A Study of Artifacts on Melanoma Classification under Diffusion-Based Perturbations

Qixuan Jin QIXUANJ@MIT.EDU
Massachusetts Institute of Technology, United States

Abstract

In melanoma classification, deep learning models have been shown to rely on non-medical artifacts (e.g., surgical markings) rather than clinically relevant features (e.g., lesion asymmetry), compromising their generalizability. In this work, we investigate the impact of artifacts on melanoma classification under two settings: (1) input disruptions, such as bounding boxes and frequency-based filtering, which isolate artifacts by region or frequency, and (2) a novel diffusion-based perturbation method that selectively introduces isolated artifacts into images, generating controlled pairs for direct comparison. We systematically analyze artifact biases in three benchmark datasets: ISIC 2018, HAM10000, and PH2. Our findings reveal that whole-image training outperforms lesion-only or background-only approaches, low-frequency features are essential for melanoma prediction, and classifiers are more sensitive to perturbations for the artifacts of ink markings, rulers, and patches. These results emphasize the need for systematic artifact assessment and provide insights for improving the robustness of melanoma classification models.

Data and Code Availability This paper uses the ISIC 2018 dataset (Codella et al., 2019) and the HAM10000 dataset (Tschandl et al., 2018), which are both available on the ISIC repository. The paper uses the PH2 dataset (Mendonça et al., 2013), which is available at their website. Our GitHub repository is at https://github.com/QixuanJin99/dermoscopic_artifacts/tree/main.

Institutional Review Board (IRB) Our research does not require IRB approval.

1. Introduction

Dermoscopy is a widely used imaging technique for melanoma diagnosis, enhancing the visualization of subsurface skin structures (Sonthalia et al., 2019). Clinicians rely on diagnostic frameworks such as the ABCD rule (Nachbar et al., 1994) and the Seven-Point Checklist (Argenziano et al., 1998; Di Leo et al., 2010) to systematically assess lesion characteristics. While machine learning models have achieved comparable performance to dermatologists in melanoma classification (Haenssle et al., 2018), these models often rely on non-medical artifacts, such as surgical markings and rulers, rather than clinically relevant features (Winkler et al., 2019; Bevan and Atapour-Abarghouei, 2021).

Spurious correlations can degrade model generalization in real world settings where the spurious feature is absent in new data where artifacts are absent (Boland et al., 2024; Singla and Feizi, 2021; Minderer et al., 2020). In medical settings, ensuring generalizability is critical, as systematic biases across hospitals and populations can significantly impact real-world performance (Vellido, 2020; Salahuddin et al., 2022). Therefore, determining the impact of artifact reliance is essential for developing robust melanoma classification models. Dermatological datasets provide an ideal setting for studying artifact-driven biases, as diagnostic decisions rely on medical images and non-medical features can mislead deep learning classifiers.

In a prior work, Bissoto et al. (2020) manually annotated seven artifacts commonly present in dermoscopic images (Figure 1), finding that melanoma classifiers rely on both clinically relevant lesion features and non-medical artifacts for prediction. We make us of this setting to investigate the impact of artifacts in additional dermatological datasets, and new perturbation settings.

First, we assess whether background artifacts are predictive of melanoma, and how well they generalize under specific dataset disruptions, such as bounding box constraints and low- and high-pass filtering. We evaluate performance of ResNet-50 (He et al., 2016) models on melonoma classification using the ISIC

2018 (Codella et al., 2019), PH2 (Mendonça et al., 2013), and HAM10000 (Tschandl et al., 2018) benchmark datasets. Models trained on ISIC and PH2 are each evaluated on the remaining datasets to measure transfer performance. We find that non-medical artifacts do not inherently degrade generalization in this setting, as some artifacts transfer between datasets. For instance, artifacts near the outer edges of images retain their predictive value when transferred from ISIC to PH2. Our findings are in strong contrast to other work in the spurious correlation literature, and indicate that the effect of non-medical features on transfer performance should be assessed by model development and deployment teams prior to widespread use.

Second, we isolate the impact of specific artifacts on classification performance by finetuning Stable Diffusion 1.5 (Rombach et al., 2022) with Dreambooth (Ruiz et al., 2023) and LoRA (Hu et al., 2021) to inpaint isolated artifacts onto existing dermoscopic images. Our diffusion models are capable of generating the common non-medical artifacts of dark corners, gel bubbles, ink markings, colored patches, and rulers. To the best of our knowledge, no prior work has utilized diffusion models specifically for adding non-medical artifacts onto dermoscopic images. We evaluate the transfer performance of classifiers trained on PH2 and ISIC when applied to PH2 images perturbed by these artifacts. Our findings reveal that classifiers are especially sensitive to diffusion-based perturbations involving ink markings, rulers, and patches. Given that certain artifacts have a greater impact on transfer performance, systematic artifact assessment is essential for developing targeted mitigation strategies before real-world deployment.

We present the following contributions:

1. Comprehensively evaluated melanoma classification performance under different input disruptions and transfer across three public benchmark datasets.

2. Expanded upon manual annotations of common non-medical dermoscopic artifacts with the labeling of the PH2 dataset. We will release this dataset for public research.

3. Developed diffusion models capable of selectively augmenting dermoscopic images with specific artifacts (e.g. dark corners, gel bubbles, ink, patches, and rulers), providing a flexible framework for future researchers to dynamically augment their datasets.

2. Related Work

2.1. Machine Learning for Melanoma Classification

Machine learning methods for melanoma classification on dermoscopic images generally fall into two categories: 1) Enhancing existing medical algorithms 2) image-based classification with vision models. Kawahara et al. (2018) uses a multitask deep convolutional neural network to predict the seven-point checklist criteria and thus perform the skin lesion diagnosis. Almaraz-Damian et al. (2020) fuses handcrafted ABCD features with vision model embeddings for improved performance. Despite these methods' medical grounding, dermatologist-labeled features like border irregularity and asymmetry have inherent degrees of subjectivity and high-quality ground truth labels are costly to obtain at scale.

In parallel, a great variety of convolutional neural networks (CNN) and vision transformers (ViT) models have been adapted from natural image settings to predict melanoma in dermoscopic images with high accuracy (Arshed et al., 2023; Budhiman et al., 2019; Gamage et al., 2024). These deep image classifiers have demonstrated comparable or improved performance over the medical-augmented algorithms and clinicians (Haenssle et al., 2018). Specifically, the CNNs demonstrated higher specificity with comparable sensitivity to dermatologists, with slightly better performance overall.

In this work, we focus on melanoma classification with vision-based models, and explore how these models can utilize non-dermoscopic features outside the domain of medical algorithms to predict melanoma with better accuracy.

2.2. Artifacts in Dermoscopic Images

Non-medical artifacts have long been investigated for spurious association with disease prognosis in dermatology. Past work illustrates how a market-approved CNN classifier spuriously associates large dark corner artifacts with malignancy, leading to decreased performance for images with this artifact (Sies et al., 2021). Surgical ink markings have also been found to be positively associated with melanoma predictions (Winkler et al., 2019). We take inspiration

from the prior work Bissoto et al. (2020), which defined a set of non-medical artifacts present in the ISIC dataset (dark corners, hair, gel border, gel bubbles, ink markings, ruler, and patches) and demonstrated how model-based debiasing methods cannot fully remove dependency on artifacts.

2.3. Denoising Diffusion Probabilistic Models

Denoising diffusion probabilistic models (diffusion models) demonstrate state-of-the-art performance for text-to-image generation (Dhariwal and Nichol, 2021). In particular, Stable Diffusion (Rombach et al., 2022) has been widely used for art creation (Liao et al., 2022) and image editing (Yang et al., 2023; Brooks et al., 2023). Methods such as Textual Inversion (Gal et al., 2022) and Dreambooth (Ruiz et al., 2023) allow pretrained diffusion models to learn specific, new concepts given a few sample images. Low-Rank Adaptation (LoRA) is a parameter-efficient fine-tuning method that modifies a set of low-rank matrices with frozen pretrained model weights to reduce computational intensity (Hu et al., 2021). In this work, we use a Stable Diffusion backbone and fine-tune the model with Dreambooth and LoRA to learn a specific artifact's visual features with a custom token.

Diffusion models have also been adapted for dermatology applications. Derm-T2IM is a Stable Diffusion model fine-tuned with DreamBooth and LoRA to generate images of benign and malignant skin lesions (Farooq et al., 2024). Additionally, diffusion models have been used to augment dermatology images, including those not captured with a dermatoscope, to improve representation for underrepresented groups (Sagers et al., 2022). However, to the best of our knowledge, there are no publicly available Stable Diffusion models designed to generate non-medical artifacts in dermoscopic images. This work aims to bridge that gap by developing a tailored diffusion model for artifact generation in dermoscopy.

3. Method

3.1. Datasets and Preprocessing

International Skin Imaging Collaboration (ISIC) is a collection of dermoscopic images used for classification, segmentation, and detection tasks (Codella et al., 2019). We specifically use the training split (n=2,594) from the ISIC 2018 Task 1-2, as only this split is labeled with artifacts by Bissoto et al. (2020). The HAM10000 dataset is a dataset of 10,015 dermoscopic images collected from diverse sources (Tschandl et al., 2018). The PH2 dataset is a collection of 200 dermoscopic images collected at Dermatology Service of Hospital Pedro Hispano, Portugal with human-labeled dermoscopic attributes (Mendonça et al., 2013).

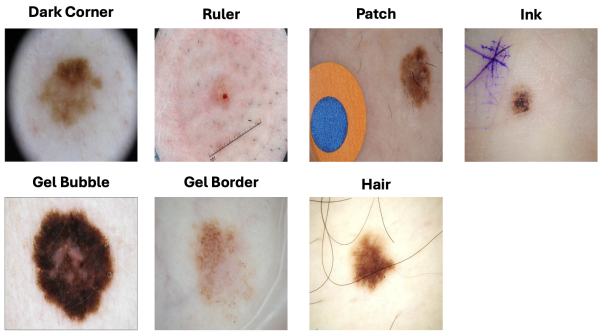


Figure 1: The types of dermoscopic artifacts.

Since we wish to investigate the robustness of image classifiers under different input image augmentations, we preprocess all datasets under 12 settings ("dataset modes"). The original image is mode "whole". With lesion segmentations, we create a dataset with only the skin lesion ("lesion") or only the background ("background"). We follow prior work (Bissoto et al., 2020) and compute the outer boundaries of the segmentation and convert the mask into a bounding box that removes the lesion boundary information. For "bbox70" and "bbox90", the bounding box is expanded until approximately 70% or 90% of the image is masked out. We additionally evaluate high/low frequency features with a high/low-pass Gaussian filter with $\sigma = 1$. We use the filters on "whole", "lesion", and "background" images to obtain the remaining six dataset modes.

For melanoma classification, we use the benign and malignant lesion labels in the ISIC 2018 dataset. Based on the ISIC definition, we divide the HAM10000 into benign ('Benign Keratosis-like Lesions', 'Dermatofibroma', 'Melanocytic Nevi', 'Vascular Lesions') and malignant ('Actinic Keratoses', 'Basal Cell Carcinoma', 'Melanoma').

3.2. Classification Models

We frame the problem of melanoma classification as a binary classification task. Prior work (Gazioğlu and Kamaşak, 2021) shows the robustness of both

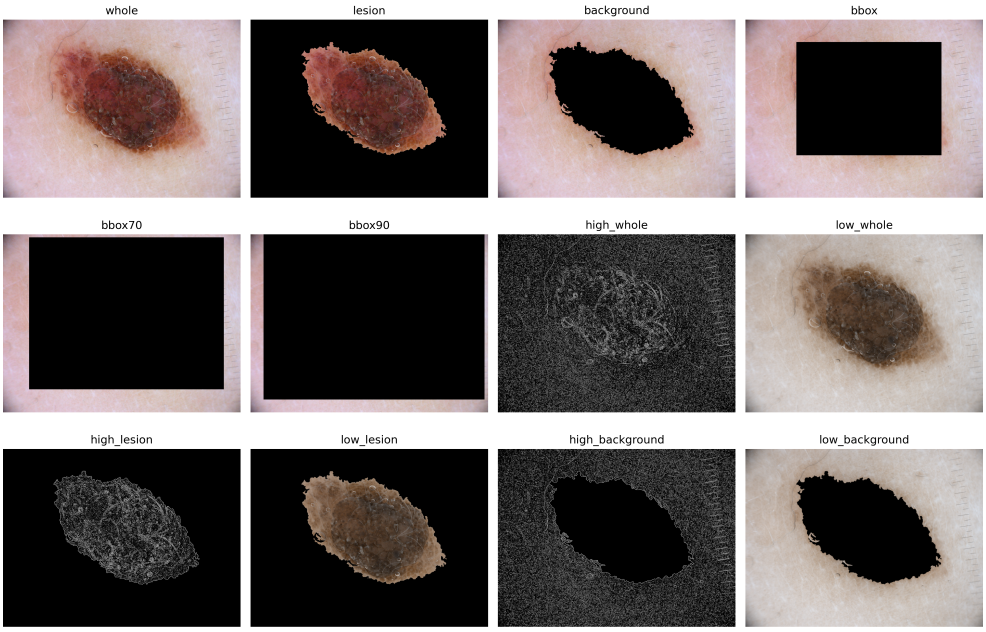


Figure 2: Visualization of the dataset modes we use to investigate the classifier's robustness under different data preprocessing settings.

ResNet50 (He et al., 2016) and DenseNet121 (Huang et al., 2017) under certain noise settings. We focus on transferring ImageNet (Deng et al., 2009) pretrained ResNet50 classifier models to the dermoscopic domain. We train with an Adam optimizer at a learning rate of 1e−4 for 10 epochs and a batch size of 32. All images are resized to 224×224 and normalized to the ImageNet distribution. We finetune from pretrained ImageNet weights.

3.3. Diffusion Models for Artifact Generation and Classifier Explainability

To investigate the trained classifier's robustness to perturbations of specific artifacts, we propose an explainability procedure that breaks the association between artifact and prediction label through the indiscriminate addition of the artifact to all images. To automate the addition of artifacts in a realistic manner, we train diffusion models for each artifact. Specifically, Dreambooth (Ruiz et al., 2023) is a method that can fine-tune text-to-image diffusion models to associate specific visual concepts with the provided text tokens given a set of example images. To prevent the association of existing visual concepts with our artifact concepts, we pick our Dreambooth tokens from a list of rare single tokens (2kpr, 2024). We define the tokens of {patches: olis, dark corner: lun, ruler: dits, ink: httr, gel bubble: sown}. Since we are learning single visual concepts, we can speed up the training process with Low-Rank Adaptation (LoRA), in which we tune low-rank matrices instead of the full diffusion model (Hu et al., 2021).

For the base model, we use Stable Diffusion 1.5 (Rombach et al., 2022) with the HuggingFace implementation ("runwayml/stable-diffusion-v1-5") and code from the diffusers library (von Platen et al., 2022). We finetune both the UNet (Ronneberger et al., 2015) and text encoder with LoRA matrices of rank 64 and $\alpha = 32$. We finetune for 3 to 5 epochs with a constant learning rate of 1e−4, a batch size of 2, and an image resolution of 512. For text conditioning, we use the instance prompt template: "a dermoscopic image of token class". The token is the rare artifact token and the class is either "benign" or "malignant" according to the melanoma label. To prevent large deviations from the melanoma class, we train with prior preservation with a prior loss weight of 0.3 and 200 generated class images. The class prompt follows the same template but excludes the artifact token. For compute, we train using seed=0, mixed

precision (fp16) on a single NVIDIA A100 GPU, with training completing in approximately one hour.

To generate artifacts for the perturbation method, we load the fine-tuned LoRA adapter into the inpainting pipeline. We generate images at a guidance scale of 10 and tune the strength parameter manually through visualizations. We decide on the following strength parameters $\{\text{patches} : 0.85, \text{dark corner} : 0.75, \text{ruler} : 0.65, \text{ink} : 0.7, \text{gel bubble} : 0.6\}$. In our perturbation analysis, we augment the PH2 dataset. We generate 5 augmented images per original image, and selected for the image that is most predictive of the artifact given artifact classifiers (ResNet50) trained on the ISIC ground truth artifact labels.

3.4. Evaluation

We use 5-fold cross-validation on each dataset to train and evaluate the classifier using non-overlapping train and test splits. We use area under the receiver operating curve (AUROC) (Hanley and McNeil, 1982) as the primary performance metric, with error bounds computed as the standard deviation across the 5 folds. We define evaluation results as "source" when the test split is drawn from the same dataset used for training, while "target" refers to evaluations on an out-of-domain test split.

4. Experiments

4.1. Melanoma Classification and Transfer Performance under Dataset Disruptions

We evaluate the performance of a classifier for the classification of benign and malignant melanoma images across two dimensions: (1) the effect of dataset perturbation by training the classifier on the train set and evaluating it on the test set from the same source dataset, and (2) the generalization ability of the perturbed classifier when evaluated on test sets from out-of-domain target datasets. Given that PH2-trained classifiers exhibit poor generalization across all settings, we primarily focus on ISIC-trained classifiers when analyzing transfer performance.

First, we note that the classifier trained on whole images consistently outperforms classifiers trained on either lesion-only or background-only perturbed images, both within the source dataset and in transfer to target datasets (Table 1 and Table 2). When evaluated on the source dataset, the whole-image classifier achieves the highest AUROC (ISIC: 0.792, PH2: 0.975), outperforming the lesion-only (ISIC: 0.746, PH2: 0.952) and background-only classifiers (ISIC: 0.752, PH2: 0.884). This trend extends to cross-dataset generalization. For the ISIC-trained models, the whole-image classifier exhibits superior transfer performance (HAM10000: 0.721, PH2: 0.858) compared to the lesion-only (HAM10000: 0.719, PH2: 0.818) and background-only classifiers (HAM10000: 0.714, PH2: 0.777). These results suggest that lesion and background features are both predictive, and the combination of the two improves model performance.

Despite the clinical intuition that lesion features should be the most predictive (Williams et al., 2021; Longo et al., 2023; Tschandl et al., 2019), background-only classifiers remain surprisingly effective. In the well-known ABCD rule (Nachbar et al., 1994), features such as asymmetry and irregular borders are defined by the lesion *boundary*, a feature that is retained in our background-only perturbation. To evaluate the significance of this boundary, we use a bounding box to remove the lesion shape while preserving the surrounding background. In ISIC, performance drops from 0.752 (background) to 0.717 (lesion) when the lesion boundary is removed. As the bounding box expands to cover 70% and 90% of the image, performance further declines to 0.681 and 0.628, respectively, indicating that the model relies on more than just the outer edges of the image. A similar performance drop is observed in PH2, despite its small sample size ($n = 200$ samples) and larger error bounds (Table 2). These findings highlight the importance of lesion boundary features in melanoma classification and also suggests that the surrounding background contains predictive information.

Importantly, we observe that bbox classifiers trained on the ISIC dataset maintain strong performance when transferred to target datasets (ISIC source: 0.717, HAM10000 target: 0.687, PH2 target: 0.821). The ability of bbox-based models to generalize suggests that certain background features associated with melanoma are consistently preserved across datasets. This finding suggests that non-medical artifacts do not inherently degrade generalization, as some artifacts may be transferable between datasets. The impact of training models with non-medical artifacts on transfer performance should therefore be assessed on a case-by-case basis.

Beyond localizing predictive regions, we examine whether the classifier relies on features at specific frequency ranges. Using high-pass and low-pass filters, we extract high- and low-frequency signals to assess their contribution to classification performance.

Domain **AUROC**	Source **ISIC**	Target **HAM10000**	Target **PH2**
whole	0.792 ± 0.024	0.721 ± 0.061	0.858 ± 0.045
lesion	0.746 ± 0.040	0.719 ± 0.024	0.818 ± 0.099
background	0.752 ± 0.017	0.714 ± 0.038	0.777 ± 0.092
bbox	0.717 ± 0.024	0.687 ± 0.040	0.821 ± 0.051
bbox70	0.681 ± 0.026	0.686 ± 0.047	0.805 ± 0.013
bbox90	0.628 ± 0.016	0.616 ± 0.022	0.679 ± 0.064
whole (high freq)	0.530 ± 0.020	0.307 ± 0.027	0.442 ± 0.124
lesion (high freq)	0.590 ± 0.042	0.575 ± 0.077	0.502 ± 0.204
background (high freq)	0.523 ± 0.037	0.500 ± 0.127	0.682 ± 0.186
whole (low freq)	0.765 ± 0.017	0.741 ± 0.030	0.748 ± 0.052
lesion (low freq)	0.717 ± 0.028	0.685 ± 0.029	0.856 ± 0.086
background (low freq)	0.704 ± 0.035	0.624 ± 0.042	0.817 ± 0.078

Table 1: Performance in AUROC with standard deviations for the melanoma classifier trained on the ISIC dataset and evaluated on the ISIC, HAM10000, and PH2 datasets.

In ISIC, the classifier struggles to learn from high-frequency signals but maintains strong performance with only low-frequency information, showing an AUROC drop of less than 0.05 compared to the original image. Additionally, low-frequency models generalize well across datasets (ISIC source: 0.765, HAM10000 target: 0.741, PH2 target: 0.748). In PH2, both high- and low-frequency models perform well, likely due to overfitting on the small training set. These results demonstrate that the model primarily relies on low-frequency signals for melanoma classification, suggesting that coarse patterns contribute more to prediction than fine-grained details.

4.2. Artifact Prevalences and Correlations

Given that the background contributes to melanoma classification, we examine whether the presence of specific non-medical artifacts may explain this predictive signal.

Table 3 shows that artifact prevalence varies across datasets. Dark corners are significantly more common in PH2 (94%) than in ISIC (37%), while rulers and patches are entirely absent from PH2. To assess potential associations between artifacts and melanoma, we compute the Pearson correlation between individual artifacts and the melanoma label in ISIC and PH2 (Figure 3). We find that no single artifact exhibits a strong direct correlation with melanoma. However, several artifacts are correlated with each other. In ISIC, rulers frequently co-occur with dark corners (corr = 0.31) and gel borders (corr = 0.34), while ink markings are often found alongside gel borders (corr = 0.42) and rulers (corr = 0.40). Since PH2 lacks rulers and contains very few ink-marked images, these inter-artifact correlations observed in ISIC do not transfer to PH2. These findings suggest that while no individual artifact strongly predicts melanoma, the combined presence of artifacts may introduce patterns that influence model predictions.

4.3. Training Diffusion Models for Artifact Generation

We first fine-tune Stable Diffusion 1.5 on the dermoscopic image domain. Then, using DreamBooth and LoRA, we further fine-tune this base dermoscopic diffusion model to generate specific artifact types, leveraging ground truth labels from the ISIC dataset.

Due to challenges in disentangling artifact concepts and achieving stable training convergence, we exclude results for the artifacts "hair" and "gel border". For the remaining five artifacts, we use the trained diffusion models to augment the PH2 dataset by introducing these artifacts through masked inpainting. We carefully tune the strength hyperparameter to achieve a balance between preserving the original image and effectively integrating the artifact (Figure 4).

Domain **AUROC**	Source **PH2**	Target **ISIC**	Target **HAM10000**
whole	0.975 ± 0.012	0.635 ± 0.050	0.638 ± 0.064
lesion	0.952 ± 0.047	0.623 ± 0.028	0.624 ± 0.034
background	0.884 ± 0.088	0.557 ± 0.080	0.505 ± 0.067
bbox	0.880 ± 0.080	0.565 ± 0.036	0.514 ± 0.041
bbox70	0.845 ± 0.046	0.566 ± 0.049	0.570 ± 0.016
bbox90	0.865 ± 0.081	0.544 ± 0.026	0.433 ± 0.030
whole (high freq)	0.903 ± 0.070	0.509 ± 0.023	0.609 ± 0.025
lesion (high freq)	0.877 ± 0.064	0.664 ± 0.055	0.625 ± 0.049
background (high freq)	0.908 ± 0.112	0.557 ± 0.049	0.493 ± 0.034
whole (low freq)	0.925 ± 0.041	0.638 ± 0.055	0.721 ± 0.047
lesion (low freq)	0.935 ± 0.087	0.582 ± 0.048	0.610 ± 0.079
background (low freq)	0.890 ± 0.071	0.590 ± 0.093	0.538 ± 0.056

Table 2: Performance in AUROC with standard deviations for the ResNet50 melanoma classifier trained on the PH2 dataset and evaluated on the ISIC, HAM10000, and PH2 datasets.

	PH2	**ISIC**
dark corner	0.940	0.370
hair	0.415	0.585
gel border	0.150	0.262
gel bubble	0.440	0.488
ruler	-	0.495
ink	0.010	0.170
patches	-	0.072

Table 3: Prevalence of artifacts in the PH2 and ISIC datasets. ISIC labels are from Bissoto et al. (2020) and PH2 labels are our manual labels.

4.4. Diffusion-Perturbed Explainability of Classifier Dependency of Artifacts

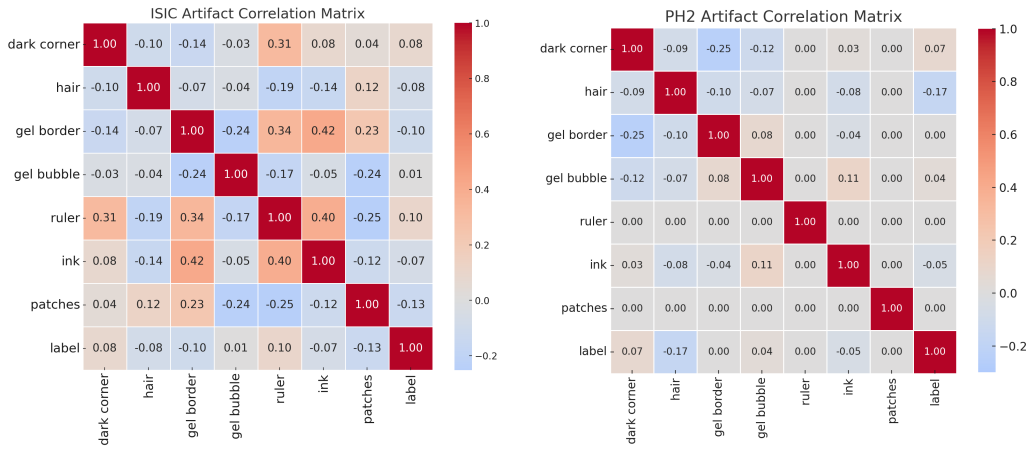
We evaluate the performance of classifiers trained on PH2 or ISIC data and tested on diffusion-perturbed PH2 datasets, where a specific artifact is added to each image to disrupt any spurious associations between melanoma and the artifact.

First, we find that classifiers trained on limited image regions are more sensitive to single-artifact perturbations after transfer. Specifically, the ISIC-trained classifier fails to generalize across all perturbations in the bbox90 setting, where 90% of the image is covered (Table 4). While the model can still learn weak associations in this heavily occluded setting, these associations do not hold when transferred to a different dataset.

Second, we observe that classifiers are more disrupted by the introduction of certain artifacts than others. Ink markings, rulers, and patches significantly degrade the performance of both PH2- and ISIC-trained classifiers, whereas the addition of dark corners and gel bubbles leads to a smaller decline.

For PH2-trained classifiers, this discrepancy may be attributed to artifact prevalence, where ink, rulers, and patches appear infrequently or are entirely absent. The lack of exposure to these artifacts during training leads to poor generalization when these artifacts are introduced in the perturbed datasets. However, classifiers trained on whole images or background-only images still exhibit reasonably strong transfer performance. While artifacts introduce noise, the whole-image and background settings retain enough diagnostic information for melanoma classification.

For ISIC-trained classifiers, the disruption caused by ink, rulers, and patches likely arises from spurious associations formed during training that do not transfer well to PH2. Despite this, whole-image and background-only models remain robust under perturbation. Whole-image classifiers likely generalize better because they preserve medically relevant features. Background-only models transfers well, suggesting that either the remaining unaltered artifact features continue to provide predictive value or that consistent background features contribute to melanoma classification across both datasets.



(a) ISIC (b) PH2

Figure 3: The correlation between the artifacts with themselves and with the melanoma label across the datasets. ISIC labels are from Bissoto et al. (2020) and PH2 labels are our manual labels.

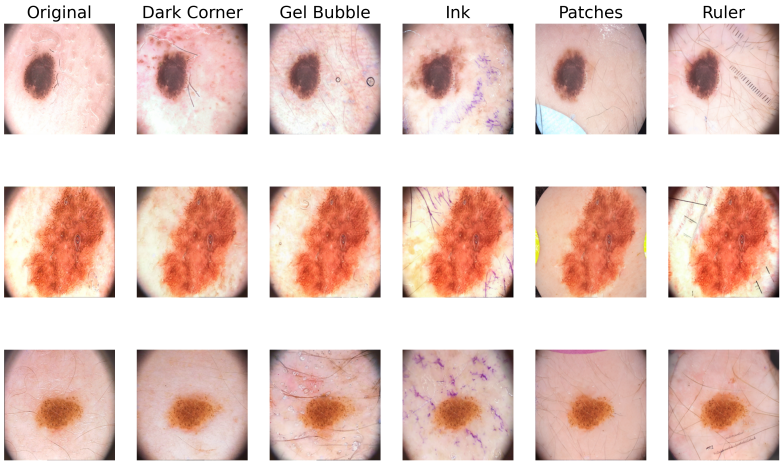


Figure 4: Visualization of three images perturbed by the artifact-specific diffusion model for the augmentation of the PH2 dataset.

These findings emphasize the importance of training classifiers on more comprehensive image regions to improve robustness to artifact perturbations. Models trained with limited image context struggle to generalize when artifacts are introduced in new datasets. Additionally, artifacts that are rare or absent during training can significantly degrade performance when encountered in out-of-domain datasets. The specific spurious artifacts affecting transfer will depend the datasets involved, underscoring the need for targeted mitigation strategies.

Dataset Mode	Artifact	PH2	ISIC	Dataset Mode	Artifact	PH2	ISIC
whole	**original**	0.975	0.858	**whole (high freq)**	**original**	0.903	**0.442**
	dark corner	0.978	0.816		**dark corner**	**0.487**	**0.430**
	gel bubble	0.973	0.841		**gel bubble**	**0.414**	**0.522**
	ink	**0.959**	0.788		**ink**	**0.484**	**0.490**
	patches	0.976	0.848		**patches**	**0.427**	**0.417**
	ruler	0.966	**0.752**		**ruler**	**0.335**	**0.466**
background	**original**	0.884	0.777	**whole (low freq)**	**original**	0.935	0.856
	dark corner	0.888	0.759		**dark corner**	0.836	0.873
	gel bubble	0.893	0.759		**gel bubble**	0.860	0.866
	ink	0.877	**0.729**		**ink**	**0.781**	**0.803**
	patches	**0.860**	0.757		**patches**	0.818	0.874
	ruler	0.890	0.765		**ruler**	0.855	0.837
bbox	**original**	0.880	0.821	**background (high freq)**	**original**	0.877	**0.502**
	dark corner	0.881	0.703		**dark corner**	0.842	**0.376**
	gel bubble	0.868	0.693		**gel bubble**	0.859	**0.555**
	ink	**0.873**	**0.584**		**ink**	**0.840**	**0.474**
	patches	0.896	0.756		**patches**	0.850	**0.479**
	ruler	0.880	0.642		**ruler**	0.847	**0.595**
bbox70	**original**	0.845	0.805	**background (low freq)**	**original**	0.890	0.817
	dark corner	0.832	0.773		**dark corner**	0.874	0.794
	gel bubble	0.852	0.787		**gel bubble**	0.850	0.778
	ink	0.826	**0.667**		**ink**	0.893	**0.768**
	patches	**0.810**	0.711		**patches**	**0.803**	0.790
	ruler	0.856	0.783		**ruler**	0.889	0.817
bbox90	**original**	0.865	0.679				
	dark corner	0.875	**0.573**				
	gel bubble	0.824	**0.595**				
	ink	**0.806**	**0.546**				
	patches	0.809	**0.531**				
	ruler	0.817	**0.498**				

Table 4: Performance in AUROC for ResNet50 classifiers evaluated on the PH2 dataset augmented with Dreambooth artifacts. PH2 and ISIC columns indicate the *training* dataset for the melanoma classifier. The models with low performance (below 0.5 AUROC) and models with the worst performance in each artifact subgroup are bolded.

5. Discussion

5.1. Takeaways

In this work, we present a comprehensive evaluation of the robustness of melanoma classification models under different dataset disruptions and diffusion-based perturbations of isolated non-medical artifacts. For researchers or clinicians hoping to implement deep learning-based models for melanoma classification in practice, we offer the following suggestions: (1) completely removing background information is not an effective solution for mitigating artifacts, as classifiers trained on whole images are more robust under both dataset shift and artifact perturbations; (2) the identification of spurious artifacts requiring correction is dataset-specific and should be systemat-

ically assessed; and (3) mild image compression during data preprocessing is unlikely to degrade classifier performance, as models primarily rely on low-frequency features.

5.2. Limitations

Our proposed diffusion-based perturbation explainability method has several inherent limitations. First, it requires prior knowledge of the artifact to be analyzed, as well as example images containing the artifact to train the diffusion model. However, this requirement aligns with the broader challenges of developing medical deep learning models, in which defining and verifying what counts as a "non-medical" feature necessitates clinical expertise. For example, irregular lesion borders are a clinically significant feature in melanoma classification, as melanoma often presents with jagged or poorly defined edges. However, a non-clinical expert might misinterpret this as a spurious feature, assuming the model is relying on segmentation artifacts rather than a meaningful diagnostic signal.

Second, we experienced difficulties with finetuning the diffusion model for the artifacts of "hair" and "gel border", suggesting that some artifacts are inherently more challenging to learn. While we used standard Dreambooth and LoRA finetuning techniques, more specialized diffusion-based methods may help address these limitations. Lastly, the diffusion model may unintentionally introduce additional diffusion-specific artifacts. To improve the quality of artifact perturbations, highly targeted artifact classification or detection models could be implemented to filter the generated images. We leave this direction for future work.

6. Conclusion

Past work have shown that non-medical artifacts, such as surgical markings and dark corners, can introduce biases that hinder model generalization in melanoma classification. In this study, we systematically evaluated the impact of common artifacts and introduced a diffusion-based perturbation method to isolate and analyze their effects. Our findings show that whole-image classifiers outperform lesion-only or background-only models, emphasizing the importance of preserving contextual information. Additionally, melanoma classifiers rely primarily on low-frequency features, making them less sensitive to mild image compression. Through controlled artifact perturbations, we identify ink markings, rulers, and patches as key sources of classifier sensitivity, while other artifacts, such as dark corners and gel bubbles, have a lesser impact. These results underscore the need for dataset-specific artifact analysis and mitigation strategies to improve the robustness of deep learning models in melanoma classification.

7. Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 2389810.

References

2kpr. Dreambooth Tokens, 2024. URL https://github.com/2kpr/dreambooth-tokens. Accessed: 2025-02-11.

Jose-Agustin Almaraz-Damian, Volodymyr Ponomaryov, Sergiy Sadovnychiy, and Heydy Castillejos-Fernandez. Melanoma and nevus skin lesion classification using handcraft and deep learning feature fusion via mutual information measures. *Entropy*, 22(4):484, 2020.

Giuseppe Argenziano, Gabriella Fabbrocini, Paolo Carli, Vincenzo De Giorgi, Elena Sammarco, and Mario Delfino. Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions: comparison of the abcd rule of dermatoscopy and a new 7-point checklist based on pattern analysis. *Archives of dermatology*, 134(12):1563–1570, 1998.

Muhammad Asad Arshed, Shahzad Mumtaz, Muhammad Ibrahim, Saeed Ahmed, Muhammad Tahir, and Muhammad Shafi. Multi-class skin cancer classification using vision transformer networks and convolutional neural network-based pre-trained models. *Information*, 14(7):415, 2023.

Peter J Bevan and Amir Atapour-Abarghouei. Skin deep unlearning: Artefact and instrument debiasing in the context of melanoma classification. *arXiv preprint arXiv:2109.09818*, 2021.

Alceu Bissoto, Eduardo Valle, and Sandra Avila. Debiasing skin lesion datasets and models? not so fast. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 740–741, 2020.

Christopher Boland, Keith A Goatman, Sotirios A Tsaftaris, and Sonia Dahdouh. There are no shortcuts to anywhere worth going: Identifying shortcuts in deep learning models for medical image analysis. In *Medical Imaging with Deep Learning*, 2024.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

Arief Budhiman, Suyanto Suyanto, and Anditya Arifianto. Melanoma cancer classification using resnet with data augmentation. In *2019 international seminar on research of information technology and intelligent systems (ISRITI)*, pages 17–20. IEEE, 2019.

Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.

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.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Giuseppe Di Leo, Alfredo Paolillo, Paolo Sommella, and Gabriella Fabbrocini. Automatic diagnosis of melanoma: a software system based on the 7-point check-list. In *2010 43rd Hawaii international conference on system sciences*, pages 1–10. IEEE, 2010.

Muhammad Ali Farooq, Wang Yao, Michael Schukat, Mark A Little, and Peter Corcoran. Derm-t2im: Harnessing synthetic skin lesion data via stable diffusion models for enhanced skin disease classification using vit and cnn. *arXiv preprint arXiv:2401.05159*, 2024.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Lahiru Gamage, Uditha Isuranga, Dulani Meedeniya, Senuri De Silva, and Pratheepan Yogarajah. Melanoma skin cancer identification with explainability utilizing mask guided technique. *Electronics*, 13(4):680, 2024.

Bilge S Akkoca Gazioğlu and Mustafa E Kamaşak. Effects of objects and image quality on melanoma classification using deep neural networks. *Biomedical Signal Processing and Control*, 67:102530, 2021.

Holger A Haenssle, Christine Fink, Roland Schneiderbauer, Ferdinand Toberer, Timo Buhl, Andreas Blum, Aadi Kalloo, A Ben Hadj Hassen, Luc Thomas, Alexander Enk, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of oncology*, 29(8):1836–1842, 2018.

James A. Hanley and Barbara J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE journal of biomedical and health informatics*, 23(2):538–546, 2018.

Peiyuan Liao, Xiuyu Li, Xihui Liu, and Kurt Keutzer. The artbench dataset: Benchmarking generative models with artworks. *arXiv preprint arXiv:2206.11404*, 2022.

Caterina Longo, Riccardo Pampena, Elvira Moscarella, Johanna Chester, Michela Starace, Elisa Cinotti, Bianca Maria Piraccini, Giuseppe Argenziano, Ketty Peris, and Giovanni Pellacani. Dermoscopy of melanoma according to different body sites: Head and neck, trunk, limbs, nail, mucosal and acral. *Journal of the European Academy of Dermatology and Venereology*, 37(9): 1718–1730, 2023.

Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. Ph 2-a dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5437–5440. IEEE, 2013.

Matthias Minderer, Olivier Bachem, Neil Houlsby, and Michael Tschannen. Automatic shortcut removal for self-supervised representation learning. In *International Conference on Machine Learning*, pages 6927–6937. PMLR, 2020.

Franz Nachbar, Wilhelm Stolz, Tanja Merkle, Armand B Cognetta, Thomas Vogt, Michael Landthaler, Peter Bilek, Otto Braun-Falco, and Gerd Plewig. The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, 30(4):551–559, 1994.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

Luke W Sagers, James A Diao, Matthew Groh, Pranav Rajpurkar, Adewole S Adamson, and Arjun K Manrai. Improving dermatology classifiers across populations using images generated by large diffusion models. *arXiv preprint arXiv:2211.13352*, 2022.

Zohaib Salahuddin, Henry C Woodruff, Avishek Chatterjee, and Philippe Lambin. Transparency of

deep neural networks for medical image analysis: A review of interpretability methods. *Computers in biology and medicine*, 140:105111, 2022.

Katharina Sies, Julia K Winkler, Christine Fink, Felicitas Bardehle, Ferdinand Toberer, Felix KF Kommoss, Timo Buhl, Alexander Enk, Albert Rosenberger, and Holger A Haenssle. Dark corner artefact and diagnostic performance of a market-approved neural network for skin cancer classification. *JDDG: Journal der Deutschen Dermatologischen Gesellschaft*, 19(6):842–850, 2021.

Sahil Singla and Soheil Feizi. Salient imagenet: How to discover spurious features in deep learning? *arXiv preprint arXiv:2110.04301*, 2021.

Sidharth Sonthalia, Sara Yumeen, and Feroze Kaliyadan. Dermoscopy overview and extradiagnostic applications. 2019.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

Philipp Tschandl, Noel Codella, Bengü Nisa Akay, Giuseppe Argenziano, Ralph P Braun, Horacio Cabo, David Gutman, Allan Halpern, Brian Helba, Rainer Hofmann-Wellenhof, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *The lancet oncology*, 20(7):938–947, 2019.

Alfredo Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, 32(24):18069–18083, 2020.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Kashif Rasul, William Falcon, Sayak Paul, and the Hugging Face Team. Diffusers: State-of-the-art pretrained diffusion models, 2022. URL https://github.com/huggingface/diffusers.

Natalie M Williams, Kristina D Rojas, John M Reynolds, Deukwoo Kwon, Jackie Shum-Tien, and Natalia Jaimes. Assessment of diagnostic accuracy of dermoscopic structures and patterns used in melanoma detection: a systematic review and meta-analysis. *JAMA dermatology*, 157(9):1078–1088, 2021.

Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155(10):1135–1141, 2019.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023.

Appendix A. DenseNet121 Transfer Performance under Dataset and Diffusion-based Perturbations

We repeat the dataset disruption and artifact perturbation experiments using DenseNet121 (Huang et al., 2017), another widely used model architecture for melanoma classification.

Overall, we observe similar trends to our ResNet50 results. Due to dataset size limitations, PH2 classifiers generally do not transfer well to other datasets. In ISIC, the whole-image classifier outperforms both the lesion-only and background-only classifiers, not only within the in-domain ISIC dataset but also when transferred to out-of-domain datasets such as HAM10000 and PH2 (Table 5). Additionally, ISIC bounding box models exhibit reasonable transferability across datasets, reinforcing the idea that certain artifacts in the outer image regions contribute to a degree of generalization. Finally, we find that low-frequency models outperform high-frequency models, consistent with our observations from ResNet50.

Appendix B. Stacked Diffusion Perturbations of Two Artifact Effect on Classifier Performance

In this section, we study the effect of stacking diffusion-based artifact perturbations. Given the already generated set of artifact-perturbed images from Section 4.4, we stack a second artifact perturbation through using our fine-tuned diffusion models to inpaint a different, second artifact. For instance, a heatmap entry with row "ink" and column "ruler" means that we use the ink-specific diffusion model to inpaint the artifact of "ink" on top of images already perturbed with "ruler".

In this section, we analyze the impact of stacking multiple diffusion-based artifact perturbations. Building on the artifact-perturbed images generated in Section 4.4, we apply a second perturbation using our fine-tuned diffusion models to inpaint a different artifact on top of the existing perturbation. For example, in the heatmap, an entry at the intersection of row "ink" and column "ruler" indicates that the ink-specific diffusion model was used to overlay an "ink" artifact onto images that had already been perturbed with "ruler."

For the PH2-trained classifiers (Figure 5), we see that the whole-image classifier remains robust even under two artifact perturbations. When we examine the bounding box classifiers, we see that combinations of patches and dark corner (AUROC bbox: 0.832) and patches and gel bubbles (AUROC bbox70: 0.786, bbox90: 0.776) suffer from the greatest drop in performance within each dataset mode. This increased sensitivity to patches is consistent with our findings from the single-artifact perturbation experiments.

For the ISIC-trained classifiers (Figure 6), whole-image and low-frequency classifiers are robust to artifact perturbations, whereas high-frequency classifiers fail to generalize. This finding is also consistent with our single-artifact results. Interestingly, we observe that both the bbox and bbox90 classifiers struggle to transfer under two artifact perturbations. This differs from the single-artifact scenario, where the bbox classifier showed better generalization. These findings suggest that as artifact perturbations become more complex, classifier generalization diminishes.

Domain **AUROC**	Source **ISIC**	Target **HAM10000**	Target **PH2**
whole	0.804 ± 0.029	0.781 ± 0.033	0.890 ± 0.082
lesion	0.758 ± 0.051	0.725 ± 0.027	0.810 ± 0.090
background	0.735 ± 0.033	0.733 ± 0.019	0.829 ± 0.076
bbox	0.720 ± 0.017	0.673 ± 0.038	0.810 ± 0.149
bbox70	0.661 ± 0.019	0.674 ± 0.057	0.762 ± 0.134
bbox90	0.634 ± 0.060	0.647 ± 0.047	0.680 ± 0.113
whole (high freq)	0.687 ± 0.034	0.567 ± 0.047	0.741 ± 0.033
lesion (high freq)	0.696 ± 0.014	0.624 ± 0.037	0.768 ± 0.152
background (high freq)	0.717 ± 0.009	0.623 ± 0.026	0.799 ± 0.068
whole (low freq)	0.807 ± 0.026	0.756 ± 0.018	0.927 ± 0.051
lesion (low freq)	0.790 ± 0.036	0.714 ± 0.047	0.900 ± 0.078
background (low freq)	0.723 ± 0.021	0.683 ± 0.013	0.849 ± 0.104

Table 5: Performance in AUROC with standard deviations for the DenseNet121 melanoma classifier trained on the ISIC dataset and evaluated on the ISIC, HAM10000, and PH2 datasets.

Domain **AUROC**	Source **PH2**	Target **ISIC**	Target **HAM10000**
whole	0.933 ± 0.042	0.573 ± 0.054	0.440 ± 0.045
lesion	0.956 ± 0.030	0.623 ± 0.038	0.700 ± 0.032
background	0.905 ± 0.054	0.624 ± 0.031	0.521 ± 0.052
bbox	0.888 ± 0.096	0.622 ± 0.015	0.554 ± 0.051
bbox70	0.871 ± 0.088	0.562 ± 0.056	0.456 ± 0.045
bbox90	0.832 ± 0.063	0.576 ± 0.023	0.552 ± 0.040
whole (high freq)	0.804 ± 0.100	0.511 ± 0.008	0.420 ± 0.062
lesion (high freq)	0.911 ± 0.104	0.595 ± 0.043	0.359 ± 0.022
background (high freq)	0.928 ± 0.078	0.564 ± 0.022	0.333 ± 0.014
whole (low freq)	0.930 ± 0.040	0.575 ± 0.054	0.551 ± 0.022
lesion (low freq)	0.936 ± 0.057	0.633 ± 0.038	0.723 ± 0.013
background (low freq)	0.869 ± 0.101	0.589 ± 0.031	0.536 ± 0.052

Table 6: Performance in AUROC with standard deviations for the DenseNet121 melanoma classifier trained on the PH2 dataset and evaluated on the ISIC, HAM10000, and PH2 datasets.

Dataset Mode	Artifact	PH2	ISIC	Dataset Mode	Artifact	PH2	ISIC
whole	**original**	0.933	0.890	**whole (high freq)**	**original**	0.804	0.741
	dark corner	0.947	0.846		**dark corner**	0.466	0.704
	gel bubble	0.948	0.888		**gel bubble**	0.505	0.738
	ink	0.946	0.837		**ink**	0.452	0.775
	patches	0.937	0.841		**patches**	0.351	0.798
	ruler	0.946	0.876		**ruler**	0.447	0.790
background	**original**	0.905	0.829	**whole (low freq)**	**original**	0.930	0.927
	dark corner	0.906	0.805		**dark corner**	0.928	0.893
	gel bubble	0.920	0.793		**gel bubble**	0.909	0.906
	ink	0.926	0.725		**ink**	0.903	0.848
	patches	0.904	0.767		**patches**	0.922	0.871
	ruler	0.920	0.812		**ruler**	0.915	0.888
bbox	**original**	0.888	0.810	**background (high freq)**	**original**	0.928	0.799
	dark corner	0.909	0.735		**dark corner**	0.865	0.744
	gel bubble	0.891	0.739		**gel bubble**	0.925	0.695
	ink	0.894	0.676		**ink**	0.827	0.712
	patches	0.919	0.689		**patches**	0.791	0.778
	ruler	0.910	0.734		**ruler**	0.864	0.799
bbox70	**original**	0.871	0.762	**background (low freq)**	**original**	0.869	0.849
	dark corner	0.861	0.695		**dark corner**	0.881	0.816
	gel bubble	0.878	0.753		**gel bubble**	0.869	0.802
	ink	0.854	0.613		**ink**	0.881	0.804
	patches	0.856	0.638		**patches**	0.877	0.841
	ruler	0.865	0.744		**ruler**	0.870	0.879
bbox90	**original**	0.832	0.680				
	dark corner	0.813	0.617				
	gel bubble	0.819	0.630				
	ink	0.819	0.592				
	patches	0.774	0.595				
	ruler	0.821	0.624				

Table 7: AUROC performance for DenseNet121 classifiers evaluated on the PH2 dataset augmented with Dreambooth artifacts. PH2 and ISIC columns indicate the *training* dataset for the melanoma classifier.

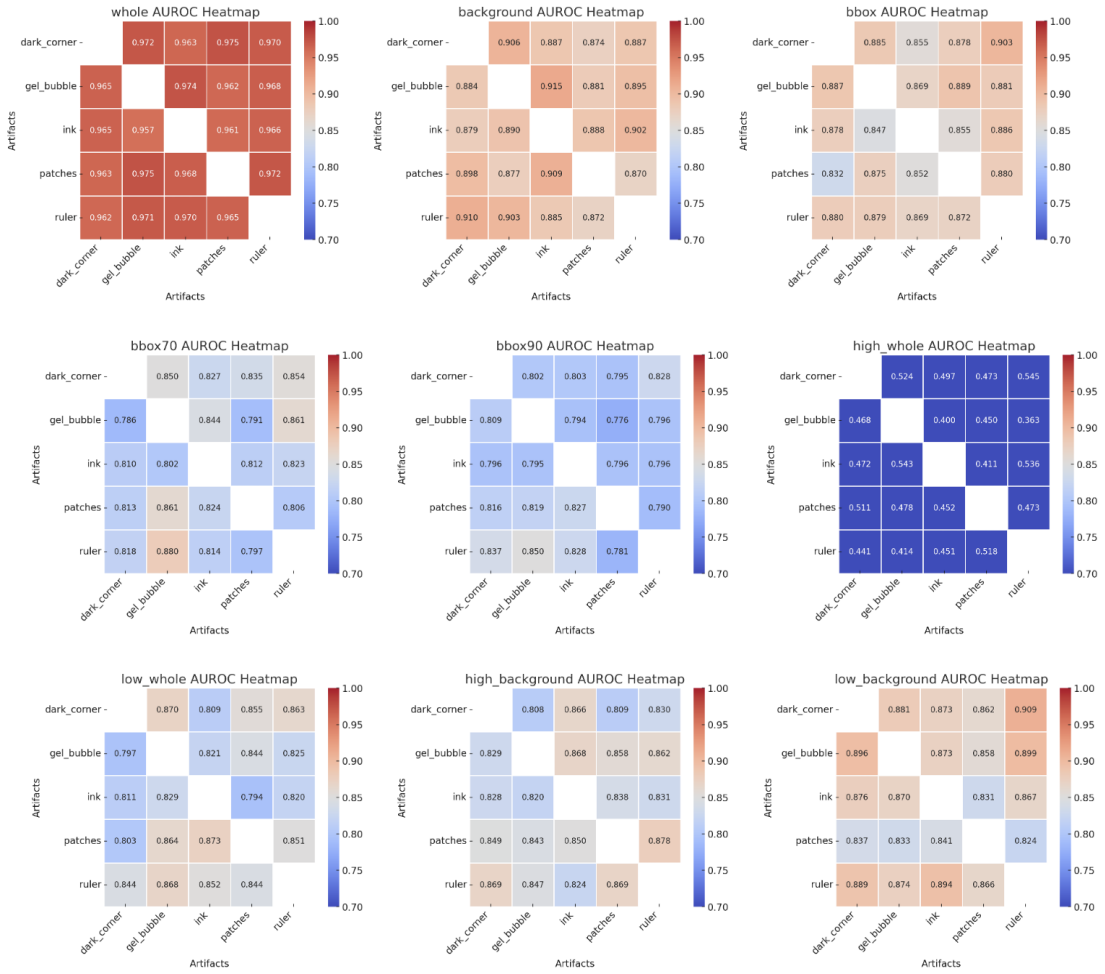


Figure 5: Heatmaps of test AUROC performances of the PH2 classifier under different dataset modes with two stacked artifact perturbations of the PH2 dataset. The x-axis is the first added artifact, and the y-axis is the second added artifact. Due to the stochasticity of diffusion generation, the matrix is not symmetric.

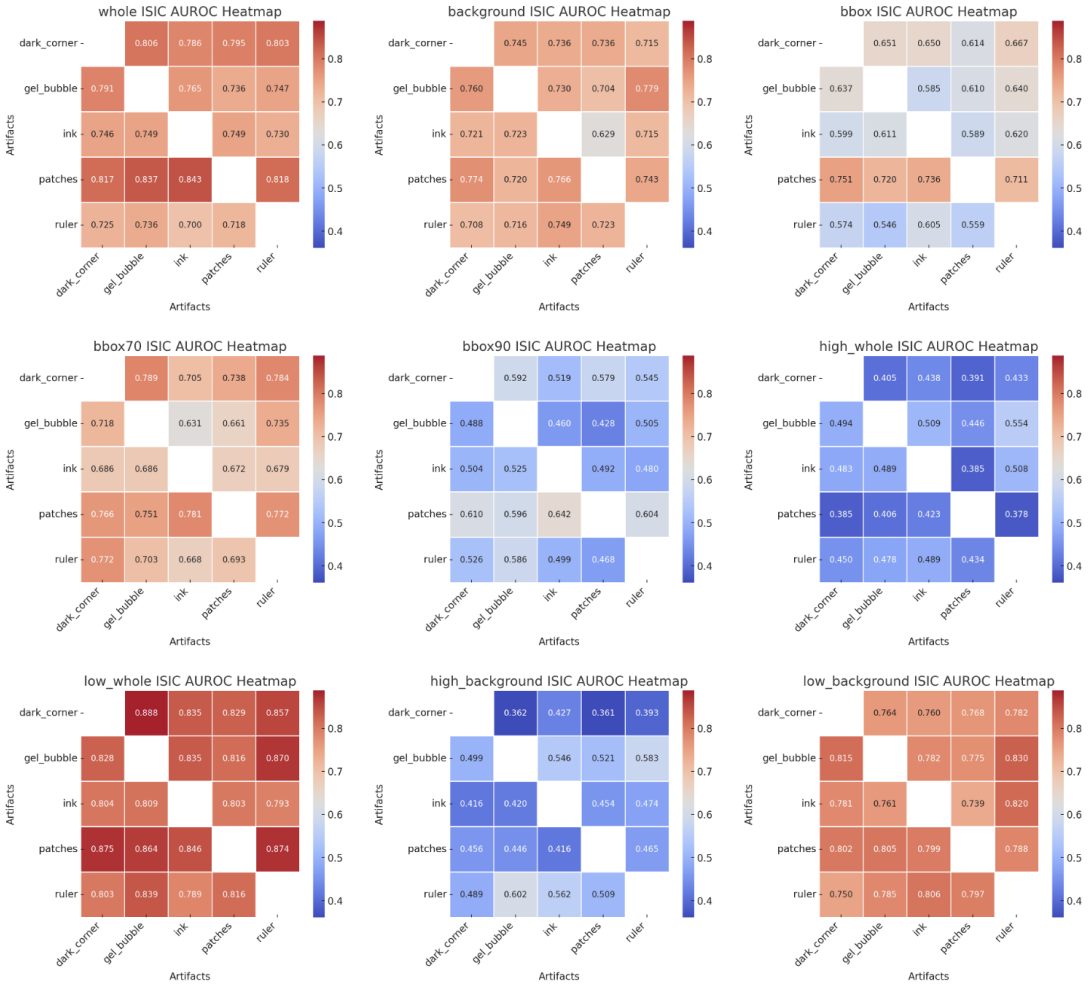


Figure 6: Heatmaps of test AUROC performances of the ISIC classifier under different dataset modes with two stacked artifact perturbations of the PH2 dataset. The x-axis is the first added artifact, and the y-axis is the second added artifact. Due to the stochasticity of diffusion generation, the matrix is not symmetric. Note that the scale of the colorbars for ISIC classifier is different from the PH2 classifiers, so colors cannot be directly compared across figures.