# Explainable Image Classification with Improved Trustworthiness for Tissue Characterisation

Alfie Roddan[1(✉)], Chi Xu[1], Serine Ajlouni[2], Irini Kakaletri[3], Patra Charalampaki[2,4], and Stamatia Giannarou[1]

[1] The Hamlyn Centre for Robotic Surgery, Department of Surgery and Cancer, Imperial College London, London, UK
{a.roddan21,chi.xu20,stamatia.giannarou}@imperial.ac.uk
[2] Medical Faculty, University Witten Herdecke, Witten , Germany
[3] Medical Faculty, Rheinische Friedrich Wilhelms University of Bonn, Bonn, Germany
[4] Department of Neurosurgery, Cologne Medical Center, Cologne, Germany

**Abstract.** The deployment of Machine Learning models intraoperatively for tissue characterisation can assist decision making and guide safe tumour resections. For the surgeon to trust the model, explainability of the generated predictions needs to be provided. For image classification models, pixel attribution (PA) and risk estimation are popular methods to infer explainability. However, the former method lacks trustworthiness while the latter can not provide visual explanation of the model's attention. In this paper, we propose the first approach which incorporates risk estimation into a PA method for improved and more trustworthy image classification explainability. The proposed method iteratively applies a classification model with a PA method to create a volume of PA maps. We introduce a method to generate an enhanced PA map by estimating the expectation values of the pixel-wise distributions. In addition, the coefficient of variation (CV) is used to estimate pixel-wise risk of this enhanced PA map. Hence, the proposed method not only provides an improved PA map but also produces an estimation of risk on the output PA values. Performance evaluation on probe-based Confocal Laser Endomicroscopy (pCLE) data verifies that our improved explainability method outperforms the state-of-the-art.

**Keywords:** Explainability · Uncertainty · MC Dropout

## 1 Introduction

When using a Machine Learning (ML) model during intraoperative tissue characterisation, it is vital that the surgeon is able to assess how reliable a model's prediction is [8]. For the surgeon to trust the output predictions of the model, the model must be able to explain itself reliably in a clinical scenario [2]. To assess an


H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14221, pp. 575–585, 2023.
https://doi.org/10.1007/978-3-031-43895-0_54

explainability method we consider five metrics of performance: speed, usability, generalisability, trustworthiness and ability to localise semantic features. The explanation of a model's predictions is trustworthy if small perturbations in the input or model parameters, results in a similar output explanation. One form of explainability in the image classification domain is pixel attribution (PA) mapping. PA maps aim to highlight the "most important" pixels to the classification. PA maps can be used to visually highlight whether a model is poorly extracting semantic features [32] and/or that the model is misinformed due to spurious correlations within the data that it was trained on [16]. To efficiently process image data, these methods mainly rely on Convolutional Neural Networks (CNNs) and achieve state-of-the-art (SOTA) performance. One of the first PA methods proposed for CNNs was class activation maps (CAM) [33]. CAM uses one forward pass of the model to find the channels in the last convolutional layer that contributed most to the prediction. One of CAM's limitations is its reliance on global average pooling (GAP) [21] after the last convolutional layer as it dramatically reduces the number of architectures that can use CAM. To improve on this, Grad-CAM [30] generalises to all CNN architectures which are differentiable from the output logit layer to the chosen convolutional layer. However, Grad-CAM often lacks sharpness in object localisation, as noted and improved on in Grad-CAM++ [6] and SmoothGrad-CAM++ [24]. These extensions of Grad-CAM have good semantic feature localisation but they are unable to be deployed for use in surgery [5]. Both Score-CAM [31] and Recipro-CAM [5] also generalise to all CNN architectures but are deployable. Score-CAM improves on object localisation within the visual PA map without losing the class specific capabilities of Grad-CAM by masking out regions of the image and measuring the change in the output score. This is similar to perturbation methods like RISE [26], LIME [28] and other perturbation techniques [3,32]. On the other hand, Recipro-CAM focuses on the speed of PA map computation whilst maintaining comparable SOTA performance. By utilising the CNN's receptive field, Recipro-CAM generates a number of spatial masks and then measures the effect on the output score much like Score-CAM.

Despite being speedy, easy to deploy and able to localise semantic features, the above methods lack trustworthiness due to the training strategy of their underlying model. Deep learning (DL) models trained with empirical risk minimisation (ERM) are overconfident in prediction [12] and vulnerable to adversarial attacks [13]. Bayesian Neural Networks (BNNs) [23] bring improved regularisation and output uncertainty estimates. Unfortunately, the non-linearity and number of variables within NNs make Bayesian inference a computationally intensive task. For this reason, variational methods [15,18] are used to approximate Bayesian inference. More recently, the variational method Bayes by Backprop [4] used Dropout [19] to approximate Bayesian inference. Dropout is a regularisation technique which has also been noted to improve salient feature extraction. Although Bayes by Backprop is trustworthy, it often fails to scale to the complex architectures of SOTA models. To improve on this lack of generalisability, another variational method called Monte Carlo (MC) Dropout [12]

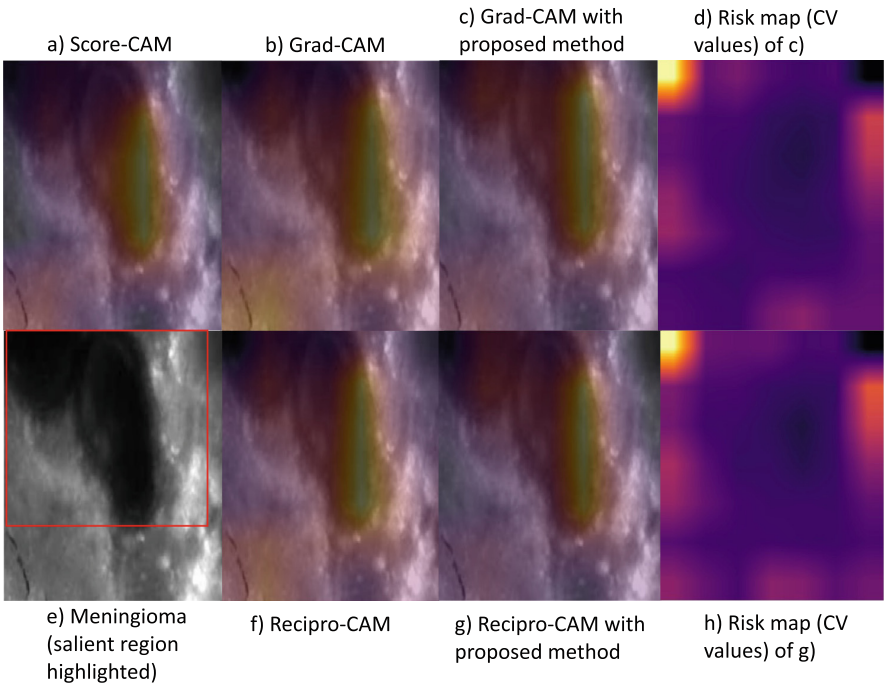


**Fig. 1.** PA maps generated using ResNet18 on meningioma pCLE data. a) Score-CAM PA map b) Grad-CAM PA map c) Grad-CAM PA map with our method applied d) Risk map (CV values) of c) e) Meningioma with the salient region highlighted with red bounding box f) Recipro-CAM PA map g) Recipro-CAM PA map with our method applied h) Risk map (CV values) of f). Yellow represents the highest PA value and black the lowest. (Color figure online)

proposes that a model trained with Dropout is equivalent to a probabilistic deep Gaussian process [7,11]. With this assumption, an estimated output distribution is computed after a number of forward passes with Dropout have been applied. This output distribution is used in practice to indicate risk in the model's predictions. Surgeons in practice can use this risk during diagnosis to trust the model for decision making [14]. Using Dropout to perturb a model is a computationally cheap method of model averaging [19]. It is worth noting though that this method's validity as a Bayesian Inference approximation was later questioned [10]. However, this does not affect the use of this method for risk estimation. So far, model explainability and risk estimation have mostly been used separately to assess models' suitability for surgical applications. DistDeepSHAP [20] computed the uncertainty of Shapley values to show uncertainty in explainability maps. However, DistDeepSHAP is a model-agnostic interpretability method that shows the global effect of perturbing inputs, instead of providing an insight to the model's learned representations. The aim of this paper is to show that the fusion of MC Dropout and PA methods leads to improved explainability.

In this paper, we propose the first approach which incorporates risk estimation into a PA method. A classification model is trained with Dropout and a PA method is used to generate a PA map. At test time, the classification model is employed with the Dropout enabled. In this work, we propose to repeat this process for a number of iterations creating a volume of PA maps. This volume is used to generate a pixel-wise distribution of PA values from which we can infer risk. More specifically, we introduce a method to generate an enhanced PA map by estimating the expectation values of the pixel-wise distributions. In addition, the coefficient of variation (CV) is used to estimate pixel-wise risk of this enhanced PA map. This provides an improved explanation of the model's prediction by clearly presenting to the surgeon which salient areas to trust in the model's enhanced PA map. In this work, we focus on the explainability of the classification of brain tumours using probe-based Confocal Laser Endomicroscopy (pCLE) data. Performance evaluation on pCLE data shows that our improved explainability method outperforms the SOTA.

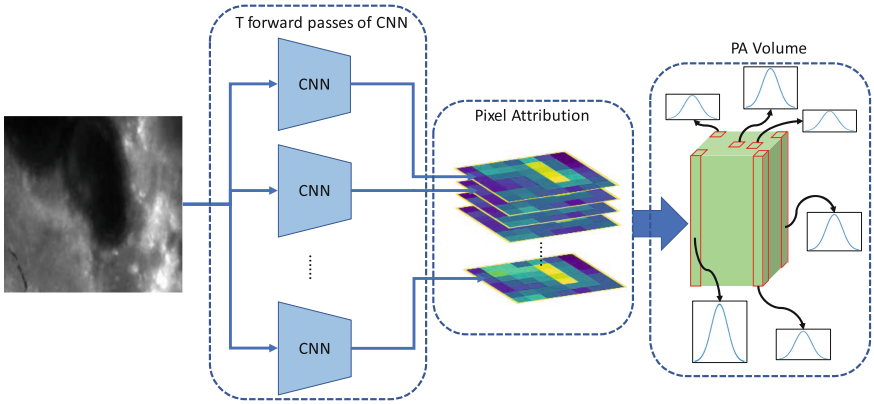


**Fig. 2.** Outline of the proposed method. A PA volume is generated using T forward passes of a CNN model with Dropout applied.

## 2 Methodology

The aim of the proposed method is to produce an improved PA map of a classification model, while providing risk estimation of the model's explainability to enhance trustworthiness in decision making during intraoperative tissue characterisation.

In our method, any CNN classification model trained with Dropout can be used. Let $\hat{\boldsymbol{Y}}$ be the output logits of the CNN model, where Dropout is enabled at test time, with input image $\boldsymbol{X} \in \mathbb{R}^{height \times width \times channels}$. Any PA method can be used to generate a PA map using the output logits $S = f_s(\hat{\boldsymbol{Y}}) \in \mathbb{R}^{height \times width}$ where $f_s(.)$ is the PA method. We propose to repeat the above process for $T$ iterations to create a volume of PA maps $\boldsymbol{S} = \{S_1, ..., S_T\} \in \mathbb{R}^{height \times width \times T}$. A

visual representation of how the volume is generated is show in Fig. 2. The aim is to use this volume to generate a pixel-wise distribution of PA values from which we can infer risk. To achieve this, we compute the expectation and variance values of the volume along the third dimension as:

$$\begin{aligned} \mathbb{E}(\boldsymbol{S}_{i,j}) &\approx \frac{1}{T}\sum_{t=1}^{T} f_s(\hat{\boldsymbol{Y}}_t)_{i,j} \\ Var(\boldsymbol{S}_{i,j}) &\approx \frac{1}{T}\sum_{t=1}^{T} f_s(\hat{\boldsymbol{Y}}_t)_{i,j}^T f_s(\hat{\boldsymbol{Y}}_t)_{i,j} - \mathbb{E}(\boldsymbol{S}_{i,j})^T \mathbb{E}(\boldsymbol{S}_{i,j}), \end{aligned} \tag{1}$$

where, $i, j$ represent the pixel's row and column coordinates, respectively. The expectation $\mathbb{E}(\boldsymbol{S}_{i,j})$ of each pixel $(i, j)$ is used to generate an enhanced PA map of size $height \times width$. The intuition is that the above distribution of PA values can produce less noisy and risky estimations of a pixel's contribution to the final explainability map compared to a single estimate.

As well as advancing SOTA PA methods, our method also estimates the trustworthiness of the enhanced PA map generated above. For risk estimation, it is important to consider that different pixels in the PA map correspond to different semantic features which contribute differently to the output logits. This makes the pixel-wise distributions have different scales. For this reason, the coefficient of variation (CV) is used to estimate pixel-wise risk, as it allows us to compare pixel-wise variances despite their different scales. This is mathematically defined as:

$$S_{i,j}^{cv} = \frac{\sqrt{Var(\boldsymbol{S}_{i,j})}}{\mathbb{E}(\boldsymbol{S}_{i,j})} = \frac{std(\boldsymbol{S}_{i,j})}{\mathbb{E}(\boldsymbol{S}_{i,j})}. \tag{2}$$

Our proposed method improves trustworthiness of explainability as it allows visualisation of both the explainability of the classification model (provided by the enhanced PA map) together with the pixel-wise risk of this map (provided by the CV map). For instance, salient areas on the PA map should not be trusted unless the CV values are low. An example of the enhanced PA and risk maps generated with the proposed method are shown in Fig. 3. This shows that the proposed method not only improves explainability but also provides associated risk information which improves trustworthiness.

## 3 Experiments and Analysis

**Dataset.** The developed explainability framework has been validated on an in vivo and ex vivo pCLE dataset of meningioma, glioblastoma and metastases of an invasive ductal carcinoma (IDC). All studies on human subjects were performed according to the requirements of the local ethic committee and in agreement with the Declaration of Helsinki (No. CLE-001 Nr: 2014480). The Cellvizio©by Mauna Kea Technologies, Paris, France has been used in combination with the mini laser probe CystoFlex©UHD-R. The distinguishing characteristic of the meningioma is the psammoma body with concentric circles that show various

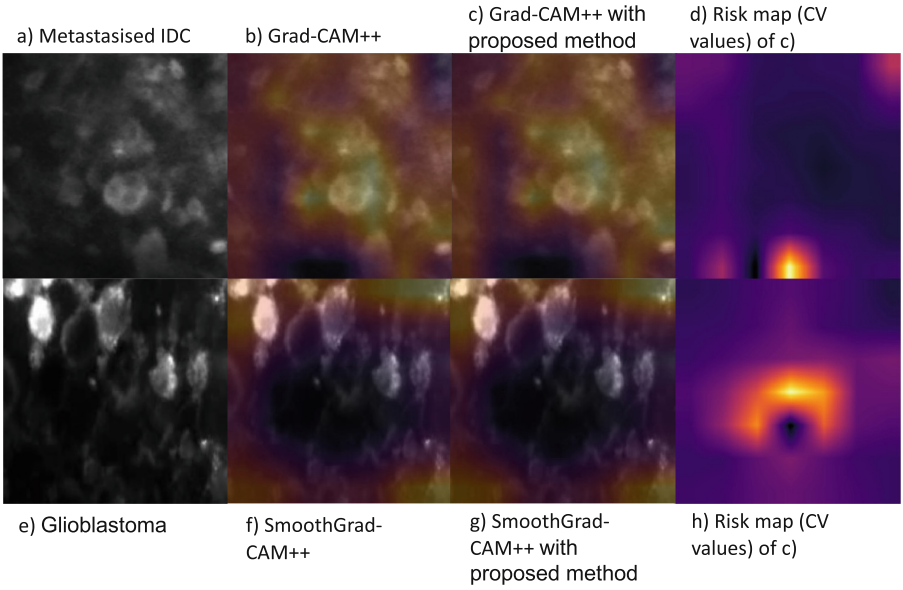


**Fig. 3.** PA maps generated using ResNet18 on pCLE data. a) Metastasised IDC b) Grad-CAM++ PA map on a) c) Grad-CAM++ PA map with our method applied on a) d) Risk map (CV values) of c) e) Glioblastoma f) SmoothGrad-CAM++ PA map on e) g) SmoothGrad-CAM++ PA map using our proposed method on e) h) Risk map (CV values) of g). Yellow represents the highest PA value and black the lowest. (Color figure online)

degrees of calcification. Regarding glioblastomas, the pCLE images allow for the visualization of the characteristic hypercellularity, evidence of irregular nuclei with mitotic activities or multinuclear appearance with irregular cell shape. When examining metastases of an IDC, the tumor presents as egg-shaped cells with uniform evenly spaced nuclei. Our dataset includes 38 meningioma videos, 24 glioblastoma and 6 IDC. Each pCLE video represents one tumour type and corresponds to a different patient. The data has been curated to remove noisy images and similar frames. This resulted in a training dataset of 2500 frames per class (7500 frames in total) and a testing dataset of the same size. The dataset is split into a training and testing subset, with the division done on the patient level.

**Implementation.** To implement the DL models we use the open-source framework PyTorch [25] and a NVIDIA Geforce RTX 3090 graphics card for parallel computation. To show our method generalises we trained two lightweight models: ResNet-18 [17] with a learning rate of 0.01 and MobileNetV2 [29] with a learning rate of 0.001. Both were trained using the Adam-W [22] optimiser with a weight decay of 0.01 and Dropout probability 0.1. We report the model's Top-1 accuracy for Resnet18 as 94.0% and for MobileNet as 86.6%. At test time, we

set $T = 100$ to create a fair distribution of PA maps. PA methods were implemented with the help of TorchCAM [9] and ReciproCAM was implemented using the authors' source code.

**Evaluation Metrics.** Evaluating a PA method is not a trivial task because a PA map may not need to be inline with what a human deems "reasonable" [1]. Segmentation scores like intersection over union (IoU) may be used with caution to compare thresholded PA maps to ground truth maps with annotated salient regions. By doing so, we can measure how informed the model is about a particular class. To quantify how misinformed a model is, we can estimate at its average drop [6]:

$$AverageDrop(f_s, \hat{\boldsymbol{Y}}, \boldsymbol{X}) = \frac{max(0, \hat{\boldsymbol{Y}}(\boldsymbol{X}) - \hat{\boldsymbol{Y}}(\hat{\boldsymbol{X}}))}{\hat{\boldsymbol{Y}}(\boldsymbol{X})}, \tag{3}$$

where, $\hat{\boldsymbol{X}} = \boldsymbol{X} \odot f_s(\hat{\boldsymbol{Y}}(\boldsymbol{X})$. The above equation measures the effect on the output score of the classification model if we only include the pixels which the PA method scored highly. A minimum average drop is desired.

As average drop was found to not be sufficient on its own, the unified method ADCC [27] was introduced which is the harmonic mean of average drop, coherency and complexity, defined as:

$$\begin{aligned} ADCC(f_s(\hat{\boldsymbol{Y}})) = 3(&\frac{1}{Coherency(f_s(\hat{\boldsymbol{Y}}))} \\ &+ \frac{1}{1 - Complexity(f_s(\hat{\boldsymbol{Y}}))} \\ &+ \frac{1}{1 - AverageDrop(f_s, \hat{\boldsymbol{Y}}, \boldsymbol{X})})^{-1}. \end{aligned} \tag{4}$$

Coherency is the Pearson Correlation Coefficient which ensures that the remaining pixels after dropping are still important, defined as:

$$Coherency(f_s(\hat{\boldsymbol{Y}})) = \frac{Cov(f_s(\hat{\boldsymbol{Y}}(\hat{\boldsymbol{X}})), f_s(\hat{\boldsymbol{Y}}))}{\sigma(f_s(\hat{\boldsymbol{Y}}(\hat{\boldsymbol{X}}))\sigma(f_s(\hat{\boldsymbol{Y}}))}, \tag{5}$$

where, $Cov(.,.)$ is the covariance and $\sigma$ is the standard deviation. A higher coherency is better. Complexity is the L1 norm of the output PA map.

$$Complexity(f_s(\hat{\boldsymbol{Y}}))) = ||f_s(\hat{\boldsymbol{Y}}))||_1. \tag{6}$$

Complexity is used to measure how cluttered a PA map is. For a good PA map, complexity should be a minimum. As it has been shown in the literature, the metrics in Eqs. (3), (5) and (6), can not be used individually to evaluate a PA method [27]. ADCC combined with computation time gives us a reliable overall metric of how a PA method is performing.

**Table 1.** Performance evaluation study based on the ADCC and time metrics. Coh is Coherence, Comp is Complexity, AD is average drop and each of these are reported for completeness. Time(s) is the average time to compute one PA map using a batch size of one. All metrics are run over the validation set and averaged. ScoreCAM with the proposed method takes >5 s per batch so was omitted due to resource constraints.

| Architecture | PA method | Coh ↑ | Comp ↓ | AD ↓ | ADCC ↑ | Time(s) ↓ |
|---|---|---|---|---|---|---|
| ResNet18 | Standard - single iteration | | | | | |
| | Grad-CAM | 90.1 | 32.7 | 10.1 | 76.6 | **0.006** |
| | Grad-CAM++ | 90.6 | 33.1 | 10.6 | 76.2 | **0.006** |
| | SmoothGradCAM++ | 88.3 | 27.6 | 14.3 | 74.8 | **0.065** |
| | Score-CAM | 90.0 | 32.3 | 5.9 | **80.5** | **0.124** |
| | Recipro-CAM | 91.0 | 41.2 | 10.0 | 72.8 | **0.007** |
| | Proposed method | | | | | |
| | Grad-CAM | 97.0 | 34.2 | 11.8 | **77.7** | 0.079 |
| | Grad-CAM++ | 93.4 | 32.6 | 12.5 | **78.2** | 0.081 |
| | SmoothGradCAM++ | 92.4 | 30.7 | 17.0 | **75.8** | 0.463 |
| | Score-CAM | – | - - | – | – | – |
| | Recipro-CAM | 92.2 | 37.9 | 11.3 | **76.1** | 0.420 |
| MoblieNetV2 | Standard – single iteration | | | | | |
| | Grad-CAM | 82.9 | 21.3 | 73.8 | 29.3 | **0.010** |
| | Grad-CAM++ | 86.2 | 30.0 | 66.9 | 37.8 | **0.010** |
| | SmoothGradCAM++ | 77.7 | 18.1 | 76.2 | 24.5 | **0.072** |
| | Score-CAM | 62.5 | 33.9 | 56.3 | **43.9** | **0.324** |
| | Recipro-CAM | 85.8 | 32.3 | 67.1 | 35.8 | **0.008** |
| | Proposed method | | | | | |
| | Grad-CAM | 90.0 | 27.0 | 59.4 | **48.0** | 0.103 |
| | Grad-CAM++ | 91.4 | 35.9 | 41.5 | **59.7** | 0.105 |
| | SmoothGradCAM++ | 89.8 | 22.1 | 71.0 | **37.5** | 0.322 |
| | Score-CAM | – | – | – | – | – |
| | Recipro-CAM | 90.6 | 33.7 | 48.3 | **55.9** | 0.674 |

**Performance Evaluation.** The proposed method has been compared to combinations of ResNet18 and MobileNetV2 with SOTA PA methods. At test time, Dropout is not enabled for these standard methods, it is only enabled for our method. In Table 1, we show that our method outperforms all the compared CNN-PA method combinations on ADCC. The Dropout version of ScoreCAM is too computationally expensive and therefore is not included in our comparison. We believe that the better performance of our method is because of the random dropping of features taking place during Dropout at test time which helps to suppress noise in the estimated enhanced PA map. The combination of Recipro-CAM with our proposed method improves performance (increases

ADCC) at the expense of increasing the computational complexity. We believe that this could be reduced using a batched implementation of Recipro-CAM. We attribute slow down in SmoothGradCAM++ when Dropout is applied during test time to the perturbations it adds on top of the PA method. Our validation study shows that Grad-CAM, Grad-CAM++ and Recipro-CAM are often leading in terms of speed as expected from the literature. In Fig. 1, we can see our proposed method reduces noise in the PA map around the salient region. The distinguishing characteristic of the meningioma is the psammoma body which is highlighted by all the PA methods. Risk estimations from Eq. (2) are also displayed and provide an added visualisation for a surgeon to trust the model. As it can be seen, areas of low CV match the areas of high PA values which verifies the trustworthiness of our method. We believe that the proposed explainability method could be used to support the surgeon intraoperatively in diagnosis and decision making during tumour resection. The enhanced PA map extracted with our method highlights the areas which were the most important to the model's prediction. When these areas correlate with clinically relevant areas, it shows that the model has learned to robustly classify the different tissue classes. Hence, it can be trusted by the surgeon for diagnosis.

## 4 Conclusion

In this work we have introduced the first combination of risk in an explainability method. Using our proposed framework we not only improve on all the tested SOTA PA method's ADCC performances but also produce an estimation of risk on the output PA values. The proposed method can clearly present to the surgeon areas of the explainability map that are more trustworthy. From this work we hope to encourage trust between the surgeon and DL models. For future work, we plan to reducing the computation time of our method and deploy the proposed framework for use in surgery.

**Acknowledgement.** This work was supported by the Engineering and Physical Sciences Research Council (EP/T51780X/1) and Intel R&D UK. Dr Giannarou is supported by the Royal Society (URF\R\201014).

## References

1. Adebayo, J., et al.: Sanity Checks for Saliency Maps
2. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I.: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC **20**(1) (2020)
3. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Towards better understanding of gradient-based attribution methods for Deep Neural Networks (Dec 2017)
4. Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D.: Weight Uncertainty in Neural Networks (May 2015)
5. Byun, S.Y., Lee, W.: Recipro-CAM: gradient-free reciprocal class activation map (Sep 2022)

6. Chattopadhyay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: improved Visual Explanations for Deep Convolutional Networks (Oct 2017)
7. Damianou, A.C., Lawrence, N.D.: Deep Gaussian Processes (Nov 2012)
8. Diprose, W.K., Buist, N., Hua, N., Thurier, Q., Shand, G., Robinson, R.: Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. J. Am. Med. Inform. Association **27**(4) (2020)
9. Fernandez, F.G.: TorchCAM: class activation explorer (2020)
10. Folgoc, L.L., et al.: Is MC Dropout Bayesian? (Oct 2021)
11. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Appendix (June 2015)
12. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning (June 2015)
13. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples (Dec 2014)
14. Gordon, L., Grantcharov, T., Rudzicz, F.: Explainable artificial intelligence for safe intraoperative decision support. JAMA Surg. **154**(11), 1064–1065 (2019)
15. Graves, A.: Practical Variational Inference for Neural Networks
16. Hagos, M.T., Curran, K.M., Mac Namee, B.: Identifying Spurious Correlations and Correcting them with an Explanation-based Learning (Nov 2022)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition (Dec 2015)
18. Hinton, G.E., van Camp, D.: Keeping neural networks simple by minimizing the description length of the weights, pp. 5–13 (1993)
19. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors (July 2012)
20. Li, X., Zhou, Y., Dvornek, N.C., Gu, Y., Ventola, P., Duncan, J.S.: Efficient Shapley Explanation for Features Importance Estimation Under Uncertainty (2020)
21. Lin, M., Chen, Q., Yan, S.: Network In Network (Dec 2013)
22. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization (Nov 2017)
23. Neal, R.M.: Bayesian Learning for Neural Networks, vol. 118 (1996)
24. Omeiza, D., Speakman, S., Cintas, C., Weldermariam, K.: Smooth Grad-CAM++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models (Aug 2019)
25. Paszke, A., etal.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019)
26. Petsiuk, V., Das, A., Saenko, K.: RISE: Randomized Input Sampling for Explanation of Black-box Models (June 2018)
27. Poppi, S., Cornia, M., Baraldi, L., Cucchiara, R.: Revisiting The Evaluation of Class Activation Mapping for Explainability: A Novel Metric and Experimental Analysis (April 2021)
28. Ribeiro, M.T., Singh, S., Guestrin, C.: Why Should I Trust You?Explaining the Predictions of Any Classifier (Feb 2016)
29. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: MobileNetV2: Inverted Residuals and Linear Bottlenecks (Jan 2018)
30. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision 2017-October, pp. 618–626 (Dec 2017)

31. Wang, H., et al.: Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks (Oct 2019)
32. Zeiler, M.D., Fergus, R.: Visualizing and Understanding Convolutional Networks (Nov 2013)
33. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning Deep Features for Discriminative Localization