, whereas the average percentage of prototype importance has increased. However, it is important to note, that these results are from less prototypes, and intersecting the candidate lists for both measures creates an empty list of identified shortcut prototypes. Based on this is is safe to assume that the measured overlap and shortcut importance are spurious and are not causally linked to the prototype. A manual examination of the prototype patches confirms this. Therefore the resulting model can be seen as free of (known) shortcuts, but of suboptimal performance.

Reinitialising the classification layer. We propose to reinitialise the classification layer with random values ($X \sim \mathcal{N}(0, 0.1)$) and train the classification layer with all available prototypes from scratch, with a reinitialised scheduler and classifier [5]. Thereby the model has access to other prototypes that may have been overshadowed by the shortcut prototypes before (as PIP-Net is optimised for sparse explanations, less important prototypes are ignored). After every epoch we run the analysis for shortcut prototypes again, to check if any of the prototypes are depicting shortcuts. If some are found, the classification layer weights, classifier and scheduler are reinitialised and the weights of the shortcut prototypes are set to zero. Thus, it can be ensured that no new shortcut prototypes are found by the model which replace the ones already identified. The results from runs with this configuration can be seen in Table 4, under the method no-SP$_{/w}$. The performance of the new models (accuracies 0.947, 0.943 and 0.960 respectively) are equal to the models performance without shortcuts (accuracy 0.946), and do not have the same performance drop as with the no-SP$_{/wo}$ method. Figure 6 shows that there is actually no dependence on shortcuts anymore as the overlap between the average IOU and average shortcut importance distribution vanished after retraining.

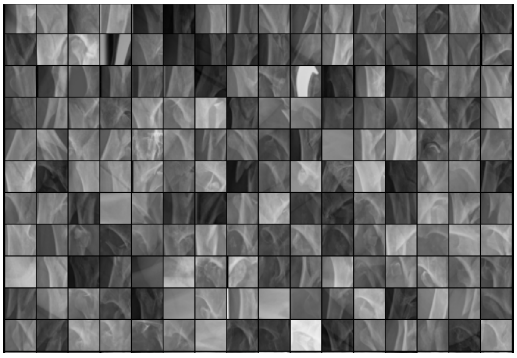
Focusing on ROI prototypes. When ROI prototypes are known the most intuitive next step is to restrict the classification for their relevant class to them. Thereby all scattered prototypes that are not related to the region of interest are ignored. To achieve this the weights of all prototypes for the class "fracture" (the only one which has ROI information) are set to 0. Prototype weights for the class "no fracture" are not touched, as for this class the ROI is not defined, and therefore it is not possible to auto-

[4] 15 epochs, batch size 64, learning rate 0.05

[5] 15 epochs, batch size 64, learning rate 0.05



(a) In one HIP-CF$^{0.5}$ baseline run five prototypes were determined to be shortcut prototypes, the most activated 10 examples can be seen above



(b) In one HIP-TF baseline run, only one prototype was determined to be a ROI prototype. A sample of the corresponding image patches can be seen above

Figure 5: Examples for shortcut and ROI prototypes in trained models identified through the measures described in Section 4

Class	**Dataset**	**# prototypes of class**	**# of ident. prototypes**	$\overline{IOU}$	$\overline{PI}$
fracture	HIP-TF	2.33 (0.57)	1.50 (0.71)	0.53 (0.17)	0.74 (0.23)
	HIP-CF	33.67 (0.58)	1.33 (0.58)	0.78 (0.20)	0.82 (0.05)
no fracture	HIP-CF	54.33 (3.05)	0	-	-
	HIP-CF$^{50\%}$	49.67 (0.58)	4.33 (0.58)	0.49 (0.08)	0.91 (0.00)
	HIP-CF$^{70\%}$	50.00 (3.00)	8.33 (2.52)	0.50 (0.05)	0.93 (0.03)
	HIP-CF$^{100\%}$	36.33 (1.15)	14.33 (1.53)	0.45 (0.02)	0.99 (0.00)

Table 3: Measures introduced in section 5, for the baseline runs shown in table 2. While the runs looked very similar in the baseline runs, the measures give indications about the differing alignment with domain knowledge. Values are average over three runs, the standard deviation is in parentheses.

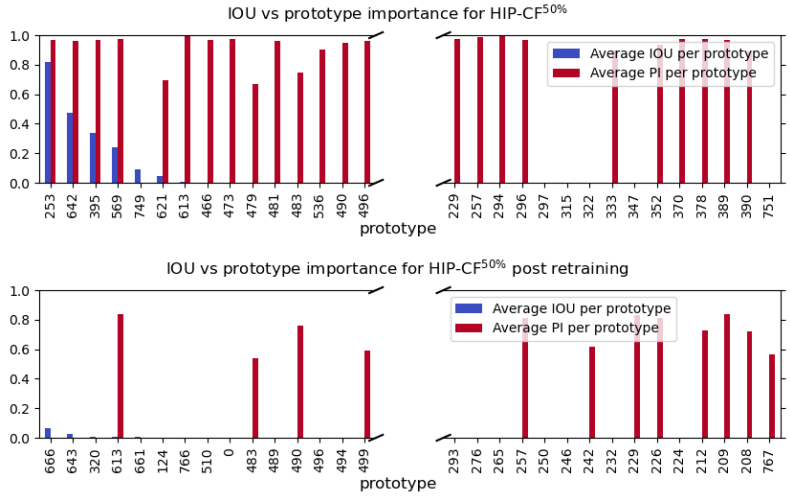


Figure 6: Average Intersection over Union and average prototype importance per prototype for a model trained on HIP-CF$^{50\%}$. The upper image shows the distribution before retraining, the lower after retraining. The overlap between the distributions diminishes. Observations in the middle of the graph are omitted for readability. Graphs for other datasets are in Appendix C

Method	Dataset	Accuracy	acc(no fracture)	acc(fracture)	$\overline{\text{IOU}}$	$\overline{\text{Imp}}$	# ident. Prototypes
no-SP$_{/wo}$	HIP-CF$^{50\%}$	0.943 (0.003)	0.938 (0.006)	0.947 (0.006)	0.124 (0.098)	0.929 (0.015)	0
	HIP-CF$^{70\%}$	0.939 (0.001)	0,925 (0.006)	0,953 (0.008)	0.173 (0.187)	0.932 (0.014)	0
	HIP-CF$^{100\%}$	**0.902** (0.020)	**0.821** (0,042)	0.983 (0.005)	0.094 (0.012)	0.972 (0.026)	0
no-SP$_{/w}$	HIP-CF$^{50\%}$	0.947 (0.000)	0.945 (0.003)	0.948 (0.002)	0.038 (0.010)	0.694 (0.111)	0
	HIP-CF$^{70\%}$	0.943 (0.002)	0.938 (0.005)	0.947 (0.009)	0.065 (0.066)	0.829 (0.102)	0
	HIP-CF$^{100\%}$	0.960 (0.001)	0.958 (0.001)	0.961 (0.003)	0.073 (0.015)	0.789 (0.015)	0
only-DK$_{/wo}$	HIP-CF	**0.888** (0,018)	0.938 (0.017)	**0.838** (0.052)	0.916 (0.131)	0.876 (0.015)	1.333 (0.577)
	HIP-TF	0.942 (0.012)	0.948 (0.028)	0.937 (0.003)	0.673 (0.076)	0.668 (0.145)	2.000 (1.000)

Table 4: Results after prototype removal without re-initialisation (no-SP$_{/wo}$), after prototype removal with re-initialisation (no-SP$_{/w}$), and after relying only on domain-knowledge prototypes without re-initialisation (only-DK$_{/wo}$). Values are average over three runs, the standard deviation is in parentheses.

matically classify a prototype to be aligned with domain knowledge. The results after retraining the classification layer [6] are in Table 4 under the method only-DK*wo*. As the HIP-TF model only had a few prototypes in the "fracture" class, all of which were closely related to the ROI, the performance metrics have not changed after retraining (accuracy changed from 0.944 to 0.942). The HIP-CF model however, which does not have reliable ROI prototypes, has a highly decreased accuracy (accuracy changed from 0.946 to 0.888). Based on these results it becomes clear that this simple solution is not sufficient to provide a model that aligns with domain knowledge based on information about the region of interest.

7 Increasing Prototype Coverage

To achieve a higher model performance we need to ensure a higher coverage of ROIs with prototypes. Ideally the ROI of the image is covered in prototypes, while no prototypes are activated in the rest of the image. To achieve this a new loss is necessary, that takes the coverage of the ROI into account.

Loss proposal. The new coverage loss needs to take the position of the ROI into account, and compare it with the prototype activation maps in the latent features which are an output of the CNN backbone (see figure 1 for reference). The latent feature dimensions of ConvNeXt [29] (the backbone for these experiments) are $(h \times w \times d) = (13 \times 13 \times 768)$, meaning that every image is represented as 768 probability maps (one for each possible prototype) before being pooled and used as input to the classification layer. The ROI maps that have been created for a subset of the datasets (see section 4), have to be converted to a binary mask of size $(h \times w) = (13 \times 13)$, which indicate the position of the ROI. An example can be seen in Figure 7

The probability maps are normalised with softmax such that $\sum_d^D z_{h,w,d} = 1$, creating almost one-hot encoded tensors. This means that for every patch position exactly one prototype is activated already. We assume that the prototypes that are activated in the ROI are already useful and varied, but do not weigh into the classification layer. Then the weights of the classification layer should be used to weigh the activation maps. Using this we propose the loss term L_C as

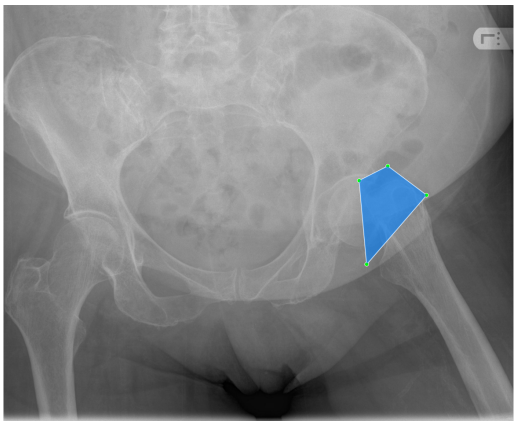
$$L_C = \sum_h^H \sum_w^W |\min(1, \sum_d^D z_{h,w,d} \cdot \omega_d) - mask_{h,w}|$$

where $z_{h,w,d}$ is probability of prototype d at patch h, w in the latent feature representation,

[6] 15 epochs, batch size 64, learning rate 0.05)

$$\begin{matrix}
0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&1&1&0&0\\
0&0&0&0&0&0&0&0&0&1&1&0&0\\
0&0&0&0&0&0&0&0&0&1&1&0&0\\
0&0&0&0&0&0&0&0&0&1&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0&0&0&0&0
\end{matrix}$$

(a) ROI activation map



(b) Origin image of the activation map

Figure 7: One hot encoding the presence of the ROI of an image into a 13×13 matrix.

ω_d is the weight of the classification layer associated with prototype d. Summing up the latent features weighted with the classification weights results in a sparse matrix, which is clamped to have values $[0, 1]$. All absolute deviations from the binary mask are summed to get the L_C loss term. Applying the loss to the HIP-TF and HIP-CF datasets does not have the desired effect, but reduces the relevant prototypes for the "fracture" class to 0.

Discussion of proposed loss. One reason for the loss performance can be the weight of L_C as part of the whole loss term. Fine-tuning the individual loss term weights can change the focus of the model training.

However, we found there is one prototype that is activated in (nearly) all images, which encodes a variety of concepts, but nothing specific for one class (see figure 8). This can be caused by the L_A loss (see Section 3 and [22] for further information). Nauta et al. [22] proposed the L_T loss to prevent this trivial solution. As their paper was developed on datasets multi class datasets (CUB [31] with 200 bird species and Stanford Cars [32] with 196 car types) this trivial solution might be because the datasets in this thesis are binary. As the model additionally optimises for a sparse explanation, there is no incentive to provide a wide variety of prototypes to classify the two classes. As the purpose of L_T is to prevent the trivial solution, giving this term more weight in the Loss term might solve this issue. Additionally a loss term that takes the patch similarities into account could be used to prevent prototypes from encoding multiple concepts. [33] (unpublished) proposed

$$L_A(p) = ||un(p') - un(p'')||_2^2$$

as an alignment loss for PIP-Net. Where p' and p'' indicate representations of two different views of the same image and un is a function normalising p to unit L_2-norm and $||\cdot||_2^2$ is the squared L_2-norm. While this loss was ultimately discarded, a similar one might be able to prevent unspecific prototypes.

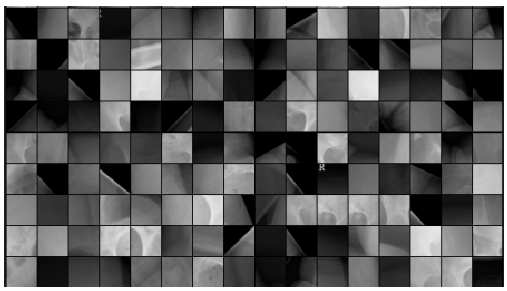


Figure 8: Extract from the unspecific prototype that was activated across (nearly) all images. No common concept can be seen across the instances of the prototype.

With the observation of unspecific prototypes the assumption of useful and varied prototypes in the ROI is violated, which prevents a coverage loss from being a viable improvement of the model. Therefore further research is needed to create a loss that focuses PIP-Net on ROIs.

8 Limitations

The solutions to the research questions in this thesis are dependent on assumptions, limiting their applicability for differing applications. The most important are discussed in this section.

Known prototype locations. The methods discussed in this thesis are highly dependent on the knowing the prototype location of a PIP-Net. However, Xu et.al. [30] have published a sanity check for the prototype location, which shows that the simplified placement of the pro-

totypes might no be accurate. The receptive field of the CNN backbone is quite broad, and all pixels of the image are therefore influencing every prototypes. The methods proposed here are ignoring the receptive field and thus the prototype placement might be faulty. While we have taken great care in also manually checking the prototype visualisation, to make sure that the experimentation results are sound, this issue should be addressed in future research.

Known shortcut and ROI locations. All methods proposed here assume that shortcut and ROI locations are known in the original image, an assumption that is likely violated in real life applications. As the shortcuts are added artificially, the exact locations were known and it was possible to have an exact counterfactual image with no shortcut. In real life the definition of shortcut can be more fuzzy. It could be a background colour, some writing in the image, or other features, such as a patient wearing a diaper. Therefore locating the exact locations of a shortcut in an image might not be feasible or even possible. Second, creating a counterfactual might be very work intensive or impossible. Painting out the background or removing parts of the image requires a high level of domain knowledge and computation power. Therefore the methods of calculating the influence of shortcuts as done in this thesis might not be applicable. However, the experiments rectifying the shortcuts influence are not based on the IOU or shortcut importance measures, but on a list of shortcut prototypes. This means that it is possible to work together with a domain expert, who can point out prototypes that should be ignored. While this assumes expert involvement (which is inherently expensive), it means that the model adaption methods are more widely applicable.

Region of interest is localised. The ROI locations are also made on the assumption that a region of interest is relatively easy to determined and is located on the image in a clustered way and not distributed or fuzzy. While this is possible for a localised trauma like a hip fracture, this might not be true for all kind of applications. Therefore, the methods applied here in creating a counterfactual, with the ROI painted out, might not be feasible in other applications. Also, it is necessary to access domain expert knowledge to determine the ROIs in the first place, which is expensive or might not be accessible. As has been shown it might not be enough in this case to just identify the prototypes that are located on the ROI to focus the model on them. The ROI information is integral to the creation of the loss function as the binary mask is derived from them directly. To limit the costs of the ROI creation the methods proposed here are tested on relatively small samples of labelled images. However, this might also lead to the results being overfitted on the small subsample, especially if it does not capture edge cases in the overall dataset.

The loss function Finally, the loss function proposed in section 6 has not been sufficiently tested or implemented yet, but should be seen as an inspiration for future research. The assumption that a higher coverage of the regions of interest is beneficial is yet to be proven, but seems intuitive. However, as discussed Section 7, due to unspecific prototypes it might not be enough to implement a coverage loss, but also other parts of PIP-Net need to be adjusted to make it a feasible solution.

9 Summary

We presented two measures which in combination can evaluate a PIP-Net based on its alignment with domain knowledge. The measures are Intersection over Union, an overlap calculation between a PIP-Net prototype and an interested region, and prototype importance, by looking at the difference in classification weights when comparing counterfactual images. Combining these two measures results in a prototype list, which depict an interested region. The measures provide a way to discriminate between different PIP-Net models with similar performance metrics based on their domain knowledge alignment.

Based on these performance metrics we further proposed methods to use the gained information to improve the models adherence to domain knowledge. For the datasets used in this thesis is sufficient to reinitialise the classification layer and set known shortcut prototype weights to zero, until no shortcut prototypes are used in the classification anymore. With this method the performance of the original (shortcut less) model performance could be reached. For the adherence to a region of interest the reinitialisation of the classification layer has not been shown to be sufficient to recreate the performance of the original model. Therefore we proposed a new loss term that could provide a higher number of prototypes that refer to the ROI. However, we found that due to the presence of unspecified prototypes more changes to PIP-Net need to be conducted in order to make this approach feasible. More research in this area is needed.

References

[1] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Medicine*, vol. 15, p. e1002683, Nov. 2018.

[2] B. Allen, S. Agarwal, L. Coombs, C. Wald, and K. Dreyer, "2020 ACR Data Science Institute Artificial Intelligence Survey," *Journal of the American College of Radiology*, vol. 18, pp. 1153–1159, Aug. 2021. Publisher: Elsevier B.V.

[3] C. D. Becker, E. Kotter, L. Fournier, and L. Martí-Bonmatí, "Current practical experience with artificial intelligence in clinical radiology: a survey of the European Society of Radiology," *Insights into Imaging*, vol. 13, Dec. 2022. Publisher: Springer Science and Business Media Deutschland GmbH.

[4] C. Destruction, "Geoff Hinton: On Radiology," Nov. 2016.

[5] Karla Adam, "The U.K. used an algorithm to estimate exam results. The calculations favored elites.," *Washington Post*, 2020.

[6] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statistics Surveys*, vol. 16, Jan. 2022.

[7] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, Feb. 2019.

[8] B. B. Babic, S. Gerke, T. Evgeniou, and I. Glenn Cohen, "Beware explanations from AI in health care the benefits of explainable artificial intelligence are not what they appear," *Science*, vol. 373, pp. 284–286, July 2021. Publisher: American Association for the Advancement of Science.

[9] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity Checks for Saliency Maps," in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, Curran Associates, Inc., 2018.

[10] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This Looks Like That: Deep Learning for Interpretable Image Recognition," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[11] D. Rymarczyk, Struski, M. Górszczak, K. Lewandowska, J. Tabor, and B. Zieliński, "Interpretable Image Classification with Differentiable Prototypes Assignment," Sept. 2022. arXiv:2112.02902 [cs].

[12] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, p. 1096, Mar. 2019.

[13] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, Aug. 2018. arXiv: 1802.01933 Publisher: Association for Computing Machinery.

[14] P. S Corbet, "The dragonflies of british columbia, by robert a. cannings & kathleen m. stuart," *Notulae odonatologicae*, vol. 1, no. 2, pp. 32–34, 1978.

[15] A. A. Freitas, "Comprehensible classification models: a position paper," *ACM SIGKDD Explorations Newsletter*, vol. 15, pp. 1–10, Mar. 2014.

[16] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens, "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models," *Decision Support Systems*, vol. 51, pp. 141–154, Apr. 2011.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin, ""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*, (San Francisco California USA), pp. 1135–1144, ACM, Aug. 2016.

[18] M. Nauta, A. Jutte, J. C. Provoost, and C. Seifert, "This Looks Like That, Because ... Explaining Prototypes for Interpretable Image Recognition," 2020.

[19] J. Wang, H. Liu, X. Wang, and L. Jing, "Interpretable Image Recognition by Constructing Transparent Embedding Space," *2021 IEEE/CVF International Conference*

on Computer Vision (ICCV), vol. null, pp. 875–884, 2021.

[20] D. Rymarczyk, Struski, J. Tabor, and B. Zieliński, "ProtoPShare: Prototypical Parts Sharing for Similarity Discovery in Interpretable Image Classification," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, (Virtual Event Singapore), pp. 1420–1430, ACM, Aug. 2021.

[21] M. Nauta, R. v. Bree, and C. Seifert, "Neural Prototype Trees for Interpretable Fine-grained Image Recognition," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. null, pp. 14928–14938, 2020.

[22] Meike Nauta, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert, "PIP-Net: Patch-Based Intuitive Prototypes for Interpretable Image Classification," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog*, pp. 2744–2753, June 2023.

[23] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, pp. 665–673, Nov. 2020.

[24] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias There's software used across the country to predict future criminals. And it's biased against blacks.," *Pro Publica*, May 2016.

[25] M. Nauta, *Explainable AI and interpretable computer vision : from oversight to insight*. PhD, University of Twente, Enschede, The Netherlands, May 2023. ISBN: 9789036555753.

[26] J. Borowski, C. M. Funke, K. Stosio, W. Brendel, T. S. A. Wallis, and M. Bethge, "The Notorious Difficulty of Comparing Human and Machine Perception," in *2019 Conference on Cognitive Computational Neuroscience*, (Berlin, Germany), Cognitive Computational Neuroscience, 2019.

[27] T. Makino, S. Jastrzkebski, W. Oleszkiewicz, C. Chacko, R. Ehrenpreis, N. Samreen, C. Chhor, E. Kim, J. Lee, K. Pysarenko, B. Reig, H. Toth, D. Awal, L. Du, A. Kim, J. Park, D. K. Sodickson, L. Heacock, L. Moy, K. Cho, and K. J. Geras, "Differences between human and machine perception in medical diagnosis," *Sci. Rep.*, vol. 12, p. 6877, Apr. 2022.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[29] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

[30] R. Xu-Darme, G. Quénot, Z. Chihani, and M.-C. Rousset, "Sanity checks and improvements for patch visualisation in prototype-based image classification," Jan. 2023. arXiv:2302.08508 [cs].

[31] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona, "Caltech-ucsd birds 200," Tech. Rep. CNS-TR-201, Caltech, 2010.

[32] T. Kramberger and B. Potočnik, "Lsun-stanford car dataset: enhancing large-scale car image datasets using deep learning for usage in gan training," *Applied Sciences*, vol. 10, no. 14, p. 4913, 2020.

[33] "PIP-Net: Patch-Based Intuitive Prototypes for Interpretable Image Classification," tech. rep.

A Non-technical summary

When we are deploying machine learning models into a medical context it is important to increase trust in the technology so it can actually be used in the daily practice. With traditional methods like deep learning the output of the model comes out of a black box and cannot be explained or interpreted. PIPNet, the basis of this thesis, is a model that combines a black box with an interpretable model. It creates so called "prototypes" (recurring features in images) and finds them in the images, at locations called patches. Each prototype is associated with a class and weighs into the classification decision the model has to make. When looking at the output the practitioner can see on which patches the model has based its decision. For example when looking at a hip fracture, we would like to see that all the patches are clustered around the area of the fracture, and not seemingly randomly strewn around the image.

This thesis aims to improve the model based on the location of the patches. It assumes two scenarios: 1. We DO NOT want the model to look at so called "shortcuts". In real applications this could be some writing in the background, a patient wearing another medical device, etc. In the context of this thesis we added artificial shortcuts to the image, to be able to control their influences better. 2. We DO want the model to look at the Region of Interest (ROI). In the case of hip fractures this is the fractured bone. For this thesis we have marked the area accordingly.

To achieve this goal the thesis introduces two measures, Intersection over Union (IoU) and Prototype importance. IoU measures how much patch is overlapping with an area (either ROI or Shortcut) and Prototype importance measures, how much the influence of the found prototype changes when we cover up the area. Combining these two measures helps us to identify whether a model is adhering to domain knowledge and it shows us which prototypes are depicting shortcuts or ROIs. With this knowledge we can then improve the model. For example by removing shortcut prototypes, or by restricting it to ROI prototypes. In the thesis we showed that it is possible to retrain the model without the shortcuts and to receive the same accuracy as before. However, we also showed that it is not always possible to restrict the model on only the regions of interest. Therefore we propose a new loss (a new priority in the model training process) that takes ROI coverage into account. The reasoning behind this is that we want more prototypes that are depicting the ROI than we already have, as this could make the classification more confident. However, because of technical reasons and time restrictions the new loss is not sufficiently tested and not currently improving the model performance.

For a practitioner these results mean, that is is possible to react to feedback and to improve the model based on domain knowledge. If a radiologist is shown the model results, they can indicate which parts of the classification make sense to them and which do not. They could also indicate where the model should have looked instead. After collecting this feedback, the model can be retrained to fit more closely to the specification. Over time, we can then hopefully get a model that is trustworthy and that can be used in the daily work flow.

B Region of Interest - Annotation

For the purpose of this thesis we annotated 2114 images with trochantic fractures, and 200 images with column fractures. The annotations were done with the labelme software version 5.3.0a0, from the labelme github. Labelme only takes standard image compressions as input, therefore the images were converted from dicom to png or jpg. The annotations take the form of polygons, which surround the fractured area of the image. Each image has one and only one region of interest, that is marked with the tag "fracture". The points of the ROI are stored in json files with the same name as the original image. To secure the accuracy of the the annotations of the trochanteric fracture dataset two approaches were used to evaluate them: 1. Another researcher (Jeroen Geerdink) annotated a subset of images (256). The overlap of the polygons for images in both datasets were compared. On average 87% of my annotations were overlapping with Jeroens and Jeroens polygons were overlapping 69% with mine, which is in accordance with the difference in our annotation methods. Jeroen tends to make the annotation area bigger. For 2% of the observations there was a disagreement in the placement of the ROI. 2. 10% of the dataset were also sent to Han Hegeman (trauma surgeon) to evaluate from a medical perspective. He rejected about 5% of the annotations. The reasons for rejections also included incomplete annotations, mostly ROIs are missing lower placed fraction areas.

Overall the annotations are of reasonably high quality.

As the model of the trochanteric fractures is not exhibiting a high spread of prototypes, the images unfortunately could not be used for model adjustments as originally planned.

Therefore I also annoated 200 images from the column fracture dataset. For time reasons the same quality controls were not undertaken as for the first dataset. However, as in this dataset

the fractures are more varied, and harder to spot it is likely that these dataset annotations are not of the same quality as the other.

C Further Graphs

Graphs depicting the effect of retraining with linear layer reinitialisation based on the datasets HIP-CF$^{70\%}$ and HIP-CF$^{100\%}$ are in figure 9

Figure C screenshot showing an example of a prototype (prototype 653) being activated across most of the image. It is exemplary for other images, which have been omitted for space reasons. The screenshots hints that the alignment loss leads to a trivial prototype attibution for many patches in the image.

D Shortcut unlearning

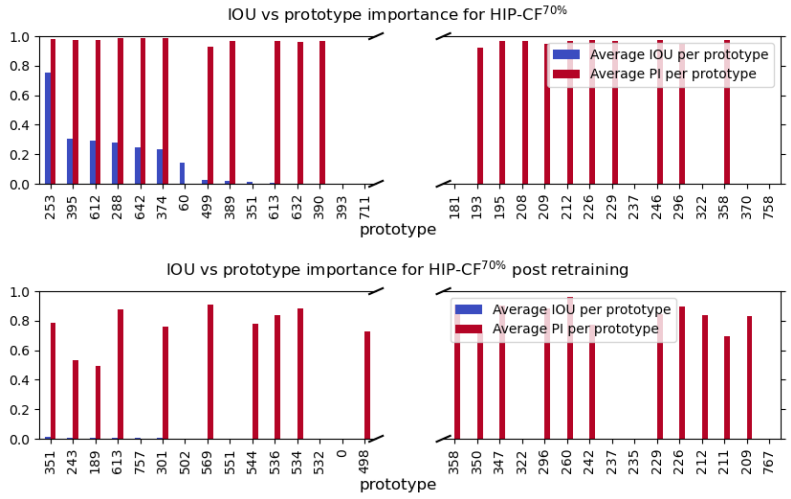
A (failed) attempt to unlearn shortcut was to use the same technique that is currently used to teach the model to ignore the transformations applied to the input images on the subject of shortcuts as well. To do this, we inserted a second pretraining step after the first one, but changing the input from one "normal" image and one transformed image to one "normal" image and one "shortcut" image instead. The alignment loss forces the model to create the same prototype distributions across both images, meaning it is encourage to ignore the shortcuts. A proposed architecture can be seen in Figure 13. Comparing the visualized prototypes before and after the shortcut unlearning step gave some encouraging results. Some "shortcut prototypes" seem to have become "shortcut agnostic", displaying areas with and without shortcut in the same prototype. However, as the method is assuming access to a whole dataset with shortcuts and without is quite unrealistic (unless the shortcuts are added artificially like in this case), there is a very limited practical value to the method. Therefore the attempt was abandoned after a few weeks in favour of more realistic solutions.

E Practical Application as a Service in ZGT

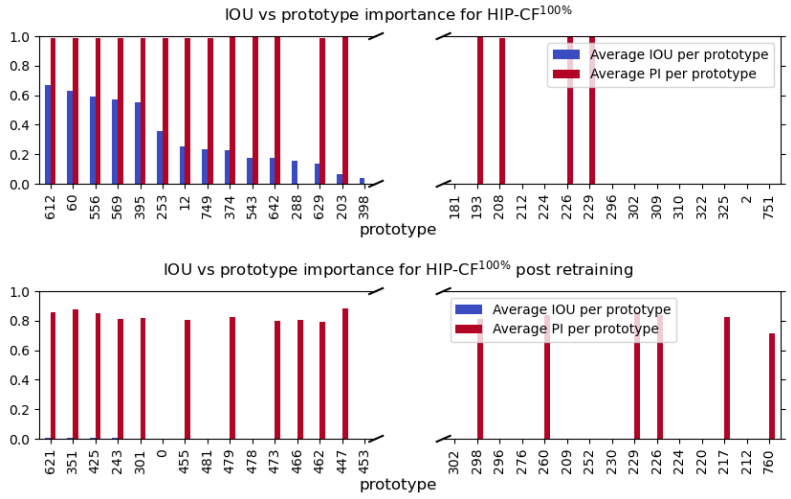
The PIP-Net Model (trained on the HIP-TF dataset) will be deployed as a service with the support Pukka-J, who have been sponsoring this thesis. MSc. Quang-Hung Nguyen, is creating a backend which is able to deploy the model and store the results. He is also working on a front end, which the radiologist can use to see the output from PIP-Net. The current development state of the application can be seen in figure E. The aim of the application is to get feedback on the prediction quality. As this thesis showed, it is possible to amend PIP-Net so it conforms closer to domain knowledge. It is unfeasible to label a lot of images next to the normal hospital workflow. However, the visual application makes it possible to quickly give feedback on the quality of the prediction. Currently this consist of a RADPEER score (developed by the American College of Radiology to standardise the peer review process), and a free input box for input.

In a future version of the application it should be possible to select which prototypes are helpful, and which are not. The prototypes which are not helpful can then be ignored in the classification layer, like shown in this thesis. Through knowing which prototypes ARE helpful could be possible to deduce regions of interest, and focus the prediction more on these areas.

Hopefully the deployment of PIP-Net in this application will lead to an improvement of the model and increased trust in the AI's capabilities.



(a) Comparison of IOU and PI values for a model trained on HIP-CF^{7}0%



(b) Comparison of IOU and PI values for a model trained on HIP-CF^{1}00%

Figure 9: Average Intersection over Union and average prototype importance per prototype for models trained on HIP-CF^{7}0% and HIP-CF^{1}00%. The upper image shows the distribution before retraining, the lower after retraining. The overlap between the distributions diminishes.

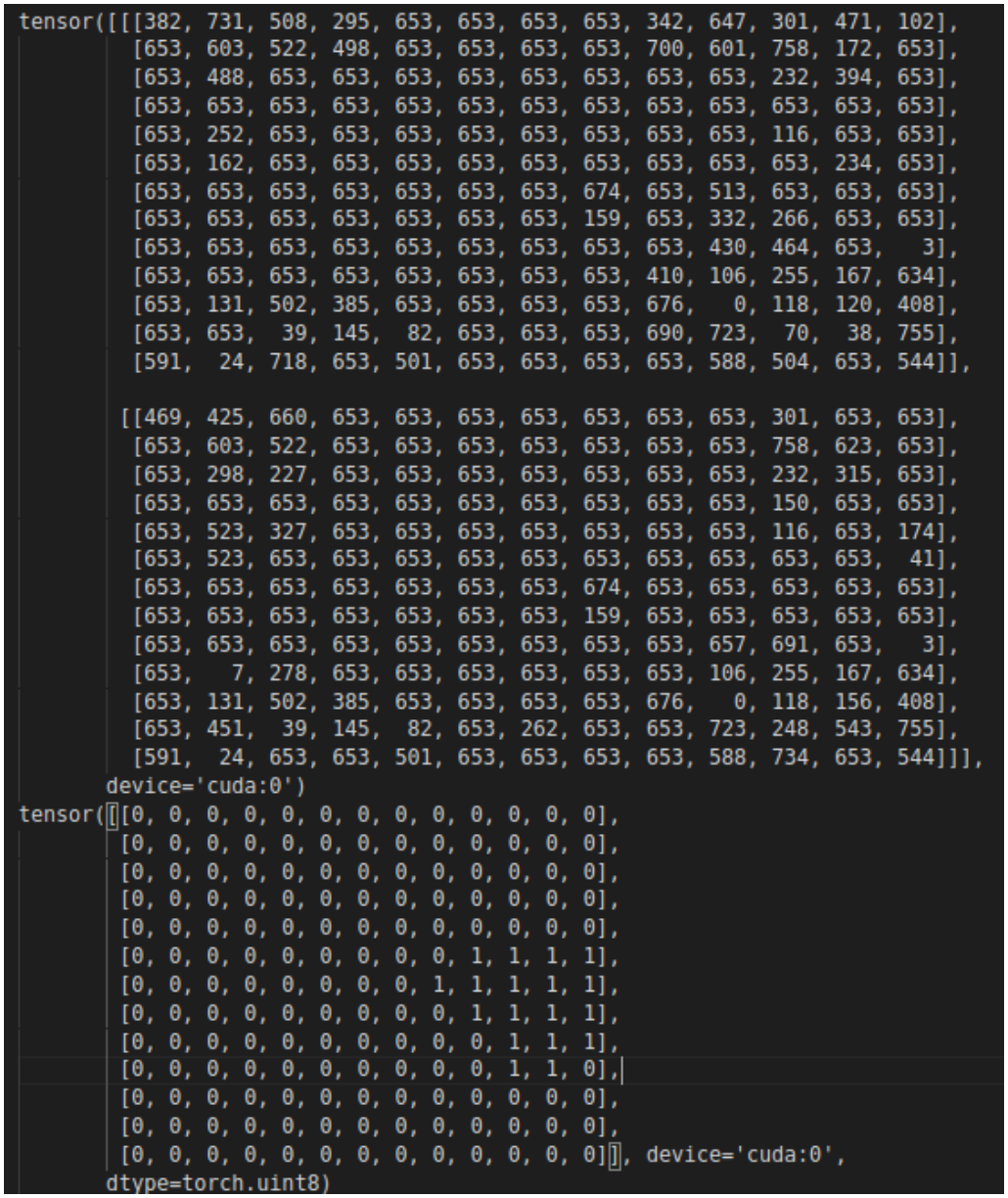


Figure 10: Screenshot of most activated prototypes in an image

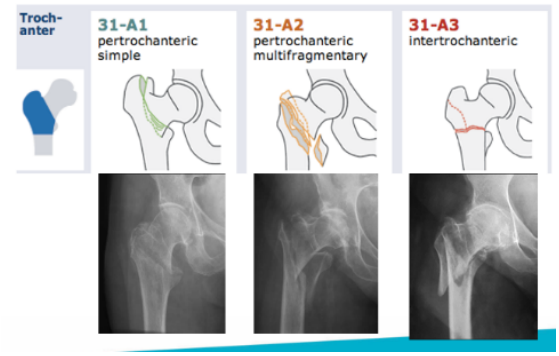


Figure 11: Examples of pertronchanteric fractures

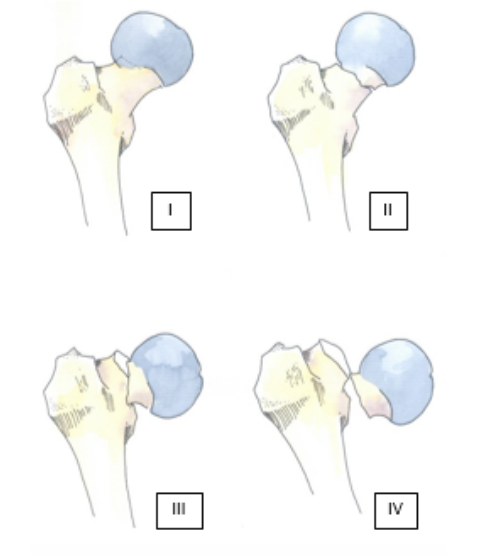


Figure 12: Examples of column fractures

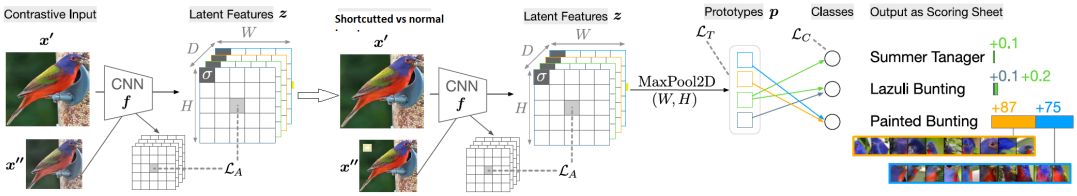


Figure 13: Architecture of the shortcut unlearning idea

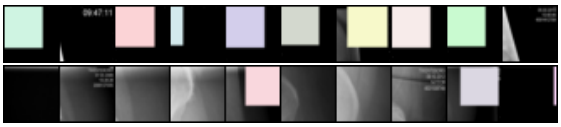
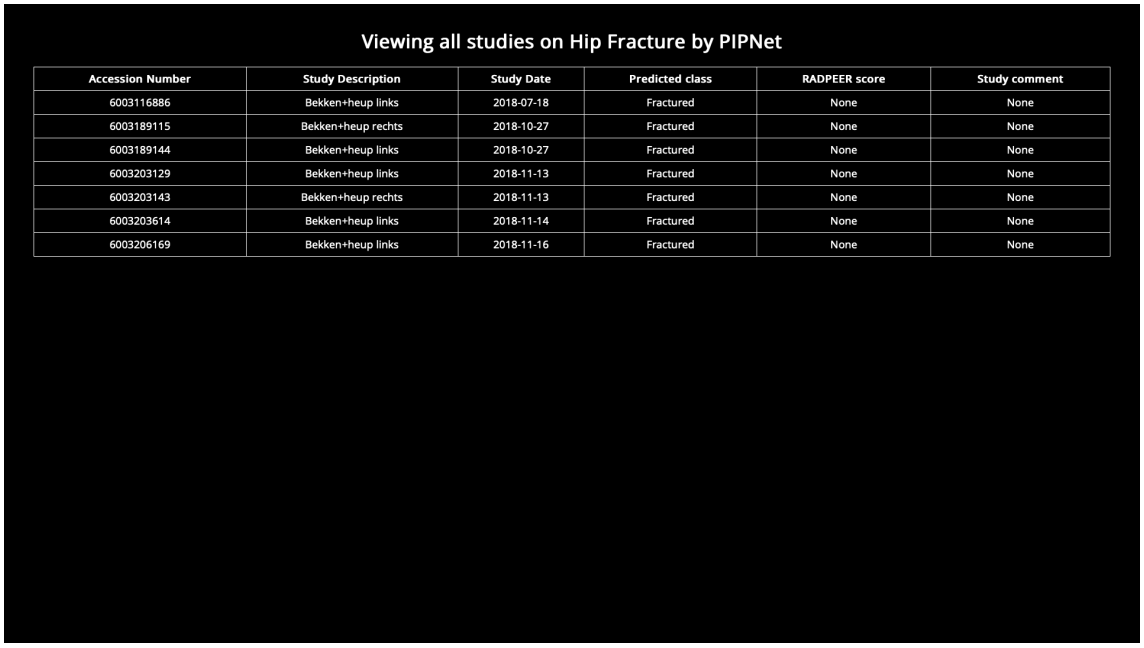


Figure 14: Prototype comparison before and after a shortcut unlearning step, it can be seen that the prototype that was only depicting shortcut before now depict both other image features and shortcuts at the same time



(a) Overview of the images and their predicted labels

(b) A detailed view of the X-ray and the found prototypes, the radiologist can give a RADPEER score and comment on the prediction quality.

Figure 15: Screenshots from the ZGT application front-end of the PIP-Net deployment