DOI 10.5162/iCCC2026/P28

Explainability in ConvNeXt V2 for Contrast-Enhanced Spectral Mammography

Walid Ghorbel, Alireza Siyavashi, Christian Herglotz
Chair of Computer Engineering, Brandenburgische Technische Universität Cottbus–Senftenberg
Cottbus, Germany
E-Mail-Adresse: ghorbwal@b-tu.de, alireza.siyavashi@b-tu.de, christian.herglotz@b-tu.de

Abstract

Deep learning models can support breast cancer diagnosis in mammography, but their predictions are not explainable enough to count on their decision. Class activation mapping (CAM) techniques highlight image regions that contribute to a model's decision, yet existing approaches frequently produce diffuse saliency maps that include large areas with limited influence on the prediction. In this work, we employ Hybrid-CAM, a visualization method that combines global channel importance with high-resolution spatial activations, to generate more precise and faithful explanations of model behaviour. We show that even a comparatively simple convolutional neural network, when paired with an appropriate visualization technique, can yield substantially improved explainability.
Quantitative insertion–deletion metrics, together with qualitative visual inspection, demonstrate that Hybrid-CAM produces more focused and reliable explanations, particularly at intermediate network layers, by concentrating on regions that truly drive the model's predictions.

1 Introduction

Breast imaging plays a central role in detecting and assessing lesions, often guiding key diagnostic and treatment decisions [1]. In recent years, Artificial Intelligence (AI) has been increasingly used to support this process by analyzing medical images [2]. These models learn complex diagnostic patterns from large collections of medical images, enabled by their rich hierarchical representations [2].
Although they often achieve high performance in lesion detection and assessment, they usually provide only a final prediction without indicating which image regions contributed to it [1], [3]. This lack of transparency is especially problematic in breast imaging, where subtle or low-contrast abnormalities may influence clinical decisions [3].
Deep learning methods are also applied to Contrast-Enhanced Spectral Mammography (CESM), a radiographic modality designed to visualize cancerous lesions more clearly, to support lesion detection and diagnostic evaluation [4]. However, their predictions still do not indicate which image regions drive the outcome [3].
Clear visual explanations, complemented by quantitative assessment, are therefore essential for understanding model behavior and for analyzing potential errors [5].
To understand how these models make predictions, it is useful to consider how they process images. Most current methods for breast image analysis use convolutional neural networks (CNNs), which extract visual patterns through a series of layers. Each layer applies small learned filters that scan the image and produce feature maps that indicate where specific patterns occur [2] as shown in Figure 1.
In 2016, Zhou et al. introduced Class Activation Mapping (CAM), a technique that uses the final-layer feature maps to identify the image regions that contributed most to a given prediction [6]. CAM generates a spatial heatmap that highlights areas with high relevance for the model's decision, providing an intuitive visual representation of which regions influenced the outcome.

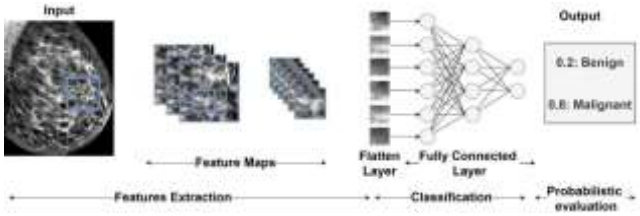


Figure 1: Overview of a typical CNN.

Several variants of CAM build on this idea. Gradient-weighted CAM (Grad-CAM) estimates how the model's output would change if a pattern detected in a deep feature map were strengthened, and uses this as an importance weight in the heatmap [7]. Because it operates on deep, low-resolution feature maps, the resulting heatmaps often highlight broad regions.
Score-CAM measures how the model's prediction changes when different parts of the input image are masked or revealed [8]. Regions that raise the class score when kept visible are considered important. Since each mask is tested independently, Score-CAM does not account for how multiple areas may jointly influence the prediction.
To address these limitations, we introduce Hybrid-CAM [9]. It starts from a Grad-CAM map and then evaluates which of the highlighted regions still support a high model prediction when the salient regions are selectively preserved in the image. Areas that continue to support the prediction are preserved, while diffuse or less relevant areas are suppressed. The resulting map isolates the areas that were most influential for the model's decision.
The main contributions of this work are as follows:

- We introduce Hybrid-CAM, a refinement of Grad-CAM that reduces over-highlighting and provides clearer, decision-relevant explanations.
- We provide a quantitative and qualitative comparison with Grad-CAM, Score-CAM, and Hybrid-CAM on a ConvNeXtV2-Tiny model [10]

DOI 10.5162/iCCC2026/P28

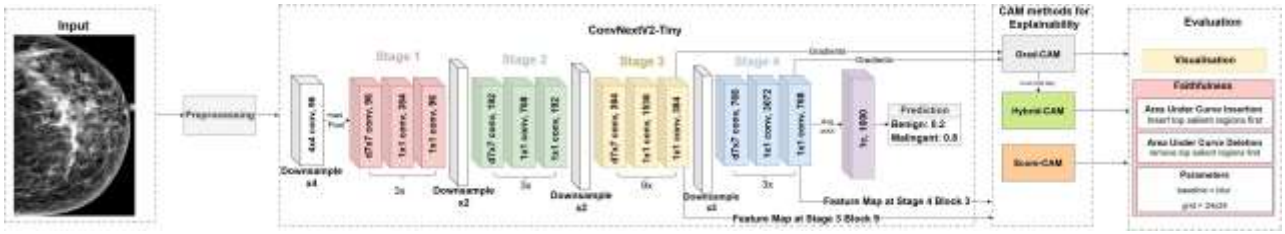


Figure 2: Overall workflow from input image and preprocessing, through ConvNeXtV2-Tiny feature extraction, to CAM generation (Grad-CAM, Hybrid-CAM, Score-CAM) and faithfulness evaluation.

using insertion-deletion metrics [11] on a public CDD-CESM dataset [12].

The remainder of this paper is organized as follows: Section 2 describes the dataset, backbone architecture, and the proposed Hybrid-CAM method. Section 3 outlines the evaluation protocol, and Sections 4 and 5 reports quantitative and qualitative results. In Section 6, we conclude the results and discuss future work.

2 Methodology

In this section, we outline how the fine-tuned ConvNeXtV2-Tiny model was prepared for explainability analysis on low-energy CESM images. After preprocessing, the images pass through the network, and feature maps are extracted from two layers: Stage 3 Block 9, the deepest layer that still preserves a clear spatial layout of the breast, and Stage 4 Block 3, the final block before classification. These maps are used to compute Grad-CAM, Score-CAM, and Hybrid-CAM. The resulting explanations are then evaluated with insertion–deletion metrics and visualized on representative test cases.

2.1 Dataset and Preprocessing

Experiments were conducted on the low-energy subset of the public CDD-CESM dataset [12]. A ConvNeXtV2-Tiny model pretrained on ImageNet-22K was fine-tuned for benign–malignant classification [10], achieving 92.2% specificity, 56.1% sensitivity, and 82.1% precision. CAM analyses were performed on a held-out test set of 141 images, following the pipeline illustrated in Fig. 2. All test images were processed using the same validation pipeline used during model training.

2.2 Backbone Network: ConvNeXtV2-Tiny

ConvNeXtV2-Tiny is used as the backbone in this study. It is a convolutional network that integrates several design elements inspired by transformer models, while retaining the spatially organized processing characteristic of CNNs [10]. This combination allows the model to capture both local details and broader patterns, which are crucial in breast imaging, where the appearance of a lesion often depends on its surrounding tissue. Figure 2 illustrates the four stages of the ConvNeXtV2 backbone, where earlier stages emphasize low-level image structures such as edges or subtle textures, while deeper stages capture complex diagnostic patterns. Each stage contains several ConvNeXt blocks, which act as small processing units that extract and refine patterns using a 7×7 depthwise convolution to examine a larger local area, followed by two 1×1 convolutions that combine and refine the extracted features. Across stages, the number of feature maps increases, allowing the model to capture more complex patterns. For the explainability analysis, we use the last block of Stage 3 and the last block of Stage 4 of the ConvNeXtV2-Tiny model, as shown in Fig. 2. At these depths, the feature maps retain sufficient spatial resolution to indicate where the model is focusing, while also encoding higher-level information that strongly influences the benign-malignant decision. These characteristics make these blocks well-suited for CAM-based explanation and for assessing whether the model relies on clinically relevant regions.

2.3 Grad-CAM

CAM methods produce spatial heatmaps that highlight which regions the model considers relevant for a prediction [6]. Grad-CAM does this by using the gradients of the class score with respect to the feature maps of a selected layer [7]. Let $\boldsymbol{A_k}$ denote the $\boldsymbol{k}$-th feature map at this layer, with spatial dimensions $\boldsymbol{H \times W}$. A feature map is the 2D activation map produced by one channel of a convolutional layer, where each value $\boldsymbol{A_k(i,j)}$ indicates how strongly this channel responds to a pattern at the location $\boldsymbol{(i,j)}$. Let $\boldsymbol{y_c}$ denote the model score for class $\boldsymbol{c}$ before the final softmax. The gradient $\boldsymbol{\partial y_c / \partial A_k(i,j)}$ describes how much the class score would change if the activation $\boldsymbol{A_k(i,j)}$ were increased, and thus reflect its relevance for class c. Grad-CAM then computes a single importance weight for each feature map by averaging these gradients over all spatial positions: $\boldsymbol{w_k} = \sum_{i,j} \frac{\partial y_c}{\partial A_{k(i,j)}}$

A positive w_k means that increasing feature map k tends to increase the prediction for class c, whereas a negative w_k indicates the opposite.

The Grad-CAM heatmap is then obtained as:

$$CAM(i,j) = ReLU\left(\sum_k w_k A_k(i,j)\right)$$

where the summation aggregates contributions from all feature maps. The ReLU (rectified linear unit) function sets negative values to zero, so that only features that support the prediction remain visible in the map. The heatmap is finally upsampled and normalized for visualization.

2.4 Hybrid-CAM

Grad-CAM shows where the model focuses, but it does not reveal how strongly these highlighted regions influence the

prediction when those regions are actually perturbed or removed. Hybrid-CAM addresses this by introducing controlled perturbations: it preserves only the most salient areas of the Grad-CAM map, replaces the remaining pixels with a baseline, and measures how the model's output changes. Given an input image x and its Grad-CAM map $A(i,j)$, pixels are ranked by saliency. For each area level $p \in \{1,2,3,5,10,20,35,50,70,90\}\%$, indicating that only the top p of $A(i,j)$ is retained, a binary mask M_k preserves those locations while the remaining pixels are replaced with a blurred baseline x_{blur} (Gaussian blur, σ =35):

$$x_p = M_p \odot x + (1 - M_p) \odot x_{blur}$$

where $\odot$ denotes elementwise multiplication across spatial locations.

Let $y(\cdot)$ denote the model's output score for class c, i.e., the raw value produced by the final linear layer before probabilities are computed. This score can take any real value, with higher values indicating stronger evidence for the class c. For each masked input x_p we compute

$$\Delta(p) = max\left(0,\, y(x_p) - y(x_{blur})\right)$$

which measures how strongly the preserved region increases the prediction compared to the blurred baseline.

These effects are projected back to the spatial domain via

$$S(i,j) = \sum_p w(p)\ \Delta(p)\, M_p(i,j), \quad w(p) = \frac{1}{\sqrt{p}}$$

Where $S(i,j)$ reflects the accumulated contribution of the preserved regions across different mask sizes. The factor $w(p) = 1/\sqrt{p}$ slightly down-weights masks with larger area levels p, which would otherwise add more to $S(i,j)$ simply because they cover more pixels. This keeps the influence of small, focused masks and larger masks on a comparable scale.

The resulting map is combined with the Grad-CAM map to obtain the final Hybrid-CAM heatmap:

$$H(i,j) = norm\big(A(i,j) \odot S(i,j)\big)$$

H highlights locations that are strongly emphasized by Grad-CAM and that still increase the class score when they are preserved in the perturbed inputs x_p. This focuses the heatmap on regions that make a robust positive contribution to the prediction, rather than on broad areas with weak or diffuse responses.

3 Faithfulness Evaluation

We assess explanation quality using the insertion–deletion (I/D) framework from RISE (Randomized Input Sampling for Explanation) [11], a widely used faithfulness metric for saliency maps [5]. Each image is divided into a $g \times g$ grid, meaning it is split into g rows and g columns of equally sized patches. For every patch, we compute a relevance score as the average value of the explanation map $A(i,j)$ inside that patch. Sorting these patch scores then gives an ordering from most to least relevant. Masked patches are replaced with a Gaussian-blurred baseline image σ =35. For a fraction p of inserted or removed patches, we measure the model's score, denoted $P_{\text{ins}}(p)$ and $P_{\text{del}}(p)$. Faithfulness is summarized by the area under these curves:

$$\text{AUC}_{\text{ins/del}} = \int_0^1 P_{\text{ins/del}}(p)\ dp$$

A high AUC_{ins} indicates that relevant regions are added early, whereas a low AUC_{del} indicates that removing highlighted regions strongly reduces the prediction. We vary the explanation layer (Stage 3 Block 9 and Stage 4 Block 3). The perturbation grid was fixed to $g = 24 \times 24$, and all metrics used the model's predicted class as the target to evaluate faithfulness to its own decision.

4 Quantitative Results

	Score-CAM $\text{AUC}_{\text{ins}}/\text{AUC}_{\text{del}}$	Grad-CAM $\text{AUC}_{\text{ins}}/\text{AUC}_{\text{del}}$	Hybrid-CAM $\text{AUC}_{\text{ins}}/\text{AUC}_{\text{del}}$
TP	0.407/0.955	0.958/0.361	0.963/0.350
TN	0.941/0.830	0.931/0.727	0.942/0.742

Table 1: Average Insertion and Deletion AUC for Score-CAM, Grad-CAM, and Hybrid-CAM at Stage 3 Block 9, computed with a blur baseline (σ = 35) and a 24×24 grid. Values are averaged over TP (32 cases) and TN (83 cases).

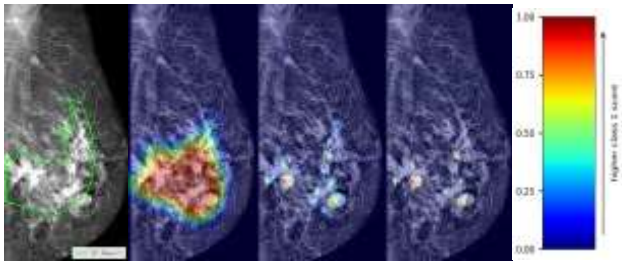
	Score-CAM $\text{AUC}_{\text{ins}}/\text{AUC}_{\text{del}}$	Grad-CAM $\text{AUC}_{\text{ins}}/\text{AUC}_{\text{del}}$	Hybrid-CAM $\text{AUC}_{\text{ins}}/\text{AUC}_{\text{del}}$
TP	0.911/0.254	0.593/0.737	0.612/0.735
TN	0.957/0.879	0.906/0.888	0.827/0.928

Table 2: Average Insertion and Deletion AUC at Stage 4 Block 3.

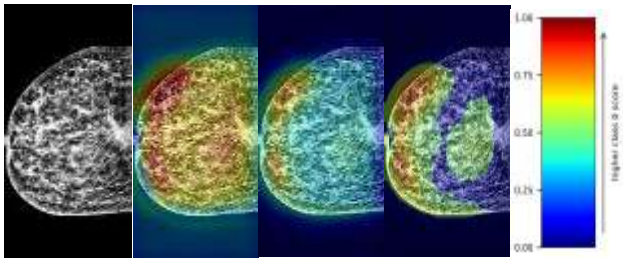
For true positives (TP) in Stage 3 Block 9 as shown in Table 1, Hybrid-CAM achieves the strongest results (0.963/0.350), closely followed by Grad-CAM (0.958/0.361). Both methods clearly affect the model's prediction when their highlighted regions are modified, suggesting that they capture the area that are most relevant for the decision. Score-CAM performs weakly (0.407/0.955), meaning that modifying its highlighted regions has little impact on the output. Table 2 shows the results in Stage 4 Block 3, the scores for Hybrid-CAM and Grad-CAM become weaker, with lower insertion and higher deletion values. This indicates that changes in the highlighted regions have a less direct effect on the prediction at this depth. For true negatives, all methods show higher deletion values, which is expected because benign predictions do not rely on a single dominant region but rather on the overall absence of suspicious patterns. Overall, Hybrid-CAM provides slightly more faithful explanations than Grad-CAM at Stage-3 Block-9. In contrast, Score-CAM performs best at the deepest layer (Stage-4 Block-3).

5 Qualitative Results

To complement the quantitative findings, we qualitatively inspected the explanations for a representative true-positive and true-negative example Fig. 3a and Fig. 3b. Hybrid-CAM and Grad-CAM are shown at Stage 3 Block 9, where they exhibited the strongest quantitative effect. Score-CAM is shown at Stage 4 Block 3. In these visualizations, the color scale indicates the relative contribution to the predicted class score: for malignant cases, the maps show contributions to the class-1 score, while for benign cases, they show contributions to the class-0 score.



Original, Score-CAM, Grad-CAM, Hybrid-CAM
(a) True positive



Original, Score-CAM, Grad-CAM, Hybrid-CAM
(b) True negative

Figure 3: Qualitative comparison of Score-CAM, Grad-CAM, and Hybrid-CAM for a TP and TN example. Each block shows the original image and the corresponding saliency maps, with a separate color bar indicating the score scale for the respective class.

Fig. 3a shows a true-positive mammogram case. Both Hybrid-CAM and Grad-CAM highlight a small number of distinct regions that contribute to the malignant prediction. Grad-CAM displays several separate activation points, whereas Hybrid-CAM produces a more compact pattern in similar locations. Score-CAM shows a broader activation pattern in this deeper layer, which matches its stronger quantitative influence at Stage 4 Block 3.

Fig. 3b illustrates a true-negative example. The highlighted regions mainly contribute to the benign prediction. Hybrid-CAM and Grad-CAM show low-intensity responses, often around breast edges and broader texture patterns, suggesting that the model does not rely on a single dominant region for this decision. Score-CAM again produces a wider activation pattern at Stage 4 Block 3, consistent with the more diffuse behavior observed in the quantitative evaluation for this layer.

Overall, the qualitative examples illustrate different activation behaviors within the selected layers, with Hybrid-CAM and Grad-CAM showing more compact patterns at Stage 3 Block 9 and Score-CAM appears more diffuse in Stage 4 Block 3 used for visualization.

6 Conclusion

This work investigated explainability in a CNN-based classifier using Class Activation Mapping techniques to explain the model decision using saliency maps. We introduced Hybrid-CAM, a refinement of Grad-CAM that incorporates controlled perturbations to assess whether highlighted regions truly affect the model output. Quantitative insertion-deletion metrics and qualitative examples showed that Hybrid-CAM produces more focused and faithful explanations than Grad-CAM and Score-CAM, particularly at intermediate network layers. These results indicate that evaluating the influence of highlighted regions can provide clearer insight into how deep learning models form their predictions in breast imaging.

In future work, we will take the main feature of ConvNextv2 and modify each of these features to find out how harsh the result will change. It will clarify for the novice medical practitioner to learn to identify cancerous lesions.

7 References

[1] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, “Deep Learning to Improve Breast Cancer Detection on Screening Mammography,” *Sci Rep*, vol. 9, no. 1, p. 12495, Aug. 2019, doi: 10.1038/s41598-019-48995-4.

[2] L. Alzubaidi *et al.*, “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,” *J Big Data*, vol. 8, no. 1, p. 53, 2021, doi: 10.1186/s40537-021-00444-8.

[3] D. K. Gurmessa and W. Jimma, “Explainable machine learning for breast cancer diagnosis from mammography and ultrasound images: a systematic review,” *BMJ Health Care Inform*, vol. 31, no. 1, p. e100954, Feb. 2024, doi: 10.1136/bmjhci-2023-100954.

[4] C. Y. Y. Wong, S. Y. S. Lee, and R. D. Mahmood, “Contrast-enhanced spectral mammography,” *Singapore Med J*, vol. 65, no. 3, pp. 195–201, Mar. 2024, doi: 10.4103/singaporemedj.SMJ-2021-268.

[5] T. Gomez, T. Fréour, and H. Mouchère, “Metrics for saliency map evaluation of deep learning explanation methods,” 2022, *arXiv*. doi: 10.48550/ARXIV.2201.13291.

[6] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning Deep Features for Discriminative Localization,” 2015, *arXiv*. doi: 10.48550/ARXIV.1512.04150.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,” 2016, doi: 10.48550/ARXIV.1610.02391.

[8] H. Wang *et al.*, “Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks,” 2019, *arXiv*. doi: 10.48550/ARXIV.1910.01279.

[9] Christian Herglotz, Alireza Siyavashi, “Toward an Energy-Efficient and Explainable Neural Network Architecture for Detection of Breast Cancer in Mammography,” *Brandenburgische Technische Universität Cottbus–Senftenberg*, 2023.

[10] S. Woo *et al.*, “ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders,” 2023, *arXiv*. doi: 10.48550/ARXIV.2301.00808.

[11] V. Petsiuk, A. Das, and K. Saenko, “RISE: Randomized Input Sampling for Explanation of Black-box Models,” 2018, *arXiv*. doi: 10.48550/ARXIV.1806.07421.

[12] R. Khaled *et al.*, “Categorized Digital Database for Low energy and Subtracted Contrast Enhanced Spectral Mammography images.” The Cancer Imaging Archive, 2021. doi: 10.7937/29KW-AE92.