Colormaps Matter: Evaluating Their Impact on CNN-Based Clinical Thermal Imaging

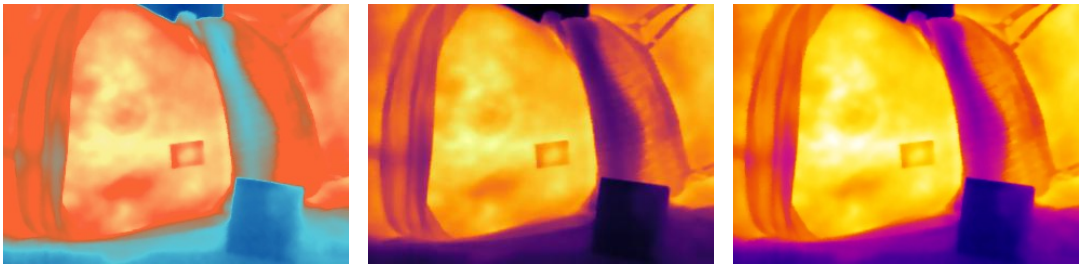
Allison Ng[1] ALLISON.NG@EMORY.EDU
Miriam Asare-Baiden[1] MIRIAM.ASARE-BAIDEN@EMORY.EDU
Sharon Eve Sonenblum[2] SHARONEVE@EMORY.EDU
Kathleen Jordan[2] K.A.JORDAN@EMORY.EDU
Judy Wawira Gichoya[3] JUDYWAWIRA@EMORY.EDU
Vicki Stover Hertzberg[2] VHERTZB@EMORY.EDU
Joyce C Ho[1] JOYCE.C.HO@EMORY.EDU
[1] *Department of Computer Science, Emory University*
[2] *Nell Hodgson Woodruff School of Nursing, Emory University*
[3] *Department of Radiology and Imaging Science, Emory School of Medicine*

Abstract

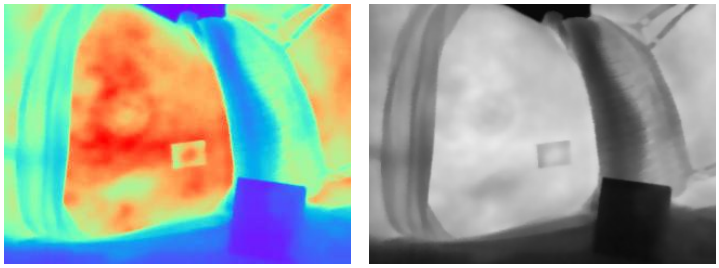
Thermal imaging is an increasingly used modality for clinical deep learning, yet a key preprocessing step in convolutional neural network (CNN)-based pipelines has gone largely unexamined. Each thermal image encodes scalar temperature values rendered into RGB via a colormap, determining how temperature information is visually represented. While existing work shows colormap selection influences human interpretation of thermal images, its impact on CNN-based classification remains unstudied. We address this gap through an evaluation of five colormaps across three pretrained CNN architectures and two clinical prediction tasks. Our results show that colormap choice produces consistent differences in predictive performance and substantially shifts the spatial regions CNNs attend to when making predictions. These findings suggest that colormap selection must be treated and reported as a factor shaping model behavior, with direct implications for how CNN-based pipelines are developed, validated, and deployed in clinical thermal imaging.

1. Introduction

Changes in skin-surface temperature serve as a reliable biomarker across dozens of conditions and therapeutic areas (Kesztyüs et al., 2023), from tumor-associated angiogenesis in breast tissue to inflammation or ischemia preceding pressure injury (Arora et al., 2008; Tzen et al., 2025). Infrared thermography, a non-invasive, non-contact modality, captures this signal directly, making it a natural candidate for automated analysis using computer vision techniques (Kaczmarek and Nowakowski, 2016). Deep learning methods, specifically Convolutional Neural Networks (CNNs), have achieved clinician-comparable performance when applied to other imaging modalities, including dermoscopy, fundus imaging, x-ray, and histopathology (Esteva et al., 2017; Litjens et al., 2017; Rajpurkar et al., 2018). Thermal images, however, pose a distinct challenge: unlike optical modalities where images are captured directly in RGB (Red, Green, Blue), thermal images encode scalar temperature values that must first be mapped into a visible representation. Perhaps as a result, CNN adoption for thermal imaging has remained comparatively limited, though recent work has



(*a*) Arctic (*b*) Inferno (*c*) Ironbow



(*d*) Rainbow (*e*) White Hot

Figure 1: Comparison of different colormaps applied to the same thermal measurements for the erythema classification task. Each colormap produces a distinct visual representation of temperature patterns.

shown promising results in breast cancer detection and pressure injury monitoring (Baron et al., 2023; Attallah, 2025; Chi et al., 2025; Wang et al., 2021).

As CNN-based approaches for thermal images have gained traction, a subtle but consequential upstream decision has gone largely unexamined: colormap selection. The thermography community has recognized that colormap choice shapes human interpretation of thermal images, affecting detection accuracy, segmentation quality, and diagnostic confidence (Tian et al., 2025; Colley et al., 2025; Agrawal and Karar, 2018; Ammer, 2015). Colormap selection has accordingly been treated as a visualization decision, often defaulted to extraction software presets such as the Ironbow palette from FLIR, a leading thermography manufacturer. But as these same color-encoded images serve as input to CNN-based classifiers, colormap choice can no longer be considered a purely visual concern.

CNNs pretrained on large natural image datasets such as ImageNet are sensitive to color. Prior work has demonstrated that color distortions and distribution shifts can meaningfully degrade CNN classification performance (De and Pedersen, 2022; Geirhos et al., 2018), and that color-specific units exist within pretrained networks at varying depths (Engilberge et al., 2017). When scalar thermal data is rendered into RGB space via different colormaps, the resulting images present different color distributions to these pretrained models. Whether these color sensitivities translate into downstream effects when CNN-based classifiers are deployed for medical thermal image analysis, however, remains untested.

To address this gap, we conduct a controlled empirical study evaluating five colormaps shown in Figure 1, across three widely used pretrained CNN architectures (ResNet50, MobileNetV2, InceptionV3) and two clinical thermal imaging tasks, breast cancer detection

and erythema classification. Our results demonstrate that colormap choice produces consistent differences in predictive performance and reveal substantial instability in the spatial regions models attend to when making predictions, as measured through Grad-CAM activation maps. We further show that colormap mismatch between training and deployment can degrade performance to near-chance levels. Our findings suggest that colormap selection carries consequences beyond the human-centered visualization convention. Since colormaps will continue to be designed for human perceptual needs, model training pipelines should treat colormap as a pipeline parameter whose interaction with model behavior is validated and reported rather than freely re-optimized. More broadly, this work functions as a pipeline audit of an upstream preprocessing step with direct implications for clinical deployment risk.

Generalizable Insights about Machine Learning in the Context of Healthcare

While this work focuses on two clinical thermal imaging tasks, the findings have broader implications. Within thermography, colormap selection is relevant to any CNN-based pipeline operating on rendered images, including tasks such as segmentation and detection. The fact that colormap choice affects not only predictive performance but also model attention has relevance for clinical deployment, where interpretability and trust in model behavior are essential. Our cross-palette generalization results further underscore this risk in practice, showing that colormap mismatch between training and deployment can severely degrade performance, a concern for multi-site settings where institutions or camera manufacturers may default to different palettes. Thermography has demonstrated clinical utility across a wide range of conditions that have yet to see CNN-based approaches applied (Kesztyüs et al., 2023). Establishing pipeline decision practices, such as colormap selection, may help enable more reliable CNN adoption across thermal imaging applications more broadly.

2. Related Work

2.1. Deep Learning for Medical Imaging

CNNs are widely adopted for image-based medical diagnosis, including mammography, dermoscopy, and histopathology (Litjens et al., 2017; Chen et al., 2025). This success is largely driven by the availability of architectures pretrained on large-scale natural image datasets such as ImageNet (Deng et al., 2009), including ResNet (He et al., 2016b), Inception (Szegedy et al., 2016), and MobileNet (Sandler et al., 2018), which can be fine-tuned for medical tasks with relatively limited labeled data. Transfer learning from these architectures has been widely adopted in healthcare applications, including dermatology and cancer diagnosis (Esteva et al., 2017; Miotto et al., 2018), and more recently in thermal imaging contexts (Wang et al., 2021, 2026; Jiang et al., 2022; Attallah, 2025; Chi et al., 2025). More recently, domain-specific alternatives such as RadImageNet (Mei et al., 2022), pretrained on medical imaging data rather than natural images, have been proposed to reduce the domain gap inherent to ImageNet transfer learning.

2.2. Thermal Imaging in Clinical Applications

Infrared thermography offers a fast, non-contact, radiation-free method to measure temperature differences on the skin surface, with the capacity for arbitrary repetition of recordings

(Kaczmarek and Nowakowski, 2016). Deviations in surface temperature can arise from many underlying causes, including inflammation, malignancy, and infection, making thermography broadly applicable across clinical settings (Lahiri et al., 2012). Kesztyüs et al. (2023) identified the use of thermography across 38 distinct health conditions spanning 13 therapeutic areas, yet CNN-based approaches remain largely underexplored.

2.2.1. Breast Cancer Detection

Breast cancer remains one of the leading causes of cancer-related mortality among women worldwide, making early detection critical (Kim et al., 2025). Mammography is the current gold standard for screening, yet its sensitivity drops for young women and women with dense breasts (Hollingsworth, 2019). Thermography offers an alternative for women with dense breasts and can serve as an effective approach in combination with ultrasound for low-income countries (Arora et al., 2008; Goñi-Arana et al., 2024). Furthermore, applying CNN-based approaches to thermal images has shown promising classification performance in distinguishing healthy from malignant cases (Ekici and Jawzal, 2020; Husaini et al., 2020; Baffa and Lattari, 2018; Chi et al., 2025; Attallah, 2025).

2.2.2. Erythema and Pressure Injury Detection

Pressure injuries remain a common and costly adverse event in hospitalized patients, and are linked to increased morbidity and mortality (Li et al., 2020). Early detection is therefore critical to prevention, and thermography has emerged as a promising tool for this purpose. Temperature differentials measured up to 48 hours before diagnosis have been associated with increased pressure injury risk (Cai et al., 2021). Yet evidence on the overall accuracy of thermographic imaging for early detection remains limited (Baron et al., 2023), motivating recent work that applies CNN-based methods to thermal images with encouraging results (Wang et al., 2021; Jiang et al., 2022).

Erythema, characterized by redness of the skin due to increased blood flow, is the earliest clinically recognized sign of pressure injury (European Pressure Ulcer Advisory Panel et al., 2019). Nonblanchable erythema, when the redness persists under light pressure, indicates that tissue damage has already occurred, and has been shown to predict progression to more severe pressure injury, highlighting the clinical importance of early identification and timely intervention (Shi et al., 2020). Despite this importance, visual skin assessment for erythema is less reliable in patients with darker skin tones (Gunowa et al., 2018). However, temperature changes do not differ significantly across skin tone groups (Bates-Jensen et al., 2024; Jordan et al., 2025). CNN-based methods for erythema detection have been evaluated across imaging protocol variations such as lighting, distance, positioning, and camera type, and found to be less sensitive to these variations than optical imaging (Asare-Baiden et al., 2026). Notably, colormap selection has not been examined.

2.3. Colormaps in Thermal Visualization

Each pixel in a thermal image represents a single scalar temperature value, rendered into a visible image using any number of predefined colormaps (or color palettes) that determine how temperature differences are visually encoded. Prior work has examined how this choice affects human perception and task performance (see Table 1), consistently showing that

Table 1: Prior work on impact of colormaps on human perception and task performance when using thermal images.

Study	Task	White Hot	Ironbow	Rainbow
Ammer (2015)	Structural fidelity	★	✓	✓
Agrawal and Karar (2018)	Target detection	★	✓	✓
Shaikh et al. (2019)	Segmentation & clustering	★		✓
Colley et al. (2025)	Human detection	★	✓	✓
Tian et al. (2025)	Contour recognition	✓	✓	✓

★ = best performing; ✓= evaluated

colormap choice influences detection accuracy, segmentation quality, and cognitive workload (Tian et al., 2025; Colley et al., 2025; Agrawal and Karar, 2018; Shaikh et al., 2019; Ammer, 2015). Notably, White Hot (or grayscale) performed well across most tasks, though Tian et al. (2025) found Ironbow and Rainbow outperformed it at lower resolutions.

While the above work examines colormap effects on human interpretation in thermal imaging, early evidence from other modalities suggests that color encoding more broadly influences CNN-based classification. Colorizing grayscale mammograms using task-specific transfer functions has been shown to improve CNN-based breast cancer classification compared to grayscale inputs (Hussein et al., 2025), and similar effects have been observed applying color mapping to chest X-ray images for COVID-19 detection (Freire et al., 2021).

2.4. Color Sensitivity and Input Representation in CNNs

CNNs pretrained on ImageNet inherit strong priors over natural image color distributions, influencing how models respond to inputs with differing color characteristics (Geirhos et al., 2018). Analysis of learned CNN representations has confirmed the existence of color-specific units within networks, with sensitivity varying across network depth (Engilberge et al., 2017), and color distortions or shifts in color distribution can significantly degrade classification performance across multiple state-of-the-art architectures (De and Pedersen, 2022). In medical imaging, this sensitivity has motivated color normalization to address stain variation in histopathology, which can otherwise degrade model generalization across labs (Litjens et al., 2017), and color space transformations such as RGB to HSV (Hue, Saturation, Value) or CIELAB have similarly been shown to affect CNN classification performance in medical contexts (Velastegui and Pedersen, 2021). Recent work has proposed architectures equivariant to hue, saturation, and luminance variation to improve generalization under color shift (Yang et al., 2025). These findings collectively suggest that color encoding is not a neutral choice.

3. Experimental Setup

We address the colormap selection gap in clinical thermal imaging pipelines through a controlled evaluation study (Figure 2) designed to isolate colormap as the sole source of variation across two clinical thermal imaging tasks, evaluating five colormaps across three pretrained CNN architectures. Code and experimental details are available on GitHub.[1]

1. https://github.com/ngxallison/Colormaps-CNN

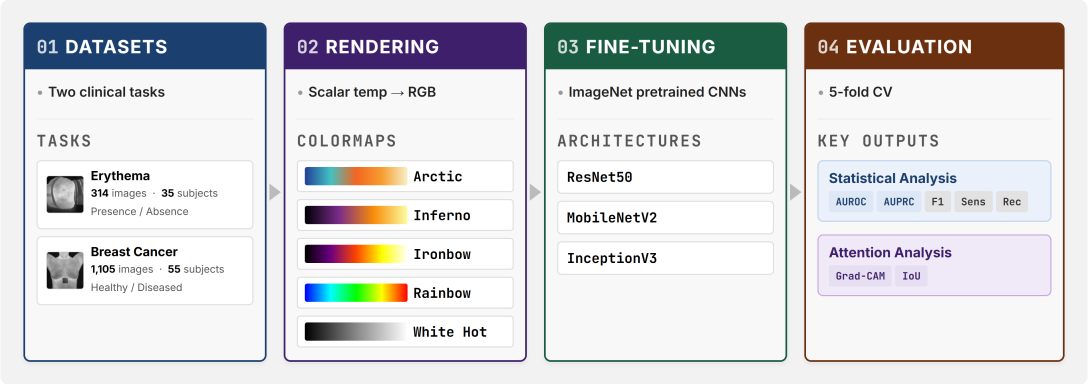


Figure 2: An illustration of our experimental pipeline.

3.1. Datasets

We evaluate colormap sensitivity on two classification tasks using two independent datasets.

Erythema Classification. We use a private dataset of thermal images collected from 35 healthy adult volunteers (ages 18–73), recruited across the Monk Skin Tone scale. The study was approved by the Emory University Institutional Review Board (eIRB number 00005999). Erythema was experimentally induced via a cupping protocol applied to the posterior superior iliac spines, as described in Jordan et al. (2025). Thermal images were captured immediately after cupping and at one-minute intervals over seven minutes, yielding multiple images per subject. Labels indicate the presence or absence of erythema based on colorimeter measurements. This dataset contains 314 total images with 196 labeled with erythema. Jordan et al. (2025) previously characterized the temperature and erythema response patterns, finding that erythema, but not temperature, differed significantly across skin tone groups.

Breast Cancer Detection. We use the publicly available DMR-IR database Silva et al. (2024), which contains thermal images from 55 patients classified as healthy (19) or diseased (36). DMR-IR contains 1105 total images with 34.48% healthy.[2] DMR-IR has been used as a benchmark for existing CNN-based breast cancer thermography research (Attallah, 2025; Chi et al., 2025; Baffa and Lattari, 2018) and its public availability makes it a natural choice for reproducible evaluation.

3.2. Colormaps

Each pixel in a thermal image represents a scalar temperature value that can be rendered into a visible RGB image using a colormap, so the same temperature data produces visually distinct images depending on the colormap applied. We evaluate five colormaps spanning a range of perceptual and practical characteristics (Figure 1). Three were selected for their prevalence in prior work on colormap effects on human interpretation of thermal images (Table 1): White Hot, which renders temperature as grayscale intensity and performed well across human perception tasks; Ironbow, a FLIR-native palette widely used clinically; and

2. The dataset was downloaded from Kaggle at https://www.kaggle.com/datasets/asdeepak/thermal-images-for-breast-cancer-diagnosis-dmrir.

Rainbow, a commonly used but perceptually non-uniform map known to introduce visual artifacts (Tian et al., 2025). We also include Arctic and Inferno. Arctic, a FLIR-native map, combines color contrast for identifying heat sources with subtle shading for detecting slight temperature differences, suiting tasks requiring both coarse and fine thermal discrimination. Inferno is a perceptually uniform map with monotonically increasing luminance from black to bright yellow, common in scientific visualization. Although Ironbow and Inferno may appear visually similar, they differ in luminance range and color structure, making them an informative pair to evaluate. Implementation details are provided in Appendix A.1.

3.3. CNN Architectures

We evaluated three widely used ImageNet-pretrained CNN architectures: ResNet50 (He et al., 2016a), MobileNetV2 (Sandler et al., 2018), and InceptionV3 (Szegedy et al., 2016). These were selected for their prevalence in medical imaging pipelines, strong performance across clinical tasks, and architectural diversity (residual, depthwise separable, and inception-based designs, respectively) (Chen et al., 2025). Since no single architecture is optimal across all tasks, these three CNNs serve as a representative sample of the pretrained models practitioners are likely to deploy in clinical thermal imaging workflows.

3.4. Evaluation Methodology

3.4.1. Training Procedure

For each CNN architecture, the final classification layer was replaced with a randomly initialized linear layer with two output nodes; all preceding layers were initialized with ImageNet-pretrained weights and fine-tuned during training. Models were trained using cross-entropy loss with Adam, with early stopping based on validation AUROC. Full training hyperparameters are provided in Appendix A.2.

3.4.2. Evaluation Setup

To prevent data leakage, we employed 5-fold stratified group cross-validation at the subject level, ensuring all images from a given subject appear exclusively in either the training or test set, stratified by class label to maintain balanced distributions across folds. Given the small number of folds, we report standard deviation rather than confidence intervals; per-fold subject counts are provided in Appendix A.2. All other experimental variables (model architecture, training procedure, dataset splits, random seeds) were held constant to isolate colormap as the sole source of variation (5 colormaps $\times$ 3 architectures $\times$ 2 tasks), with a separate model trained and evaluated exclusively on images rendered with each colormap, without exposure to cross-colormap inputs at any stage. For each architecture and task, we identified the best-performing colormap by mean fold-level AUROC and applied Wilcoxon signed-rank tests against the remaining four, reported without multiple comparison correction given the exploratory nature of this analysis.

3.4.3. Evaluation Metrics

Predictive Performance Metrics. Model performance was evaluated using AUROC and Area Under the Precision-Recall Curve (AUPRC) as primary metrics, alongside sensitivity,

specificity, and F1-score as secondary metrics. Metrics were computed for each fold and reported as mean ± standard deviation across folds. AUROC and AUPRC were prioritized as widely adopted threshold-independent metrics in medical image classification (Kocak et al., 2025). AUPRC is potentially more informative as it emphasizes performance on the positive class (i.e., erythema presence and cancer diagnosis respectively). Secondary metrics are reported to provide a fuller picture of model behavior but are interpreted with the caveat that they depend on a fixed decision threshold and may vary accordingly.

Cross-Palette Generalization. Colormap conventions vary across camera manufacturers and imaging software, raising the risk that a model trained at one clinical site may be deployed on images rendered with a different colormap. To assess this risk, we evaluate whether models retain predictive performance when applied to a colormap different from the one used in training, without retraining or fine-tuning. For both classification tasks, models trained exclusively on Ironbow and, separately, White Hot were evaluated on images rendered with each of the four remaining colormaps.

Interpretability Metrics. To assess how colormap choice influences model attention, we compute Gradient-weighted Class Activation Mapping (Grad-CAM) activation maps (Selvaraju et al., 2017) for each model and colormap condition, applied to the final convolutional layer. We conduct both qualitative analysis through visual inspection of representative activation maps and quantitative analysis through Intersection-over-Union (IoU). For visualization purposes, the Grad-CAM activation maps are displayed using the standard red-blue heatmap (red = high; blue = low) and overlaid onto a grayscale (White Hot) version of the input image. Colormap-specific hue information is removed to ensure the attention regions are visually comparable without interference from the underlying palette.

For IoU computation, activation maps are normalized to $[0, 1]$ and binarized at a fixed threshold ($\tau = 0.5$), retaining the top activations as the predicted attention region. For the erythema dataset, where ground truth segmentation masks are available, IoU is computed between this binary mask and human-provided segmentation masks to assess whether models attend to clinically relevant regions. For both datasets, we additionally compute IoU relative to the White Hot baseline to quantify how colormap choice shifts model attention across conditions. White Hot is selected as the baseline both because it is commonly used for human perception tasks and because its absence of chromatic information provides a neutral reference for isolating the effect of color encoding on model attention.

4. Results

4.1. Effect of Colormap on Classification Performance

4.1.1. Erythema Classification

Table 2 summarizes AUROC and AUPRC across five colormaps for erythema detection. Despite identical underlying thermal data, colormap choice introduces consistent performance variation across all three CNN architectures. White Hot achieves the highest average AUROC (0.772), while Inferno achieves the highest average AUPRC (0.824); Arctic yields the lowest overall performance (AUROC: 0.728, AUPRC: 0.796). The gap between best- and worst-performing colormaps (0.044 AUROC, 0.028 AUPRC) is comparable in magnitude

Table 2: Erythema classification performance across colormaps. † denotes statistically significant pairwise difference from 1+ colormaps (Wilcoxon signed-rank, $p < 0.05$).

	ResNet50		MobileNetV2		InceptionV3		Average	
Colormap	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
Arctic	0.743	0.803	0.687	0.783	0.755	0.802	0.728	0.796
Inferno	0.801	0.849	0.702	0.797	0.751	0.827	0.751	**0.824**
Ironbow	0.713	0.780	**0.763**	0.790	**0.811**†	**0.864**†	0.762	0.811
Rainbow	0.806	**0.852**	0.687	0.763	0.743	0.805	0.745	0.807
White Hot	**0.821**	**0.852**†	0.730	**0.799**	0.766	0.806	**0.772**	0.819
CNN Avg	**0.777**	**0.827**	0.714	0.786	0.765	0.821	–	–

Table 3: Breast cancer classification performance across colormaps. No statistically significant pairwise differences were observed (Wilcoxon signed-rank, $p < 0.05$).

	ResNet50		MobileNetV2		InceptionV3		Average	
Colormap	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
Arctic	0.997	0.999	0.997	**0.999**	**1.000**	**1.000**	0.998	0.999
Inferno	**1.000**	**1.000**	**0.998**	**0.999**	**1.000**	**1.000**	**0.999**	**1.000**
Ironbow	0.994	0.998	**0.998**	**0.999**	0.987	0.995	0.993	0.997
Rainbow	0.995	0.998	0.992	0.997	0.997	0.999	0.995	0.998
White Hot	0.994	0.998	0.991	0.997	0.998	0.999	0.995	0.998
CNN Avg	0.996	**0.999**	0.995	0.998	**0.997**	**0.999**	–	–

to the spread across architectures (0.063 AUROC between ResNet50 and MobileNetV2), suggesting colormap selection matters as much as architecture choice.

We observe that colormap ranking is architecture-dependent. InceptionV3 performs best under Ironbow (AUROC: 0.811, AUPRC: 0.864), while ResNet50 performs strongest under White Hot (AUROC: 0.821, AUPRC: 0.852). Where significant differences emerge (denoted †), they follow the same pattern. Ironbow significantly outperforms three of four alternative colormaps for InceptionV3 on both AUROC and AUPRC, while White Hot shows a significant AUPRC advantage over Arctic for ResNet50 (Appendix B). The optimal colormap also depends on the evaluation metric. White Hot maximizes AUROC while Inferno achieves the highest AUPRC, which may be relevant when sensitivity to the positive class is prioritized. Notably, Ironbow achieves the strongest ranking performance for InceptionV3 but ranks among the weakest in sensitivity and F1, suggesting colormap choice may differentially affect ranking ability versus threshold-dependent measures (Appendix B).

4.1.2. Breast Cancer Detection

Table 3 summarizes AUROC and AUPRC across five colormaps for breast cancer classification. In contrast to erythema, performance is uniformly high across all colormap and architecture combinations, ranging from 0.987 to 1.000 AUROC and 0.995 to 1.000 AUPRC. Inferno achieves the highest average performance (AUROC: 0.999, AUPRC: 1.000) and the best performance across all three architectures individually, while Ironbow yields the lowest (AUROC: 0.993, AUPRC: 0.997). Interestingly, White Hot, one of the strongest colormaps for erythema, performs among the weakest here, suggesting colormap ranking is not consistent across tasks. No statistically significant pairwise differences are observed for any

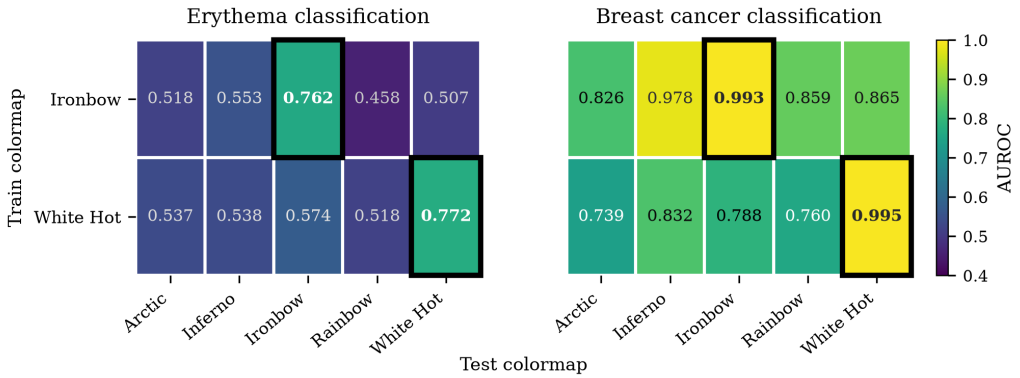


Figure 3: Cross-palette generalization AUROC (CNN-averaged) for erythema and breast cancer classification. Rows indicate the colormap used for training; columns indicate the colormap used for evaluation, without retraining. Bordered cells indicate in-distribution performance (train colormap = test colormap).

architecture, and the spread across colormaps (0.006 AUROC) is an order of magnitude smaller than for erythema (0.044 AUROC). Per-architecture metrics, including secondary measures, are provided in Appendix C.

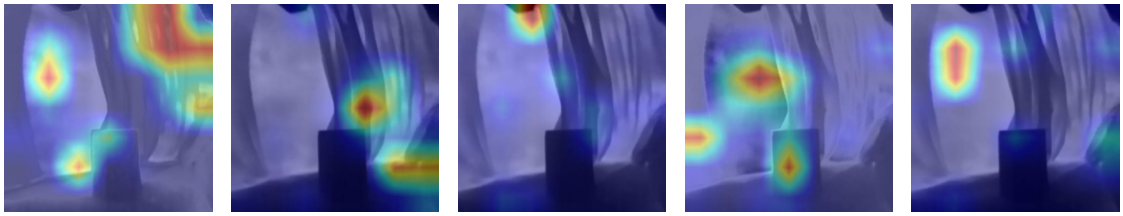
4.2. Cross-Palette Generalization

Figure 3 summarizes the cross-palette generalization experimental results for both tasks, reporting CNN-averaged AUROC when models trained on Ironbow or White Hot are evaluated on the remaining four colormaps without retraining. Performance degrades substantially for erythema classification. Models trained on Ironbow, which achieve an in-distribution AUROC of 0.762 (Table 2), drop to an average of 0.518 on Arctic, 0.553 on Inferno, 0.458 on Rainbow, and 0.507 on White Hot, an average degradation of up to 0.304 AUROC (Ironbow to Rainbow). Models trained on White Hot, with an in-distribution AUROC of 0.772, show a comparable pattern, falling to an average of 0.518 AUROC on Rainbow, a decline of 0.254. For several target colormaps, cross-palette AUROC falls near or below chance level (0.5), indicating a near-total loss of discriminative ability rather than a modest performance dip. Breast cancer classification degrades by a similar magnitude. Models trained on Ironbow (in-distribution AUROC 0.993) retain AUROC between 0.826-0.978 across target colormaps, a drop of up to 0.167. Models trained on White Hot (in-distribution AUROC 0.995) yield AUROC of 0.739-0.832, a drop of up to 0.256. This consistency across two tasks with very different baseline performance suggests that colormap mismatch degrades performance regardless of task. Full per-architecture results, including AUPRC, are provided in Appendix D.

4.3. Effect of Colormap on Model Attention

4.3.1. Erythema Classification

Figure 4 shows Grad-CAM activations for ResNet50 (the best-performing architecture) on the same erythema-positive subject and timepoint across all five colormaps. White Hot



(*a*) Arctic (*b*) Inferno (*c*) Ironbow (*d*) Rainbow (*e*) White Hot

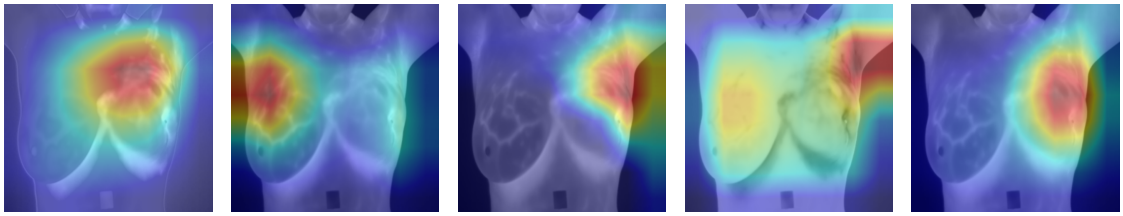
Figure 4: Grad-CAM visualizations for ResNet50 on an erythema-positive (left lower back or upper left) case across the colormaps. Images were rendered in grayscale for display only and differences in highlighted regions indicate colormap selection affects model attention.

Table 4: Erythema IoU against ground truth and White Hot baseline references.

	Ground truth			**White Hot baseline**		
Colormap	**ResNet50**	**MobileNetV2**	**InceptionV3**	**ResNet50**	**MobileNetV2**	**InceptionV3**
Arctic	0.014 ± 0.043	0.012 ± 0.038	0.020 ± 0.052	0.089 ± 0.182	0.096 ± 0.168	0.145 ± 0.215
Inferno	0.008 ± 0.038	0.016 ± 0.033	0.016 ± 0.025	**0.101 ± 0.200**	0.149 ± 0.242	**0.232 ± 0.238**
Ironbow	0.003 ± 0.015	0.022 ± 0.042	**0.028 ± 0.057**	0.069 ± 0.160	**0.162 ± 0.239**	0.141 ± 0.237
Rainbow	0.005 ± 0.028	**0.032 ± 0.061**	0.023 ± 0.054	0.042 ± 0.139	0.113 ± 0.214	0.160 ± 0.208
White Hot	**0.021 ± 0.062**	0.018 ± 0.052	0.024 ± 0.035	–	–	–

produced the highest overlap with the erythema region (IoU=0.32) and correctly predicted erythema presence. Inferno also predicted correctly but attended instead to the upper back and clothing (IoU=0), suggesting the correct prediction was not driven by the erythema region. Arctic showed partial overlap (IoU=0.05), while Ironbow and Rainbow showed none. Ground truth mask overlays for Arctic and White Hot are provided in Appendix Figure 6. These differences illustrate that colormaps produce distinct spatial activation patterns, and that correct predictions are not always driven by clinically relevant regions.

Table 4 reports IoU against ground truth segmentation masks and against the White Hot baseline. Values are uniformly low across both references, all colormaps, and all architectures, with high standard deviations suggesting considerable variability across subjects. These low ground truth values indicate that Grad-CAM attention does not reliably localize to the clinically relevant erythema region under any colormap (as illustrated qualitatively in Figure 4). Given known limitations of Grad-CAM as a localization method in medical imaging (Saporta et al., 2022), we interpret this primarily as evidence of attention instability rather than a definitive claim about spatial reasoning quality. Even within this noisy metric, colormap rankings are inconsistent across architectures. Rainbow yielded the highest ground-truth IoU for MobileNetV2 while White Hot performed among the lowest for that architecture, yet the pattern reversed for ResNet50. Against the White Hot baseline, Inferno yielded the highest IoU for InceptionV3 and ResNet50, and Ironbow for MobileNetV2, while Rainbow consistently produced the lowest overlap across all architectures. This instability suggests attention alignment must be evaluated per architecture, not assumed.



(*a*) Arctic (*b*) Inferno (*c*) Ironbow (*d*) Rainbow (*e*) White Hot

Figure 5: Grad-CAM visualizations for InceptionV3 on a breast cancer-positive case across five colormaps. Images rendered in grayscale for display only; differences in highlighted regions indicate colormap selection affects model attention.

Table 5: IoU comparison across colormaps against the White Hot baseline for breast cancer classification. Values are reported as mean ± standard deviation.

Colormap	ResNet50	MobileNetV2	InceptionV3
Arctic	**0.176 ± 0.195**	0.248 ± 0.189	0.337 ± 0.241
Inferno	0.141 ± 0.211	0.350 ± 0.330	0.189 ± 0.205
Ironbow	0.117 ± 0.180	0.320 ± 0.293	0.292 ± 0.180
Rainbow	0.175 ± 0.290	**0.385 ± 0.336**	**0.368 ± 0.224**

4.3.2. Breast Cancer Detection

Figure 5 shows Grad-CAM activations for InceptionV3 (the best-performing architecture for this task) on the same subject across all five colormaps. Arctic, Ironbow, and White Hot concentrated attention on the left breast, while Inferno shifted focus entirely to the right breast. Rainbow produced the most distributed pattern, spanning both breasts and extending into the armpit (roughly the union of the other colormaps' regions), with Ironbow showing a milder version of this same extension. These differences are striking given uniformly high classification performance across all colormaps (Table 3), indicating similar performance can coincide with substantially different spatial activation patterns.

Table 5 reports IoU against the White Hot baseline, the only available reference since no ground truth segmentation masks exist for this dataset. Values thus quantify relative attention shifts between colormaps rather than clinical alignment. IoU values are generally low across colormaps and architectures, reflecting substantial attention shifts. Rainbow yielded the highest IoU for MobileNetV2 and InceptionV3, consistent with its broad activation pattern, while Inferno yielded the lowest for ResNet50 and InceptionV3, possibly reflecting focus on a different breast side than other colormaps. High standard deviations throughout suggest attention shifts vary considerably across subjects.

5. Discussion

5.1. Colormap as a Tunable Design Parameter

CNNs pretrained on ImageNet develop sensitivity to color distributions in natural images. When scalar thermal data is mapped into RGB space via different colormaps, the resulting images align differently with learned filters in early convolutional layers, which may ex-

plain the performance variation we observe despite identical underlying temperature data. Grayscale-based colormaps such as White Hot preserve monotonic intensity relationships that may align with luminance-sensitive filters, while perceptually uniform colormaps such as Inferno introduce structured gradients that may aid detection of subtle patterns. Non-uniform colormaps such as Rainbow may introduce artificial edges that interact unpredictably with learned representations, consistent with Rainbow's low baseline IoU across architectures (Table 4), though these explanations remain correlational rather than mechanistically established.

This effect is further modulated by architecture. InceptionV3 is most sensitive to colormap choice, while MobileNetV2 is comparatively robust. This suggests architectural characteristics such as receptive field size influence how color-encoded information is processed. Colormap selection and model architecture should therefore not be treated independently. We do not propose a universal decision rule. Ironbow is optimal for InceptionV3 on erythema classification, while White Hot is optimal for ResNet50 on the same task. This inconsistency is itself informative: any fixed recommendation would be misleading in practice. We instead position colormap as a parameter requiring empirical validation per architecture and task, analogous to hyperparameter tuning. We release our code and colormap implementations to lower the barrier for practitioners to do so.

Colormap sensitivity also persists under medical-domain pretraining and temperature normalization. When using RadImageNet weights (Mei et al., 2022) instead of ImageNet, colormap choice continued to produce meaningful performance variation for erythema classification, with the best-performing colormap shifting relative to the ImageNet-pretrained setting (Appendix E.1). This suggests colormap sensitivity is not merely an artifact of the natural-to-medical domain gap. A sensitivity analysis using global rather than per-image min-max normalization (Appendix E.2) shows a similar pattern.

5.2. Implications for Clinical Deployment

Beyond predictive performance, colormap selection carries implications for clinical interpretability. The breast cancer attention shifts we observe show that colormap effects are not fully captured by predictive metrics alone, and raise the question of whether models exploit distributed discriminative features or dataset-specific shortcuts. Given the low ground-truth IoU across all conditions, we treat our Grad-CAM findings as a warning signal of attention instability. That attention shifts meaningfully with an ostensibly cosmetic preprocessing choice is itself significant, regardless of Grad-CAM's known limitations as a tool. In clinical settings where attention visualizations support human decision-making, this raises concerns beyond accuracy (Amirian et al., 2025; Sun et al., 2025): a colormap yielding strong predictive performance may simultaneously produce misleading attention maps, undermining clinician trust (Rosenbacke et al., 2024). Colormap selection should therefore be evaluated along two complementary axes, predictive performance and attention alignment, rather than optimizing for one alone, particularly for applications such as pressure injury monitoring or cancer screening where interpretability is critical.

A further deployment concern is colormap mismatch between model development and real-world use. Thermal imaging pipelines often rely on extraction software presets that differ across manufacturers and institutions (e.g., FLIR defaults to Ironbow). Our cross-

palette results show this mismatch can degrade performance to near-chance levels. This risk is subtle because visually similar colormaps can present meaningfully different color distributions. Ironbow and Inferno appear similar to a human observer yet produce notably different model behavior in our experiments. Colormap selection should therefore be treated as a reportable pipeline parameter alongside model architecture and training procedure, to reduce the risk of silent performance degradation in clinical deployment.

5.3. Limitations

This study has several limitations. First, both datasets are relatively small for CNN-based approaches. The erythema dataset derives from controlled cupping-induced erythema rather than clinical presentations, so results should not be interpreted as establishing a clinically deployable detection model. The breast cancer task's near-ceiling performance limits its informativeness for studying colormap-driven performance effects, though it remains useful for attention-shift analysis. A harder variant such as reduced resolution or fewer samples could offer a more sensitive test, but we prioritize the DMR-IR benchmark as published for comparability with prior work (Attallah, 2025; Chi et al., 2025; Baffa and Lattari, 2018). Similarly, our RadImageNet (Mei et al., 2022) robustness check (Appendix E.1) was limited to ResNet50 and InceptionV3 on the erythema task, as no domain-matched (i.e., thermal-image-pretrained) source currently exists to our knowledge and RadImageNet weights were unavailable for MobileNetV2.

Second, the five colormaps evaluated span a range of perceptual and practical characteristics but are not exhaustive. They also differ along multiple axes simultaneously, including luminance range, perceptual uniformity, and hue coverage, and disentangling these individual properties' contributions remains open, requiring a factorial design beyond the scope of our pipeline-level audit. We similarly did not explore whether colormap-specific hyperparameter tuning (e.g., learning rate) could equalize performance across colormaps, as this would confound colormap choice with training procedure; joint tuning is a natural extension. It is also an open question whether similar effects hold in segmentation or detection settings, beyond the classification tasks studied here. We do not include a single-channel raw-temperature baseline, since a true single-channel model would preclude ImageNet pretraining, which is important at this data scale (Tajbakhsh et al., 2016). Learned input adapters that preserve pretraining benefits while operating closer to the raw signal are a promising direction for future work.

Third, our interpretability analysis relies on Grad-CAM, which has known limitations as an attribution method in medical imaging contexts. Prior work shows low localization performance on chest X-rays, particularly for smaller or more complex findings (Saporta et al., 2022). Additionally, saliency maps exhibit low sensitivity and weak robustness, and may be irrelevant to model output even when classification performance appears strong (Zhang et al., 2024). The uniformly low ground truth IoU we observe across all colormaps and architectures is consistent with these known limitations, suggesting that the attention shifts we observe should be interpreted with caution. Future work should examine whether our attention findings are robust across multiple interpretability approaches.

Finally, our analysis is limited to CNN-based architectures without evaluating vision transformers (ViTs) or vision-language models (VLMs). Preliminary TinyViT (Wu et al.,

2022) experiments on both tasks (Appendix E.3) suggest colormap sensitivity is not unique to CNNs, showing performance on par with, but not exceeding, the CNN architectures, consistent with the larger data requirements typically associated with ViTs (Dosovitskiy et al., 2021; Touvron et al., 2021). However, whether this sensitivity holds under cross-palette generalization or attention-based analysis remains open for both ViTs and VLMs. ViT-adapted Grad-CAM could further extend our attention analysis to these architectures. Color encoding has also proven relevant beyond CNNs in other modalities. For example, PseudoColorViT-CXR (Ahmed, 2025) shows pseudo-color transformations improve ViT-based chest X-ray classification over grayscale inputs. Broader evaluation across larger, more diverse datasets, a wider range of colormaps, additional task types, and alternative architectures would help establish more generalizable conclusions.

6. Conclusion

We presented a systematic evaluation of colormap choice in CNN-based clinical thermal imaging pipelines. By isolating colormap as the sole source of variation across three CNN architectures, two clinical tasks, and five colormaps, we demonstrated that this seemingly incidental preprocessing choice introduces variation in both predictive performance and model attention. Our findings suggest that colormap selection must be understood and validated as a factor influencing model behavior, alongside model architecture, rather than defaulted to extraction software presets without scrutiny. This raises the possibility that model training and human visualization needs may sometimes call for different colormap choices, a tension we see as an important direction for future work rather than one this paper resolves. Moreover, the implications extend beyond performance: colormap choice influences where models attend, with consequences for clinical interpretability and human-AI interaction. These findings motivate more principled practices around colormap selection in thermal imaging pipelines. Future work should extend this analysis to larger datasets and other tasks, broader colormap spaces, and emerging architectures such as vision transformers and vision-language models, as well as to explore color-robust representations that reduce sensitivity to color encoding choices altogether.

Acknowledgments

This research was supported in part by the National Center for Advancing Translational Sciences of the National Institutes of Health (NIH) under Award Number UL1TR002378, and by the National Institute of Biomedical Imaging and Bioengineering of the NIH under Award Number R01EB036579. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Divya Agrawal and Vinod Karar. Color palette selection in thermal imaging for enhancing situation awareness during detection-recognition tasks. *2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering*, 00:1227–1232, 2018. doi: 10.1109/icrieece44171.2018.9008486.

Faisal Ahmed. Pseudocolorvit-cxr: Colormap-enhanced vision transformers for tuberculosis and pneumonia detection from grayscale chest x-ray images. *Available at SSRN 5547319*, 2025.

Soheyla Amirian, Fengyi Gao, Nickolas Littlefield, Jonathan H. Hill, Adolph J. Yates, Johannes F. Plate, Liron Pantanowitz, Hooman H. Rashidi, and Ahmad P. Tafti. State-of-the-art in responsible, explainable, and fair AI for medical image analysis. *IEEE Access*, 13:58229–58263, 2025. ISSN 2169-3536. doi: 10.1109/access.2025.3555543.

Kurt Ammer. The influence of colour scale on the accuracy of infrared thermal images. In *Infrared Imaging: A casebook in clinical medicine*, pages 4–1. IOP Publishing Bristol, UK, 2015.

Nita Arora, Deborah Martins, Diana Ruggerio, Eleni Tousimis, Alexander J. Swistel, Michael P. Osborne, and Rache M. Simmons. Effectiveness of a noninvasive digital infrared thermal imaging system in the detection of breast cancer. *The American Journal of Surgery*, 196(4):523–526, 2008. doi: 10.1016/j.amjsurg.2008.06.015.

Miriam Asare-Baiden, Sharon Eve Sonenblum, Kathleen Jordan, Glory Tomi John, Andrew Chung, Judy Wawira Gichoya, Vicki Stover Hertzberg, and Joyce C Ho. Impact of imaging protocols on thermal detection of pressure injuries: Threshold versus deep learning across skin tones. *medRxiv*, pages 2026–05, 2026.

Omneya Attallah. Harnessing infrared thermography and multi-convolutional neural networks for early breast cancer detection. *Scientific Reports*, 15(1):27464, 2025. doi: 10.1038/s41598-025-09330-2.

Matheus F O Baffa and Lucas G Lattari. Convolutional neural networks for static and dynamic breast infrared imaging classification. *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images*, pages 174–181, 2018. doi: 10.1109/sibgrapi.2018.00029.

Miriam Viviane Baron, Paulo Ricardo Hernandes Martins, Cristine Brandenburg, Janine Koepp, Isabel Cristina Reinheimer, Amanda Corrêa dos Santos, Michele Paula dos Santos, Andres Felipe Mantilla Santamaria, Thomas Miliou, and Bartira Ercília Pinheiro da Costa. Accuracy of thermographic imaging in the early detection of pressure injury: A systematic review. *Advances in Skin & Wound Care*, 36(3):158–167, 2023. ISSN 1527-7941. doi: 10.1097/01.asw.0000912000.25892.3f.

Barbara M. Bates-Jensen, Kathleen Jordan, William Jewell, and Sharon E. Sonenblum. Thermal measurement of erythema across skin tones: Implications for clinical identification of early pressure injury. *Journal of Tissue Viability*, 2024. ISSN 0965-206X. doi: 10.1016/j.jtv.2024.08.002.

Fuman Cai, Xiaoqiong Jiang, Xiangqing Hou, Duolao Wang, Yu Wang, Haisong Deng, Hailei Guo, Haishuang Wang, and Xiaomei Li. Application of infrared thermography in the early warning of pressure injury: A prospective observational study. *Journal of Clinical Nursing*, 30(3-4):559–571, 2021. ISSN 0962-1067. doi: 10.1111/jocn.15576.

Chao Chen, Nor Ashidi Mat Isa, and Xin Liu. A review of convolutional neural network based methods for medical image classification. *Computers in Biology and Medicine*, 185: 109507, 2025. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2024.109507.

Thanh Nguyen Chi, Hong Le Thi Thu, Tu Doan Quang, and David Taniar. A lightweight method for breast cancer detection using thermography images with optimized CNN feature and efficient classification. *Journal of Imaging Informatics in Medicine*, 38(3): 1434–1451, 2025. doi: 10.1007/s10278-024-01269-6.

Ashley Colley, Timo Luukkonen, and Jonna Häkkilä. Evaluating thermal color palettes for manual human detection in aerial search and rescue imagery. *Proceedings of the 24th International Conference on Mobile and Ubiquitous Multimedia*, pages 310–315, 2025. doi: 10.1145/3771882.3771886.

Kanjar De and Marius Pedersen. Effect of hue shift towards robustness of convolutional neural networks. *Electronic Imaging*, 34(15):156–1–156–6, 2022. doi: 10.2352/ei.2022.34.15.color-156.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Sami Ekici and Hushang Jawzal. Breast cancer diagnosis using thermography and convolutional neural networks. *Medical Hypotheses*, 137:109542, 2020. ISSN 0306-9877. doi: 10.1016/j.mehy.2019.109542.

Martin Engilberge, Edo Collins, and Sabine Süsstrunk. Color representation in deep neural networks. *2017 IEEE International Conference on Image Processing*, pages 2786–2790, 2017. doi: 10.1109/icip.2017.8296790.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017. ISSN 1476-4687. doi: 10.1038/nature21056.

European Pressure Ulcer Advisory Panel, National Pressure Injury Advisory Panel, and Pan Pacific Pressure Injury Alliance. Prevention and treatment of pressure ulcers/injuries: Clinical practice guideline. Technical report, EPUAP/NPIAP/PPPIA, 2019. URL http://www.internationalguideline.com.

Jonathan David Freire, Jordan Rodrigo Montenegro, Hector Andres Mejia, Franz Paul Guzman, Carlos Enrique Bustamante, Ronny Xavier Velastegui, and Lorena De Los Angeles Guachi. The impact of histogram equalization and color mapping on ResNet-34's overall

performance for COVID-19 detection. *2021 4th International Conference on Data Storage and Data Engineering*, pages 45–51, 2021. doi: 10.1145/3456146.3456154.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

Ane Goñi-Arana, Jorge Pérez-Martín, and Francisco Javier Díez. Breast thermography: a systematic review and meta-analysis. *Systematic Reviews*, 13(1):295, 2024. doi: 10.1186/s13643-024-02708-9.

Neesha Oozageer Gunowa, Marie Hutchinson, Joanne Brooke, and Debra Jackson. Pressure injuries in people with darker skin tones: A literature review. *Journal of Clinical Nursing*, 27(17-18):3266–3275, 2018. ISSN 0962-1067. doi: 10.1111/jocn.14062.

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, 2016a.

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, 2016b.

Alan B. Hollingsworth. Redefining the sensitivity of screening mammography: A review. *The American Journal of Surgery*, 218(2):411–418, 2019. ISSN 0002-9610. doi: 10.1016/j.amjsurg.2019.01.039.

Mohammed Abdulla Salim Al Husaini, Mohamed Hadi Habaebi, Shihab A. Hameed, Md. Rafiqul Islam, and Teddy Surya Gunawan. A systematic review of breast cancer detection using thermography and neural networks. *IEEE Access*, 8:208922–208937, 2020. ISSN 2169-3536. doi: 10.1109/access.2020.3038817.

Abbas Ali Hussein, Morteza Valizadeh, Mehdi Chehel Amirani, and Sedighe Mirbolouk. Breast lesion classification via colorized mammograms and transfer learning in a novel CAD framework. *Scientific Reports*, 15(1):25071, 2025. doi: 10.1038/s41598-025-10896-0.

Xiaoqiong Jiang, Yu Wang, Yuxin Wang, Min Zhou, Pan Huang, Yufan Yang, Fang Peng, Haishuang Wang, Xiaomei Li, Liping Zhang, and Fuman Cai. Application of an infrared thermography-based model to detect pressure injuries: a prospective cohort study*. *British Journal of Dermatology*, 187(4):571–579, 2022. ISSN 0007-0963. doi: 10.1111/bjd.21665.

Kathleen Jordan, Glory Tomi John, Andrew Chung, Miriam Asare-Baiden, Vicki Stover Hertzberg, Joyce C Ho, and Sharon Eve Sonenblum. Impact of skin tone and cupping on erythema and thermal imaging measurements. *Scientific Reports*, 2025.

Mariusz Kaczmarek and Antoni Nowakowski. Active IR-thermal imaging in medicine. *Journal of Nondestructive Evaluation*, 35(1):19, 2016. ISSN 0195-9298. doi: 10.1007/s10921-016-0335-y.

Dorothea Kesztyüs, Sabrina Brucher, Carolyn Wilson, and Tibor Kesztyüs. Use of infrared thermography in medical diagnosis, screening, and disease monitoring: a scoping review. *Medicina*, 59(12):2139, 2023.

Joanne Kim, Andrew Harper, Valerie McCormack, Hyuna Sung, Nehmat Houssami, Eileen Morgan, Miriam Mutebi, Gail Garvey, Isabelle Soerjomataram, and Miranda M. Fidler-Benaoudia. Global patterns and trends in breast cancer incidence and mortality across 185 countries. *Nature Medicine*, 31(4):1154–1162, 2025. ISSN 1078-8956. doi: 10.1038/s41591-025-03502-3.

Burak Kocak, Michail E. Klontzas, Arnaldo Stanzione, Aymen Meddeb, Aydın Demircioğlu, Christian Bluethgen, Keno K. Bressem, Lorenzo Ugga, Nathaniel Mercaldo, Oliver Díaz, and Renato Cuocolo. Evaluation metrics in medical imaging AI: fundamentals, pitfalls, misapplications, and recommendations. *European Journal of Radiology Artificial Intelligence*, 3:100030, 2025. ISSN 3050-5771. doi: 10.1016/j.ejrai.2025.100030.

B.B. Lahiri, S. Bagavathiappan, T. Jayakumar, and John Philip. Medical applications of infrared thermography: A review. *Infrared Physics & Technology*, 55(4):221–235, 2012. ISSN 1350-4495. doi: 10.1016/j.infrared.2012.03.007.

Zhaoyu Li, Frances Lin, Lukman Thalib, and Wendy Chaboyer. Global prevalence and incidence of pressure injuries in hospitalised adult patients: a systematic review and meta-analysis. *International journal of nursing studies*, 105:103546, 2020.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5):e210315, 2022.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T. Dudley. Deep learning for healthcare: Review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686, 2018.

Rikard Rosenbacke, Åsa Melhus, Martin McKee, and David Stuckler. How explainable artificial intelligence can increase or decrease clinicians' trust in ai applications in health care: systematic review. *JMIR AI*, 3:e53207, 2024.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven Q. H. Truong, Chanh D. T. Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G. Blankenberg, Andrew Y. Ng, Matthew P. Lungren, and Pranav Rajpurkar. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878, 2022. doi: 10.1038/s42256-022-00536-x.

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.

Shazia Shaikh, Nazneen Akhter, and Ramesh Manza. Medical image processing of thermal images in light of applied color palettes. *International Journal of Engineering and Advanced Technology*, 8:1520–1524, 2019.

C. Shi, L.J. Bonnett, J.C. Dumville, and N. Cullum. Nonblanchable erythema for predicting pressure ulcer development: a systematic review with an individual participant data meta-analysis. *British Journal of Dermatology*, 182(2):278–286, 2020. ISSN 0007-0963. doi: 10.1111/bjd.18154.

L. F. Silva, D. C. M. Saade, G. O. Sequeiros, A. C. Silva, A. C. Paiva, R. S. Bravo, and A. Conci. A New Database for Breast Research with Infrared Image. *Journal of Medical Imaging and Health Informatics*, 4(1):92–100, 2024. doi: 10.1166/jmihi.2014.1226.

Qiyang Sun, Alican Akman, and Björn W. Schuller. Explainable artificial intelligence for medical applications: A review. *ACM Transactions on Computing for Healthcare*, 6(2): 1–31, 2025. doi: 10.1145/3709367.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

Xinru Tian, Yunfeng Xie, and Xiaoteng Tang. Effect of color palette and text encoding on infrared thermal imaging perception with various screen resolution accuracy. *Displays*, 87:102939, 2025. ISSN 0141-9382. doi: 10.1016/j.displa.2024.102939.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR, 18–24 Jul 2021.

Yi-Ting Tzen, Barbara Delmore, Kath M Bogie, Sharon Eve Sonenblum, David Newton, Deanna Vargo, Jamie Ronin, Amy Hester, Carroll Gillespie, Ann Tescher, et al. State-of-the-art review of current technology in pressure injury early detection. *Advances in Skin & Wound Care*, 38(10):E128–E137, 2025.

Ronny Velastegui and Marius Pedersen. The impact of using different color spaces in histological image classification using convolutional neural networks. In *2021 9th European Workshop on Visual Information Processing*, pages 1–6. IEEE, 2021.

Yu Wang, Xiaoqiong Jiang, Kangyuan Yu, Fuqian Shi, Longjiang Qin, Hui Zhou, and Fuman Cai. Infrared thermal images classification for pressure injury prevention incorporating the convolutional neural networks. *IEEE Access*, 9:15181–15190, 2021.

Yu Wang, Xiaoqiong Jiang, Yuqi Wang, Xiaoxiao Han, Guangyao Xi, Fuqian Shi, Nilanjan Dey, and Fuman Cai. Early detection of pressure injuries via infrared thermography and ConvNeXt. *Biomedical Signal Processing and Control*, 116:109572, 2026. ISSN 1746-8094. doi: 10.1016/j.bspc.2026.109572.

Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *European conference on computer vision*, pages 68–85. Springer, 2022.

Yulong Yang, Felix O'Mahony, and Christine Allen-Blanchette. Learning color equivariant representations. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jiajin Zhang, Hanqing Chao, Giridhar Dasegowda, Ge Wang, Mannudeep K Kalra, and Pingkun Yan. Revisiting the trustworthiness of saliency methods in radiology AI. *Radiology: Artificial Intelligence*, 6(1):e220221, 2024. doi: 10.1148/ryai.220221.

Appendix A. Implementation Details

A.1. Colormap Rendering

All colormaps were implemented as lookup tables (LUTs) mapping normalized scalar temperature values to uint8 RGB. Prior to colormap application, each image was normalized to [0, 1] using per-image min-max scaling based on its own temperature range. Three colormaps were sourced from `matplotlib`'s standard library: Inferno (`matplotlib.cm.inferno`), Rainbow (`matplotlib.cm.rainbow`), and White Hot (`matplotlib.cm.gray`). Arctic and Ironbow were implemented using custom LUTs approximating FLIR's proprietary palettes, sourced from MickTheMechanic's FLIR-style thermal color palettes GitHub repository.[3] These approximate rather than exactly reproduce FLIR's native palettes, which are proprietary. Researchers using FLIR camera software exports directly may observe subtle differences from the colormaps used here.

A.2. Training and Evaluation Details

After colormap application, images were resized to match each architecture's native input resolution (224 × 224 for ResNet50 and MobileNetV2, 299 × 299 for InceptionV3) using bilinear interpolation. Models were optimized with Adam (learning rate 0.001, batch size 32, maximum 100 epochs), with early stopping based on validation AUROC (patience of 20 epochs) using a validation set carved out from each training fold. All experiments used fixed random seeds and deterministic settings for reproducibility. On average, each fold contained approximately 7 subjects in the erythema held-out test set (28 train / 7 test) and 11 patients in the breast cancer held-out test set (44 train / 11 test).

Appendix B. Additional Erythema Classification Results

Table 6 reports per-architecture AUROC and AUPRC with standard deviations across colormaps. For ResNet50, White Hot achieves the strongest overall performance, with a significant AUPRC advantage over Arctic (p=0.0313). However, differences on AUROC do not reach significance. For MobileNetV2, no colormap consistently dominates and no pairwise differences reach significance, suggesting this architecture is comparatively robust to colormap choice. For InceptionV3, Ironbow substantially outperforms most alternatives, with significant AUROC and AUPRC differences against Arctic, Inferno, and Rainbow (p=0.0313), making it the most colormap-sensitive architecture.

Figure 6 visualizes the ground truth erythema mask, binarized Grad-CAM activation mask, and their intersection for the subject shown in Figure 4 under Arctic and White Hot colormaps. White Hot produces an activation region with the largest overlap with the ground truth erythema region (IoU=0.32), while Arctic shows partial overlap (IoU=0.05) with activation split between the erythema region and a secondary area. These examples illustrate the difficulty of ground truth alignment even for the best-performing colormaps on this subject, consistent with the uniformly low mean IoU values reported in Table 4.

Table 7 reports sensitivity, specificity, and F1 across the five colormaps and architectures. These threshold-based metrics reveal patterns that diverge from the AUROC and AUPRC

3. https://github.com/MickTheMechanic/FLIR-style-thermal-color-palettes/tree/main

Table 6: Erythema classification performance (AUROC, AUPRC) across architectures, with Wilcoxon signed-rank p-values reported against the top-performing colormap.

(a) ResNet

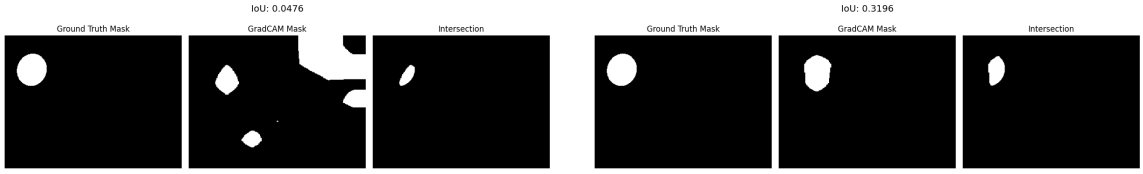
Colormap	**AUROC**	**p-value**	**AUPRC**	**p-value**
Arctic	0.743 ± 0.071	0.0625	0.803 ± 0.094	**0.0313**
Inferno	0.801 ± 0.048	0.1563	0.849 ± 0.071	0.5938
Ironbow	0.713 ± 0.092	0.0625	0.780 ± 0.115	0.0625
Rainbow	0.806 ± 0.064	0.3125	0.852 ± 0.092	0.7813
White Hot	$\mathbf{0.821 \pm 0.036}$	—	$\mathbf{0.852 \pm 0.076}$	—

(b) MobileNetV2

Colormap	**AUROC**	**p-value**	**AUPRC**	**p-value**
Arctic	0.687 ± 0.051	0.2188	0.783 ± 0.083	0.4063
Inferno	0.702 ± 0.076	0.1563	0.797 ± 0.087	0.5000
Ironbow	0.763 ± 0.093	1.0000	0.790 ± 0.120	0.4063
Rainbow	0.687 ± 0.138	0.1563	0.763 ± 0.116	0.1563
White Hot	$\mathbf{0.730 \pm 0.068}$	—	$\mathbf{0.799 \pm 0.106}$	—

(c) InceptionV3

Colormap	**AUROC**	**p-value**	**AUPRC**	**p-value**
Arctic	0.755 ± 0.046	**0.0313**	0.802 ± 0.098	**0.0313**
Inferno	0.751 ± 0.070	**0.0313**	0.827 ± 0.054	0.0625
Ironbow	$\mathbf{0.811 \pm 0.037}$	—	$\mathbf{0.864 \pm 0.056}$	—
Rainbow	0.743 ± 0.052	**0.0313**	0.805 ± 0.085	**0.0313**
White Hot	0.766 ± 0.075	0.2188	0.806 ± 0.103	0.1563



(a) Arctic (b) White Hot

Figure 6: Ground truth erythema mask, binarized Grad-CAM mask, and their intersection (left to right) for the subject in Figure 4, under Arctic and White Hot colormaps.

results. Notably, Ironbow under InceptionV3, which achieves the highest AUROC (0.811) and AUPRC (0.864), yields the lowest sensitivity (0.437) and F1 (0.403) of any architecture-colormap combination. This indicates strong ranking performance without corresponding discriminative ability at a fixed threshold. Conversely, Arctic under ResNet50 achieves the highest sensitivity (0.865) despite below-average AUROC (0.743), reflecting a high-recall, low-precision operating point. These divergences suggest that colormaps may influence ranking ability and threshold-based discrimination differently, and that metric selection is consequential when evaluating or selecting colormaps for deployment.

Table 8 reports ground truth IoU stratified by true label. Values remain uniformly low across both classes and architectures, consistent with Table 4. For the positive class, higher IoU with the erythema region is expected if the model is attending to clinically relevant

Table 7: Secondary performance metrics on erythema classification.

	ResNet50			MobileNetV2			InceptionV3		
Colormap	**Sens**	**Spec**	**F1**	**Sens**	**Spec**	**F1**	**Sens**	**Spec**	**F1**
Arctic	0.865	0.446	0.774	0.621	0.599	0.644	0.596	0.658	0.593
Inferno	0.611	0.608	0.566	0.898	0.284	0.764	0.739	0.583	0.674
Ironbow	0.669	0.559	0.662	0.629	0.744	0.650	0.437	0.704	0.403
Rainbow	0.603	0.713	0.642	0.650	0.647	0.665	0.831	0.427	0.742
White Hot	0.700	0.631	0.678	0.391	0.811	0.438	0.796	0.577	0.766

Table 8: IoU comparison against ground truth for erythema classification, separated by negative and positive classes. Values are reported as mean ± standard deviation.

	ResNet50		MobileNetV2		InceptionV3	
Colormap	**Neg**	**Pos**	**Neg**	**Pos**	**Neg**	**Pos**
Arctic	0.015 ± 0.043	0.014 ± 0.043	0.015 ± 0.042	0.009 ± 0.034	0.038 ± 0.060	0.010 ± 0.044
Inferno	0.008 ± 0.038	0.008 ± 0.037	0.007 ± 0.024	0.021 ± 0.037	0.007 ± 0.020	0.022 ± 0.026
Ironbow	0.004 ± 0.018	0.002 ± 0.013	0.012 ± 0.030	0.028 ± 0.047	0.045 ± 0.083	0.017 ± 0.028
Rainbow	0.006 ± 0.035	0.004 ± 0.022	0.024 ± 0.048	0.036 ± 0.067	0.004 ± 0.015	0.035 ± 0.065
White Hot	0.015 ± 0.072	0.025 ± 0.055	0.023 ± 0.052	0.015 ± 0.051	0.014 ± 0.030	0.030 ± 0.036

features. This pattern holds to some degree for MobileNetV2 and InceptionV3, where positive class IoU is slightly higher than negative for several colormaps including Rainbow, Inferno, and Ironbow. ResNet50 shows less consistent separation between classes. Arctic is a notable exception, with higher negative than positive IoU for InceptionV3, suggesting that even for the positive class, model attention is not driven by the erythema region.

Appendix C. Additional Breast Cancer Classification Results

Table 9 reports per-architecture AUROC and AUPRC with standard deviations across colormaps. For ResNet50, Inferno achieves the strongest overall performance; however, no pairwise differences in either AUROC or AUPRC reach statistical significance, indicating minimal sensitivity to colormap variation. Similarly, MobileNetV2 shows no statistically significant differences across colormaps despite Inferno yielding the highest AUROC, suggesting that performance remains stable across color representations. For InceptionV3, Inferno again achieves the highest overall performance, but as with the other architectures, none of the comparisons reach significance. Overall, these results indicate that breast cancer performance is highly consistent across colormaps for all three architectures, with negligible statistical evidence supporting meaningful differences between color mappings.

Table 10 reports sensitivity, specificity, and F1 for breast cancer classification across colormaps and architectures. The values are generally high and consistent with the near-perfect AUROC and AUPRC reported in Table 3. Ironbow was a notable exception, showing markedly higher specificity than sensitivity for ResNet50 (0.983 vs 0.879) and InceptionV3 (0.997 vs 0.764), suggesting the model was conservative in predicting negative cases despite strong ranking performance. Arctic tended to balance sensitivity and specificity well across architectures, while Rainbow showed the opposite pattern to Ironbow, with higher sensitivity but lower specificity for several architectures. These divergences suggest that col-

Table 9: Breast cancer classification performance (AUROC, AUPRC) across CNNs, with Wilcoxon signed-rank p-values reported against the top-performing colormap.

(*a*) ResNet50

Colormap	AUROC	p-value	AUPRC	p-value
Arctic	0.997 ± 0.005	0.2500	0.999 ± 0.002	0.2500
Inferno	$\mathbf{1.000 \pm 0.001}$	—	$\mathbf{1.000 \pm 0.001}$	—
Ironbow	0.994 ± 0.011	0.3750	0.998 ± 0.004	0.5000
Rainbow	0.995 ± 0.009	0.3125	0.998 ± 0.003	0.3125
White Hot	0.994 ± 0.011	0.3750	0.998 ± 0.004	0.5000

(*b*) MobileNetV2

Colormap	AUROC	p-value	AUPRC	p-value
Arctic	0.997 ± 0.005	0.1875	0.999 ± 0.002	0.1250
Inferno	$\mathbf{0.998 \pm 0.003}$	—	$\mathbf{0.999 \pm 0.001}$	—
Ironbow	0.998 ± 0.004	0.5000	0.999 ± 0.001	0.7500
Rainbow	0.992 ± 0.016	0.5000	0.997 ± 0.006	0.2500
White Hot	0.991 ± 0.011	0.2500	0.997 ± 0.004	0.2500

(*c*) InceptionV3

Colormap	AUROC	p-value	AUPRC	p-value
Arctic	$\mathbf{1.000 \pm 0.001}$	0.5000	$\mathbf{1.000 \pm 0.000}$	0.5000
Inferno	$\mathbf{1.000 \pm 0.000}$	—	$\mathbf{1.000 \pm 0.000}$	—
Ironbow	0.987 ± 0.025	0.2500	0.995 ± 0.009	0.2500
Rainbow	0.997 ± 0.003	0.1250	0.999 ± 0.001	0.1250
White Hot	0.998 ± 0.002	0.2500	0.999 ± 0.001	0.2500

Table 10: Secondary performance metrics on breast cancer classification.

	ResNet50			MobileNetV2			InceptionV3		
Colormap	Sens	Spec	F1	Sens	Spec	F1	Sens	Spec	F1
Arctic	0.946	0.941	0.951	0.953	0.913	0.957	0.978	0.953	0.980
Inferno	0.927	0.928	0.935	0.891	0.947	0.922	0.925	0.935	0.937
Ironbow	0.879	0.983	0.924	0.953	0.941	0.963	0.764	0.997	0.852
Rainbow	0.944	0.875	0.942	0.969	0.986	0.980	0.976	0.833	0.950
White Hot	0.887	0.864	0.898	0.878	0.998	0.931	0.904	0.962	0.933

ormap choice can shift the operating characteristics of a model even when overall ranking performance is uniformly high.

Appendix D. Cross-palette Generalization Detailed Results

The detailed performance including both AUROC and AUPRC for the cross-generalization palette experiments for both classification tasks and the two source colormaps (Ironbow and White Hot) are reported in Tables 11-12.

Table 11 reports erythema classification performance for models trained on Ironbow or White Hot and evaluated on the four remaining colormaps. InceptionV3 is consistently the most robust architecture under this source colormap, retaining the highest AUROC against every target (0.511-0.632), while ResNet50 degrades most severely (0.369-0.476). Averaged across architectures, Inferno yields the smallest degradation (0.553 AUROC) and Rainbow experiences the largest performance drop (0.458), though all four remain well

Table 11: Cross-palette generalization performance for erythema classification. Models trained on Ironbow or White Hot, tested on unseen target colormaps.

	ResNet50		MobileNetV2		InceptionV3		Average	
Test Colormap	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
Trained on Ironbow								
Arctic	0.369	0.544	0.552	0.668	0.632	0.742	0.518	0.651
Inferno	0.476	0.621	0.554	0.659	0.630	0.725	0.553	0.668
Rainbow	0.411	0.588	0.451	0.615	0.511	0.675	0.458	0.626
White Hot	0.389	0.562	0.512	0.632	0.620	0.730	0.507	0.641
CNN Avg	0.411	0.579	0.517	0.644	0.598	0.718	–	–
Trained on White Hot								
Arctic	0.588	0.727	0.530	0.692	0.492	0.642	0.537	0.687
Inferno	0.646	0.762	0.507	0.652	0.461	0.622	0.538	0.679
Ironbow	0.639	0.731	0.578	0.703	0.504	0.641	0.574	0.692
Rainbow	0.566	0.679	0.502	0.666	0.486	0.627	0.518	0.657
CNN Avg	0.610	0.725	0.529	0.678	0.486	0.633	–	–

below the 0.762 in-distribution average reported in Table 2. For models trained on White Hot, the architecture ranking reverses relative to the Ironbow-trained setting: ResNet50 is now the most robust architecture (0.566-0.646 AUROC across targets) and InceptionV3 the least (0.461-0.504). This reversal indicates that architectural robustness to colormap shift depends not on model alone, but also on the interaction between the source and target colormap pair. Averaged across architectures, Ironbow experiences the least performance degradation (0.574 AUROC).

Table 12 reports the corresponding breakdown for breast cancer classification based on the Ironbow and White Hot colormaps. Degradation is lower than erythema across all architecture and colormap pairs. For both colormaps, Inferno is the strongest-generalizing target (0.978 AUROC from Ironbow, 0.832 from White Hot), while Arctic is the weakest (0.826 and 0.739, respectively). InceptionV3 is the most robust architecture when trained on Ironbow (0.921-0.972 AUROC across targets), while MobileNetV2 is most robust when trained on White Hot (0.818-0.922), again illustrating that no single architecture is uniformly robust to colormap shift independent of the training colormap.

Appendix E. Robustness and Sensitivity Analysis

We conduct three additional experiments to assess whether colormap sensitivity is robust to alternative pipeline choices to our primary experimental configuration. Appendix E.1 evaluates whether colormap sensitivity persists under domain-specific (RadImageNet) rather than natural-image (ImageNet) pretraining. Appendix E.2 evaluates whether it persists under global rather than per-image temperature normalization. Appendix E.3 evaluates whether it extends beyond CNN-based architectures to a transformer-based model.

E.1. RadImageNet Results

We repeated the erythema classification experiments using RadImageNet-pretrained weights (Mei et al., 2022) for ResNet50 and InceptionV3. MobileNetV2 is excluded, as no RadIm-

Table 12: Cross-palette generalization performance for breast cancer classification. Models trained on Ironbow or White Hot, tested on four target colormaps not seen during training.

	ResNet50		MobileNetV2		InceptionV3		Average	
Test Colormap	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
Trained on Ironbow								
Arctic	0.714	0.800	0.815	0.922	0.948	0.972	0.826	0.898
Inferno	0.974	0.989	0.991	0.994	0.971	0.989	0.978	0.991
Rainbow	0.778	0.854	0.877	0.914	0.921	0.963	0.859	0.911
Whitehot	0.776	0.872	0.952	0.981	0.868	0.935	0.865	0.929
CNN Avg	0.810	0.879	0.909	0.953	0.927	0.965	–	–
Trained on White Hot								
Arctic	0.571	0.737	0.888	0.947	0.759	0.811	0.739	0.832
Inferno	0.756	0.852	0.922	0.948	0.819	0.854	0.832	0.885
Rainbow	0.632	0.733	0.818	0.862	0.831	0.895	0.760	0.830
Ironbow	0.721	0.855	0.915	0.945	0.730	0.780	0.788	0.860
CNN Avg	0.670	0.794	0.886	0.925	0.785	0.835	–	–

Table 13: Predictive performance across the five colormaps for erythema classification using RadImageNet (RIN) pretrained weights for ResNet50 and InceptionV3. MobileNetV2 is excluded as no RIN checkpoint is available. No statistically significant pairwise differences were found (Wilcoxon signed-rank, $p < 0.05$).

	ResNet50 (RIN)		InceptionV3 (RIN)		Average	
Colormap	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
Arctic	**0.755**	**0.807**	0.762	0.830	**0.758**	**0.819**
Inferno	0.650	0.740	**0.763**	**0.833**	0.707	0.786
Ironbow	0.709	0.746	0.747	0.813	0.728	0.779
Rainbow	0.639	0.737	0.737	0.803	0.688	0.770
Whitehot	0.654	0.735	0.667	0.795	0.661	0.765
CNN Avg	0.681	0.753	**0.735**	**0.815**	–	–

ageNet checkpoint is available for this architecture. Table 13 reports the results. Colormap sensitivity persists under RadImageNet pretraining, with AUROC spreads of 0.105 for ResNet50 and 0.096 for InceptionV3, though no pairwise differences reach statistical significance at this sample size. Notably, Arctic is the best-performing colormap for RadImageNet-pretrained ResNet50 (AUROC 0.755), a ranking not observed under ImageNet pretraining (Table 2, where White Hot performs best for ResNet50). Overall performance under RadImageNet pretraining is somewhat lower than under ImageNet pretraining for both architectures (e.g., CNN-average AUROC of 0.681 vs. 0.777 for ResNet50), despite RadImageNet's reduced domain gap to medical imaging, suggesting that the scale and diversity of ImageNet's pretraining corpus may outweigh domain relevance for this task. This shift in optimal colormap, together with the persistence of colormap sensitivity regardless of pretraining source, suggests that colormap sensitivity is not solely an artifact of ImageNet's natural-image color priors, and reinforces that optimal colormap choice should be validated.

Table 14: Predictive performance across the colormaps for erythema and breast cancer classification using global thermal normalization and ImageNet pre-trained weights.

	ResNet50		MobileNetV2		InceptionV3		Average	
Colormap	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
Erythema Classification								
Arctic	0.771	0.835	0.671	0.773	**0.817**	**0.857**	0.753	**0.822**
Inferno	**0.817**	**0.867**	0.672	0.748	0.780	0.840	**0.756**	0.818
Ironbow	0.771	0.823	**0.756**	**0.797**	0.732	0.800	0.753	0.807
Rainbow	0.750	0.808	0.633	0.725	0.784	0.835	0.722	0.789
White Hot	0.733	0.793	0.646	0.739	0.760	0.833	0.713	0.788
CNN Avg	0.768	0.825	0.676	0.756	**0.775**	**0.833**	–	–
Breast Cancer Classification								
Arctic	0.999	0.999	0.990	0.997	1.000	1.000	0.996	0.999
Inferno	0.999	1.000	**1.000**	**1.000**	1.000	1.000	1.000	1.000
Ironbow	0.994	0.998	**1.000**	**1.000**	**1.000**	**1.000**	0.998	0.999
Rainbow	0.997	0.999	1.000	1.000	1.000	1.000	0.999	1.000
White Hot	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
CNN Avg	0.998	0.999	0.998	0.999	**1.000**	**1.000**	–	–

E.2. Per-Image versus Global Normalization

Table 14 reports erythema and breast cancer classification performance, respectively, under global (dataset-level) min-max normalization instead of the per-image normalization used elsewhere. For global normalization, a single minimum and maximum were applied across the entire dataset, selected by visual inspection of the dataset's temperature histogram to exclude extreme outlier values while retaining the clinically relevant temperature range. For the erythema dataset, this yielded a global range of [16,36]°C; for the breast cancer dataset, [22.13, 32.09]°C.

Overall patterns of colormap sensitivity are broadly preserved under global normalization for erythema classification. For example, Rainbow yields 0.750 AUROC under global normalization for ResNet50 versus 0.806 under per-image normalization, and 0.784 versus 0.743 for InceptionV3. While absolute values shift modestly by colormap and architecture, the relative ranking of colormaps and the overall magnitude of colormap-driven variation remain consistent across normalization schemes, suggesting our central findings are not an artifact of the per-image normalization choice.

Similar trends are observed with the breast cancer classification. Performance remains near-ceiling across all colormaps and architectures (AUROC 0.990–1.000, AUPRC 0.997–1.000), with White Hot achieving the highest average performance (AUROC 1.000, AUPRC 1.000) and Arctic the lowest (AUROC 0.996, AUPRC 0.999), a narrower spread than observed under per-image normalization. This consistency across normalization schemes reinforces that the breast cancer task's near-ceiling performance, rather than the normalization strategy, limits its sensitivity for detecting colormap-driven effects.

E.3. TinyVIT

We evaluate TinyViT (Wu et al., 2022), a compact, distillation-pretrained vision transformer better suited to our limited data regime than larger ViT variants. Table 15 reports

Table 15: TinyViT classification performance across colormaps for erythema and breast cancer tasks. No pairwise comparisons reached significance for either task.

	Erythema		Breast Cancer	
Colormap	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
Arctic	0.735	0.806	0.990	0.996
Inferno	0.786	0.832	0.989	0.996
Ironbow	**0.803**	0.838	0.990	0.996
Rainbow	0.749	0.822	0.991	0.997
White Hot	0.774	**0.838**	**0.991**	**0.997**
TinyViT Avg	0.769	0.827	0.990	0.996

classification performance for TinyViT (Wu et al., 2022) across the five colormaps for both tasks, using ImageNet-pretrained weights. For erythema classification, colormap sensitivity persists under this architecture. AUROC ranges from 0.735 (Arctic) to 0.803 (Ironbow), a spread of 0.068, comparable in magnitude to the spread observed across the three CNN architectures (Table 2). No pairwise comparisons reached significance for either Ironbow or White Hot. For breast cancer classification, performance is uniformly high across all colormaps (AUROC 0.989–0.991, AUPRC 0.996–0.997), mirroring the near-ceiling, low-variance pattern observed for the CNN architectures (Table 3), and no pairwise differences reach significance.

Averaged across colormaps, TinyViT's overall performance (erythema AUROC 0.769; breast cancer AUROC 0.990) is comparable to, but does not exceed, the best CNN architectures on either task. This is consistent with vision transformers typically requiring larger training sets to reach comparable performance to CNNs (Dosovitskiy et al., 2021; Touvron et al., 2021). The results suggest that colormap-driven performance variation is not specific to CNNs alone and may generalize across architecture families, but necessitates further validation across additional transformer architectures, including cross-palette generalization and attention-based analysis for these models.