MICCAI

This actually looks like that: Proto-BagNets for local and global interpretability-by-design

Kerol Djoumessi[1,2(✉)][0009−0004−1548−9758], Bubacarr Bah[3], Laura Kühlewein[4], Philipp Berens[1,2(✉)][0000−0002−0199−4727], and Lisa Koch[1,2,5(✉)][0000−0003−4377−7074]

[1] Hertie Institute for AI in Brain Health, University of Tübingen, Germany
{kerol.djoumessi-donteu, philipp.berens}@uni-tuebingen.de
[2] Tübingen AI Center, University of Tübingen, Germany
[3] Medical Research Council Unit The Gambia at London School of Hygiene and Tropical Medicine
[4] University Eye Clinic, University of Tübingen, Germany
[5] Department of Diabetes, Endocrinology, Nutritional Medicine and Metabolism UDEM, Inselspital, Bern University Hospital, University of Bern, Switzerland
{lisa.koch}@unibe.ch

Abstract. Interpretability is a key requirement for the use of machine learning models in high-stakes applications, including medical diagnosis. Explaining black-box models mostly relies on post-hoc methods that do not faithfully reflect the model's behavior. As a remedy, prototype-based networks have been proposed, but their interpretability is limited as they have been shown to provide coarse, unreliable, and imprecise explanations. In this work, we introduce Proto-BagNets[6], an interpretable-by-design prototype-based model that combines the advantages of bag-of-local feature models and prototype learning to provide meaningful, coherent, and relevant prototypical parts needed for accurate and interpretable image classification tasks. We evaluated the Proto-BagNet for drusen detection on publicly available retinal OCT data. The Proto-BagNet performed comparably to the state-of-the-art interpretable and non-interpretable models while providing faithful, accurate, and clinically meaningful local and global explanations.

Keywords: Interpretability-by-design · Optical Coherence Tomography · Part-prototype networks

1 Introduction

For adopting deep learning models in safety-critical applications such as medical diagnosis, it is crucial that users can understand why a model produced a specific output [16]. This form of interpretability is usually obtained either through post-hoc explanations of black-box models [20] or through architectural design [4,13]. Post-hoc methods [17,20] interpret an approximation of the

---

[6] Code available at https://github.com/kdjoumessi/Proto-BagNets

true decision mechanism [11] through saliency maps. These highlight the most discriminating regions in the input, but often provide inaccurate and unfaithful explanations [1, 16]. To remedy this, several approaches have been proposed with structurally built-in interpretability, such as bag-of-local-features models (BagNets) [3], concept-based models [15], and prototype-based models [4].

The BagNet [3] is an implicitly patch-based interpretable-by-design model with a small receptive field, where predictions solely rely on local evidence. Its recent modification [13] provides sparse and fine-grained local class activation maps, but does not allow humans to gain a global understanding of the model's decision. Concept-based models [15] follow a case-based reasoning process where high-level representations of the data (concepts) are learned and used to classify new images. Prototype-based networks [4] can be seen as a special case of concept-based models, in which learned concepts are replaced by the representative training image parts (prototypes) to improve interpretability. In these models, similarities to the learned prototypes are used to classify new examples. Explanations can be obtained during inference by highlighting, for a query image, its prototypical parts most similar to each learned prototype, thus providing both local explanations thanks to the similarity map and global explanations through the visualization of the learned prototypes. ProtoPNet [4], the first prototype-based network, has gained considerable attention due to its easy-to-understand architecture and high-level reasoning process close to that of humans in solving complex tasks. Although numerous variants have been proposed to improve its performance and interpretability [2, 7, 10, 14], applications in medical imaging remain relatively limited. This may in part be due to the fact that the interpretability of prototype-based models is more limited than appears at first glance, as it has been shown that they do not actually provide faithful explanations [7, 19].

We propose Proto-BagNet, an interpretable-by-design prototype-based model that combines the local and fine-grained interpretability of BagNet with the global interpretability of prototype learning. We integrated recent advances in training prototype-based models and proposed an additional prototype diversity constraint. We evaluated our model for detecting drusen lesions on Optical Coherence Tomography (OCT) images and showed that the Proto-BagNet preserves high predictive performance while providing faithful, clinically meaningful, and precise explanations. Our model explanations accurately localized drusen both in the learned prototypes and query test images.

2   Developing a faithful prototype-based network

2.1   Baseline ProtoPNet model

We built on the ProtoPNet [4] as a baseline, which consists of three main components: a backbone feature extractor f, a prototype layer g_p, and a classification layer h. Given an input image $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ (with height H, width W, and number of channels C), the backbone first extracts a meaningful feature representation $\mathbf{Z} = f(\mathbf{X}) \in \mathbb{R}^{M \times N \times D}$. The prototype layer $g_{\mathbf{p}}$ consists of $b = m \times c$
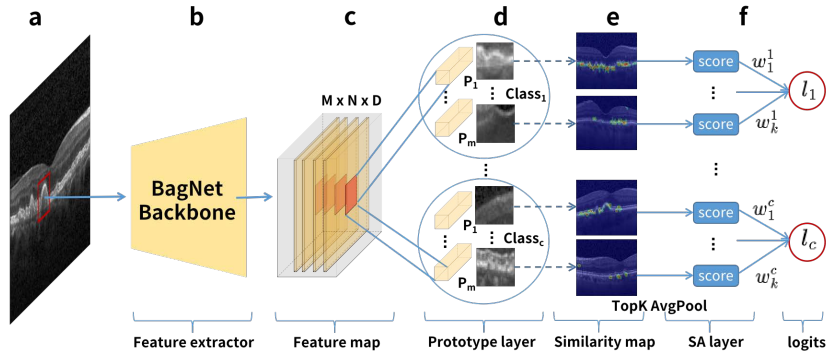
Fig. 1. Architecture of the Proto-BagNet. **(a)** Example OCT B-scan image. The red patch illustrates the small receptive field of **(b)** the BagNet backbone. **(c)** Feature map and **(d)** Prototype layer with m prototypes per class. **(e)** Resulting similarity maps from each prototype to the input. **(f)** The soft aggregation layer aggregates the average top-k scores from each similarity map into their allocated categories for classification.

learnable prototypes $\mathbf{P} = \{\mathbf{p}_j^i \in \mathbb{R}^{H_P \times W_P \times D}\}_{j=1}^m$ (typically $H_P = W_P = 1$) where m denotes the number of prototypes per class, c the number of classes, and $\mathbf{p}_j^i$ the j-th prototype of class i. For each prototype $\mathbf{p}_j^i$, the prototype layer computes a similarity map $M_{\mathbf{P}_j^i}^{\mathbf{x}} = \text{Sim}(\mathbf{Z}, \mathbf{p}_j^i) \in \mathbb{R}^{M \times N}$. The similarity score between a prototype $\mathbf{p}_j^i$ and the feature vector $\mathbf{z}^{(h,w)} \in \mathbf{Z}$ is defined as $s_{i,j}^{(h,w)} = \log\left((d_{i,j}^{h,w} + 1)/(d_{i,j}^{h,w} + \epsilon)\right)$, where $d_{i,j}^{h,w} = \|\mathbf{p}_j^i - \mathbf{z}^{(h,w)}\|_2^2$. The similarity maps contain positive scores indicating where and to what extent prototypes are present in an image. ProtoPNet uses the highest value of the similarity map $g_{\mathbf{p}_j^i} = \max(M_{\mathbf{p}_j^i}^{\mathbf{x}})$ as the final similarity score between $\mathbf{p}_j^i$ and $\mathbf{X}$, indicating how strong the prototype $\mathbf{p}_j^i$ is present in $\mathbf{X}$. Finally, the b similarity scores from the prototype layer g_p are aggregated in the fully connected layer h to generate the final classification logits. To make the prototypes visualizable as specific proto-typical parts of a sample, the learned prototypes are replaced with the closest feature representation from real training images to ensure interpretability.

ProtoPNet explains its predictions for a given image ("local explanation") by (1) visualizing the similarity map for each prototype $\mathbf{p}_j^i$ and (2) by computing the smallest bounding box enclosing the $95th$ percentile of all similarity values [4], providing the corresponding cropped region as the most similar part to the learned prototype to imply *'this part of the input looks like that learned proto-type'*. The same approach is used to provide explanations of the concepts learned by the model ("global explanations") by cropping the prototypes from the most similar training image. However, ProtoPNet provides only coarse-grained similarity maps due to the large receptive field size of the model [7]. Furthermore, the explanation is not faithful to the model, as the cropped area does not correspond to the model's receptive field. As a result, ProtoPNet provides both imprecise

local and global explanations of its decisions. These issues are likely to be shared by all prototype-based models derived from ProtoPNet [19].

2.2   Enhancing interpretability with the BagNet backbone

As prototypes are learned in the feature space, the backbone feature extractor plays a crucial role for interpretability. It implicitly determines the size of the learned prototypes through its receptive field and thus the size of the explanation. The ProtoPNet and its variants use classical architectures such as ResNet-50 [4, 10, 14], resulting in large receptive fields (e.g. 427×427 for a ResNet-50 backbone) with variable explanation size. Here, we propose to replace the feature extractor with a BagNet architecture [3, 13] (Fig. 1b), leading to a model we called Proto-BagNet (Fig. 1). The feature map $\mathbf{Z} = f(\mathbf{X})$ is extracted from the BagNet's penultimate layer (Fig. 1c, typically $D = 2048$). This architecture leads to a small fixed receptive field and prototypes of size $r \times r$ independent of the input size. It also allows for higher-resolution feature and prototype similarity maps, and can therefore provide localized, fine-grained explanations.

2.3   Integrating recent advances in training prototype-based models

In addition, we implemented recent advances in prototype-based networks training [2, 10, 14]. (1) To prevent prototypes of one class from contributing to the prediction of other classes, we replaced the fully connected classification layer of ProtoPNet with a soft aggregation (SA) layer (Fig. 1f) [10], which aggregates the prototypes' similarity scores only in their assigned classes, setting weights between classes to zero. (2) To enable the model to consider multiple image regions for the classification task instead of considering only the region with the highest score of each similarity map as in ProtoPNet, we considered the top-k scores through average pooling as $g_{\mathbf{p}^{j^i}} = \mathrm{AvgPool}\big(\mathrm{topk}(M_{\mathbf{p}_j^-})\big)$ [2]. Thus, our similarity map indicates to what extent a prototype is present on average in the k most similar prototypical parts of the input. (3) We regularized the prototype layer by adding a sparsity constraint to each similarity map as in [13,14], to constrain activation to discriminative input regions. (4) Finally, as we noticed redundant prototypes (often extracted from the same training image, see Suppl. Fig. 1), we introduced a dissimilarity loss (see below) to prevent the network from learning duplicate prototypes while promoting their coherence and uniqueness. Thus, the total loss function was:

$$\mathcal{L} = \mathcal{L}_{ce} + \underbrace{\lambda_{clst}\mathcal{L}_{clst} + \lambda_{sep}\mathcal{L}_{sep}}_{\text{ProtoPNet}} + \lambda_{L1,c}\mathcal{L}_{L1,c} + \lambda_{L1,s}\mathcal{L}_{L1,s} - \lambda_{diss}\sum_{p_i,p_j}\|\mathbf{p}_i - \mathbf{p}_j\|^2$$

Here, $\mathcal{L}_{\mathrm{ce}}$ is the cross-entropy loss; $\mathcal{L}_{clst}$ and $\mathcal{L}_{sep}$ the cluster and separation losses from ProtoPNet [4], $\mathcal{L}_{L1,c}$ is the ℓ_1 regularization of the classification layer as in [4]; $\mathcal{L}_{L1,s}$ regularizes the similarity maps [13,14]. Finally, $\sum_{p_i,p_j}\|\mathbf{p}_i - \mathbf{p}_j\|^2$ is our proposed dissimilarity loss with $\mathbf{p}_i, \mathbf{p}_j \in \mathbf{P}, i \neq j$.

Our architectural changes alongside with the modified loss function led to an interpretable-by-design prototype-based model (Proto-BagNet) that is easier to interpret, relying on the small receptive field of the model for predictions and explanations. In addition, Proto-BagNet provides accurate local and global explanations (see Sec. 3.3, and 3.4) in the form of *'this part of the input actually looks like that learned prototype'*.

3   Results

3.1   Dataset

We used a publicly available, anonymized dataset [12] consisting of retinal OCT B-scans from patients with various diseases (drusen, DME, CNV). We focused on the binary task of drusen detection and filtered out images with DME and CNV diagnoses, as well as low-resolution images (width < 496). To counter the class imbalance, we removed half of the healthy images, leading to a dataset of $34,962$ images ($8,616$ drusen, $26,346$ healthy). We split the resulting dataset into training (80%) and validation (20%) sets, preserving the imbalance proportion (73% vs 27%) and ensuring that all B-scans from each patient were assigned to the same set. We then used the separate test set included in the dataset for evaluation, consisting of 250 healthy and 248 drusen images (51% vs 49%), reflecting the high variability of drusen prevalence according to age group [18]. All images were resized to 496×496 and normalized by the mean and standard deviation of the training set. To evaluate the relevance of the learned prototypes, an experienced in-house ophthalmologist provided detailed drusen annotations on a selection of 40 test images.

3.2   Proto-BagNet yields good accuracy on drusen detection

We first evaluated the classification performance of our method for a clinically relevant binary task of detecting patient's OCT-B scans with drusen lesions (lipid deposits under the retina [6]), characteristic of age-related macular degeneration and diabetic retinopathy [6,12]. For Proto-BagNet, we configured the backbone feature extractor (BagNet model) to a receptive field size $r = 33$ as in [13]. Hyperparameters including regularization coefficients, data augmentation, and the number of prototypes were optimized on the validation dataset using a grid search, while λ_{sep} and λ_{clst} were set as in [4]. Based on the validation performance, we set $k = 5$ considering the average top-5 and used $m = 5$ prototypes per class, which lead to a total of $b = 10$ prototypes.

We compared Proto-BagNet against ProtoPNet with a ResNet-50 backbone [4] and with non-prototype classification networks such as a dense BagNet [13] and ResNet-50 [9]. We followed the same training procedure for ProtoPNet (with $\lambda_{L1,s} = \lambda_{diss} = 0$, $K=1$) and Proto-BagNet (with $\lambda_{L1,s} = 4 \cdot 10^{-2}$, $\lambda_{diss} = 5 \cdot 10^{-3}$, $K=5$), as well as for dense BagNet and ResNet-50. Our Proto-BagNet performed comparably to the state-of-the-art models (Tab. 1, see confidence intervals in Suppl. Tab. 1), showing that our modifications towards better interpretability did not substantially impair classification performance.
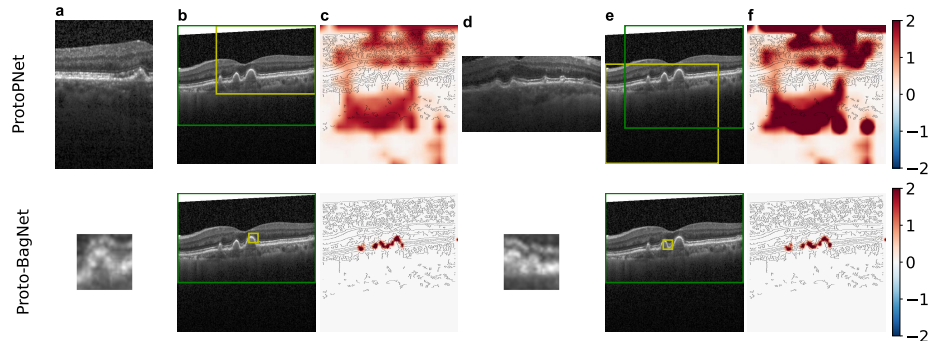
Fig. 2. Example explanations of ProtoPNet and Proto-BagNet. **(a,d)** show two learned prototypes with the highest classification weights. Proto-BagNet's prototypes were magnified for visualization only. **(b,e)** show bounding boxes around regions of highest activation using the visualization technique provided by ProtoPNet (green) and the model's receptive field (yellow). **(c,f)** show prototypes activations of the query image.

3.3   Proto-BagNet provides understandable localized explanations

To qualitatively assess explanations provided by our model, we visualized the two learned prototypes with the highest classification weights for our Proto-BagNet as well as ProtoPNet (Fig. 2a,d). ProtoPNet learned very large prototypes covering almost the entire retina, while Proto-BagNet learned prototypes of small regions of interest with fixed sizes corresponding to its receptive field. For a query image, we then displayed bounding boxes (Fig. 2b,e) around the most similar prototypical part to the learned prototypes. For both models, we first computed the explanation of the prototypical part as the receptive field around the location of the highest prototype similarity (yellow boxes). Due to the large receptive field in ProtoPNet, the explanations were not informative, while the Proto-BagNet yielded small localized patches of the same size as the prototypes themselves. We then computed the explanation as the bounding box around the 95th percentile of the similarity map as in [4], which is not faithful to the model's predictions, as it usually leads to large and similar explanations (green bounding boxes, Fig. 2b,e). In both models, this again led to large bounding boxes around the entire retina, indicating that such prototype explanations may

Table 1. Classification performance for drusen detection on validation and test sets.

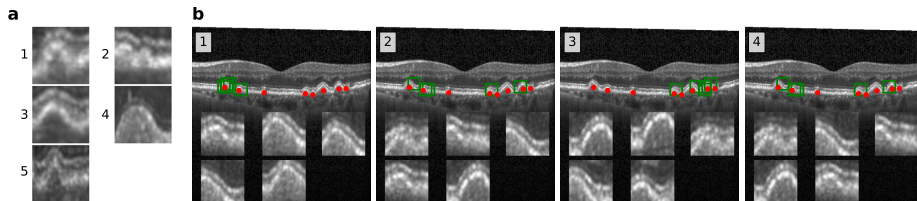|	Validation set				Test set			
	Accuracy	AUC	Recall	Precision	Accuracy	AUC	Recall	Precision
ResNet-50	0.991	0.999	0.982	0.986	0.994	0.999	0.992	0.996
dense BagNet	0.990	0.999	0.978	0.985	0.988	0.999	0.976	0.999
ProtoPNet	0.987	0.996	0.975	0.974	0.998	0.999	0.996	0.999
Proto-BagNet	0.978	0.990	0.935	0.981	0.968	0.992	0.940	0.996

Fig. 3. We show **(a)** the five learned disease prototypes and **(b)** suspicious regions (green boxes, enlarged below) extracted from each prototype similarity map on an example image. Drusen (annotated with red markers) are detected with high precision.

not be useful when evidence may be spread in an image (see similarity maps in Fig. 2c,f), as is often the case in medical images. To summarize, the coarse explanations provided by ProtoPNet were not informative as already highlighted in [7, 19], while the small receptive field of Proto-BagNet leads to fine-grained and localized explanations.

3.4   Proto-BagNet learns meaningful and relevant prototypes

We assessed the clinical meaning and relevance [5, 8] of the prototypes learned by Proto-BagNet by evaluating (1) their interpretability and (2) their coherence as the precision of their corresponding similarity maps at localizing drusen.

We first evaluated the interpretability of the learned prototypes by showing the prototypes without additional context (Fig. 3a) to a clinical expert and asking her if she could understand the concepts encoded in each of them. She described each prototype despite their low resolution as: *(1) soft drusen*; *(2) two drusen in transformation*; *(3) typical drusen*; *(4) drusen with RPE [7] thinning and dense substance inside*; and *(5) drusen with RPE thinning probably in transformation*, showing that the learned prototype where semantically meaningful even when seen in isolation, and diverse due to the enforced dissimilarity (Suppl. Fig. 1). Next, we obtained annotations of the 5 training images from which the learned prototypes were extracted (Suppl. Fig. 2). We found that all learned prototypes (i.e. 100%) were extracted from regions labeled as drusen, confirming that Proto-BagNet learns interpretable and meaningful prototypes.

Subsequently, we evaluated the relevance of the learned prototypes on a subset of 40 test images where an ophthalmologist annotated the presence of drusen. On this subset, we calculated the precision of the prototype similarity maps at localizing drusen lesions to assess whether the highlighted prototypical parts contain similar concepts (drusen related) to those encoded by the learned prototypes. We extracted the $k = 5$ prototypical parts as the most discriminative regions similar to each disease prototype (Fig. 3b, more examples in Suppl. Figs. 3, 4), which were also used in the classification mechanism.
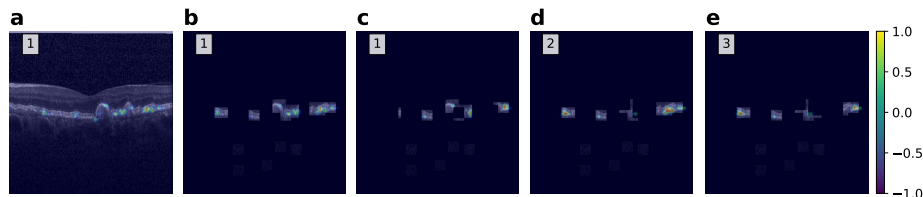
---

[7] Retinal Pigment Epithelium

Fig. 4. **(a)** Example drusen image with overlaid prototype similarity map. **(b)** Occluded image keeping only the top five regions most similar to each prototype. **(c-e)** Images resulting from occluding the most five regions similar to a prototype (1,2,3).

From these, we calculated the precision as the proportion of the prototypical parts (green boxes, magnified on the bottom) which contained annotated drusen lesions (red markers) [13]. The top-k prototypical parts highlighted by the prototype heatmaps contained drusen lesions with high precision ranging from $0.83\,(k=2)$ to $0.87\,(k=1)$ depending on $k \in \{1,\ldots,5\}$, with $p = 0.84{\pm}0.2$ (mean $\pm$ SD) for $k=5$.

3.5   Proto-BagNet relies on a faithful decision-making process

Finally, we verified that Proto-BagNet makes decisions based solely on the visual explanation it provides (i.e. on the k input prototypical parts most similar to each prototype). To assess the faithfulness of our model, we applied the model to the test set (example image in Fig. 4a) and then masked all image regions except the top-k prototypical parts identified by the model (Fig. 4b). We reapplied the model to these occluded images and compared the classification output to the output on the original data. The distribution of predicted logits on original images $(0.03 \pm 0.06,$ and $0.96 \pm 0.15)$ was similar to that on occluded images $(0.03 \pm 0.07,$ and $0.96 \pm 0.15)$, respectively, for healthy and diseased images. The AUC after occlusion was almost similar (0.9918 vs 0.9916) to that obtained without occlusion. We conclude that Proto-BagNet really makes decisions based only on the prototypical parts, and that the explanations provided by the extracted regions are faithful representations of these prototypical parts.

We quantified the importance of each prototype by additionally masking its k prototypical parts in the occluded test images (Fig. 4c-e) and measuring the change in classifier predictions. The predicted drusen probability on healthy images increased by 0.27 ± 0.14 on average when removing healthy prototypical parts (i.e. the region similar to healthy prototypes), while it decreased by 0.42 ± 0.17 when removing disease prototypical parts on drusen images. To summarize, the prototypical parts extracted by the model are indeed evidence for healthy and diseased tissue, respectively, as can be seen by the changes in classifier predictions when occluding them.

4   Discussion and Conclusion

In this work, we proposed Proto-BagNet, an interpretable-by-design prototype-based model that provides faithful and highly localized explanations as well as global interpretability through meaningful prototypes. We evaluated the interpretability of our model through feedback by an ophthalmologist who identified diverse and clinically relevant concepts in the learned prototypes. Furthermore, the model explanations precisely detected drusen lesions in the images. We evaluated the Proto-BagNet for drusen classification on OCT, a solved task in terms of predictive performance (Tab. 1), which lends itself for studying interpretable models due to identifiable and well-characterized regions of interest. However, we noticed that some design choices (e.g., SA layer, k-values) and introduced loss components (dissimilarity and sparsity) enhance interpretability but compete with classification performance (Tab. 1), suggesting that determining the ideal tradeoff might depend on the specific clinical setting. Additionally, the appropriate receptive field size may vary depending on the clinical task and image resolution. In our case, the drusen are small ($< 63\mu m$ [18]) and fit into a patch of size 33×33 and can be changed for other tasks to inject clinical knowledge and adjust for resolution. In a next step, we believe our approach could also be applied to more challenging tasks, and could be useful for the discovery of unknown relevant concepts. In summary, our work may enable prototype-based networks to take a more central stage for realistic task settings, as a promising alternative to post-hoc explanations of black-box models, in particular on medical images.

Disclosure of Interests. The authors declare no competing interests.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. Advances in neural information processing systems **31** (2018)
2. Barnett, A.J., Schwartz, F.R., Tao, C., Chen, C., Ren, Y., Lo, J.Y., Rudin, C.: A case-based interpretable deep learning model for classification of mass lesions in digital mammography. Nature Machine Intelligence **3**(12), 1061–1070 (2021)
3. Brendel, W., Bethge, M.: Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. In: Proc. International Conference on Learning Representations (ICLR) (2019)

4. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.K.: This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems **32** (2019)

5. Davoodi, O., Mohammadizadehsamakosh, S., Komeili, M.: On the interpretability of part-prototype based classifiers: a human centric analysis. Scientific Reports **13**(1), 23088 (2023)

6. Fleckenstein, M., Keenan, T.D., Guymer, R.H., Chakravarthy, U., Schmitz-Valckenberg, S., Klaver, C.C., Wong, W.T., Chew, E.Y.: Age-related macular degeneration. Nature reviews Disease primers **7**(1), 31 (2021)

7. Gautam, S., Höhne, M.M.C., Hansen, S., Jenssen, R., Kampffmeyer, M.: This looks more like that: Enhancing self-explaining models by prototypical relevance propagation. Pattern Recognition **136**, 109172 (2023)

8. Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. Advances in neural information processing systems **32** (2019)

9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)

10. Huang, Q., Xue, M., Huang, W., Zhang, H., Song, J., Jing, Y., Song, M.: Evaluation and improvement of interpretability for self-explainable part-prototype networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2011–2020 (2023)

11. Jacovi, A., Goldberg, Y.: Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 4198–4205 (Jul 2020)

12. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. cell **172**(5), 1122–1131 (2018)

13. Kerol, D., Ilanchezian, I., Kühlewein, L., Faber, H., Baumgartner, C.F., Bah, B., Berens, P., Koch, L.M.: Sparse activations for interpretable disease grading. In: Medical Imaging with Deep Learning (2023)

14. Kim, E., Kim, S., Seo, M., Yoon, S.: Xprotonet: diagnosis in chest radiography with global and local explanations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15719–15728 (2021)

15. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International conference on machine learning. pp. 5338–5348. PMLR (2020)

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

17. Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Grad-Cam, B.: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV). pp. 618–626

18. Silvestri, G., Sillery, E., Henderson, D., Brogan, P., Silvestri, V.: Prevalence of drusen and drusen size in young adults. Investigative Ophthalmology & Visual Science **46**(13), 3298–3298 (2005)

19. Xu-Darme, R., Quénot, G., Chihani, Z., Rousset, M.: Sanity checks for patch visualisation in prototype-based image classification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023. IEEE (2023)

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