Interpretable deep learning in medical image

University of Minho, Portugal

Abstract. Deep learning has proven to be highly effective in medical diagnosis, often surpassing human experts in accuracy. However, the black-box nature of these algorithms has prevented their widespread clinical use. Recent studies have focused on developing methods for explaining the decision-making process of these models, with a particular emphasis on medical imaging tasks. These efforts have led to the development of various interpretability methods that seek to reveal the features that influence the model's decisions the most. The lack of interpretability, transparency, and trust associated with deep learning models has been a significant barrier to their adoption in clinical practice. To address this issue, several studies have proposed interpretability methods for medical diagnosis based on deep learning models. While deep learning models have shown great promise in medical diagnosis and treatment decisions, their adoption into clinical practice requires greater interpretability, transparency, fairness and accountability. Future research should focus on developing guidelines for the use of interpretability methods in clinical settings and further improving the interpretability of deep learning models.

Keywords: Deep Learning · Interpretability · Deep Neural Networks · Medical Image.

1 Introduction

Using AI in computer-aided diagnostics (CAD) has the potential to make the diagnosis process more efficient and accessible to a wider population. Deep learning, which is a leading method of AI, has been used for medical imaging tasks such as Alzheimer's classification[30], lung cancer detection[24], and retinal disease detection[60]. Despite the remarkable results achieved, AI-based methods have not yet been significantly deployed in clinics due to the black-box nature of deep learning algorithms and computational costs.

To gain trust from physicians, regulators and patients, a medical diagnosis system needs to be transparent, understandable and explainable. This means that the complete logic of how a decision is made should be explained to all involved parties. Regulations like the European General Data Protection Regulation (GDPR) require retraceability of decisions, making it harder for the use of black-box models in healthcare and other businesses.[22]

Interpretability is essential for the safe, ethical, fair and trustworthy use of AI and a critical enabler for its deployment in the real world. By showing what a model looked at while making a decision, myths about AI can be broken, and trust can be built among end-users.[64]

This paper describes studies related to the interpretability of deep learning models in the context of medical imaging. The methods under study are described in the following section. However, since for each situation it is necessary to evaluate them, there is a section 3 that reviews various interpretability methods applied to different medical imaging modalities. Section 4 conducts a discussion of the results presented above, as well as a summary of future work and current trends in interpretable deep learning models in medical imaging analysis.

2 Methods

Several models have been proposed in the literature to classify different interpretability methods. Generally, DL models and methods are not absolute, they can vary greatly depending on the characteristics of the problem at hand.

Concept learning models are used in Deep Learning to explain outcomes in terms of interpretable human concepts, such as semantic features. These models use concepts generated by experts during training to improve performance and enable clinical interventions. Examples of concept learning models include Conceptual Alignment Deep Neural Networks and Capsule Networks.[35][41][62]

Concerning case-based models for prediction in images make predictions by comparing features extracted from an input image with discriminative class prototypes. The ProtoPNet is an example of a case-based model that does not sacrifice performance over black-box deep learning models and is used in classifying different diseases in chest X-ray images. However, it is important to note that similarity in latent space does not always translate into similarity in terms of features interpretable by humans.[7][33][39][43][65]

The counterfactual explanation technique generates images by minimally perturbing the original image to cause a maximum change in the classifier's prediction and change the predicted class of the original image. This helps to identify diseased areas and understand the changes needed to change the classifier's prediction. Counterfactual images are synthesized using Generative Adversarial Networks (GANs) or by perturbing the latent space of an autoencoder. These techniques have been validated in various medical applications, including detection of tumors in metastatic lymph nodes, prediction of diabetic macular edema in retinal images, and classification of MRI images for Alzheimer's.[3][12][38][56][57][72][63][66][68]

In what concerns concept attribution, this explains how deep neural networks work in medical images. Methods such as TCAV, ACE and Radiomics are used to identify and quantify concepts in medical images. RCV is used to identify continuous concepts that are important for the classification task. These methods are applied in different medical diagnostic cases, such as breast cancer classification and liver tumor segmentation.[11][17][21][27][32][77]

Deep neural networks can provide explanations in text form, called language description, which consists of justifications learned in a supervised or unsupervised manner. Training these networks is challenging due to the complexity of natural language and the lack of structured medical reports for training. Some methods have been proposed to provide textual descriptions of medical diagnoses, but these methods add annotated costs because they require reformatting the medical reports.[10][37][72][79][80]

The latent space in convolutional neural networks (CNNs) consists of a compressed representation of the input image and aims to model salient factors of variation in the data independently. Autoencoders are used for nonlinear dimensionality reduction and consist of an encoder and a decoder. The goal is to capture the salient features of the data that can be understood by humans. There are several techniques, such as Variational Autoencoders (VAE) to visualize and interpret the latent space of CNNs, which can be used for interpretability, model validation and detection of biases in the data.[5][8][11][14][16][75][76]

The explanation of deep learning (DL) models using attribution map, highlights the relevant regions of the input image for prediction. There are different gradient-based interpretability methods, such as Class Activation Map (CAM), Multi Layer Class Activation Map (MLCAM), Gradient Class Activation Map (Grad-CAM), Guided Grad-CAM and Grad-CAM++. Other methods include Integrated Gradient (IG), Saliency Maps, SmoothGrad, and Guided Turn Propagation (GBP). These methods have been used in different medical image analysis tasks such as breast cancer detection, oral cancer, diabetic retinopathy, multiple sclerosis, and lung adenocarcinoma. Interpretability comes with a performance cost, but can improve the accuracy of medical diagnosis with the help of physicians. Note that the perturbation-based methods investigate the effect of changing different parts of the input image on model predictions, and also that the occlusion method produces an assignment map by systematically perturbing the image to observe the effect on the output, so multiple inferences need to be made for the same input image.[11][14][16][75][76][19][26][31][41][42][46][54]

The CAM-based interpretability methods generates class activation maps to highlight discriminant regions in images for classification tasks. However, it is restricted to a specific architecture that includes a global pooling layer and can only visualize the last convolutional layer. To overcome these limitations, Grad-CAM was proposed, which is a generalization of CAM applicable to a variety of convolutional neural network models and can generate a coarse location map. Grad-CAM can also be combined with the Guided Backpropagation method to generate a more refined feature location map, known as Guided Grad-CAM. In addition, Score-CAM is a gradient-based technique that combines CAM and perturbation-based methods to generate a comprehensible location map.[37][80][59][72][78]

Regarding LRP, this is a method based on pixel decomposition of nonlinear classifiers that generates a heat map by evaluating a relevance score. The relevance is calculated on each layer backwards from the last layer, and LRP highlights the relevant information in the high-dimensional fMRI data. It should also be emphasized that perturbation methods can alter parts of the image that are not clinically understandable, so significant perturbations are required.[26][31][41][42][46][54]

Attention modules in CNNs improve classification performance and generate a detailed attribution map centered on key parts of the image. The method uses a compatibility score calculation between local and global features to calculate attention weights that are applied to local features. The attention results are concatenated and fed to a fully connected layer to perform classification, and attention modules are most effective when used relatively late in the CNN pipeline. In addition, other methods of explanation in neural networks besides Grad-CAM are mentioned, such as DeepLIFT, which assigns importance to neurons based on the activation of each neuron relative to its reference activation. [31][41][42][46][54]

Deep SHAP is an adaptation of DeepLIFT that approximates SHAP values, while Deep Taylor applies layer-by-layer Taylor decomposition. As for CDEP, it incorporates explanation errors into the loss function and Expected Gradients uses an assignment priority to generate desirable assignment maps. Finally, PatternNet and PatternAttribution are explanation methods for linear models that provide correct signal visualizations.[26][31][41][42]

Using anatomical prior information and segmentation networks in medical images can improve the interpretability of deep learning models by detecting important features of the task. The internal network representation of the neural network is important for interpretability, and methods such as Activation Maximization and Network Dissection exist to understand the organization of information in segmentation models. Natekar et al. applied these methods to understand brain tumor segmentation models.[18][48][67][4][45]

Relative to SIBNet, this is a deep learning model training approach to generate spatially distinct and coherent saliency maps that can be used on unsupervised data. It uses class distinction loss and spatial coherence terms to regulate the spatial distribution of the generated saliency maps. This approach is applied in medical image segmentation tasks, where a separate classification model is used to generate saliency maps for each class label. The encoder block of the classification model is used as a pre-trained encoder for a UNet model, which is trained with weighted cross entropy loss and data loss. The result is called UNetSIBNet and uses the inductive biased guidance of SIBNet.[25]

The perturbation-based interpretability methods include sensitivity analysis and sensitivity concealment. The former measures the impact of uncertainty in the input variables on the final prediction, while the latter hides parts of the input and measures the change in the output.[52][53]

Surrogate interpretability methods involve training interpretable models to mimic the behavior of black box models. Local interpretable model-agnostic explanation (LIME) and Knowledge Distillation (KD) are two surrogate interpretability methods. LIME can be used to explain a model by approximating its behavior locally using a regression model weighted by a cosine distance model.[47][80][72][20]

There are three main types of intrinsic interpretability methods: attention mechanisms, rule-based extraction and decision trees. While all of these methods provide interpretability and make the model transparent, the accuracy of rule-based extraction methods and decision trees may not be accurate enough.[15][44][61][74]

3 Results

In the previous section, methods that make deep neural networks more understandable in medical imaging applications were discussed. However, it is difficult to determine which method is most suitable for a given application, as explanations can be subjective. In addition, the uncertainty present in Deep Learning models also impacts interpretability, which makes the evaluation of interpretability methods challenging. Of note, deep learning models have outperformed traditional methods in the medical field with high efficiency and accuracy. Thus, the applications of interpretability in disease diagnosis are reviewed here, with the aim of introducing the current state of research in the medical field.[51]

In this way, some evaluation metrics are still verified for each case study. These are useful for evaluating the performance of disease diagnosis. In contrast, there are no standard evaluation metrics for interpretability and most of the performance is judged by humans. Therefore, these evaluation metrics are important for evaluating the interpretability performance of deep learning models in the medical domain in order to provide benchmarks for readers. The evaluation metrics in both cases presented are Accuracy, Specificity, Sensitivity and AUC, through the values of true positive (TP), false positive (FP), true negative (TN) and false negative (FN).[69]

3.1 Diabetic Retinopathy

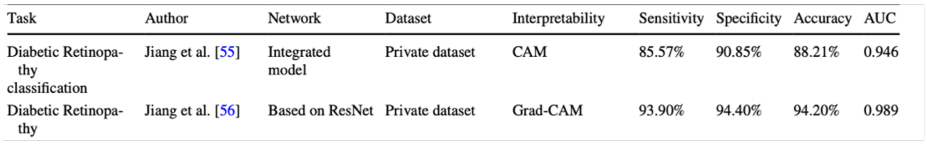
Diabetic retinopathy is a chronic eye disease caused by diabetes that can lead to blindness. To detect diabetic retinopathy, Kind et al. [34] applied the R-CNN ResNet101 fast architecture to classify retinal fundus images as healthy or not. To better diagnose the severity of the disease, the researchers designed several interpretable models, exploiting the excellent performance of deep learning. For example:

De et ai. [71] created an interpretable diabetic retinopathy classifier that uses modified LRP to score significant pixels in the image and evaluate their contribution to disease severity classification.

Kumar et ai. [36] developed an interpretable system called CLEAR-DR to classify diabetic retinopathy and provides a visual map to interpret the decision-making process.

Jiang et al. created an ensemble model with three different models to classify diabetic retinopathy. They used the Adaboost algorithm to combine the results of the models and the CAM technique to generate class activation maps that explain the classification. They also proposed a multirotule deep learning model that does not require a specific structure and obtained better results than the ensemble model that are shown in Figure 1. In addition, an EfficientNet architecture was proposed to detect different types of diabetic retinopathy with faster and more accurate experimental results.[29][28][9]

More qualitatively, evaluation grounded in the presence of a physician is needed to determine whether end users are satisfied with the explanations provided. Thus, the System Causability Scale (SCS) was used, which assesses the interpretability of methods in medical contexts.[23]



Task	Author	Network	Dataset	Interpretability	Sensitivity	Specificity	Accuracy	AUC
Diabetic Retinopathy classification	Jiang et al. [55]	Integrated model	Private dataset	CAM	85.57%	90.85%	88.21%	0.946
Diabetic Retinopathy	Jiang et al. [56]	Based on ResNet	Private dataset	Grad-CAM	93.90%	94.40%	94.20%	0.989

Fig. 1. The experimental results of diagnosing diabetic retinopathy

Sayres et al. [55] studied the impact of the DL algorithm and found that when using explanations generated by AI algorithms to diagnose Diabetic Retinopathy, integrated gradient explanations can introduce bias and lead to over-diagnosis. Thus, human-AI collaboration using explanations has been shown to be more accurate. Therefore, it is crucial to evaluate interpretability methods in the clinical context to determine their usefulness to clinicians. Below you can see in Figure 3 the impact of the CAM method on this diagnosis.

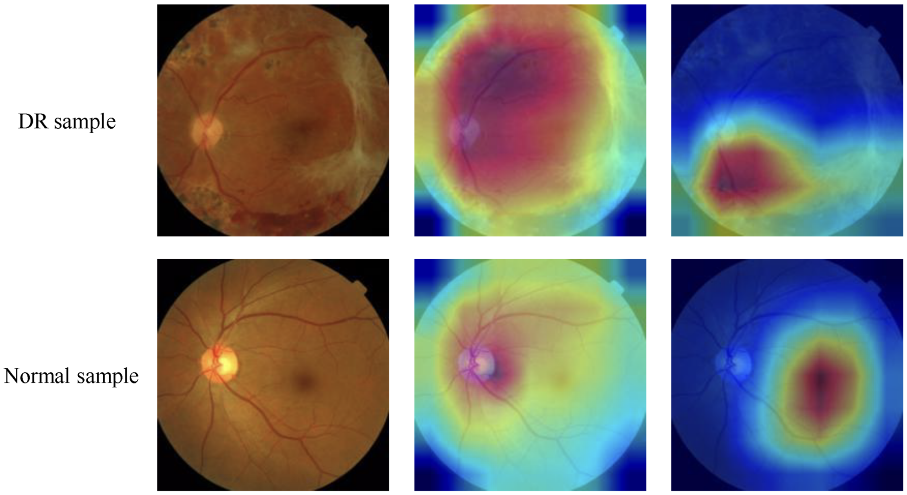
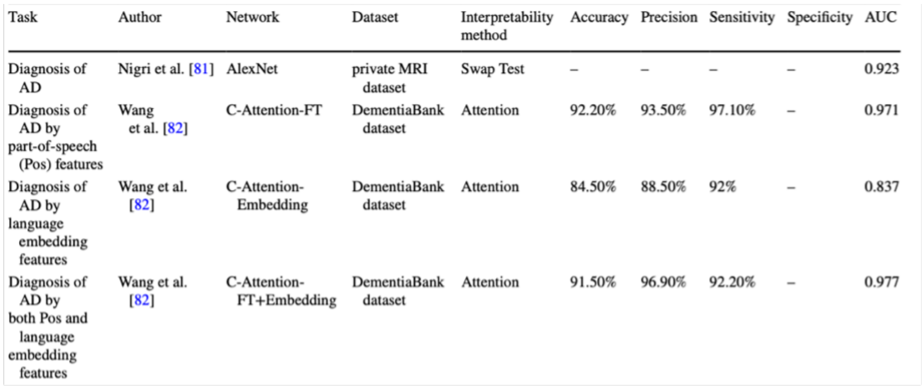


Fig. 2. The highlighted regions of the DR decision. From the left to right: the input, highlighting without CAM-Attention and highlighting with CAM-Attention.

3.2 Alzheimer

Alzheimer's disease is an age-related brain disease that causes memory loss and cognitive dysfunction, leading patients to gradually lose the ability to work and live independently. To accurately diagnose such a disease, different methods have been explored, such as Guided Grad-CAM and occlusion to generate complementary maps for the classification of Alzheimer's disease pathologies.

However, Nigri et al. proposed the exchange test method to interpret the prediction results, which proved to be more suitable for medical images, as shown in Figure 2. In addition, Wang et al. presented explainable models based on language skills to automatically detect the disease. Another multimodal multilayer model proposes using the SHAP method to provide explanations for each layer of the model. Two argumentation-based methods have also been proposed to diagnose Alzheimer's disease, with positive results.[69]



Task	Author	Network	Dataset	Interpretability method	Accuracy	Precision	Sensitivity	Specificity	AUC
Diagnosis of AD	Nigri et al. [81]	AlexNet	private MRI dataset	Swap Test	–	–	–	–	0.923
Diagnosis of AD by part-of-speech (Pos) features	Wang et al. [82]	C-Attention-FT	DementiaBank dataset	Attention	92.20%	93.50%	97.10%	–	0.971
Diagnosis of AD by language embedding features	Wang et al. [82]	C-Attention-Embedding	DementiaBank dataset	Attention	84.50%	88.50%	92%	–	0.837
Diagnosis of AD by both Pos and language embedding features	Wang et al. [82]	C-Attention-FT+Embedding	DementiaBank dataset	Attention	91.50%	96.90%	92.20%	–	0.977

Fig. 3. Part of experimental results for diagnosing alzheimer

Different metrics and evaluation strategies have been proposed to overcome the lack of explanatory truth. However, it is not possible to arrive at metrics that are applicable to all interpretability methods [55]. Thus, other strategies to identify and correctly diagnose Alzheimer's disease, namely, by biomarkers and attribution maps, will still be addressed.

Imaging Biomarker can be described as any quantifiable feature or structure in an image that can be used for the diagnosis and prognosis of a disease.[73] DL models automatically detect patterns in data and learn complex functions that are not understandable to humans. These functions can learn to use established biomarkers from the data along with unknown imaging biomarkers for diagnosis. DL model explanations can be validated using known imaging biomarkers to show that the DL model is using clinically relevant features for diagnosis. Such validation can promote confidence in the utility of DL models in clinical practice.

Boehle et al. [6] used LRP to explain deep neural networks for the diagnosis of Alzheimer's disease. Imaging biomarkers can be used for both qualitative and quantitative evaluation of interpretability methods. Thus, quantitatively it was shown that the areas of high relevance correlate well with hippocampus volume, which was identified as a key biomarker of Alzheimer's disease. The high variability in the heat maps between different disease cases showed that LRP can be used to identify biomarkers for different stages of Alzheimer's disease.

Attribution Maps are a way to explain how the machine learning model is making decisions by showing the part of the input image that is considered important. However, it is important to note that these maps do not explain how the model uses the relevant information, and so it is essential to evaluate the robustness of attribution maps produced by different methods quantitatively as well as qualitatively.[50] Eitel and Ritter performed tests to evaluate the robustness of assignment maps for Alzheimer's disease classification and found that some methods, such as SmoothGrad were more sensitive to sample size, whereas the LIME method is sensitive to the number of perturbed instances as well as random seeds for superpixel generation.[2]

4 Discussion

Interpretability methods in deep learning have been used to identify new imaging biomarkers for various diseases. Layer-wise Relevance Propagation and a novel architecture utilizing an encoder-decoder and a CNN classifier were used to identify possible biomarkers in different fields. These imaging biomarkers can be used for risk assessment and diagnosis in clinical practice after validation. However, further improvements in interpretability methods are needed to better understand DL models and detect new imaging biomarkers.[49][58][40][13][51]

The importance of inductive bias in machine and deep learning is discussed, particularly in medical image analysis where smaller datasets and real-world confounders make it necessary to inject domain knowledge. While convolutional neural networks (CNNs) are a successful example of inductive bias for image recognition, visual image transformer networks have demonstrated high performance levels but require more training data. Hybrid approaches combining CNNs and visual image transformer networks are emerging. The Saliency Inductive Bias Network (SIBNet) approach improves model performance and interpretability through experimentation with two common medical imaging problems of classification and segmentation, several ablation and robustness tests.

Post-hoc interpretability methods are only approximations and do not accurately illustrate the correct model behavior, therefore compromise trust in the explanations. There is a need to evaluate the post-hoc explanations carefully before they can be used in the clinical workflow. Both quantitative and qualitative evaluation of post-hoc interpretability methods should ensure the robustness and faithfulness of the explanations. The insights and explanations provided by the interpretability methods can help discover new imaging biomarkers. Applications grounded evaluations should be carried out in presence of a clinician to ensure the utility of the explanations and eliminate concerns related to bias.[1][50][51]

Transparency of deep neural networks is an essential clinical, legal, and ethical requirement. Thus, it is crucial to analyze the applications of interpretability in disease diagnosis and classified them by disease category according to the body part.[69]

Concerning to eye diseases, most researchers have used CAM-based and attention-based methods to provide visual interpretability by generating heatmaps. Since the CAM technique only produces a rough localization, researchers tried to combine it with other methods. To accurately localize and identify smaller lesion areas, attention-based methods were also fully employed to enhance regions of interest and suppress irrelevant regions. In addition, all the models can not only implement disease diagnosis accurately but also provide interpretability.

Relating to lung diseases, it is known that most researchers prefer to exploit the Grad-CAM technique to provide an intuitive interpretation.[69] Also, there are some researchers that exploit attention mechanisms to strengthen the regions of interest. In short, both approaches are competent to enhance model transparency and yield satisfactory results.[69] Analyzing several interpretable applications of brain diseases, researchers have tried to exploit different approaches to analyze models and provide interpretability, since brain diseases affect the human neurological system. Extensive experiments have found that these methods achieve promising outcomes.[69] For heart diseases, VAE was introduced to build interpretable models, which is an effective way to observe the beat changes of EGG to diagnose the disease. For skin diseases, most researchers are using attention-based models to diagnose the diseases, which can strengthen the interest regions to enhance the interpretability. For breast diseases, there are a lot of options, from estimating breast density to predicting cancer and then predicting breast cancer recurrence.

In addition to the above-mentioned, there are various interpretable applications with remarkable performance. All of them can improve diagnostic efficiency and facilitate the development of deep learning in the medical field.[70][49][69]

5 Future Works

Models that can be easily understood and explained, such as case-based models and concept learning models, show promise as interpretability methods for use in clinical settings. These models perform similarly to more complex models like black-box CNNs, but are designed to be interpretable. To ensure their reliability, it's important to conduct tests on the attribution maps used by these models. By using multiple interpretability methods, it is possible to gain a better understanding of how specific regions highlighted by attribution maps contribute to the final predictions. To improve performance and interpretability, multi-modal data, which includes images, texts, and genomics data, can be utilized. One proposed solution is the use of Graph Neural Networks, which can establish causal links between different types of data using graph structures. To determine the most suitable interpretability methods, a combination of human-centered evaluations and quantitative functionality-based evaluations is needed.[69]

6 Conclusion

In the medical field, deep learning models have exceeded the traditional methods with high efficiency and accuracy, so several applications have been designed to help doctors make more accurate decisions. However, it is essential to mention that the evaluation of the interpretability methods is a critical step to validate the utility of the generated explanations and develop insights.[51]

Concept learning models and case-based models, which are interpretable by design, have achieved performance at par with black-box models in medical imaging applications. This has debunked the myth of compromise between performance and interpretability. It depends on the researcher's ability to discover the patterns in an interpretable way while at the same time allowing for flexibility to fit the data accurately.[51][50]

Interpretability of the deep neural networks should be considered as important as the performance. To promote trust in the DL solutions for medical image analysis tasks, it is important to involve clinicians in the model design process so that the algorithm becomes interpretable inherently, using multi-modal data (text, images) and annotating clinical concepts for diagnosis.

Multiple interpretability methods can be combined to understand the predictions of DL models. Interpretability of DL networks for segmentation is a difficult problem because attribution of a single pixel holds little significance. Hence, there are comparatively few interpretability methods available for segmentation networks. An additional model for interpretability can lead to an overly complicated decision system because two models may require troubleshooting instead of one.[51]

References

1. Babic, B.B., Gerke, S., Evgeniou, T., Cohen, I.G.: Beware explanations from ai in health care. Science **373**, 284–286 (7 2021). https://doi.org/10.1126/SCIENCE.ABG1834, https://www.science.org/doi/10.1126/science.abg1834
2. Bansal, N., Agarwal, C., Nguyen, A.: Sam: The sensitivity of attribution methods to hyperparameters. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops **2020-June**, 11–21 (6 2020). https://doi.org/10.1109/CVPRW50498.2020.00009
3. Bass, C., da Silva, M., Sudre, C., Tudosiu, P.D., Smith, S.M., Robinson, E.C.: Icam: Interpretable classification via disentangled representations and feature attribution mapping. Advances in Neural Information Processing Systems **2020-December** (6 2020), https://arxiv.org/abs/2006.08287v2
4. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 **2017-January**, 3319–3327 (4 2017). https://doi.org/10.1109/CVPR.2017.354, https://arxiv.org/abs/1704.05796v1
5. Biffi, C., Cerrolaza, J.J., Tarroni, G., Bai, W., Marvao, A.D., Oktay, O., Ledig, C., Folgoc, L.L., Kamnitsas, K., Doumou, G., Duan, J., Prasad, S.K., Cook, S.A., O'regan, D.P., Rueckert, D.: Explainable anatomical shape analysis through deep hierarchical generative models (2020). https://doi.org/10.5281/zenodo.3247898, https://github.com/UK-Digital-Heart-Project/lvae
6. Böhle, M., Eitel, F., Weygandt, M., Ritter, K.: Layer-wise relevance propagation for explaining deep neural network decisions in mri-based alzheimer's disease classification. Frontiers in Aging Neuroscience **10**, 194 (7 2019). https://doi.org/10.3389/FNAGI.2019.00194/BIBTEX
7. Chen, C., Li, O., Tao, C., Barnett, A.J., Su, J., Rudin, C.: This looks like that: Deep learning for interpretable image recognition. Advances in Neural Information Processing Systems **32** (6 2018), https://arxiv.org/abs/1806.10574v5
8. Chen, Z., Bei, Y., Rudin, C.: Concept whitening for interpretable image recognition. Nature Machine Intelligence **2**, 772–782 (12 2020). https://doi.org/10.1038/S42256-020-00265-Z
9. Chetoui, M., Akhloufi, M.A.: Explainable diabetic retinopathy using efficientnet. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS **2020-July**, 1966–1969 (7 2020). https://doi.org/10.1109/EMBC44109.2020.9175664
10. Chowdhury, A., Santamaria-Pang, A., Kubricht, J.R., Tu, P.: Emergent symbolic language based deep medical image classification (8 2020), https://arxiv.org/abs/2008.09860v1
11. Clough, J.R., Oksuz, I., Puyol-Antón, E., Ruijsink, B., King, A.P., Schnabel, J.A.: Global and local interpretability for cardiac mri classification. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11767 LNCS**, 656–664 (2019). https://doi.org/10.1007/978-3-030-32251-9_72
12. Cohen, J.P., Brooks, R., En, S., Zucker, E., Pareek, A., Lungren, M.P., Chaudhari, A.: Gifsplanation via latent shift: A simple autoencoder approach to counterfactual generation for chest x-rays. Proceedings of Machine Learning Research (2 2021), https://arxiv.org/abs/2102.09475v2

13. Ding, Y., Sohn, J.H., Kawczynski, M.G., Trivedi, H., Harnish, R., Jenkins, N.W., Lituiev, D., Copeland, T.P., Aboian, M.S., Aparici, C.M., Behr, S.C., Flavell, R.R., Huang, S.Y., Zalocusky, K.A., Nardo, L., Seo, Y., Hawkins, R.A., Pampaloni, M.H., Hadley, D., Franc, B.L.: A deep learning model to predict a diagnosis of alzheimer disease by using 18 f-fdg pet of the brain. Radiology **290**, 456–464 (3 2019), https://pubs.rsna.org/doi/10.1148/radiol.2018180958
14. Dinsdale, N.K., Jenkinson, M., Namburete, A.I.: Deep learning-based unlearning of dataset bias for mri harmonisation and confound removal. NeuroImage **228**, 117689 (3 2021). https://doi.org/10.1016/J.NEUROIMAGE.2020.117689
15. Du, M., Liu, N., Hu, X.: Techniques for interpretable machine learning. Communications of the ACM **63**, 68–77 (7 2018). https://doi.org/10.1145/3359786, https://arxiv.org/abs/1808.00033v3
16. Faust, K., Xie, Q., Han, D., Goyle, K., Volynskaya, Z., Djuric, U., Diamandis, P.: Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction. BMC Bioinformatics **19**, 1–15 (5 2018). https://doi.org/10.1186/S12859-018-2184-4/FIGURES/6, https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2184-4
17. Gamble, P., Jaroensri, R., Wang, H., Tan, F., Moran, M., Brown, T., Flament-Auvigne, I., Rakha, E.A., Toss, M., Dabbs, D.J., Regitnig, P., Olson, N., Wren, J.H., Robinson, C., Corrado, G.S., Peng, L.H., Liu, Y., Mermel, C.H., Steiner, D.F., Chen, P.H.C.: Determining breast cancer biomarker status and associated morphological features using deep learning. Communications Medicine **1** (7 2021). https://doi.org/10.1038/S43856-021-00013-3
18. Geirhos, R., Michaelis, C., Wichmann, F.A., Rubisch, P., Bethge, M., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. 7th International Conference on Learning Representations, ICLR 2019 (11 2018), https://arxiv.org/abs/1811.12231v3
19. Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J.H., Harrington, R.A., Liang, D.H., Ashley, E.A., Zou, J.Y.: Deep learning interpretation of echocardiograms. NPJ digital medicine **3** (12 2020). https://doi.org/10.1038/S41746-019-0216-8, https://pubmed.ncbi.nlm.nih.gov/31993508/
20. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**, 1789–1819 (6 2020). https://doi.org/10.1007/s11263-021-01453-z, http://arxiv.org/abs/2006.05525 http://dx.doi.org/10.1007/s11263-021-01453-z
21. Graziani, M., Andrearczyk, V., Müller, H.: Regression concept vectors for bidirectional explanations in histopathology. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11038 LNCS**, 124–132 (4 2019). https://doi.org/10.1007/978-3-030-02628-8_14, https://arxiv.org/abs/1904.04520v1
22. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable ai systems for the medical domain? (12 2017), https://arxiv.org/abs/1712.09923v1
23. Holzinger, A., Carrington, A., Müller, H.: Measuring the quality of explanations: The system causability scale (scs): Comparing human and machine explanations. KI - Kunstliche Intelligenz **34**, 193–198 (6 2020). https://doi.org/10.1007/S13218-020-00636-Z/TABLES/1, https://link.springer.com/article/10.1007/s13218-020-00636-z

24. Hua, K.L., Hsu, C.H., Hidayati, S.C., Cheng, W.H., Chen, Y.J.: Computer-aided classification of lung nodules on computed tomography images via deep learning technique. OncoTargets and therapy **8**, 2015 (8 2015). https://doi.org/10.2147/OTT.S80733, /pmc/articles/PMC4531007/ /pmc/articles/PMC4531007/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4531007/
25. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods 2020 18:2 **18**, 203–211 (12 2020). https://doi.org/10.1038/s41592-020-01008-z, https://www.nature.com/articles/s41592-020-01008-z
26. Izadyyazdanabadi, M., Belykh, E., Cavallo, C., Zhao, X., Gandhi, S., Moreira, L.B., Eschbacher, J., Nakaji, P., Preul, M.C., Yang, Y.: Weakly-supervised learning-based feature localization in confocal laser endomicroscopy glioma images. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11071 LNCS**, 300–308 (4 2018). https://doi.org/10.1007/978-3-030-00934-2_34, https://arxiv.org/abs/1804.09428v2
27. Janik, A., Dodd, J., Ifrim, G., Sankaran, K., Curran, K.: Medical imaging 2021: Image processing **11596** (2021)
28. Jiang, H., Xu, J., Shi, R., Yang, K., Zhang, D., Gao, M., Ma, H., Qian, W.: A multi-label deep learning model with interpretable grad-cam for diabetic retinopathy classification. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS **2020-July**, 1560–1563 (7 2020). https://doi.org/10.1109/EMBC44109.2020.9175884
29. Jiang, H., Yang, K., Gao, M., Zhang, D., Ma, H., Qian, W.: An interpretable ensemble deep learning model for diabetic retinopathy disease classification. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS pp. 2045–2048 (7 2019). https://doi.org/10.1109/EMBC.2019.8857160
30. Jo, T., Nho, K., Saykin, A.J.: Deep learning in alzheimer's disease: Diagnostic classification and prognostic prediction using neuroimaging data. Frontiers in Aging Neuroscience **11**, 220 (8 2019). https://doi.org/10.3389/FNAGI.2019.00220/BIBTEX
31. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M.K., Pei, J., Ting, M., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., Shi, A., Zhang, R., Zheng, L., Hou, R., Shi, W., Fu, X., Duan, Y., Huu, V.A., Wen, C., Zhang, E.D., Zhang, C.L., Li, O., Wang, X., Singer, M.A., Sun, X., Xu, J., Tafreshi, A., Lewis, M.A., Xia, H., Zhang, K.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**, 1122–1131.e9 (2 2018). https://doi.org/10.1016/J.CELL.2018.02.010, https://pubmed.ncbi.nlm.nih.gov/29474911/
32. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav) (7 2018), https://proceedings.mlr.press/v80/kim18d.html
33. Kim, E., Kim, S., Seo, M., Yoon, S.: Xprotonet: Diagnosis in chest radiography with global and local explanations (2021)
34. Kind, A., Azzopardi, G.: An explainable ai-based computer aided detection system for diabetic retinopathy using retinal fundus images. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelli-

gence and Lecture Notes in Bioinformatics) **11678 LNCS**, 457–468 (2019). https://doi.org/10.1007/978-3-030-29888-3_37/COVER

35. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models (11 2020), https://proceedings.mlr.press/v119/koh20a.html
36. Kumar, D., Taylor, G.W., Wong, A.: Discovery radiomics with clear-dr: Interpretable computer aided diagnosis of diabetic retinopathy. IEEE Access **7**, 25891–25896 (2019). https://doi.org/10.1109/ACCESS.2019.2893635
37. Lee, H., Kim, S.T., Ro, Y.M.: Generation of multimodal justification using visual word constraint model for explainable computer-aided diagnosis. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11797 LNCS**, 21–29 (6 2019). https://doi.org/10.1007/978-3-030-33850-3_3, https://arxiv.org/abs/1906.03922v1
38. Lenis, D., Major, D., Wimmer, M., Berg, A., Sluiter, G., Bühler, K.: Domain aware medical image classifier interpretation by counterfactual impact analysis. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **12261 LNCS**, 315–325 (7 2020). https://doi.org/10.1007/978-3-030-59710-8_31, https://arxiv.org/abs/2007.06312v2
39. Li, O., Liu, H., Chen, C., Rudin, C.: Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. Proceedings of the AAAI Conference on Artificial Intelligence **32**, 3530–3537 (4 2018). https://doi.org/10.1609/AAAI.V32I1.11771, https://ojs.aaai.org/index.php/AAAI/article/view/11771
40. Li, X., Dvornek, N.C., Zhou, Y., Zhuang, J., Ventola, P., Duncan, J.S.: Efficient interpretation of deep learning models using graph structure and cooperative game theory: Application to asd biomarker discovery. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11492 LNCS**, 718–730 (2019). https://doi.org/10.1007/978-3-030-20351-1_56/COVER
41. Lundberg, S.M., Allen, P.G., Lee, S.I.: A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems **30** (2017), https://github.com/slundberg/shap
42. Magesh, P.R., Myloth, R.D., Tom, R.J.: An explainable machine learning model for early detection of parkinson's disease using lime on datscan imagery. Computers in biology and medicine **126** (11 2020). https://doi.org/10.1016/J.COMPBIOMED.2020.104041, https://pubmed.ncbi.nlm.nih.gov/33074113/
43. Mohammadjafari, S., Cevik, M., Thanabalasingam, M., Basar, A., Initiative, A.D.N.: Using protopnet for interpretable alzheimer's disease classification. Proceedings of the Canadian Conference on Artificial Intelligence (6 2021). https://doi.org/10.21428/594757DB.FB59CE6C, https://caiac.pubpub.org/pub/klwhoig4/release/3
44. Mohankumar, A.K., Nema, P., Narasimhan, S., Khapra, M.M., Srinivasan, B.V., Ravindran, B.: Towards transparent and explainable attention models. Proceedings of the Annual Meeting of the Association for Computational Linguistics pp. 4206–4216 (4 2020). https://doi.org/10.18653/v1/2020.acl-main.387, https://arxiv.org/abs/2004.14243v1

45. Natekar, P., Kori, A., Krishnamurthi, G.: Demystifying brain tumor segmentation networks: Interpretability and uncertainty analysis. Frontiers in Computational Neuroscience **14**, 6 (2 2020). https://doi.org/10.3389/FNCOM.2020.00006/BIBTEX
46. Pereira, S., Meier, R., Alves, V., Reyes, M., Silva, C.A.: Automatic brain tumor grading from mri data using convolutional neural networks and quality assessment **11038** (9 2018). https://doi.org/10.1007/978-3-030-02628-8, http://arxiv.org/abs/1809.09468 http://dx.doi.org/10.1007/978-3-030-02628-8
47. Pintelas, E., Livieris, I.E., Pintelas, P.: A grey-box ensemble model exploiting black-box accuracy and white-box intrinsic interpretability. Algorithms 2020, Vol. 13, Page 17 **13**, 17 (1 2020). https://doi.org/10.3390/A13010017, https://www.mdpi.com/1999-4893/13/1/17/htm https://www.mdpi.com/1999-4893/13/1/17
48. Pisov, M., Goncharov, M., Kurochkina, N., Morozov, S., Gombolevsky, V., Chernina, V., Vladzymyrskyy, A., Zamyatina, K., Cheskova, A., Pronin, I., Shifrin, M., Belyaev, M.: Incorporating task-specific structural knowledge into cnns for brain midline shift detection. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11797 LNCS**, 30–38 (8 2019). https://doi.org/10.1007/978-3-030-33850-3_4, https://arxiv.org/abs/1908.04568v3
49. Puyol-Antón, E., Chen, C., Clough, J.R., Ruijsink, B., Sidhu, B.S., Gould, J., Porter, B., Elliott, M., Mehta, V., Rueckert, D., Rinaldi, C.A., King, A.P.: Interpretable deep models for cardiac resynchronisation therapy response prediction. Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention **2020**, 284–293 (2020). https://doi.org/10.1007/978-3-030-59710-8_28, https://pubmed.ncbi.nlm.nih.gov/34109325/
50. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 2019 1:5 **1**, 206–215 (5 2019). https://doi.org/10.1038/s42256-019-0048-x, https://www.nature.com/articles/s42256-019-0048-x
51. Salahuddin, Z., Woodruff, H.C., Chatterjee, A., Lambin, P.: Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Computers in Biology and Medicine **140**, 105111 (1 2022). https://doi.org/10.1016/J.COMPBIOMED.2021.105111
52. Samek, W., Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R.: Evaluating the visualization of what a deep neural network has learned. IEEE Transactions on Neural Networks and Learning Systems **28**, 2660–2673 (11 2017). https://doi.org/10.1109/TNNLS.2016.2599820
53. Samek, W., Montavon, G., Lapuschkin, S., Anders, C.J., Müller, K.R.: Explaining deep neural networks and beyond: A review of methods and applications. Proceedings of the IEEE **109**, 247–278 (3 2020). https://doi.org/10.1109/JPROC.2021.3060483, http://arxiv.org/abs/2003.07631 http://dx.doi.org/10.1109/JPROC.2021.3060483
54. Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., Xu, S., Barb, S., Joseph, A., Shumski, M., Smith, J., Sood, A.B., Corrado, G.S., Peng, L., Webster, D.R.: Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmol-

ogy **126**, 552–564 (4 2019). https://doi.org/10.1016/J.OPHTHA.2018.11.016, https://pubmed.ncbi.nlm.nih.gov/30553900/

55. Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., Xu, S., Barb, S., Joseph, A., Shumski, M., Smith, J., Sood, A.B., Corrado, G.S., Peng, L., Webster, D.R.: Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology **126**, 552–564 (4 2019). https://doi.org/10.1016/j.ophtha.2018.11.016, http://www.aaojournal.org/article/S0161642018315756/fulltext http://www.aaojournal.org/article/S0161642018315756/abstract https://www.aaojournal.org/article/S0161-6420(18)31575-6/abstract
56. Schutte, K., Moindrot, O., Hérent, P., Schiratti, J.B., Jégou, S.: Using stylegan for visual interpretability of deep learning models on medical images (1 2021), https://arxiv.org/abs/2101.07563v1
57. Seah, J.C., Tang, J.S., Kitchen, A., Gaillard, F., Dixon, A.F.: Chest radiographs in congestive heart failure: Visualizing neural network learning. Radiology **290**, 514–522 (3 2019). https://doi.org/10.1148/RADIOL.2018180887, https://pubmed.ncbi.nlm.nih.gov/30398431/
58. Seegerer, P., Binder, A., Saitenmacher, R., Bockmayr, M., Alber, M., Jurmeister, P., Klauschen, F., Müller, K.R.: Interpretable deep neural network to predict estrogen receptor status from haematoxylin-eosin images. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **12090 LNCS**, 16–37 (2020). https://doi.org/10.1007/978-3-030-50402-1_2
59. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision **128**, 336–359 (10 2016). https://doi.org/10.1007/s11263-019-01228-7, http://arxiv.org/abs/1610.02391 http://dx.doi.org/10.1007/s11263-019-01228-7
60. Sengupta, S., Singh, A., Leopold, H.A., Gulati, T., Lakshminarayanan, V.: Ophthalmic diagnosis using deep learning with fundus images – a critical review. Artificial Intelligence in Medicine **102**, 101758 (1 2020). https://doi.org/10.1016/J.ARTMED.2019.101758
61. Serrano, S., Smith, N.A.: Is attention interpretable? ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference pp. 2931–2951 (6 2019). https://doi.org/10.18653/v1/p19-1282, https://arxiv.org/abs/1906.03731v1
62. Shen, S., Han, S.X., Aberle, D.R., Bui, A.A., Hsu, W.: An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. Expert systems with applications **128**, 84–95 (8 2019). https://doi.org/10.1016/J.ESWA.2019.01.048, https://pubmed.ncbi.nlm.nih.gov/31296975/
63. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. 2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings (12 2013), https://arxiv.org/abs/1312.6034v2
64. Singh, A., Sengupta, S., Lakshminarayanan, V.: Explainable deep learning models in medical image analysis
65. Singh, G., Yow, K.C.: An interpretable deep learning model for covid-19 detection with chest x-ray images. IEEE access : practical innovations, open so-

lutions **9**, 85198–85208 (2021). https://doi.org/10.1109/ACCESS.2021.3087583, https://pubmed.ncbi.nlm.nih.gov/35256923/
66. Singla, S., Eslami, M., Pollack, B., Wallace, S., Batmanghelich, K.: Explaining the black-box smoothly- a counterfactual approach. Medical Image Analysis **84** (1 2021). https://doi.org/10.1016/j.media.2022.102721, https://arxiv.org/abs/2101.04230v3
67. Sun, J., Darbehani, F., Zaidi, M., Wang, B.: Saunet: Shape attentive u-net for interpretable medical image segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **12264 LNCS**, 797–806 (1 2020). https://doi.org/10.1007/978-3-030-59719-1_77, https://arxiv.org/abs/2001.07645v3
68. Tang, Y., Tang, Y., Zhu, Y., Xiao, J., Summers, R.M.: A disentangled generative model for disease decomposition in chest x-rays via normal image synthesis. Medical image analysis **67** (1 2021). https://doi.org/10.1016/J.MEDIA.2020.101839, https://pubmed.ncbi.nlm.nih.gov/33080508/
69. Teng, Q., Liu, Z., Song, Y., Han, K., Lu, Y.: A survey on the interpretability of deep learning in medical diagnosis. Multimedia Systems **28**, 2335 (12 2022). https://doi.org/10.1007/S00530-022-00960-4, /pmc/articles/PMC9243744/ /pmc/articles/PMC9243744/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9243744/
70. Thomas, S.M., Lefevre, J.G., Baxter, G., Hamilton, N.A.: Interpretable deep learning systems for multi-class segmentation and classification of non-melanoma skin cancer. Medical Image Analysis **68**, 101915 (2 2021). https://doi.org/10.1016/J.MEDIA.2020.101915
71. de la Torre, J., Valls, A., Puig, D.: A deep learning interpretable classifier for diabetic retinopathy disease grading. Neurocomputing **396**, 465–476 (7 2020). https://doi.org/10.1016/J.NEUCOM.2018.07.102
72. Wang, J., Gou, L., Zhang, W., Yang, H., Shen, H.W.: Deepvid: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation. IEEE transactions on visualization and computer graphics **25**, 2168–2180 (6 2019). https://doi.org/10.1109/TVCG.2019.2903943, https://pubmed.ncbi.nlm.nih.gov/30892211/
73. Weaver, O., Leung, J.W.: Biomarkers and imaging of breast cancer. https://doi.org/10.2214/AJR.17.18708 **210**, 271–278 (11 2017). https://doi.org/10.2214/AJR.17.18708, www.ajronline.org
74. Wiegreffe, S., Pinter, Y.: Attention is not not explanation. EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference pp. 11–20 (8 2019). https://doi.org/10.18653/v1/d19-1002, https://arxiv.org/abs/1908.04626v2
75. Yang, G., Ye, Q., Xia, J.: Unbox the black-box for the medical explainable ai via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. Information Fusion **77**, 29–52 (1 2022). https://doi.org/10.1016/J.INFFUS.2021.07.016
76. Yang, J., Dvornek, N.C., Zhang, F., Zhuang, J., Chapiro, J., Lin, M., Duncan, J.S.: Domain-agnostic learning with anatomy-consistent embedding for cross-modality liver segmentation. ... IEEE International Conference on Computer Vision workshops. IEEE International Conference on Computer Vision **2019**, 323–331 (10 2019). https://doi.org/10.1109/ICCVW.2019.00043, /pmc/articles/PMC8528125/ /pmc/articles/PMC8528125/?report=abstract https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8528125/

77. Yeche, H., Harrison, J., Berthier, T.: Ubs: A dimension-agnostic metric for concept vector interpretability applied to radiomics. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11797 LNCS**, 12–20 (2019). https://doi.org/10.1007/978-3-030-33850-3_2
78. Zhang, Q., Rao, L., Yang, Y.: Group-cam: Group score-weighted visual explanations for deep convolutional networks (3 2021), https://arxiv.org/abs/2103.13859v4
79. Zhang, Z., Chen, P., McGough, M., Xing, F., Wang, C., Bui, M., Xie, Y., Sapkota, M., Cui, L., Dhillon, J., Ahmad, N., Khalil, F.K., Dickinson, S.I., Shi, X., Liu, F., Su, H., Cai, J., Yang, L.: Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nature Machine Intelligence **1**, 236–245 (5 2019). https://doi.org/10.1038/S42256-019-0052-1
80. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2016-December**, 2921–2929 (12 2016). https://doi.org/10.1109/CVPR.2016.319