ProtoAL: Interpretable Deep Active Learning with Prototypes for Medical Imaging

Iury B de A Santos[1], André C P L F Carvalho[1,*]

[1]Instituto de Ciências Matemáticas e de Computação (ICMC), University of São Paulo (USP), São Carlos, Brazil

Abstract

The adoption of deep learning algorithms in the medical imaging area is a prominent research issue, with high potential for advancing AI-based Computer-aided diagnosis solutions. However, current solutions face challenges due to a lack of interpretability features and high data demands, prompting recent efforts to address these issues. In this study, we propose the ProtoAL model, where we integrate an interpretable deep network into the deep active learning framework. This approach aims to address both challenges by focusing on the medical imaging context and utilizing an inherently interpretable model based on prototypes. We evaluated ProtoAL on the Messidor dataset, achieving results as 0.81 in F1-score and accuracy, and an area under the precision-recall curve of 0.79 while utilizing only 76.54% of the available labeled data. The yielded results were inline with baselines investigated, while providing interpretability prototypes requiring less training data. These capabilities can enhances the practical usability of a deep learning model in the medical field, providing a means of trust calibration in domain experts and a suitable solution for learning in the data scarcity context often found.

Keywords

Deep Active Learning, Interpretability, Medical Imaging

1. Introduction

Deep learning (DL) is an area of machine learning (ML) that has been rapidly growing in recent years. Employed for training of deep artificial neural network architectures, its use has been extensively explored, due to its impressive results in various areas, such as bioinformatics [1], natural language processing [2, 3], and image processing [4, 5]. Some of the main applications in image processing are in medicine, where AI-based Computer-aided diagnosis (AI-CAD) solutions often use DL. In these cases, DL models support medical diagnoses based on images such as magnetic resonance imaging (MRI), X-rays, computed tomography and conventional photographs.

Despite considerable community interest, practical application of AI-CAD solutions encounters obstacles, including the lack of interpretability features in models. These models are often perceived as black boxes, making it challenging for humans to understand their internal reasoning, which raises trust issues among experts and regulatory concerns [6, 7, 8]. Current AI-CAD solutions often lack robustness, being susceptible to biases during training and failing

EXPLIMED - First Workshop on Explainable Artificial Intelligence for the medical domain - 19-20 October 2024, Santiago de Compostela, Spain

*Corresponding author.

iuryandrade@usp.br (I. B. d. A. Santos); andre@icmc.usp.br (A. C. P. L. F. Carvalho)

0000-0001-7234-6877 (I. B. d. A. Santos); 0000-0002-4765-6459 (A. C. P. L. F. Carvalho)

to provide experts with confidence estimations or limitations regarding the results. This poses challenges, especially in healthcare settings, where less experienced professionals may overly rely on computational models [6, 9, 10].

ProtoPNet, introduced by Chen et al. [11], is a deep neural network (DNN) architecture aiming to enhance interpretability features within DNN models. During inference, ProtoPNet showcases prototypes that share similar features with the input image in a "bag-of-features" format. Several studies have applied the ProtoPNet architecture in medical image analysis. For instance, Mohammadjafari et al. [12] used ProtoPNet to classify MRI brain scans as either healthy or indicative of Alzheimer's disease. Vaseli et al. [13] introduced ProtoASNet, a modified version of ProtoPNet tailored for handling spatio-temporal data and integrating aleatory uncertainty estimation into prototypes in the context of aortic stenosis. Wei et al. [14] presented the MProtoNet model, specifically designed for tumor brain classification tasks in multi-parametric MRI (mpMRI) analysis.

The use of DL is also affected by limited availability of large datasets, especially in supervised learning. Besides data being abundant, labeling requires expert input, resulting in high demands in costs and time. To address this issues, deep active learning (DAL) emerges as a feasible approach, extending active l earning (AL) [15] concepts to work with DNNs. DAL assumes that a model with comparable results can be achieved using less but carefully selected training instances. Several works have investigated DAL in medicine, with satisfactory outcomes with less data compared with models trained on the entire dataset [16, 17, 18, 19, 20]. Smailagic et al. [21] proposed the O-medal method, which employs online training and eliminates the necessity of the complete model re-training at each DAL cycle. This method presents a more feasible scheme within the context of DNNs.

Limited research exists on the intersection of DAL and interpretable models. Phillips et al. [22] used LIME interpretability to aid experts in understanding query batches and tracking uncertainty bias. Das et al. [23] employed AL for anomaly detection, providing experts with explanations of model decisions using tree-based ensembles. They introduced a novel formalism for compact descriptions to enhance diversity and generate interpretable rule sets. Liu et al. [24] proposed a DAL approach based on model interpretability, aiming to maximize representativeness in unlabeled data by selecting examples from different linear separable regions of an interpretable DNN. Furthermore, Mondal and Ganguly [25] trained an explainer model alongside a classifier to select new instances based on dissimilarity of explanations compared to already labeled instances. Except for Das et al. [23], none of these works applied the approaches with medical data. Das et al. [23] only explored the use in tabular data.

The main contribution of our work is ProtoAL [1], a novel model capable of integrating an interpretable DNN into the DAL framework, specifically designed for medical image analysis. ProtoAL achieves interpretability through prototypes, providing explanations based on image patches from the training dataset, aligning closely with clinical practices in a visually intuitive manner. Additionally, we investigated how the DAL framework impacts the objective of reducing the required training instances to achieve results comparable to conventionally trained models. To our knowledge, ProtoAL is the first model to integrate an intrinsically interpretable DNN through visual prototypes into the DAL framework. To assess its effectiveness, ProtoAL was

[1]ProtoAL source code is available at https://github.com/IuryBAS/ProtoAL_paper

compared against baselines using robust predictive performance measures, such as the area under the precision-recall curve (AUPRC).

2. Methods

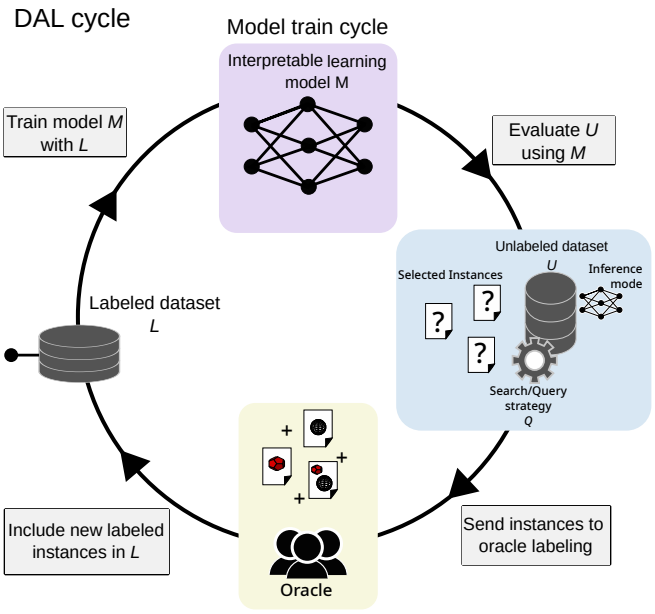

Figure 1: Schematic view illustrating the DAL and model training cycles. In the DAL cycle, labeled instances are added to $\mathcal{L}$ by selecting unlabeled instances from $\mathcal{U}$ using a search strategy. Meanwhile, the learning model M (ProtoPNet) undergoes training iterations within each DAL cycle.

Our model, ProtoAL, aims to integrate interpretable DL into the DAL framework, with focus on the AI-CAD to medical imaging. The DAL framework allows train a learning model, M, using a training set, $\mathcal{L}$, consisting of selected instances based on a search strategy Q, which criteria typically target uncertainty instances or aim to enhance diversity. These selected instances are part of a large, unlabeled dataset $\mathcal{U}$, and are labeled by an oracle before being added to $\mathcal{L}$. The oracle can operate online, where domain experts label the selected instances; or in a "simulated" manner, where ground-truth labels are hidden and only revealed for the selected instances. The DAL framework assumes that selected instances offer pertinent information about the problem, which enables a model achieve comparable performance to one trained on a fully random instance dataset while needing fewer examples. This is particularly beneficial in medical context, where large datasets of unlabeled data exists, but labeled datasets suitable for DL training are scarce, allowing reduce the expenses associated with expert labeling. The DAL cycle is repeated until a stop condition is reached, as a satisfactory performance, budget constraints or depletion of the $\mathcal{U}$ dataset.

Figure 1 illustrates our workflow, divided in two cycles: the outer DAL cycle, referring to

the process described above; and the training of the interpretable model itself, with its inner workings (detailed in the Subsection 2.2). The proposed ProtoAL model uses the ProtoPNet as the base model $\mathcal{M}$, integrating an intrinsically-oriented DNN into the DAL framework. This allows the interpretable capabilities of the ProtoPNet to be incorporated into the DAL framework, requiring fewer training instances and taking advantage of the more informative and significant ones. With ProtoAL, the adoption of models like ProtoPNet in the medical context becomes more feasable, where large labeled datasets are scarce.

2.1. Deep Active Learning

We use the O-medal method [21] to define our DAL framework. The O-medal operate similarly to the framework described above, but differs by avoiding the need to retrain the model from scratch at each DAL iteration. It offers a more reliable, better fitting, and computationally efficient approach. Also, $\mathcal{L}$ does not include all previously labeled instances. Instead, it consists of newly labeled instances and a partition p of the previously labeled data.

The $\mathcal{Q}$ strategy was based on uncertainty estimation. A common approach involves using Monte Carlo Dropout (*MC Dropout*)[26], with the Dropout technique [27] working as a Bayesian approximation. During the inference of the $\mathcal{U}$ instances, T forward steps are performed, and averaging is applied across each instance. The n most uncertain instances are labeled and added to the $\mathcal{L}$ set [26]. Throughout this, all layers except the dropout layers remain frozen. As oracle, we adopted the simulated strategy, only revealing the instances labels when selected to labeling.

2.2. Prototypical Neural Network

The ProtoPNet [11] is an interpretability-oriented DNN that uses prototypes to explain the reasoning of the learning model. Unlike explainability in *posthoc* approaches, the ProtoPNet proposes an intrinsically oriented method for achieving interpretable DNN, trough imposed constraints on the learning process, which considers the explanations during training.

In summary, the ProtoPNet model compromises a backbone neural architecture f, such as VGGNet [28], ResNet [29], or any other selected architecture, extended by a prototypical layer $g_{\mathbf{p}}$ and a fully connected layer h. The input feature extraction is realized by f, while $g_{\mathbf{p}}$ learns m prototypes $\mathbf{P} = \{\mathbf{p}_j\}_{j=1}^{m}$ of $H_1 \times W_1 \times D$ shape, with H, W, and D representing height, width and dimensions of the prototypes [11]. The prototype activation pattern acts as a patch representing some prototypical image patch in the original pixel space, where p can be understood as a latent representation of some prototypical part. The squared L^2 distances between the input patches and the prototypes of p are inverted into similarity scores, forming an activation map where the values indicate how strongly a prototypical part is present in the image. The activation map is reduced using global max-pooling to a single similarity score, indicating how much a prototypical part is present in some input image patche. The m similarity scores produced by $g_{\mathbf{p}}$ are multiplied by the outputs of h and normalized using softmax, yielding the predicted probabilities of the image [11].

The main advantage of ProtoPNet is provide visual explanations of the predictions, which it optimizes and learns alongside the classifier. These prototypes are related to instances from the training set, representing real cases and attributes.

3. Experiments

3.1. Dataset

The experiments were carried out using the Messidor dataset [30][2] of diabetic retinopathy (DR). It compromises 1200 color images of the eyes fundus, obtained from three different ophthalmologic departments. Experts evaluated each image and classified it based on the retinopathy grade and risk of macular edema. Retinopathy grade ranges from 0 (normal) to 3, considering the number of microaneurysms, hemorrhages and the presence of neovascularization. Corrections were made by correcting mislabeling and removing duplicated images files according to the instructions in the dataset download page[3].

Following the preprocessing outlined in Smailagic et al. [21], we grouped the retinopathy grades as healthy (DR = 0) or diseased (DR $\geq$ 1). The risk of macular edema feature was not used in the experiments. The images were resized to 512×512, and data were augmented by randomly applying rotations up to 15 degrees, horizontal flips and scaling in the range [0.9, 1]. The train, validation and test sets were composed of 759, 190 and 238 instances, respectively.

3.2. Baselines

We compared the ProtoAL model in two perspectives, to observe both the interpretability and DAL framework factors. For such, we adopted three baselines, targeting distinct contexts: (*i)* A **vanilla ResNet-18** model, trained conventionally and without interpretability features, with access to the entire training dataset from beginning; (*ii*) the **ProtoPNet standalone baseline**, with the aim of evaluate the performance of the interpretability model used in ProtoAL without incorporating it into the DAL framework; and (*iii*) the **ProtoAL with random search strategy**, utilizing as search strategy the selection of random instances from $\mathcal{U}$. This baseline allows to observe the impact of the MC Dropout as search strategy in the performance of the ProtoAL model. The ResNet-18 and ProtoPNet were pretrained on the ImageNet dataset [31], both when used as baselines and backbone in the ProtoAL.

3.3. Implementation details

As mentioned earlier, the ProtoAL model follows the structure of the O-medal method, while employing a ProtoPNet as DNN model. The model hyper-parameters were optimized using grid search[32]. This was deemed the most appropriate approach given the computational constraints associated with exhaustive search. The runs were executed by varying seed values (0, 1, 2, 5, 10, 12, 42, 123, 1234, 12345), and hyper-parameters for both the DAL method and ProtoPNet model, as MC Dropout steps (10, 30, 50), instances to label per DAL iteration (10, 20, 30), batch size (32, 64) and epochs per DAL iteration (5, 10, 20).

We conducted runs using both random (ProtoAL-Random) and MC Dropout (ProtoAL-MC) as search strategy. The stop condition for the DAL cycle was determined as no remaining instances left to be labeled in $\mathcal{U}$. The number of runs was dynamically adjusted based on the labeled instances per DAL iteration. The fixed hyper-parameters follows the Smailagic et al. [21] work.

[2]Kindly provided by the Messidor program partners (see https://www.adcis.net/en/third-party/messidor/).
[3]https://www.adcis.net/en/third-party/messidor/

The $\mathcal{L}$ set initially consisted of 100 randomly selected instances. The percentage of previously labeled examples forming $\mathcal{L}$ was set to 0.875.

A ResNet-18 was used as the backbone DNN of the ProtoAL, with prototype layers featuring 256 channels. We settled 12 prototypes, with 6 prototypes allocated for each class (healthy and diseased). During training, ProtoPNet undergoes warm-up for 5 epochs during the first DAL iteration. After warm-up, joint training is performed for e epochs (epochs per DAL iteration hyper-parameter), succeeded by a projection (push) step and last layer optimization for 15 steps. Excepting for warm-up, this cycle repeats for all DAL iterations. We used the Adam optimizer, with learning rates as outlined in Chen et al. [11]. Learning rate was decreased exponentially per epoch. All experiments were run on Nvidia Tesla V100 GPU, with 32GB VRAM. The DNN models were implemented with PyTorch 2.0.1 framework and Python 3.10.

4. Experimental Results

During the grid search, we trained a total of 540 ProtoAL models. We selected the best model configuration using AUPRC, based on validation set results. The best run was achieved by a model trained with a batch size of 32, employing 10 iterations of MC Dropout per instance, and selecting 30 new examples at each DAL iteration. For the joint optimization phase, the ProtoPNet model underwent 10 training epochs. Table 1 presents a comparison of results between ProtoAL with MC dropout (ProtoAL-MC) and the baselines evaluated on the test set.

The ProtoAL-MC achieved an AUPRC of 0.79, along with an F1-Score and accuracy of 0.81. The ProtoAL-Random, utilizing random search strategy, achieved 0.77 in AUPRC and 0.79 in both F1-Score and accuracy. These results demonstrate that the MC Dropout search strategy yielded superior results, particularly evident in the AUPRC metric.

ProtoPNet and ResNet-18 obtained comparable or superior results to ProtoAL in specific metrics. It is worth noting that both baselines had access to the entire training dataset from the start, being more exposed to the instances. It can be observed that ProtoPNet exhibits a decline

Table 1
Evaluation test results of the ProtoAL-MC and baselines in relation to AUPRC, F1-Score, Precision, Recall and Accuracy

Model	AUPRC	F1-Score	Precision	Recall	Accuracy
ResNet-18	0.8462	0.8699	0.9067	0.8359	0.8655
ProtoPNet	0.7900	0.8273	0.8512	0.8046	0.8193
ProtoAL-Random	0.7773	0.7999	0.8571	0.75	0.7983
ProtoAL-MC	0.7935	0.8181	0.8684	0.7734	0.8151

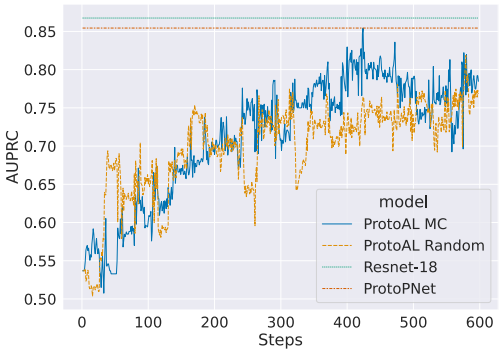

Figure 2
Comparisons of ProtoAL-MC model and the baselines, evaluated on the validation set

in performance compared to ResNet-18, possibly due to the added complexity of optimization from the inclusion of interpretability features. Consequently, ProtoAL's performance remains consistent and similar to that of the ProtoPNet baseline.

The ProtoAL-MC model achieved results comparable to the ProtoPNet baseline in the 17th DAL iteration out of 23 total, with the $\mathcal{L}$ dataset consisting of 512 instances, having labeled 581 instances and leaving 178 instances in the $\mathcal{U}$ set. By this stage, ProtoAL-MC required only 76.54% of the training instances, out of the initial total of 759 examples, to achieve comparable results to models trained with all available examples. Figure 2 presents a comparison between ProtoAL-MC and the baseline models across DAL cycle steps when evaluated on the validation set, with ProtoAL-MC achieving its best performance in approximately 400, out of 600 steps. The steps are in respect to the total of model training steps (joint, push and last only phases) during all DAL iterations. Figure 3 displays prototypes demonstrating correct classification of a image with DR for qualitative evaluation. It shows the three most similar prototypes for both healthy and diseased cases, along with prototype weight, similarity scores, and prediction value.

Through prototypes, a domain expert can observe the most similar instances for both disease and healthy cases, along with their respective relevant regions in the input image. As interpretability criteria, experts can rely on visual inspection, and prototype similarity scores and weights, assessing whether the model's inferences are meaningful and relevant. For instance, a expert can verify whether a highly relevant prototype corresponds to a meaningful region of the

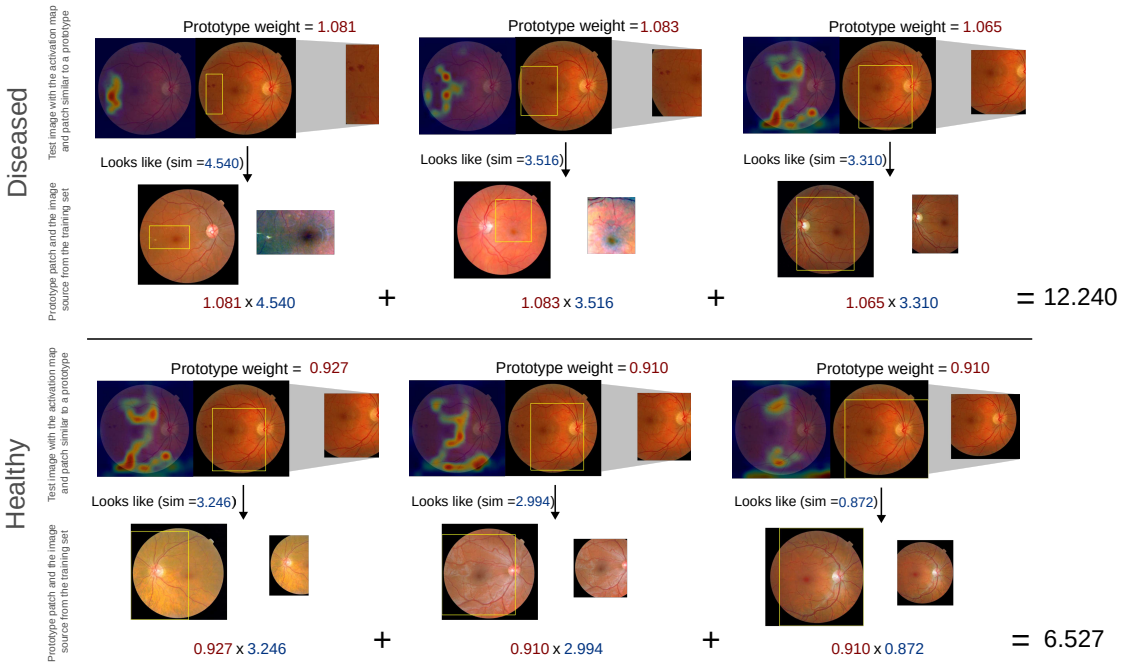

Figure 3: The most similar prototypes for both diseased and healthy conditions for an accurately classified image with DR. It includes similarity scores, prototype weights, and prediction values. For each prototype, the relevant activated region from the input image and its corresponding prototype from a training image are shown

input image, assess if it contains features consistent with the predicted output, and observe the similarity score it achieves. With such information, an expert has the tools to more accurately calibrate their confidence and trust in the model's prediction.

5. Discussion and conclusion

Our model, ProtoAL, integrates an interpretable DNN model with prototypes into a DAL framework. Specifically tailored for the AI-CAD context of medical imaging, the aim was to enhance the reliability of AI-CAD solutions in practical medical contexts while exploiting the capacity to comprehend the model decisions and training in limited labeled datasets. Quantitative results presented in Table 1 demonstrate the success of providing a interpretable model while utilizing a reduced amount of training data, addressing two key challenges in the adoption of AI in medical contexts: lack of interpretability and scarcity of labeled datasets. The qualitative results in Figure 3 illustrate prototypes relevant to the network's inference reasoning for a diseased input image, providing visual explanations that domain experts can interpret.

ProtoAL offers interpretability features lacking in the ResNet-18 baseline, which enhance the practical usability of ProtoAL as an AI-CAD solution while maintaining a performance level similar to that of the ProtoPNet model, albeit with reduced training data demands. However, preliminary experiments with ProtoPNet using a greater number of prototypes resulted in the repetition of similar or identical prototypes, which is undesirable as it does not promote prototype diversity. Furthermore, our work is limited to quantitative evaluation using standard metrics such as AUPRC and F1-Score, which do not assess the interpretability characteristics of the model.

Future works could explore how to enhance integration of interpretability features within the DAL framework, including leveraging information from prototypes to refine search strategies. Moreover, investigate how to promote prototype diversity and integrate domain experts in evaluation of the results would provide valuable insights and a important qualitative analysis.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and grant #2022/05788-4, São Paulo Research Foundation (FAPESP).

References

[1] A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, A. Bridgland, et al., Improved protein structure prediction using potentials from deep learning, Nature 577 (2020) 706–710.

[2] D. W. Otter, J. R. Medina, J. K. Kalita, A survey of the usages of deep learning for natural language processing, IEEE trans. on neural networks and learning systems 32 (2020) 604–624.

[3] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, J. Gao, Deep learning–based text classification: a comprehensive review, ACM Computing Surveys (CSUR) 54 (2021) 1–40.
[4] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, D. Terzopoulos, Image segmentation using deep learning: A survey, IEEE trans. on pattern analysis and machine intelligence (2021).
[5] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, R. Qu, A survey of deep learning-based object detection, IEEE access 7 (2019) 128837–128868.
[6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60–88. doi:10.1016/j.media.2017.07.005.
[7] A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, Y. Liu, E. Topol, J. Dean, R. Socher, Deep learning-enabled medical computer vision, npj Digital Medicine 4 (2021) 5. doi:10.1038/s41746-020-00376-2.
[8] A. Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care, Neural Computing and Applications 32 (2020) 18069–18083. doi:10.1007/s00521-019-04051-w.
[9] S. Gaube, H. Suresh, M. Raue, A. Merritt, S. J. Berkowitz, E. Lermer, J. F. Coughlin, J. V. Guttag, E. Colak, M. Ghassemi, Do as ai say: susceptibility in deployment of clinical decision-aids, NPJ digital medicine 4 (2021) 1–8.
[10] D. W. Bates, D. Levine, A. Syrowatka, M. Kuznetsova, K. J. T. Craig, A. Rui, G. P. Jackson, K. Rhee, The potential of artificial intelligence to improve patient safety: a scoping review, NPJ digital medicine 4 (2021) 1–8.
[11] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, J. K. Su, This looks like that: deep learning for interpretable image recognition, Advances in neural information processing systems 32 (2019).
[12] S. Mohammadjafari, M. Cevik, M. Thanabalasingam, A. Basar, Using protopnet for interpretable alzheimer's disease classification., in: Canadian Conference on AI, 2021.
[13] H. Vaseli, A. N. Gu, S. N. Ahmadi Amiri, M. Y. Tsang, A. Fung, N. Kondori, A. Saadat, P. Abolmaesumi, T. S. Tsang, Protoasnet: Dynamic prototypes for inherently interpretable and uncertainty-aware aortic stenosis classification in echocardiography, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2023, pp. 368–378.
[14] Y. Wei, R. Tam, X. Tang, Mprotonet: A case-based interpretable model for brain tumor classification with 3d multi-parametric magnetic resonance imaging, in: Medical Imaging with Deep Learning, PMLR, 2024, pp. 1798–1812.
[15] B. Settles, Active learning literature survey (2009).
[16] Z. Zhao, Z. Zeng, K. Xu, C. Chen, C. Guan, Dsal: Deeply supervised active learning from strong and weak labelers for biomedical image segmentation, IEEE journal of biomedical and health informatics 25 (2021) 3744–3751.
[17] X. Wu, C. Chen, M. Zhong, J. Wang, J. Shi, Covid-al: The diagnosis of covid-19 with deep active learning, Medical Image Analysis 68 (2021) 101913.
[18] V. Nath, D. Yang, H. R. Roth, D. Xu, Warm start active learning with proxy labels and selection via semi-supervised fine-tuning, in: International Conference on Medical Image

Computing and Computer-Assisted Intervention, Springer, 2022, pp. 297–308.
[19] S. Belharbi, I. Ben Ayed, L. McCaffrey, E. Granger, Deep active learning for joint classification & segmentation with weak annotator, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3338–3347.
[20] M. L. Di Scandalea, C. S. Perone, M. Boudreau, J. Cohen-Adad, Deep active learning for axon-myelin segmentation on histology data, arXiv preprint arXiv:1907.05143 (2019).
[21] A. Smailagic, P. Costa, A. Gaudio, K. Khandelwal, M. Mirshekari, J. Fagert, D. Walawalkar, S. Xu, A. Galdran, P. Zhang, et al., O-medal: Online active deep learning for medical image analysis, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10 (2020) e1353.
[22] R. Phillips, K. H. Chang, S. A. Friedler, Interpretable active learning, in: S. A. Friedler, C. Wilson (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency, volume 81 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 49–61. URL: https://proceedings.mlr.press/v81/phillips18a.html.
[23] S. Das, M. R. Islam, N. K. Jayakodi, J. R. Doppa, Active anomaly detection via ensembles: Insights, algorithms, and interpretability, arXiv preprint arXiv:1901.08930 (2019).
[24] Q. Liu, Z. Liu, X. Zhu, Y. Xiu, Deep active learning by model interpretability, arXiv preprint arXiv:2007.12100 (2020).
[25] I. Mondal, D. Ganguly, Alex: Active learning based enhancement of a classification model's explainability, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 3309–3312.
[26] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, PMLR, 2016, pp. 1050–1059.
[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, The journal of machine learning research 15 (2014) 1929–1958.
[28] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[30] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay, et al., Feedback on a publicly distributed image database: the messidor database, Image Analysis & Stereology 33 (2014) 231–234.
[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, International journal of computer vision 115 (2015) 211–252.
[32] S. M. LaValle, M. S. Branicky, S. R. Lindemann, On the relationship between classical grid search and probabilistic roadmaps, The International Journal of Robotics Research 23 (2004) 673–692.