Fair and Explainable Melanoma Diagnosis: Integrating Conditional Diffusion-Based Image Generation with Grad-CAM

Anna Berdichevskaia

Email: annab4@mail.tau.ac.il

Wassem Bsharat

Email: wassemb@mail.tau.ac.il

***Abstract*—Melanoma is a highly aggressive skin cancer where early detection is paramount for improving patient outcomes. Deep learning has shown promise in automating melanoma diagnosis; however, prevalent datasets like HAM10000 suffer from underrepresentation of darker skin tones, which may lead to biased performance. In this study, we propose a dual-model pipeline that combines an explainable binary classifier with a conditional image generator based on denoising diffusion probabilistic models (DDPMs) to enhance both accuracy and fairness. The classifier leverages pre-trained ResNet backbones (ResNet18, ResNet34, ResNet50, and ResNet101) and employs Grad-CAM to produce visual attributions that elucidate the model's decision-making process.**

Initially, the models were trained on HAM10000 and subsequently fine-tuned on a curated subset of the Fitzpatrick17k dataset to better capture the diversity of skin tones. The fine-tuning process resulted in substantial improvements. For example, the ResNet18 model achieved a ROC AUC of 0.9923, balanced accuracy of 0.9500, and perfect sensitivity (1.0000) on Fitzpatrick Scale 4, while the ResNet50 model demonstrated a ROC AUC of 0.9485 and balanced accuracy of 0.9706 with perfect sensitivity on Fitzpatrick Scale 3. Similar performance gains were observed with ResNet34 and ResNet101, confirming that targeted fine-tuning significantly mitigates skin tone disparities.

Our results indicate that the integration of synthetic data augmentation and explainable AI can yield robust and fair diagnostic models. The Grad-CAM attributions provide critical insights into lesion characteristics, fostering greater clinical trust. Future work will focus on further optimizing the DDPM generator, incorporating explainability directly into the training loss, and quantitatively evaluating synthetic image quality using metrics such as FID, LPIPS, and SSIM. These advancements aim to develop diagnostic systems that are both accurate and equitable across diverse patient populations.

***Index Terms*—Deep Learning, Computer Vision, XAI, Diffusion Models, Medical Imaging, Classification, Melanoma, Skin Cancer.**

I. Introduction

Melanoma is a dangerous form of skin cancer that can spread rapidly if not detected early. Deep learning models have shown promise in automating melanoma classification from dermoscopic images, offering potential to assist dermatologists with early diagnosis. However, the adoption of AI systems in clinical practice demands transparency, interpretability, and fairness — qualities often lacking in standard deep neural networks.

Explainable AI (XAI) techniques have emerged to address these challenges. Tools like Grad-CAM [1], SHAP, and saliency maps allow visualization of model attention and reasoning, helping clinicians better understand and trust automated decisions [2]. These techniques are especially important in medical AI, where interpretability is critical for safe deployment.

A major issue in dermatological AI is dataset bias. Studies have shown that skin cancer datasets significantly underrepresent darker skin tones — particularly Fitzpatrick types IV-VI [3]. This imbalance can lead to poor model performance on these subgroups and exacerbate healthcare disparities [4]. Additionally, skin tone has been shown to influence cancer risk and presentation, further motivating the need for balanced, fair models.

To address these concerns, we propose a dual-model pipeline for melanoma diagnosis. It includes an explainable classifier and a conditional image generator based on denoising diffusion probabilistic models (DDPMs). The generator is trained to synthesize dermoscopic images conditioned on metadata such as skin tone and diagnosis, with the goal of augmenting underrepresented cases. Prior work has shown that diffusion-based generation can significantly improve classifier robustness and fairness under distribution shifts [5].

A visualization of skin types and their associated melanoma risk is shown in Figure 1.

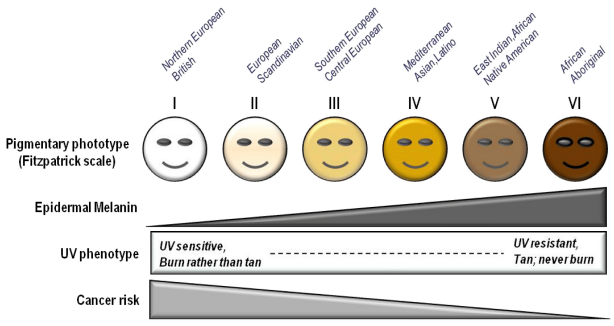


Figure 1: The Fitzpatrick scale and the risk of skin cancer [6].

II. Related Work

Previous research has applied deep learning to skin cancer classification with high accuracy on benchmark datasets. Models such as ResNet and EfficientNet have been used to identify malignant lesions from dermoscopic images [7]. However, these models typically lack interpretability and do not account for demographic disparities.

Explainability techniques like Grad-CAM and SHAP have been proposed to increase clinician trust and regulatory compliance by visualizing model attention [1, 2]. These methods help identify whether a model relies on meaningful lesion features or spurious artifacts.

Recent work also explores fairness in dermatological AI. Several studies have shown that models trained on biased datasets underperform on images of darker skin tones, which are often underrepresented [4]. This can lead to diagnostic inaccuracies and unequal healthcare outcomes.

To mitigate this, Ktena *et al.* [5] introduced a generative augmentation framework based on diffusion models. Their approach generates synthetic images conditioned on skin tone and diagnosis label, and has been shown to improve classifier fairness under distribution shift. Our project builds on this approach by integrating a similar conditional DDPM into a dual-model pipeline and evaluating its early-stage outputs for melanoma data generation.

III. Methods

A. Software and Statistics

All experiments were conducted using Python 3.9 on Google Colab Pro with NVIDIA L4 GPUs. The project relied heavily on PyTorch and PyTorch Lightning for model development, training, and management. Data preprocessing and augmentation were performed using Albumentations and custom scripts built on NumPy and Pandas. Visualization and monitoring were done with Matplotlib and TensorBoard.

The following libraries were used:

- **Core Machine Learning:** PyTorch, PyTorch Lightning, TorchMetrics, torchvision
- **Data Handling:** NumPy, Pandas, OpenCV, PIL, Albumentations
- **Model Explainability:** Captum, Grad-CAM (via `LayerGradCam`, `LayerAttribution`)
- **Training Utilities:** tqdm, DataLoader, WeightedRandomSampler, TensorBoardLogger
- **Evaluation and Analysis:** Scikit-learn, Matplotlib, Seaborn, torchmetrics (Accuracy, AUROC, Recall, Specificity)

B. Dataset

1) HAM10000: In our work, we used the publicly available HAM10000 dataset [8], which includes 10,015 dermoscopic images capturing various skin conditions from different anatomical sites. The collection features images from both genders and covers an age range of 0 to 85 years, grouped in five-year intervals. Since multiple images were frequently taken for each lesion-and sometimes a single patient contributed more than one lesion-the total image count exceeds the number of unique lesions, and that in turn is higher than the number of individual patients. Confirmatory diagnoses were obtained via excision followed by pathological examination, expert panel review, or through patient follow-up. For our study, we specifically selected all biopsy-confirmed melanoma and nevus cases from HAM10000, which amounted to 3,611 images corresponding to 1,981 unique lesions.

The data splitting procedure began by loading the metadata and separating lesion IDs into two groups based on the binary label 'benign_malignant' – with melanoma lesions (label 1) and nevi lesions (label 0). From these, 100 melanoma and 100 nevi lesions were randomly sampled to form the test set, ensuring balanced class representation. For each selected lesion in the test set, the latest image was chosen by sorting the images by their identifier and taking the final entry. The remaining lesions were then reserved for training and validation. Unique lesion IDs from the remaining dataset were stratified by their class labels and split into training (82%) and validation (18%) subsets using a fixed random seed for reproducibility. Finally, the split labels ('train', 'val', and 'test') were assigned to the metadata and saved for subsequent experiments.

2) Fitzpatrick17k: For additional testing and fine-tuning, we utilized the Fitzpatrick17k dataset [9][10], which comprises 16,577 clinical images annotated with Fitzpatrick skin type labels. This dataset spans 114 skin conditions, with each condition represented by 53 to 653 images.

To create a smaller subset for our experiments, we first filtered the metadata to retain only images labeled as either "melanoma" or "nevus" (yielding 975 images) and assigned a binary label (1 for melanoma, 0 for nevus). We then partitioned the data by Fitzpatrick skin tone, shuffling each group and splitting it into an 80% training set and a 20% test set (with a separate test split for each Fitzpatrick scale, labeled as "test[1-6]"). Finally, a custom download function—configured with a timeout and browser-mimicking HTTP headers—was used to retrieve the images from their URLs.

C. Diffusion-Based Image Generator

As part of our dual-model pipeline for melanoma diagnosis enhancement, we implemented a conditional image generator based on denoising diffusion probabilistic models (DDPMs). Inspired by Ktena et al. [5], our model generates synthetic dermoscopic images conditioned on metadata, with the goal of mitigating dataset imbalance and improving fairness.

1) Forward Diffusion Process: We adopted the standard DDPM forward process [11], in which Gaussian noise is added to clean images over a series of $T = 1000$ timesteps:

$$z_t = \sqrt{\bar{\alpha}_t} \cdot x + \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

Here, $\bar{\alpha}_t$ is computed using a cosine-based schedule. This process is fixed and used only to generate noisy training inputs.

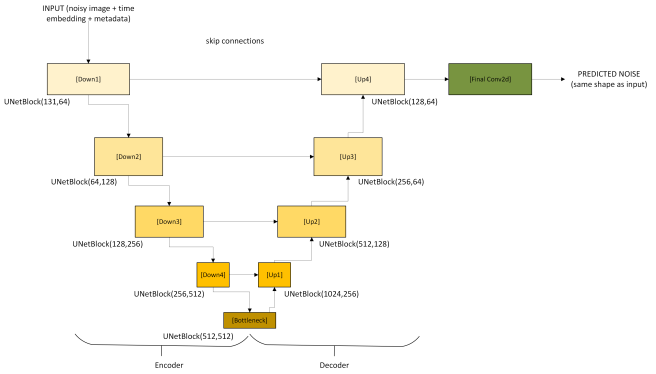


Figure 2: U-Net-based generator architecture for predicting noise from input z_t.

2) Reverse Process and Generator Architecture: Our reverse model is a deep U-Net that predicts the noise added at each timestep t. It receives as input the noisy image z_t, the timestep t, and patient metadata. Its architecture includes:

- Four downsampling and four upsampling convolutional blocks
- A central bottleneck for abstract feature processing
- Conditioning branches for timestep embeddings and metadata

These embeddings are broadcast to match the input image and concatenated before entering the encoder. The decoder combines features with skip connections to preserve spatial detail. A final 1x1 convolution outputs the predicted noise $\hat{\epsilon}$.

3) Training Objective: The model is trained to minimize the mean squared error (MSE) between predicted and true noise:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{x,t,\epsilon}\left[\|\epsilon - \hat{\epsilon}_\theta(z_t, t, \text{metadata})\|^2\right] \quad (2)$$

This objective ensures accurate denoising at any timestep.

4) Implementation Details:

- **Noise schedule**: Cosine schedule with 1,000 timesteps
- **Input**: 256×256 RGB images
- **Metadata**: 12-dimensional patient vector
- **Training**: Adam optimizer ($lr = 2 \times 10^{-4}$), 300 epochs, early stopping, gradient clipping

D. Binary classification model

The model leverages a ResNet backbone (ResNet18/34/50/101) for binary melanoma classification. Its architecture, illustrated in the Fig.3, starts with an image preprocessing pipeline that applies extensive data augmentation—including random transposition (p=0.2), vertical and horizontal flips (p=0.5), color jittering (p=0.5), CLAHE (clip limit=4.0, p=0.7), hue-saturation-value adjustments (with hue_shift_limit=10, sat_shift_limit=20, val_shift_limit=10, p=0.5), and shift-scale-rotate transformations (shift_limit=0.1, scale_limit=0.1, rotate_limit=15, border_mode=0, p=0.85)—followed by resizing to 224×224 pixels and normalization. The processed images are then fed into the chosen pre-trained ResNet for feature extraction, after which a custom classification head (comprising a dropout layer with a rate of 0.4 and a fully connected layer) produces the final probability via a sigmoid activation.

The training was performed as a minimization of a standard Binary Cross-Entropy loss function. The model was trained for 25 epochs using the Adam optimizer (with epsilon=10^{-8}) at a learning rate of 0.0001, with a batch size of 64 and a seed set to 42 for reproducibility.

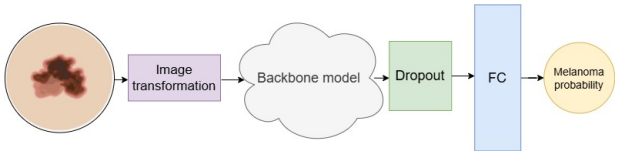


Figure 3: The architecture of the classifier.

E. Design of the XAI

To enhance the interpretability of our binary classification model, we employ a Grad-CAM based approach for generating visual attributions. During the forward pass, the model's predictions are accompanied by attribution maps that indicate which regions of the input image most strongly influenced the output. Specifically, we utilize the LayerGradCam method, applied to the last convolutional layer of the backbone network (i.e., model.base_model.layer4[-1]).

The forward function works as follows: an input image is passed through the model to obtain the prediction. Depending on whether the output exceeds a threshold of 0.5, the Grad-CAM attributions are computed with a ReLU applied directly (for outputs greater than or equal to 0.5) or, alternatively, the negative attributions are first computed and then rectified using ReLU. This modification ensures that the attributions correctly emphasize the regions responsible for positive predictions while inverting the signals when the prediction is negative. Finally, the attribution maps are interpolated to match the original input image dimensions (specified by IMAGE_SIZE), which facilitates direct overlay and visualization.

IV. Results

A. Binary classification on HAM10000 dataset

We evaluated the performance of our deep learning models on the HAM10000 dataset to distinguish between melanoma and non-melanoma cases. Four different ResNet backbones—ResNet18, ResNet34, ResNet50, and ResNet101—were trained and tested on this task, with the dataset providing a challenging and diverse set of high-resolution dermoscopic images. The models were optimized using the Adam optimizer, and the evaluation metrics included ROC AUC, Balanced Accuracy, Sensitivity, and Specificity, which together provide a comprehensive view of their diagnostic performance.

As detailed in Table I, all models exhibited strong overall performance with ROC AUC values above 0.84. In particular, ResNet101 achieved the highest ROC AUC (0.8620)

and the best specificity (0.8373), suggesting it is particularly effective at minimizing false positives. On the other hand, ResNet34 recorded the highest balanced accuracy (0.7932), while ResNet50 demonstrated the best sensitivity (0.7870), indicating its robust ability to correctly identify melanoma cases. These differences underscore the trade-offs inherent in model selection, where the choice of architecture may depend on whether minimizing false negatives or false positives is more critical in a clinical context.

Table I: Test Metrics for Different Model Backbones

Model Backbone	ROC AUC	Balanced Accuracy	Sensitivity	Specificity
ResNet18	0.8421	0.7682	0.7593	0.7771
ResNet34	0.8609	**0.7932**	0.7731	0.8133
ResNet50	0.8543	0.7851	**0.7870**	0.7831
ResNet101	**0.8620**	0.7867	0.7361	**0.8373**

B. Evaluation of XAI Quality via Fidelity Metrics

To assess the quality of our explainability approach, we employed a fidelity-based evaluation using Grad-CAM attributions. Specifically, for each image, we computed Grad-CAM attributions using a modified forward pass, where the attribution map was thresholded at various percentiles (ranging from 5% to 95%). The attribution maps were then used to generate modified inputs—by either blackening the regions deemed important—followed by measuring the absolute difference between the original model output and the output after modification. This absolute difference, defined as the fidelity, serves as an indicator of how closely the attributions align with the model's decision.

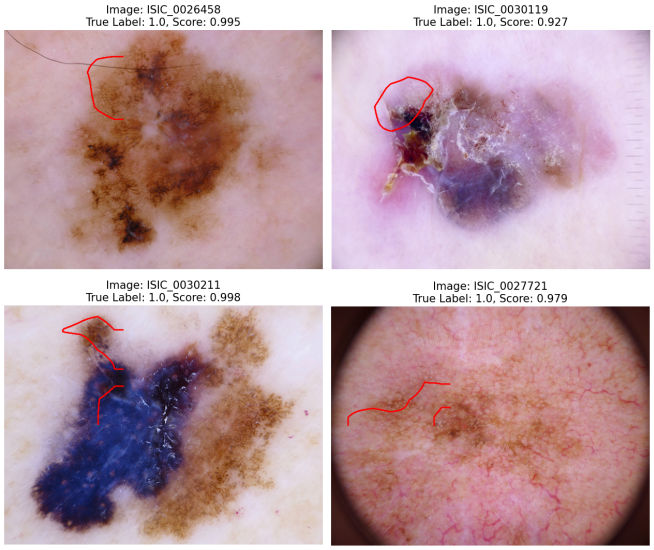
We aggregated the fidelity values across the test images to produce mean fidelity curves, boxplots, and histograms. These analyses enable us to understand the dependency of fidelity on the chosen threshold and to quantify the robustness of the explanation. In general, lower thresholds tend to produce broader masks that result in higher fidelity (i.e., a larger change in prediction), while higher thresholds yield more localized modifications and, consequently, lower fidelity values.

Figures 4 display the fidelity results for models based on ResNet18, ResNet34, ResNet50, and ResNet101, respectively. In Figure 4(a), the ResNet18 model demonstrates a moderate fidelity distribution with a clear trend as the threshold increases. The ResNet34 model, shown in Figure 4(b), exhibits slightly more stable and consistent attributions. The ResNet50 model (Figure 4(c)) tends to achieve higher average fidelity at the selected threshold, suggesting that its attributions are more strongly aligned with the prediction. Meanwhile, the ResNet101 model (Figure 4(d)) displays overall good fidelity, albeit with some variability and occasional outliers. These findings indicate that while all backbones provide meaningful explanations, there is room for improvement.

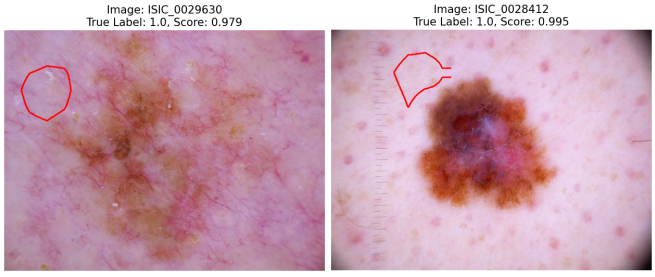
C. XAI results visualization

Figure5 presents Grad-CAM attributions for several melanoma cases correctly classified by our model. The red contours highlight the areas with the highest relevance for the final prediction. In many instances, these contours align with clinically significant features, such as irregular borders or heterogeneous pigmentation, suggesting that the model is focusing on lesion characteristics that dermatologists typically consider indicative of malignancy.

Nevertheless, we occasionally observe contours that appear random or unrelated to any discernible lesion structures. These outliers underscore the fact that while Grad-CAM can provide valuable insights into the model's decision-making process, it is not a perfect reflection of clinical reasoning. Further work may be required to refine attribution methods, ensuring more consistent alignment with meaningful visual cues in dermoscopic images.



(a) Good examples: Grad-CAM attributions correctly highlighting lesion regions.

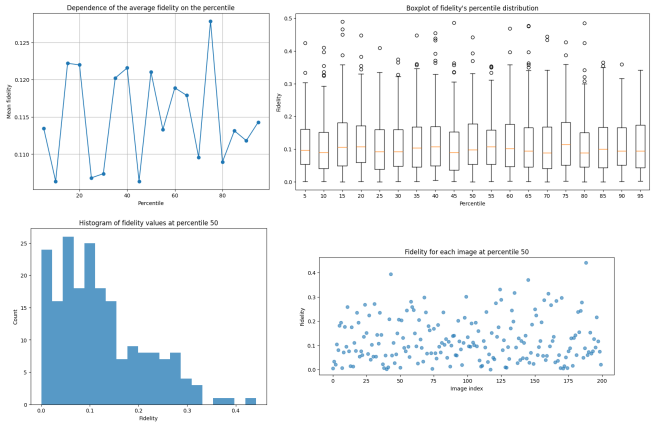


(b) Poor examples: Attributions that appear random and lack clear correlation with lesion features.

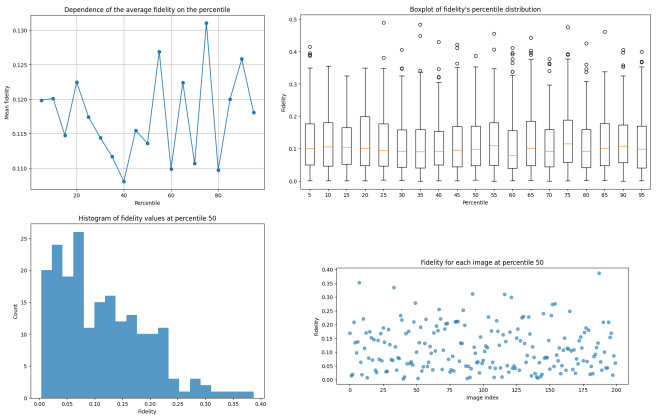
Figure 5: (a) Example Grad-CAM attributions for melanoma images, showing regions deemed most influential (outlined in red) along with true diagnosis and prediction score; (b) Poor examples where the attribution contours do not correlate with meaningful lesion structures.

D. Testing on Different Skin Tones: Fine-Tuning and Re-Evaluation

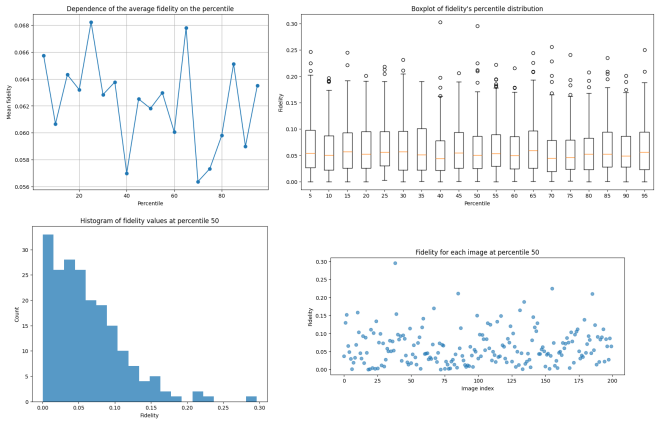
Initially, our melanoma classification models were trained exclusively on the HAM10000 dataset, which, although comprehensive in its representation of various skin diseases, is known to have limited diversity in skin tone. When evaluating



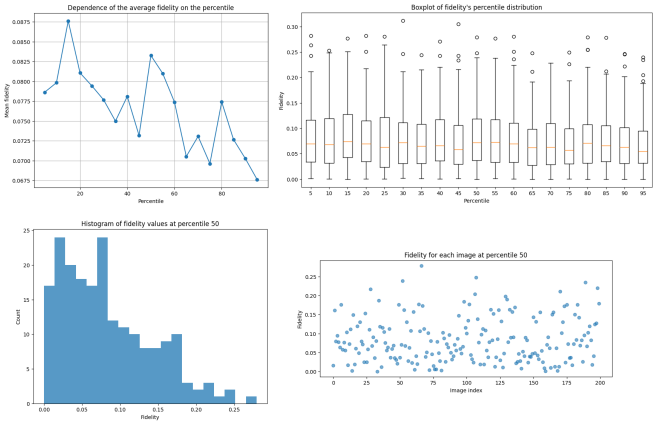
(a) ResNet18



(b) ResNet34



(c) ResNet50



(d) ResNet101

Figure 4: Fidelity evaluation of Grad-CAM attributions for different model backbones. Panel (a) shows the results for ResNet18, (b) for ResNet34, (c) for ResNet50, and (d) for ResNet101. The fidelity metric is computed as the absolute change in model output when regions of the image with high attribution values (thresholded at different percentiles) are modified.

these models on a subset of the data stratified by Fitzpatrick skin types (derived from the available HAM10000 images), we observed significant variability in performance metrics across different skin tones, as summarized in Table II. For instance, some Fitzpatrick scales exhibited lower ROC AUC and sensitivity, indicating that the models struggled to consistently capture lesion characteristics across a diverse range of skin pigmentation.

To address this limitation and improve fairness, we further fine-tuned the initially HAM10000-trained models on a curated subset of the Fitzpatrick17k dataset. This additional training phase was designed to expose the models to a broader and more representative range of skin tones. After fine-tuning, we re-evaluated the models on the same stratified test set, which now consisted of the remaining images from the Fitzpatrick17k subset. As detailed in Table III, the fine-tuning process led to notable improvements in performance across most Fitzpatrick scales. ROC AUC, balanced accuracy, and sensitivity increased markedly, suggesting that the models had better adapted to the visual nuances present in different skin types. Specificity remained high, reinforcing the robustness of the fine-tuning approach.

In summary, while the initial models trained solely on HAM10000 demonstrated promising overall performance, their diagnostic accuracy varied considerably with skin tone. The additional fine-tuning on the Fitzpatrick17k subset significantly enhanced the models' ability to generalize across diverse skin types, leading to a more equitable and reliable performance. These results underscore the importance of incorporating diverse datasets and targeted fine-tuning strategies in the development of clinical diagnostic tools.

E. Early Generation Results

Although the diffusion model has not yet fully converged, early results indicate meaningful learning. Figure 6 shows example images generated by the model after 50 epochs. The outputs exhibit recognizable lesion structures and reduced noise artifacts compared to the initial noisy inputs.

Table II: Performance Metrics by Fitzpatrick Scale for Each Model Backbone before fine-tuning

(a) ResNet18

Fitzpatrick Scale	ROC AUC	Bal. Acc.	Sensitivity	Specificity
1	0.5158	0.5459	0.2632	0.8286
2	0.5894	0.5560	0.2500	0.8621
3	0.4044	0.4743	0.1250	0.8235
4	0.7692	0.6538	0.3077	1.0000
5	0.5938	0.5000	0.0000	1.0000
6	0.6667	0.7500	0.5000	1.0000

(b) ResNet34

Fitzpatrick Scale	ROC AUC	Bal. Acc.	Sensitivity	Specificity
1	0.6677	0.6609	0.5789	0.7429
2	0.5970	0.6153	0.4375	0.7931
3	0.3971	0.5092	0.3125	0.7059
4	0.6538	0.5269	0.1538	0.9000
5	0.3750	0.4375	0.1250	0.7500
6	0.8333	0.5833	0.5000	0.6667

(c) ResNet50

Fitzpatrick Scale	ROC AUC	Bal. Acc.	Sensitivity	Specificity
1	0.6992	0.6128	0.3684	0.8571
2	0.4494	0.5420	0.1875	0.8966
3	0.4338	0.5037	0.1250	0.8824
4	0.8385	0.6538	0.3077	1.0000
5	0.1563	0.3750	0.0000	0.7500
6	0.6667	0.7500	0.5000	1.0000

(d) ResNet101

Fitzpatrick Scale	ROC AUC	Bal. Acc.	Sensitivity	Specificity
1	0.5474	0.5248	0.4211	0.6286
2	0.5075	0.4962	0.4063	0.5862
3	0.3272	0.5386	0.3125	0.7647
4	0.5154	0.5154	0.2308	0.8000
5	0.4063	0.3125	0.1250	0.5000
6	1.0000	0.8333	1.0000	0.6667

Table III: Fine-Tuned Performance Metrics by Fitzpatrick Scale for Each Model Backbone

(a) ResNet18

Fitzpatrick Scale	ROC AUC	Bal. Acc.	Sensitivity	Specificity
1	0.8812	0.8331	0.8947	0.7714
2	0.9084	0.7807	0.9063	0.6552
3	0.9412	0.9099	0.9375	0.8824
4	0.9923	0.9500	1.0000	0.9000
5	0.9063	0.8750	1.0000	0.7500
6	1.0000	1.0000	1.0000	1.0000

(b) ResNet34

Fitzpatrick Scale	ROC AUC	Bal. Acc.	Sensitivity	Specificity
1	0.9684	0.8737	0.9474	0.8000
2	0.8427	0.6945	0.9063	0.4828
3	0.9301	0.9412	1.0000	0.8824
4	0.9769	0.8500	1.0000	0.7000
5	0.7188	0.7500	1.0000	0.5000
6	1.0000	0.8333	1.0000	0.6667

(c) ResNet50

Fitzpatrick Scale	ROC AUC	Bal. Acc.	Sensitivity	Specificity
1	0.9534	0.8519	0.7895	0.9143
2	0.8459	0.8044	0.7813	0.8276
3	0.9485	0.9706	1.0000	0.9412
4	0.9769	0.9115	0.9231	0.9000
5	0.8125	0.7500	1.0000	0.5000
6	1.0000	1.0000	1.0000	1.0000

(d) ResNet101

Fitzpatrick Scale	ROC AUC	Bal. Acc.	Sensitivity	Specificity
1	0.9383	0.8000	1.0000	0.6000
2	0.8987	0.6724	1.0000	0.3448
3	1.0000	0.8824	1.0000	0.7647
4	0.9692	0.7615	0.9231	0.6000
5	0.6563	0.7500	1.0000	0.5000
6	1.0000	1.0000	1.0000	1.0000

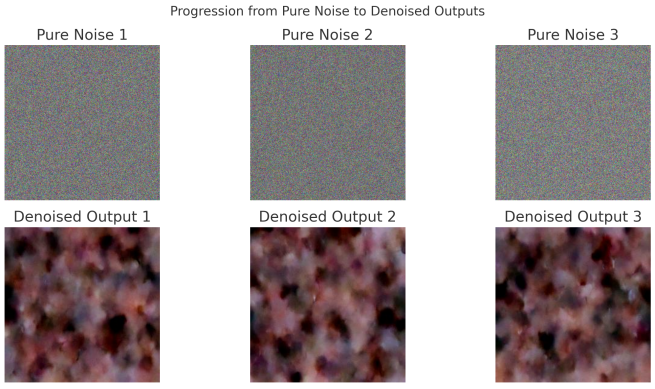


Figure 6: Synthetic melanoma images generated by the partially trained model. While not yet fully refined, lesion boundaries and skin textures begin to emerge.

Figure 7 presents the training loss (MSE) over the course of training. The steady decline confirms that the model is learning to predict the noise component as expected in the DDPM framework.

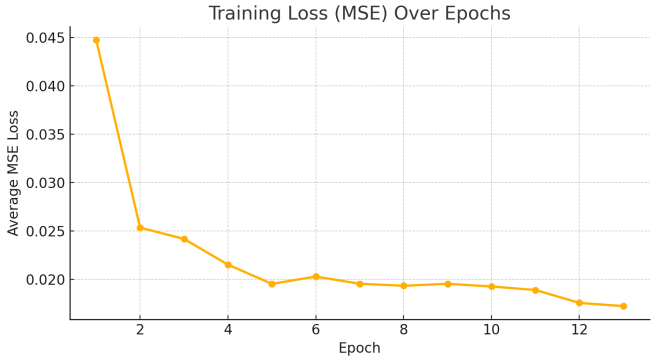


Figure 7: Training loss curve (MSE) over epochs. A consistent decline indicates effective learning, despite incomplete training.

These preliminary outputs support the viability of the approach and lay the groundwork for continued training and evaluation.

V. Discussion

Our study demonstrates that deep learning can be an effective tool for melanoma detection, achieving robust performance across various ResNet-based architectures. Overall, the

results obtained on the HAM10000 dataset, along with further fine-tuning on a curated subset of Fitzpatrick17k images, indicate that our models are capable of learning clinically relevant features. However, our evaluation also reveals variability in performance across different skin tone groups. Specifically, while the overall metrics are promising, some Fitzpatrick scales exhibit lower sensitivity and ROC AUC values. This inconsistency suggests that the current training data may not fully capture the diversity of skin pigmentation found in the real world, which can lead to disparities in diagnostic accuracy.

In parallel, we explored the generation of synthetic images with darker skin tones using a denoising diffusion-based generator. The motivation behind this approach is to augment the dataset with images that are underrepresented in the original collection, potentially reducing bias and improving model fairness. Unfortunately, our initial attempts to generate high-quality synthetic images were not entirely successful due to convergence challenges and limited training time. Nonetheless, the preliminary outputs demonstrated promising structural learning, as the generator began to capture relevant lesion characteristics. We believe that with further development—such as extended training to full convergence, incorporation of classifier-free guidance, and tuning with standard metrics like FID, LPIPS, and SSIM—synthetic image generation could become a valuable tool for creating more balanced datasets, particularly for darker skin tones.

Furthermore, our study incorporates an explainable AI (XAI) component using Grad-CAM to generate visual attributions that highlight regions of high importance in the input images. These attributions were evaluated using a fidelity metric, which quantifies the change in model output when high-attribution regions are modified. The results from the fidelity evaluation indicate that, while the attributions generally align with clinically significant features (e.g., irregular borders, heterogeneous pigmentation), there are instances where the contours appear random or inconsistent. This variability underscores the need for further refinement in our XAI methodology. Future work will investigate integrating Grad-CAM directly into the training process—potentially as part of the loss function—to encourage the model to produce more reliable and interpretable attributions. Additionally, leveraging annotated images with expert markings could further guide the model in learning to focus on the most clinically meaningful regions.

In summary, while our current findings are encouraging and demonstrate the potential of deep learning for melanoma detection, they also highlight important areas for improvement. The performance disparities across skin tones point to the need for more diverse and balanced training data, which could be addressed by future advances in synthetic image generation. Similarly, the ongoing challenges in achieving consistent and meaningful XAI outputs motivate further research into integrating explainability into the model training process. These improvements are essential for developing diagnostic systems that are both accurate and fair across all patient groups.

VI. Data and code availability

All data we used (HAM1000 and Fitzpatrick17k datasets), code with README files and saved models weights can be found on GoogleDrive.

References

[1] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 618–626. DOI: 10.1109/ICCV.2017.74.

[2] Kashif Rasul, Jonas Wassenberg, and Heiko Neumann. "Explainable deep learning in dermoscopy: A survey of techniques and evaluation". In: *International Conference on Medical Imaging with Deep Learning*. 2021.

[3] Newton M. Kinyanjui et al. "Fairness of Classifiers Across Skin Tones in Dermatology". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Ed. by Anne L. Martel et al. Cham: Springer International Publishing, 2020, pp. 320–329. ISBN: 978-3-030-59725-2.

[4] Roxana Daneshjou et al. "Predicting dermatological diagnoses and skin types using deep learning on a diverse dataset". In: *Nature Medicine* 26.6 (2020), pp. 900–908.

[5] Ira Ktena et al. "Generative models improve fairness of medical classifiers under distribution shifts". In: *Nature Medicine* 30.4 (2024), pp. 1166–1173. DOI: 10.1038/s41591-024-02838-6.

[6] John Orazio et al. "UV Radiation and the Skin". In: *International Journal of Molecular Sciences* 14.6 (2013), pp. 12222–12248. ISSN: 1422-0067. DOI: 10.3390/ijms140612222. URL: https://www.mdpi.com/1422-0067/14/6/12222.

[7] International Skin Imaging Collaboration. *ISIC 2020 Challenge on Skin Lesion Analysis Toward Melanoma Detection*. Available at https://challenge.isic-archive.com/. 2020.

[8] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". In: *Sci. Data* 5 (2018), p. 180161. DOI: 10.1038/sdata.2018.161.

[9] Matthew Groh et al. "Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1820–1828.

[10] Matthew Groh et al. "Towards transparency in dermatology image datasets with skin tone annotations by experts, crowds, and an algorithm". In: *Proceedings of the ACM on Human-Computer Interaction* 6.CSCW2 (2022), pp. 1–26.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 6840–6851.