Med-CAM: Minimal Evidence for Explaining Medical Decision Making

Anonymous Authors[1]

Abstract

Reliable and interpretable decision-making is essential in medical imaging, where diagnostic outcomes directly influence patient care. Despite advances in deep learning, most medical AI systems operate as opaque black boxes, providing little insight into *why* a particular diagnosis was reached. In this paper, we introduce **Med-CAM**, a framework for generating minimal and sharp maps as evidence-based explanations for **Med**ical decision making via **C**lassifier **A**ctivation **M**atching. **Med-CAM** trains a segmentation network from scratch to produce a mask that highlights the minimal evidence critical to model's decision for any seen or unseen image. This ensures that the explanation is both *faithful* to the network's behavior and *interpretable* to clinicians. Experiments show, unlike prior spatial explanation methods, such as Grad-CAM and attention maps, which yield only fuzzy regions of relative importance, Med-CAM with its superior spatial awareness to shapes, textures, and boundaries, delivers conclusive, evidence-based explanations that faithfully replicate the model's prediction for any given image. By explicitly constraining explanations to be compact, consistent with model activations, and diagnostic alignment, MedCAM advances transparent AI to foster clinician understanding and trust in high-stakes medical applications such as pathology and radiology.

1. Introduction

Explanations are increasingly being recognized as essential for understanding and trusting the decision-making of machine learning systems deployed in medical settings. In clinical workflows, where algorithmic predictions can directly influence diagnosis or treatment, it is not enough for a model to be accurate—it must also be interpretable and verifiable. Deep neural networks, despite their remarkable success across medical imaging tasks such as pathology, radiology, and dermatology, often operate as opaque black boxes. Their predictions arise from complex, high-dimensional computations that are inaccessible to human reasoning, leaving clinicians uncertain about whether the model's conclusions stem from meaningful pathology or from confounding artifacts.

In this context, *minimal evidence* becomes a key desideratum for trustworthy medical explanations. By isolating the smallest set of image regions sufficient to support a model's diagnostic decision, one can reveal the true basis of the model's reasoning in a form that is both human-understandable and clinically verifiable. Minimal explanations reveal the essential visual cues such as atypical cellular morphologies in histopathology or subtle opacities in radiographs, that directly drive the network's prediction. Such evidence-centered explanations not only enhance clinical interpretability but also serve as a diagnostic sanity check, allowing practitioners to verify weather that the model attends to medically relevant structures or not.

To address this need, we propose **Med-CAM**, a **Med**ical **C**lassifier **A**ctivation **M**atching framework for generating **minimal, evidence-based explanations** for medicinal decision making. Med-CAM learns a lightweight U-Net from scratch that produces a binary mask highlighting the minimal evidence necessary for a model's output. The U-Net is trained individually on a single image within seconds, optimizing for consistency between the classifier's activations on the original image and on the masked explanation. This *activation-matching* principle ensures that the explanation preserves both the internal reasoning and the diagnostic prediction of the underlying model.

Compared to existing saliency-based approaches such as Grad-CAM and attention maps, Med-CAM offers several key advantages. Grad-CAM provides gradient-weighted heatmaps that highlight regions of relative importance but fail to capture precise spatial characteristics such as boundaries, textures, or surface continuity. Med-CAM explicitly enforces activation and decision consistency, resulting in spatially aware, compact, and conclusive evidence maps that faithfully replicate the model's diagnostic behavior.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

2. Prior Work

2.1. Inversion

Inversion methods aim to reconstruct inputs that elicit specific outputs or internal activations of a neural network seeking to visualize what a model has learned rather than why it makes a particular decision. Early work reconstructed representative patterns from multilayer perceptrons through gradient-based optimization, though such visualizations were often noisy or adversarial-like (Kindermann & Linden, 1990; Jensen et al., 1999; Saad & Wunsch, 2007). Subsequent studies explored evolutionary search and constrained optimization (Wong, 2017), followed by the introduction of prior-based regularization techniques such as smoothness constraints and pretrained generative models to enhance realism and interpretability (Mahendran & Vedaldi, 2014; Yosinski et al., 2015; Mordvintsev et al., 2015; Nguyen et al., 2016; 2017). More recent methods stabilize the inversion process through learned surrogate landscapes (Liu et al., 2022a) or reframe it within logical reasoning frameworks (Suhail, 2024) while others (Suhail & Sethi, 2024) use generative models for inversion.

2.2. Explainability

While inversion captures model behavior in aggregate, explainability focuses on generating faithful rationales for individual predictions and has therefore emerged as a major research area (Ali et al., 2023; Hsieh et al., 2024; Gilpin et al., 2018), driven by the need to enhance trust, transparency, and accountability in high-stakes applications.

Post-hoc attribution methods remain the dominant paradigm. **LIME** builds local surrogate models to approximate black-box predictors (Hamilton et al., 2022), while **Anchors** (Ribeiro et al., 2018) extends this idea through high-precision, rule-based local explanations that define sufficient conditions for predictions. **Grad-CAM** highlights salient image regions by gradient-weighted class activation mapping (Selvaraju et al., 2019), whereas **DeepLIFT** (Shrikumar et al., 2019) backpropagates contribution scores relative to reference activations. The **SHAP** framework (Lundberg & Lee, 2017) unifies several additive feature-attribution approaches using Shapley values to assign consistent feature importance scores. Beyond local attributions, concept-based and structural methods attempt to align neural activations with semantically meaningful parts (Lee et al., 2025). Architectural approaches (Böhle et al., 2022) promote explicit weight–input alignment for interpretable transformations, while (Stalder et al., 2022) introduces an auxiliary explainer network that produces class-specific binary masks identifying discriminative evidence regions.

Parallel to these efforts, logic- and reasoning-based frameworks explore explanation as a formal process. (Darwiche & Hirth, 2020) introduced theoretical foundations for necessary, sufficient, and complete reasons behind classifier decisions. Similarly, (Shih et al., 2018) proposed a symbolic approach to explain Bayesian network classifiers, identifying minimal sufficient feature sets responsible for classification. Abductive reasoning techniques (Ignatiev et al., 2018) extend this logic-driven line of work by computing subset- or cardinality-minimal explanations with formal guarantees of faithfulness. Recent surveys (Zhou et al., 2021) emphasize the need for quantitative evaluation criteria—combining fidelity, stability, and human-centered assessment—to move beyond purely visual interpretability. Interactive frameworks like (Teso et al., 2022) further demonstrate how explanations can guide human-in-the-loop model refinement.

2.3. Explainability in Medical Diagnostics

In medical imaging, interpretability is not merely desirable—it is essential. (Singh et al., 2020) surveyed explainable deep learning methods for medical image analysis and highlighted major barriers to clinical adoption, including lack of interpretive rigor and standardized evaluation. (van der Velden et al., 2022) proposed a comprehensive taxonomy of explainable medical imaging methods categorized by anatomical region, modality, and transparency level. More recently, (Chaddad et al., 2025) emphasized the joint role of explainability and generalizability, showing that integrating CNNs with XAI methods such as Grad-CAM, XGradCAM, and LayerCAM enhances visual interpretability and diagnostic consistency across multiple medical datasets. (Ghasemi et al., 2024) presents a focused scoping review on explainability in breast cancer detection identifying SHAP as the most widely used and clinically interpretable method, particularly suited to feature-level analysis and biomarker attribution in survival and risk prediction.

Despite the steady progress of XAI in medicine, most existing frameworks emphasize saliency visualization or heuristic attribution without enforcing *minimality*—the principle that an explanation should contain only the smallest subset of evidence sufficient to preserve a model's decision.

Our proposed work **Med-CAM**, addresses this limitation by formulating explanation generation as an activation-matching inverse problem: we seek the minimal pre-image that preserves the classifier's internal activations and decision, thereby producing compact, faithful, and diagnostically meaningful evidence maps.

3. Methodology

Per-image explanation generation can be viewed as an *inverse problem*, aiming to recover a decision-wise conclusive and critical pre-image—a minimal subset of the input—that is sufficient to sustain a model's prediction. While the region

outside this subset acts as a distractor, containing information that does not contribute meaningfully to the decision and may even obscure it.

Med-CAM aims to generate minimal and faithful evidence maps that explain the diagnostic decisions of a frozen medical classifier f on any given image x. To achieve this, we train a lightweight U-Net that outputs a binary mask m, where the masked explanation is defined as $e = m \odot x$, and non-critical regions are suppressed. The autoencoder is optimized individually for each image, requiring only a few seconds to converge, thus adapting dynamically to case-specific diagnostic cues. Training is guided by a composite loss function that integrates *activation matching*, *output fidelity*, and *mask priors* for minimality, along with a robustness constraint ensuring clinical sufficiency and stability.

3.1. Weighted Activation Matching

The core principle of Med-CAM is that the explanation must preserve the classifier's decision process at both the output and feature levels. This is enforced through activation alignment, distributional consistency and label match between the original image x and the masked explanation e.

Activation Matching Loss: To ensure that the explanation elicits the same internal responses as the original input, we minimize $\mathcal{L}_{\text{act}} = \sum_\ell \alpha_\ell \, d\big(\phi_\ell(x), \phi_\ell(e)\big)$, where $\phi_\ell(\cdot)$ denotes the post-ReLU activations at layer ℓ, d is a distance metric (mean squared error or cosine distance), and α_ℓ controls per-layer weighting. This loss ensures that e reproduces the hierarchical activations responsible for diagnostic reasoning—such as tissue texture patterns, lesion morphology, or intensity gradients.

KL Divergence Loss: To preserve the classifier's probabilistic output distribution, we minimize the Kullback–Leibler divergence between the softmax outputs of the original and masked images: $\mathcal{L}_{\text{KL}} = D_{\text{KL}}\big(\text{softmax}(f(x)) \,\|\, \text{softmax}(f(e))\big)$. This encourages the masked explanation to produce the same confidence distribution as the original image, ensuring diagnostic consistency.

Cross-Entropy Loss: To explicitly retain the predicted class label, we minimize $\mathcal{L}_{\text{CE}} = -\log p_{f(e)}(y)$, where y is the top-1 class predicted by $f(x)$. This enforces decision-level faithfulness: the masked explanation must independently yield the same diagnosis.

3.2. Mask Priors for Minimal Evidence

To generate interpretable evidence maps that highlight diagnostically relevant regions, we impose priors on the mask m.

Area Loss for Sparsity:

$$\mathcal{L}_{\text{area}} = \|m\|_1.$$

This term penalizes large active regions, encouraging Med-CAM to use the fewest possible pixels necessary for preserving the model's decision corresponding to the principle of *minimality*.

Binarization Loss for Crispness: $\mathcal{L}_{\text{bin}} = \|m - m^2\|_1$. This drives the mask toward binary values (0 or 1), producing clear, interpretable evidence maps rather than diffuse saliency heatmaps.

Total Variation Loss for Smoothness:

$$\mathcal{L}_{\text{tv}} = \sum_{i,j} \big(|m_{i,j} - m_{i+1,j}| + |m_{i,j} - m_{i,j+1}|\big).$$

This regularizes the spatial structure of the mask, promoting contiguous evidence regions and suppressing isolated activations critical for delineating anatomical boundaries.

3.3. Abductive Robustness Constraint

Minimality alone does not ensure that the explanation is clinically sufficient. To enforce robustness, we introduce an abductive constraint ensuring that randomizing irrelevant regions does not alter the model's decision. Given a perturbed background r, we define $\tilde{e} = m \odot x + (1 - m) \odot r$, and require that $f(\tilde{e})$ yield the same label as $f(x)$ via the loss $\mathcal{L}_{\text{rob}} = -\log p_{f(\tilde{e})}(y)$. This ensures that explanations are faithful even when non-evidence regions are replaced with random or domain-shifted content—an important criterion for robustness in medical interpretation.

3.4. Training Objective

The Med-CAM U-Net is optimized using a composite loss $\mathcal{L}_{\text{EXP}}$ that enforces activation alignment, minimality, and robustness. For clarity, we group the terms as:

$$\begin{aligned}
\mathcal{L}_{\text{AM}} &= \lambda_{\text{act}} \mathcal{L}_{\text{act}} + \lambda_{\text{CE}} \mathcal{L}_{\text{CE}} + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}}, \\
\mathcal{L}_{\text{MIN}} &= \lambda_{\text{area}} \mathcal{L}_{\text{area}} + \lambda_{\text{bin}} \mathcal{L}_{\text{bin}} + \lambda_{\text{tv}} \mathcal{L}_{\text{tv}}, \\
\mathcal{L}_{\text{ROB}} &= \lambda_{\text{rob}} \mathcal{L}_{\text{rob}}.
\end{aligned}$$

The complete objective is:

$$\mathcal{L}_{\text{EXP}} = \mathcal{L}_{\text{AM}} + \mathcal{L}_{\text{MIN}} + \mathcal{L}_{\text{ROB}}.$$

With appropriate coefficient choices $\{\lambda_\cdot\}$, Med-CAM learns binary, sparse, and spatially coherent masks that retain the classifier's decision and internal activations while suppressing irrelevant context. Unlike Grad-CAM or attention maps, which provide coarse regions of relative importance, Med-CAM produces conclusive pixel-level evidence maps that isolate the minimal diagnostic cues essential to the model's prediction.

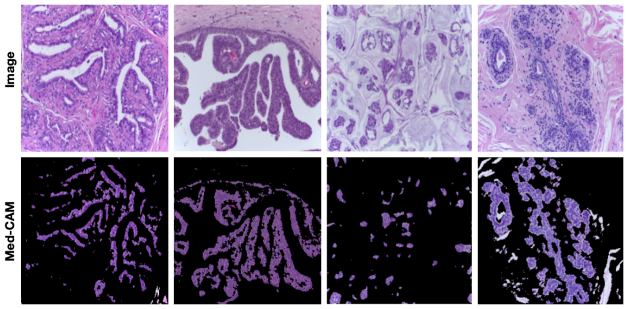

Figure 1. Med-CAM explanations on BACH H&E breast cancer histopathology slides using a ViT-16 classifier.

4. Results

Med-CAM is model-agnostic and operates on any frozen classifier and any seen or unseen image at inference time. Because the U-Net (Ronneberger et al., 2015) explainer is trained per-image in a few seconds, the framework naturally adapts to case-specific visual cues without requiring dataset-level retraining. We evaluate Med-CAM on four diverse medical imaging modalities—dermatology, histopathology, MRI, and retinal fundus—using four architectures: ViT-16 (Dosovitskiy et al., 2021) on BACH (Aresta et al., 2019)(89% test accuracy), ConvNeXt-Small (Liu et al., 2022b) on HAM10000 (Tschandl et al., 2018)(85%), ResNet-18 (He et al., 2015) on IDRiD (Porwal et al., 2018) (79%), and MobileNet-V2 (Sandler et al., 2019) on Brain Tumor MRI (82%).

Figure 1 shows an image from each BACH class (*Normal, Benign, In Situ, Invasive*) with corresponding Med-CAM explanations. Histology images exhibit highly irregular, multi-scale patterns, making explanation generation particularly challenging. Med-CAM successfully adapts to these diverse morphologies, highlighting luminal structures, glandular boundaries, and malignant epithelial clusters. Using relative weights of 10:100:10 for activation matching, minimality, and robustness, we observe that Med-CAM consistently isolates concise diagnostic evidence with improved classifier confidence.

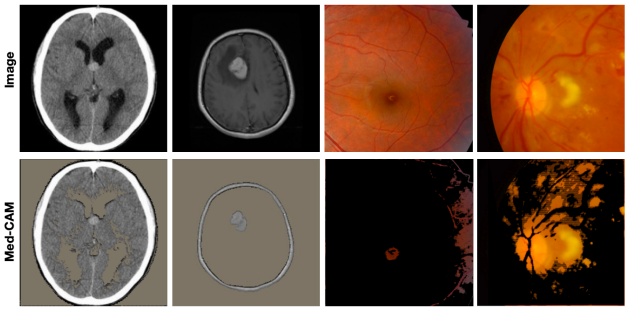

Figure 2. Med-CAM explanations on brain tumor MRI (left) and IDRiD retinal fundus images (right).

In Figure 2 for MRI, we include one healthy and one tumor-containing slice. Med-CAM highlights the outer brain boundary and central parenchyma for normal cases, and selectively isolates only the tumor core and surrounding edema for pathological cases. For IDRiD, we show images from retinopathy grades 2 and 4. Grade 2 exhibits mild lesions, and Med-CAM correspondingly identifies a very small, compact set of pixels capturing microaneurysms and early exudates. Grade 4 contains large, clinically significant lesions highlighted by Med-CAM.

Figure 3 shows results for four HAM10000 classes (*Melanoma, Nevus, Vascular Lesion, Dermatofibroma*) in which Med-CAM produces explanations that almost resemble segmentation masks–highlighting pigment networks, lesion borders, vascular blobs, and firm nodular structures with remarkable precision. With relative loss weights of 10:150:20, the evidence masks remain compact but highly informative.

5. Comparisons

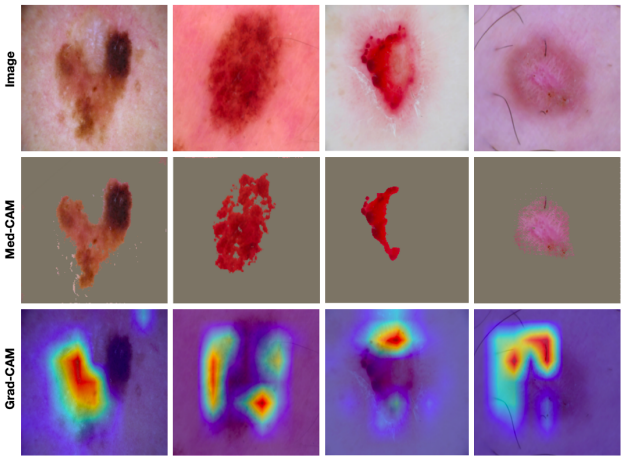

Figure 3. **Med-CAM vs Grad-CAM on HAM10000.**

We compare Med-CAM against the widely used Grad-CAM on the HAM10000 dataset (Figure 3). Grad-CAM heatmaps highlight large, smooth regions of relative importance but lack spatial awareness of fine lesion structure. Grad-CAM explanations fragment into multiple disconnected blobs despite the lesion being compact. Med-CAM, in contrast, produces crisp explanations with sharply defined boundaries that adhere to the exact lesion morphology. The method precisely captures irregular lesion outlines, pigment asymmetry, and textural cues essential for melanoma and nevus differentiation. Unlike Grad-CAM, Med-CAM's evidence masks are minimal yet guaranteed to preserve the classifier's original diagnosis. Empirically, Grad-CAM often spreads its attention across both lesion and non-lesion regions, while Med-CAM isolates only the discriminative pixels.

References

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., and Herrera, F. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805, 2023. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.101805. URL https://www.sciencedirect.com/science/article/pii/S1566253523001148.

Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S. S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., Fernandez, G., Zeineh, J., Kohl, M., Walz, C., Ludwig, F., Braunewell, S., Baust, M., Vu, Q. D., To, M. N. N., Kim, E., Kwak, J. T., Galal, S., Sanchez-Freire, V., Brancati, N., Frucci, M., Riccio, D., Wang, Y., Sun, L., Ma, K., Fang, J., Kone, I., Boulmane, L., Campilho, A., Eloy, C., Polónia, A., and Aguiar, P. Bach: Grand challenge on breast cancer histology images. *Medical Image Analysis*, 56:122–139, August 2019. ISSN 1361-8415. doi: 10.1016/j.media.2019.05.010. URL http://dx.doi.org/10.1016/j.media.2019.05.010.

Böhle, M., Fritz, M., and Schiele, B. B-cos networks: Alignment is all we need for interpretability, 2022. URL https://arxiv.org/abs/2205.10268.

Chaddad, A., Hu, Y., Wu, Y., Wen, B., and Kateb, R. Generalizable and explainable deep learning for medical image computing: An overview. *Current Opinion in Biomedical Engineering*, 33:100567, March 2025. ISSN 2468-4511. doi: 10.1016/j.cobme.2024.100567. URL http://dx.doi.org/10.1016/j.cobme.2024.100567.

Darwiche, A. and Hirth, A. On the reasons behind decisions, 2020. URL https://arxiv.org/abs/2002.09284.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Ghasemi, A., Hashtarkhani, S., Schwartz, D., and Shaban-Nejad, A. Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review, 07 2024.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M. A., and Kagal, L. Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89, 2018. URL https://api.semanticscholar.org/CorpusID:59600034.

Hamilton, N., Webb, A., Wilder, M., Hendrickson, B., Blanck, M., Nelson, E., Roemer, W., and Havens, T. C. Enhancing visualization and explainability of computer vision models with local interpretable model-agnostic explanations (lime). In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 604–611, 2022. doi: 10.1109/SSCI51031.2022.10022096.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Hsieh, W., Bi, Z., Jiang, C., Liu, J., Peng, B., Zhang, S., Pan, X., Xu, J., Wang, J., Chen, K., Feng, P., Wen, Y., Song, X., Wang, T., Liu, M., Yang, J., Li, M., Jing, B., Ren, J., Song, J., Tseng, H.-M., Zhang, Y., Yan, L. K. Q., Niu, Q., Chen, S., Wang, Y., and Liang, C. X. A comprehensive guide to explainable ai: From classical models to llms, 2024. URL https://arxiv.org/abs/2412.00800.

Ignatiev, A., Narodytska, N., and Marques-Silva, J. Abduction-based explanations for machine learning models, 2018. URL https://arxiv.org/abs/1811.10656.

Jensen, C., Reed, R., Marks, R., El-Sharkawi, M., Jung, J.-B., Miyamoto, R., Anderson, G., and Eggen, C. Inversion of feedforward neural networks: algorithms and applications. *Proceedings of the IEEE*, 87(9):1536–1549, 1999. doi: 10.1109/5.784232.

Kindermann, J. and Linden, A. Inversion of neural networks by gradient descent. *Parallel Computing*, 14(3):277–286, 1990. ISSN 0167-8191. doi: https://doi.org/10.1016/0167-8191(90)90081-J. URL https://www.sciencedirect.com/science/article/pii/016781919090081J.

Lee, J. H., Mikriukov, G., Schwalbe, G., Wermter, S., and Wolter, D. Concept-based explanations in computer vision: Where are we and where could we go? In Del Bue, A., Canton, C., Pont-Tuset, J., and Tommasi, T. (eds.), *Computer Vision – ECCV 2024 Workshops*, pp. 266–287, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-92648-8.

Liu, R., Mao, C., Tendulkar, P., Wang, H., and Vondrick, C. Landscape learning for neural network inversion, 2022a. URL https://arxiv.org/abs/2206.09027.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. A convnet for the 2020s, 2022b. URL https://arxiv.org/abs/2201.03545.

Lundberg, S. and Lee, S.-I. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

Mahendran, A. and Vedaldi, A. Understanding deep image representations by inverting them, 2014. URL https://arxiv.org/abs/1412.0035.

Mordvintsev, A., Olah, C., and Tyka, M. Inceptionism: Going deeper into neural networks, 2015. URL https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., and Clune, J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks, 2016. URL https://arxiv.org/abs/1605.09304.

Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., and Yosinski, J. Plug & play generative networks: Conditional iterative generation of images in latent space, 2017. URL https://arxiv.org/abs/1612.00005.

Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., and Meriaudeau, F. Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research. *Data*, 3(3), 2018. ISSN 2306-5729. doi: 10.3390/data3030025. URL https://www.mdpi.com/2306-5729/3/3/25.

Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: high-precision model-agnostic explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Saad, E. W. and Wunsch, D. C. Neural network explanation using inversion. *Neural Networks*, 20(1):78–93, 2007. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2006.07.005. URL https://www.sciencedirect.com/science/article/pii/S0893608006001730.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks, 2019. URL https://arxiv.org/abs/1801.04381.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128 (2):336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Shih, A., Choi, A., and Darwiche, A. A symbolic approach to explaining bayesian network classifiers. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pp. 5103–5111. AAAI Press, 2018. ISBN 9780999241127.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences, 2019. URL https://arxiv.org/abs/1704.02685.

Singh, A., Sengupta, S., and Lakshminarayanan, V. Explainable deep learning models in medical image analysis, 2020. URL https://arxiv.org/abs/2005.13799.

Stalder, S., Perraudin, N., Achanta, R., Perez-Cruz, F., and Volpi, M. What you see is what you classify: Black box attributions, 2022. URL https://arxiv.org/abs/2205.11266.

Suhail, P. Network inversion of binarised neural nets. In *The Second Tiny Papers Track at ICLR 2024*, 2024. URL https://openreview.net/forum?id=zKcB0vb7qd.

Suhail, P. and Sethi, A. Network inversion of convolutional neural nets. In *Muslims in ML Workshop co-located with NeurIPS 2024*, 2024. URL https://openreview.net/forum?id=f9sUu7U1Cp.

Teso, S., Öznur Alkan, Stammer, W., and Daly, E. Leveraging explanations in interactive machine learning: An overview, 2022. URL https://arxiv.org/abs/2207.14526.

Tschandl, P., Rosendahl, C., and Kittler, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1), August 2018. ISSN 2052-4463. doi: 10.1038/sdata.2018.161. URL http://dx.doi.org/10.1038/sdata.2018.161.

van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., and Viergever, M. A. Explainable artificial intelligence (xai) in deep learning-based medical image analysis. *Medical Image Analysis*, 79:102470, 2022. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2022.102470. URL https://www.sciencedirect.com/science/article/pii/S1361841522001177.

Wong, E. Neural network inversion beyond gradient descent. In *WOML NIPS*, 2017. URL https://api.semanticscholar.org/CorpusID:208231247.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., and Lipson, H. Understanding neural networks through deep visualization, 2015. URL https://arxiv.org/abs/1506.06579.

Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5), 2021. ISSN 2079-9292. doi: 10.3390/electronics10050593. URL https://www.mdpi.com/2079-9292/10/5/593.