TTE-CAM: Built-in Class Activation Maps for Test-Time Explainability in Pretrained Black-Box CNNs

Kerol Djoumessi[1] KEROL.DJOUMESSI-DONTEU@UNI-TUEBINGEN.DE

Philipp Berens[1,2] PHILIPP.BERENS@UNI-TUEBINGEN.DE

[1] *Hertie Institute for AI in Brain Health, University of Tübingen, Germany*

[2] *Tübingen AI Center, University of Tübingen, Germany*

Editors: Under Review for MIDL 2026

Abstract

Convolutional neural networks (CNNs) achieve state-of-the-art performance in medical image analysis yet remain opaque, limiting adoption in high-stakes clinical settings. Existing approaches face a fundamental trade-off: post-hoc methods provide unfaithful approximate explanations, while inherently interpretable architectures are faithful but often sacrifice predictive performance. We introduce TTE-CAM, a test-time framework that bridges this gap by converting pretrained black-box CNNs into self-explainable models via a convolution-based replacement of their classification head, initialized from the original weights. The resulting model preserves black-box predictive performance while delivering built-in faithful explanations competitive with post-hoc methods, both qualitatively and quantitatively.

Keywords: Test-time explainability, Built-in CAMs, Mechanistic faithfulness, CNNs.

1. Introduction

Convolutional neural networks (CNNs) achieve human-level performance across many tasks, including medical image analysis (Liu et al., 2019), yet their opaque decision processes limit interpretability and hinder adoption in high-stakes clinical settings (Ratti and Graves, 2022). Existing explainability approaches face a fundamental trade-off: post-hoc methods generate saliency maps from the model that do not directly drive the output, making them inherently unfaithful (Adebayo et al., 2018). In contrast, interpretable-by-design architectures (Rudin, 2019; Djoumessi et al., 2024) are faithful—their predictions are computed from the explanation—but they often require complex training or involve a trade-off in predictive performance. Bridging this gap by transforming high-performing black-box CNNs into self-explainable models without retraining or loss of accuracy remains an open challenge.

We propose TTE-CAM, an architectural reformulation of class activation maps (CAMs) that transforms pretrained black-box CNNs into self-explainable models by replacing the classification head with 1×1 convolution layers initialized from the original weights. This reformulation yields built-in CAMs that serve as the sole input to the final prediction, enabling linearly interpretable decisions without post-hoc overhead. Unlike SoftCAM (Djoumessi and Berens, 2025), which requires retraining, and conventional CAM-based post-hoc methods that derive explanations by weighting penultimate feature maps using different methods

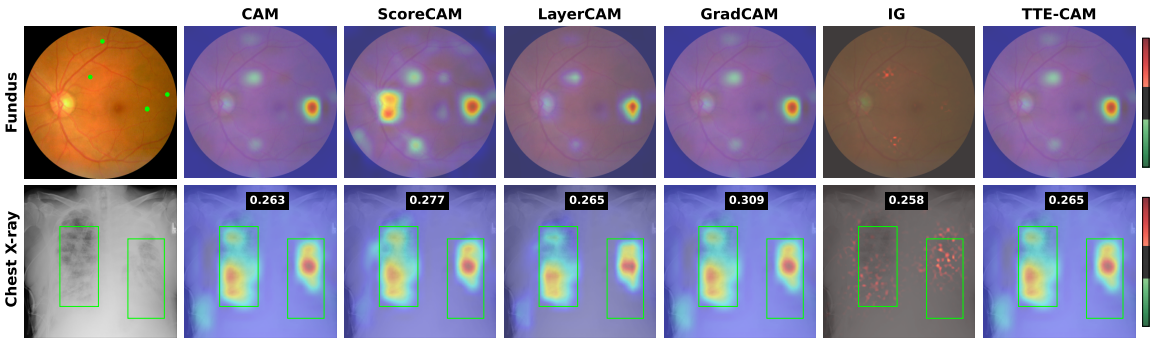

Figure 1: **Qualitative explanations comparison.** The first column shows a DR fundus with clinical annotations (green markers) and a pneumonia CXR with ground-truth bounding boxes (green). Columns 2-6 show post-hoc saliency maps; the last column shows TTE-CAM explanations. CXR scores indicate activation precision.

such as classification layer weights, gradients, or perturbations (He et al., 2022), TTE-CAM integrates CAMs directly into the architecture. This preserves predictive performance while providing faithful, built-in explanations that are competitive with post-hoc approaches, as demonstrated on two medical imaging classification tasks.

2. Materials and Methods

Datasets. TTE-CAM was evaluated on two public medical imaging datasets. The Kaggle fundus Diabetic Retinopathy (DR) dataset (Dugas et al., 2015) was used for binary classification of No DR (grade 0) versus DR (grades 1–4), while the RSNA Chest X-Ray (CXR) dataset (Shih et al., 2019) was used for pneumonia detection. For explanation evaluation, RSNA bounding box annotations and clinical annotations from 65 DR fundus images (Djoumessi et al., 2025) were used for quantitative and qualitative assessment, respectively.

Method. TTE-CAM reformulates the classification head of pretrained CNNs by removing the global average pooling (GAP) layer and replacing the fully connected layer (FCL) with a 1×1 convolutional layer comprising C filters, where C is the number of classes. Because a FCL is equivalent to a 1×1 convolution (Donteu et al., 2023), the pretrained classification weights can be transferred directly without retraining. This reformulation mirrors the original CAM architecture (Zhou et al., 2016), in which class activation maps are obtained post-hoc by weighting feature maps with classification layer weights—here integrated into the architecture. The resulting layer produces built-in CAMs that are spatially averaged to compute class scores and then passed through a softmax to obtain the final predictions.

Post-hoc baseline. TTE-CAM was compared against five post-hoc explainability methods from three families: gradient-free (CAM, ScoreCAM) (Zhou et al., 2016; Wang et al., 2020), gradient-based (GradCAM, LayerCAM) (Selvaraju et al., 2017; Jiang et al., 2021), and the backpropagation-based Integrated Gradients (IG) (Sundararajan et al., 2017).

Evaluation metrics. Predictive performance was evaluated using accuracy (Acc.) and area under the curve (AUC). Explanation quality was assessed with three metrics ($k = 10$):

		CAM	S. CAM	L. CAM	G. CAM	IG	TTE-CAM
Fundus	Topk Prec. ↑	.33 ± .29	.28 ± .22	.31 ± .26	.39 ± .28	.39 ± .28	.33 ± .28
	Topk Sens. ↓	0.629	0.668	0.644	0.629	0.605	0.629
CXR	Acti. Prec. ↑	.12 ± .09	.13 ± .10	.12 ± .09	.13 ± .10	.12 ± .09	.12 ± .09
	Topk Sens. ↓	0.953	0.959	0.955	0.953	0.963	0.953

Table 1: **Quantitative explanation comparison.** ↑ higher is better; ↓ lower is better.

top-k sensitivity (Yeh et al., 2019), measuring the relative drop in predicted probability after masking the top-k most relevant regions; *top-k localization*, quantifying the overlap between the top-k activated regions and annotated lesions; and *activation precision* (Djoumessi and Berens, 2025), measuring the fraction of activations within ground-truth bounding boxes.

3. Results

TTE-CAM was applied to a ResNet-50 (He et al., 2016) trained on each dataset, with the checkpoint achieving the best validation accuracy used at test-time[1]. Replacing the FCL with a 1×1 convolutional classifier preserved predictive performance, yielding Acc. = 0.899, AUC = 0.923 for DR and Acc. = 0.953, AUC = 0.988 for pneumonia detection.

Qualitative (Fig. 1) and quantitative (Tab. 1) results show that TTE-CAM produces explanations similar to CAM and competitive with other methods across both datasets, while being built-in by design. The sparse fundus annotations are better suited for top-k localization, whereas the denser CXR bounding boxes are better suited for activation precision.

4. Discussion and Conclusion

We show that pretrained back-box CNNs can provide built-in explanations at inference time by replacing the classification head with convolutional classifiers. TTE-CAM preserves the original predictive performance while producing explanations competitive with five post-hoc baselines spanning gradient-free, gradient-based, and backpropagation-based methods. Like post-hoc methods, it leverages pretrained weights without retraining, but generates explanations in a single forward pass, in contrast to post-hoc methods that require one forward pass per class (e.g. GradCAM) or multiple passes per class (e.g. ScoreCAM).

TTE-CAM explanations are identical to CAM and competitive with other baselines, sharing a related feature map weighting mechanism. Like all CAM-based methods, reliance on low-resolution feature maps can produce coarse explanations, limiting fine-grained localization, as observed in DR. Weight transfer constraints further restrict applicability to architectures where the final feature map channel dimension matches the classifier input size (e.g., ResNet and DenseNet), excluding models such as VGG. Future work could address these constraints and extend this mechanism to vision transformers for built-in attention map explanations.

Acknowledgments

This project was supported by the Hertie Foundation. and the German Science Foundation.

1. The code is available at https://github.com/kdjoumessi/Test-Time-Explainability

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Kerol Djoumessi and Philipp Berens. Soft-cam: Making black box models self-explainable for high-stakes decisions. *arXiv preprint arXiv:2505.17748*, 2025.

Kerol Djoumessi, Bubacarr Bah, Laura Kühlewein, Philipp Berens, and Lisa Koch. This actually looks like that: Proto-bagnets for local and global interpretability-by-design. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 718–728, 2024.

Kerol Djoumessi, Ziwei Huang, Laura Kühlewein, Annekatrin Rickmann, Natalia Simon, Lisa M Koch, and Philipp Berens. An inherently interpretable ai model improves screening speed and accuracy for early diabetic retinopathy. *PLOS Digital Health*, 4(5):e0000831, 2025.

Kerol R Djoumessi Donteu, Indu Ilanchezian, Laura Kühlewein, Hanna Faber, Christian F Baumgartner, Bubacarr Bah, Philipp Berens, and Lisa M Koch. Sparse activations for interpretable disease grading. In *Medical Imaging with Deep Learning*, 2023.

Emma Dugas, Jorge Jared, and Will Cukierski. Diabetic retinopathy detection, 2015. URL https://kaggle.com/competitions/diabetic-retinopathy-detection.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Mingwei He, Bohan Li, and Songlin Sun. A survey of class activation mapping for the interpretability of convolution neural networks. In *International Conference On Signal And Information Processing, Networking And Computers*, pages 399–407. Springer, 2022.

Peng-Tao Jiang, Chang-Bin Zhang, Qibin Hou, Ming-Ming Cheng, and Yunchao Wei. Layercam: Exploring hierarchical class activation maps for localization. *IEEE Transactions on Image Processing*, 2021.

Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The lancet digital health*, 1(6):e271–e297, 2019.

Emanuele Ratti and Mark Graves. Explainable machine learning practices: opening another black box for reliable medical ai. *AI and Ethics*, 2(4):801–814, 2022.

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

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in neural information processing systems*, 32, 2019.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.