INFORMER- Interpretability Founded Monitoring of Medical Image Deep Learning Models

Shelley Zixin Shu[1], Aurélie Pahud de Mortanges[1], Alexander Poellinger[23], Dwarikanath Mahapatra[4], and Mauricio Reyes[15]

[1] ARTORG Center for Biomedical Engineering Research, University of Bern, Murtenstrasse 50, Bern 3008, Switzerland
[2] Inselspital (Bern University Hospital), 3010 Bern, Switzerland
[3] Insel Gruppe Bern Universitätsinstitut für Diagnostische, Interventionelle und Pädiatrische Radiologie
[4] Inception Institute of Artificial Intelligence, Abu Dhabi, UAE
[5] Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, Bern, Switzerland

Abstract. Deep learning models have gained significant attention due to their promising performance in medical image tasks. However, a gap remains between experimental accuracy and real-world applications. The inherited black-box nature of the deep learning model introduces uncertainty, trustworthy issues, and difficulties in performing quality control of deployed deep learning models. While quality control methods focusing on uncertainty estimation for segmentation tasks exist, there are comparatively fewer approaches for classification, particularly in multi-label datasets. This paper addresses this gap by proposing a quality control method that bridges interpretability and uncertainty estimation through a graph-based class distinctiveness calculation. Using the CheXpert dataset, the proposed approach achieved a higher F_1 score on the bootstrapped test set compared to baselines quality control approaches based on predictive entropy and test-time augmentation.

Keywords: Interpretability · Quality Control · Multi-label Classification · Medical Images · Deep learning.

1 Introduction

Deep learning has revolutionized the field of medical imaging with improved performance and next-level inference capabilities [24,33,14,19]. However, this also comes at the expense of increased system complexity and a black-box model perception, wherein explaining the reasoning behind model predictions is a highly complex task. This makes the process of auditing or verifying the reliability of deep learning model outputs complex, yet, extremely necessary for medical imaging applications due to the high-stake nature of the healthcare sector.

Quality control of deep learning models has been explored for medical image segmentation using uncertainty estimation of model predictions [13,6,7], or

regressing directly a metric of segmentation quality [27,31]. Similarly, for medical image classification, uncertainty-based quality control approaches have been proposed [3]. However, it is known that the reliability of uncertainty estimations depends on the calibration properties of the deep learning model [13,12,8], and that modern neural networks tend to be overconfident [9,16]. While there exist several strategies to calibrate a deep learning model, it remains an area of active research and highly empirical in practice [29,28,18]. The situation becomes more complex for multi-label tasks, in which a patient can present more than one condition (multiple class labels per sample). In this scenario, it has been shown that model calibration is harder than in multi-class tasks [5].

In this paper, inspired by [17] using interpretability information to guide the training process of deep learning models, we propose an interpretability-driven graph-based quality control method. In contrast to [17], we did not train the model with a class-distinctiveness objective but assessed the class-distinctiveness to flag the model uncertainty. Specifically, we build on information from class-specific saliency maps to derive criteria to flag potentially wrong model predictions. Class-specific saliency maps are basically heatmaps representing pixel-wise attribution levels of a model classifying a sample into a specific class. The rationale to use class-specific saliency maps for quality control lies in the intuition presented in [17] where a well-trained model is characterized by distinctive class-specific saliency maps. Conversely, a poorly trained model - in our scenario, a low-confidence model - is characterized by similar class-specific saliency maps. Building on this concept, we propose a simple yet effective quality control approach to monitor classification models. In this study, we present results for the challenging scenario of multi-label classification from chest X-ray images.

In the following, we present the proposed approach, termed INFORMER, for **In**terpretability **Fo**unded Monito**r**ing of **Me**dical Image Deep Lea**R**ning Models, followed by experiments and benchmarking on the publically available CheXPert dataset [11].

2 Methodology

With INFORMER, we propose an interpretability-based method to determine whether a model's output is flagged as wrong. Fig. 1 summarises the proposed INFORMER approach. Given a trained deep learning model under inspection and test samples, a prediction alarm module computes class-specific saliency maps and aggregates their information into a score to flag potentially wrong predictions. The prediction alarm module is fine-tuned to the inspected model via a validation set (not necessarily the same one used for the inspected model). We note that the class-specific saliency maps can be computed with any available saliency map generator (e.g., LRP [1], Input $\times$ Gradient [25], GradCam [23], etc.).

In the next section, we describe how class-specific saliency maps are aggregated into a single score and fine-tuned via a validation set.

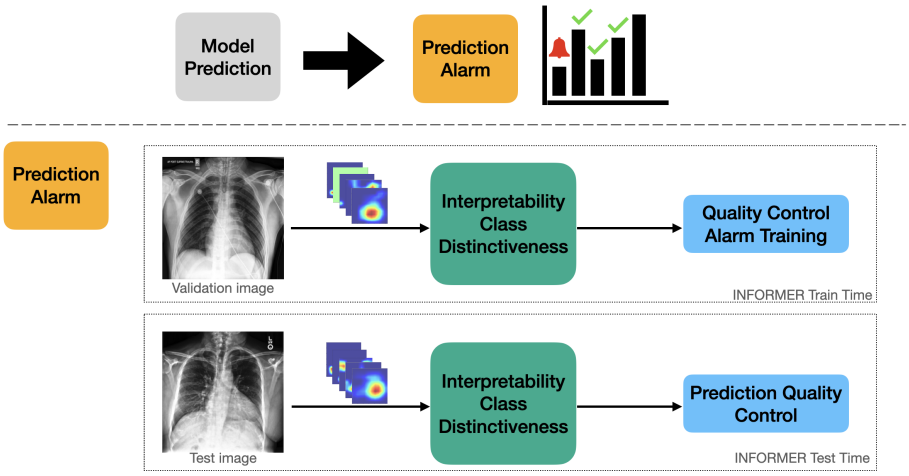

Fig. 1. Proposed **In**terpretability **Fo**unded Monito**r**ing of **Me**dical Image Deep Lea**R**ning Models (INFORMER). Upper-row: Given a trained model under inspection and test samples, a prediction alarm module detects potentially wrong predictions. Lower-row: The prediction alarm computes class-specific saliency maps, and their information is aggregated into a single flagging score, which is fine-tuned to the inspected model via a validation set (INFORMER Train Time). During test time, the fine-tuned prediction alarm is coupled to the inspected model to flag potentially wrong predictions.

Interpretability Class Distinctiveness: Given a test image I, an inspected classification model M, and a class-specific saliency map generator G, we obtain class-specific saliency maps $S_{i,i\in C} = G(I, M)$, where C is the total number of classes. We then calculate the pairwise cosine distance between saliency maps as:

$$d_{ij} = 1 - CoSim(S_i, S_k) \tag{1}$$

where $i \neq j$, $i, j \in C$ and $CoSim(.)$ is the cosine similarity, ranged [−1,1].

For the saliency map of class i, a set of pairwise comparisons with length *C-1* is calculated as $\mathbf{v_i} = \{d_{ij}\}, j \in C, j \neq i$. After computing such class distinctiveness for every class, the results are C pairwise comparison vectors, as is shown in Fig. 2. To generalize the class-distinctiveness for every class, the entries of every vector are summed to obtain the corresponding representation d_i for class i, i.e., $||v_0||_1 = d_0$. The representation vector $\mathbf{d} = \{d_i\}_{i\in C}$ encodes how difficult a given sample can be confidently classified by the inspected model for every class label.

Quality Control Alarm Similar to [3], we use thresholds to fine-tune our quality control module. The quality control module finds the best threshold τ_i for class $i \in C$ in the validation set achieving the highest F_1 after swapping flagged predictions. The flagged predictions are the ones marked to be highly uncertain.

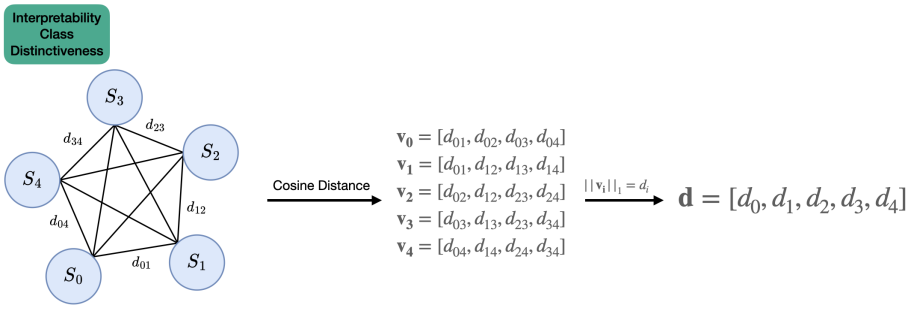

Fig. 2. Illustration of interpretability class distinctiveness: node S_i represents the saliency map for the chosen disease i for a given patient. The edge between S_i and S_k, labeled d_{ik}, denotes the cosine distance between the corresponding saliency maps calculated as Equation 1. Every disease i has a pairwise vector representation $\mathbf{v_i}$. All pairwise distances between diseases, excluding self-connections (i.e., d_{ii}), are calculated. The entries of every vector v_i are summed to obtain the generalized class distinctiveness representation d_i. The vector $\mathbf{d}$, the class-distinctiveness representation, encodes how difficult a given sample is to be confidently classified by the inspected model for every class label and is used for quality control.

After fine-tuning, each sample is passed through the fine-tuned alarm module at test time to evaluate the prediction. Every class is evaluated separately.

3 Experiments

Dataset and Evaluation Details

We used the CheXpert dataset [11] to evaluate the proposed approach. The CheXpert dataset comprises 224,316 images from 65,240 patients. This includes 223,414 training images, 234 validation images, and 668 test images. The dataset contains frontal and lateral images, but we trained the model exclusively with frontal images in this study. The labels for the training set were automatically generated based on patient reports. Images with uncertain labels were excluded from the study. If a disease is not mentioned in the report, it is considered negative for that disease in our implementation. The model evaluation focuses on the five diseases selected in the CheXPert challenge: atelectasis, cardiomegaly, consolidation, edema, and pleural effusion. Patients without any of these five pathologies were excluded from the analysis for the prevalence. Following this selection and labeling strategy, the dataset includes 90,839 training images, 128 validation images, and 284 test images. To evaluate the robustness of all tested methods, we assessed them using 30 bootstrapped test sets. A bootstrapped test set is obtained with sampling with replacement till the original size of the test set, i.e. 284 images in every bootstrapped test set.

We generated the model's saliency maps using a gradient-based interpretability method. To demonstrate the method's robustness across different interpretability approaches, we employed two methods in our experiments based: Layer-wise Relevance Propagation (lrp) [1] and Input × Gradient (ixg) [25]. These are referred to as INFORMER_lrp and INFORMER_ixg, respectively.

Benchmarking Baselines

Predictive Entropy: We use predictive entropy as a baseline approach, which is also used by G. Carneiro [3] for quality control for medical images. A higher predictive entropy indicates greater uncertainty in the prediction, which is then flagged. In multi-label classification problems, every prediction is treated as binary for a selected sample. Then, the predictive entropy for class i for one sample is calculated as

$$H(x_i) = -p_i log_2 p_i - (1 - p_i) log_2 (1 - p_i), \tag{2}$$

where p_i is the logit output after sigmoid normalisation for class i. log_2 is the log base 2 operator. Every sample has a vector with length C to represent the predictive entropy of the sample.

Test-Time Augmentation (TTA): Test-time augmentation (TTA) has been proposed as an alternative to compute uncertainty estimation. This method applies transformations to samples during the testing phase and analyses the variation in the logits across these transformations [30]. We use TTA-based uncertainty estimation as the second baseline.

Implementation

We use DenseNet121 [10], pre-trained on ImageNet from PyTorch 2.0.0+cu118 [20]. Learning rate at 0.0001, and images resized to 320 × 320 pixels. Batch size is 16. We apply data augmentation techniques during training, including affine transformations, rotation, box blur, and cropping. The model is trained for three epochs, and the final model, which achieved an average AUC of 0.806 on the validation set and 0.796 on the test set, is saved for evaluation purposes. The binary cross entropy loss is used for model training. The prediction threshold is determined based on the best Youden index from the validation set. Saliency maps are generated using Captum [15]. For TTA, we applied random rotations of 45 degrees, bluring with kernel size 10 by 10 and random cropping with a padding 8.

The quality control alarm module of INFORMER flags highly uncertain cases according to the threshold τ_i, described in the methodology section. In practice, we performed a grid-search using 0.001 intervals, covering the range from 0 to the maximum d_i values for each class. Similarly, for the two baselines, thresholds are found with a grid-search using 0.001 intervals.

Results

We evaluated our method across datasets with different disease distributions. We plotted the mean F_1 score across five classes for each fold, as shown in Fig. 3. The median F_1 score for the predictive entropy and TTA are 0.564. In comparison, the median F_1 scores for INFORMER_ixg, and INFORMER_lrp are 0.573, and 0.572, respectively. Both variants of INFORMER demonstrate improved performance in terms of mean F_1 score on the bootstrapped test sets. Notably, the first quartile of both INFORMER_ixg and INFORMER_lrp also exceeds the baseline median, indicating that both INFORMER variants perform better and more robustly than the baselines.

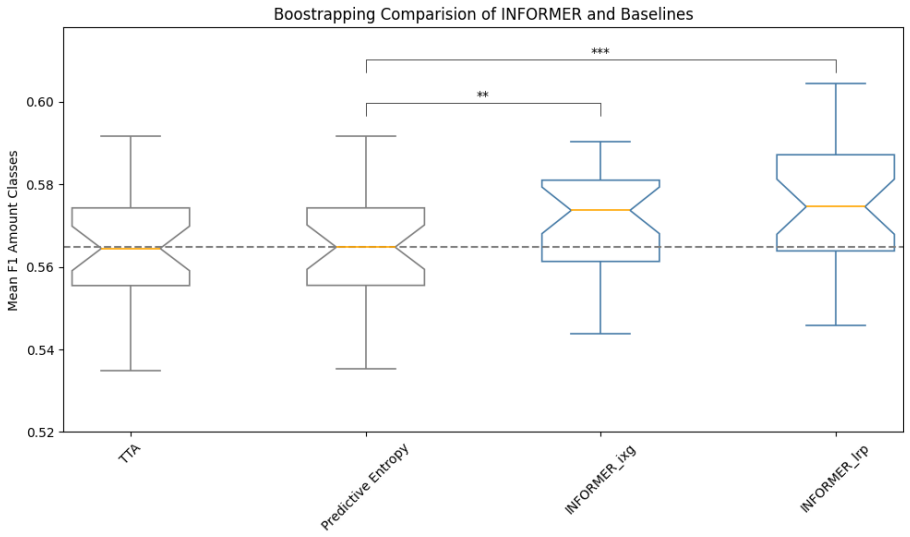

Fig. 3. Boxplot illustrating the results of the bootstrap experiment, highlighting the variability across bootstrapped test sets. The y-axis represents the mean F1 score across five diseases. Predictive Entropy and TTA are the baseline methods (colored in grey). The proposed methods are colored in blue. INFORMER_lrp refers to the proposed method with LRP-generated saliency maps and INFORMER_ixg indicates the method with Input $\times$ Gradient-generated saliency maps. The stars represent the statistical significance between the paired t-test of INFORMER variants and predictive entropy. P-value ≤ 0.01 is $**$ and P-value ≤ 0.001 is $***$.

Besides the bootstrapping experiment, we evaluated our method using the F_1 score on the original test set for direct comparison of metrics, as shown in Table 1. Overall, INFORMER outperformed both versions of Baseline. INFORMER_lrp demonstrated the best overall performance, with a mean F_1 score of 0.574. The predictive entropy and test-time augmentation have mean F_1 at 0.563 over five classes. We did not find performance differences between the INFORMER and

the Baselines observed for Edema and Pleural Effusion. This is because none of the methods detected any uncertain prediction for these classes in this test set. The most significant improvement over the baselines was seen for cardiomegaly, with a gap of 0.048 in F_1 score between INFORMER_lrp and the baselines.

Table 1. Performance measuring in F_1 comparison between INFORMER with different interpretability methods and baselines. INFORMER_ixg is INFORMER with input-times-gradient generated saliency map. INFORMER_lrp is with LRP generated saliency map. Pred. Entropy corresponds to the predictive entropy baseline, and TTA is the test-time augmentation baseline. P.Effusion is pleural effusion. Bold indicates the best result in the pathology.

Model	Atelectasis	Cardiomegaly	Consolidation	Edema	P.Effusion	$\overline{F_1}$
INFORMER$_{ixg}$	**0.648**	0.628	**0.240**	0.609	0.732	0.569
INFORMER$_{lrp}$	0.636	**0.655**	0.230	0.609	0.732	**0.574**
Pred. Entropy	0.636	0.607	0.231	0.609	0.732	0.563
TTA	0.636	0.607	0.230	0.609	0.732	0.563

Clinical Evaluation: We observed that INFORMER_ixg features a higher detection sensitivity (SENS) for atelectasis cases at the cost of a lower specificity (SPEC) compared to the baselines. However, this trade-off is preferable clinically. Atelectasis can have various causes, many of which can be relieved with chest physical therapy like breathing exercises, change of body position, or movement. These measures are well-suited for most hospitalized patients and contribute to well-being independent of the existence of actual atelectasis. Therefore, maximizing the sensitivity of atelectasis detection, even at the cost of a lower specificity, is preferable.

Regarding consolidation, the INFORMER_lrp method exceeds the baselines in sensitivity (SENS: Baselines 0.690, INFORMER_lrp 0.724, SPEC: predictive entropy 0.514, INFORMER_lrp 0.510). It has been estimated that >1.5 million unique adults are being hospitalized annually in the US for community-acquired pneumonia (CAP) [21]. Assuming that 1'500'000 patients receive a chest X-ray (CXR) on which consolidation is visible as a sign of CAP, the baseline method with a SENS of 0.690 would correctly detect 1'035'000 cases. In contrast, INFORMER_lrp method would correctly detect 1'086'000 cases. This means an additional 51'000 cases could be detected, potentially preventing the progression of undiagnosed CAP to complications like respiratory failure or sepsis. The table specificity and sensitivity of the model are in supplementary Table S1 and S2, respectively.

4 Discussion

We proposed an interpretability-based prediction quality control approach for multi-label classification problems. From experiments, we observed improve-

ments over two popular baselines used to flag potentially incorrect model predictions. The bootstrap result also demonstrates the robustness of the performance across test sets. From a clinical scenario point of view, improving the sensitivity by 0.034 can result in an estimate of approximately 51'000 additional consolidation cases that could annually be detected to prevent further progression, which we believe shows the clinical value of the obtained results.

Most multi-label studies only report AUC, as the prediction varies between the chosen thresholds. Here, we use the Youden index [32,26] to define the prediction. The result was evaluated with F_1, which considers class imbalance compared to accuracy. We can see that in the model with a mean AUC over 0.8 on the test set, the F_1 score is just over 0.65. That emphasizes the importance of quality control of model predictions towards deployment of deep learning based classification models.

Entropy and test-time augmentation are the traditional metrics used for quality control of deep learning models in the field. Our method using saliency map derived information outperformed both methods. The proposed method is independent of the model architecture, and focuses on the information inherited in the model gradients. In contrast to the baseline approaches, necessitating the model's logit output, INFORMER can be applied to a black box model in which the model logit is not available by using interpretability designed to work on black-box models, such as LIME [22]. This can be the case for enterprise-level models that need to be audited, but their inner information is unavailable.

There are some limitations to this method that are worth mentioning. We also experimented with the method with Grad-Cam [23] but the performance was worse than the reported methods. We think this outcome is due to the unfocused saliency map generation nature of Grad-Cam, as also reported by others [4,2]. Similar methods such as Grad-Cam++ [4] are also worth exploring in the future as they yield more spatially focused saliency maps. Further research on the saliency map's precision and the consistency of such evaluation with the clinicians is also needed. We also anticipate improvements by replacing the threshold-based approach used here to flag cases with other systematic methods, such as a multi-step threshold approach, or directly using the cosine distances from the pairwise comparison in conjunction with a machine learning classification model.

5 Conclusion

Black-box models perform exceptionally on medical images in state-of-the-art research. However, applying these models in the medical domain without understanding their uncertainty brings significant risks. While there has been some research on quality control for segmentation tasks, there is limited work on classification. We explore in this paper using the interpretability information in the saliency map to drive quality control for multi-label classification beyond only using the logits. The proposed method, INFORMER, outperformed existing baselines in terms of mean F_1 score by utilizing interpretability-based information.

Moreover, INFORMER demonstrated more robust performance on the bootstrap dataset, indicating its higher reliability and effectiveness in classification tasks than current baseline approaches.

Acknowledgments

We acknowledge funding by the Swiss National Science Foundation (project number 212939). We report no financial relationship or conflicts of interest.

References

1. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one **10**(7), e0130140 (2015)
2. Brandt, R., Raatjens, D., Gaydadjiev, G.: Precise benchmarking of explainable ai attribution methods. arXiv preprint arXiv:2308.03161 (2023)
3. Carneiro, G., Pu, L.Z.C.T., Singh, R., Burt, A.: Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. Medical image analysis **62**, 101653 (2020)
4. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE winter conference on applications of computer vision (WACV). pp. 839–847. IEEE (2018)
5. Cheng, J., Vasconcelos, N.: Towards calibrated multi-label deep neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27589–27599 (2024)
6. DeVries, T., Taylor, G.W.: Leveraging uncertainty estimates for predicting segmentation quality. arXiv preprint arXiv:1807.00502 (2018)
7. Eaton-Rosen, Z., Bragman, F., Bisdas, S., Ourselin, S., Cardoso, M.J.: Towards safe deep learning: accurately quantifying biomarker uncertainty in neural network predictions. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I. pp. 691–699. Springer (2018)
8. Faghani, S., Moassefi, M., Rouzrokh, P., Khosravi, B., Baffour, F.I., Ringler, M.D., Erickson, B.J.: Quantifying uncertainty in deep learning of radiologic images. Radiology **308**(2), e222217 (2023)
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
10. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
11. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
12. Jungo, A., Balsiger, F., Reyes, M.: Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Frontiers in neuroscience **14**, 501743 (2020)

13. Jungo, A., Reyes, M.: Assessing reliability and challenges of uncertainty estimations for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22. pp. 48–56. Springer (2019)
14. Kelly, B.S., Judge, C., Bollard, S.M., Clifford, S.M., Healy, G.M., Aziz, A., Mathur, P., Islam, S., Yeom, K.W., Lawlor, A., et al.: Radiology artificial intelligence: a systematic review and evaluation of methods (raise). European radiology **32**(11), 7998–8007 (2022)
15. Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., et al.: Captum: A unified and generic model interpretability library for pytorch. arXiv preprint arXiv:2009.07896 (2020)
16. Lambert, B., Forbes, F., Doyle, S., Dehaene, H., Dojat, M.: Trustworthy clinical ai solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. Artificial Intelligence in Medicine p. 102830 (2024)
17. Mahapatra, D., Poellinger, A., Reyes, M.: Interpretability-guided inductive bias for deep learning based medical image. Medical image analysis **81**, 102551 (2022)
18. Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., Lucic, M.: Revisiting the calibration of modern neural networks. Advances in Neural Information Processing Systems **34**, 15682–15694 (2021)
19. Oren, O., Gersh, B.J., Bhatt, D.L.: Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. The Lancet Digital Health **2**(9), e486–e488 (2020)
20. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
21. Ramirez, J.: Adults hospitalized with pneumonia in the united states: Incidence, epidemiology, and mortality. Clincal Infectious Diseases (2017)
22. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
23. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
24. Shen, D., Wu, G., Suk, H.I.: Deep learning in medical image analysis. Annual review of biomedical engineering **19**, 221–248 (2017)
25. Shrikumar, A., Greenside, P., Shcherbina, A., Kundaje, A.: Not just a black box: Learning important features through propagating activation differences. arXiv preprint arXiv:1605.01713 (2016)
26. Sun, S., Woerner, S., Maier, A., Koch, L.M., Baumgartner, C.F.: Inherently interpretable multi-label classification using class-specific counterfactuals. arXiv preprint arXiv:2303.00500 (2023)
27. Valindria, V.V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B.: Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. IEEE transactions on medical imaging **36**(8), 1597–1606 (2017)
28. Varoquaux, G., Cheplygina, V.: Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ digital medicine **5**(1), 48 (2022)

29. Wang, C.: Calibration in deep learning: A survey of the state-of-the-art. arXiv preprint arXiv:2308.01222 (2023)
30. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing **338**, 34–45 (2019)
31. Wang, S., Nuyts, J., Filipovic, M.: Uncertainty estimation in liver tumor segmentation using the posterior bootstrap. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 188–197. Springer (2023)
32. Youden, W.J.: Index for rating diagnostic tests. Cancer **3**(1), 32–35 (1950)
33. Zhou, S.K., Greenspan, H., Shen, D.: Deep learning for medical image analysis. Academic Press (2023)