DermDiff: Generative Diffusion Model for Mitigating Racial Biases in Dermatology Diagnosis

Nusrat Munia, Abdullah-Al-Zubaer Imran

Department of Computer Science
University of Kentucky, Lexington, KY, USA

Problem

Existing AI models for skin lesion diagnosis (such as melanoma detection) often struggle with **bias**, as they are trained on datasets that lack representation of **diverse skin tones**.

Contributions

- A generative diffusion framework to generate a diverse dataset, to *increase the diversity* of the training data and thereby, *reducing the risk of bias* in the dermatology AI system
- Extensive evaluation of the generated samples–*fidelity and diversity*, detection of skin tones–*algorithmically*, and disease diagnosis–*well-generalized*

Proposed Methods: DermDiff

- Leveraging latent diffusion model, DermDiff generates new samples of diverse skin tones conditioned on metadata-guided text prompts
- DermDiff automatically detects skin tones avoiding the new for manual, labor-intensive processes
- Combining real and synthetic data, DermDiff, once trained on diverse samples, effectively predicts bening vs malignant cases

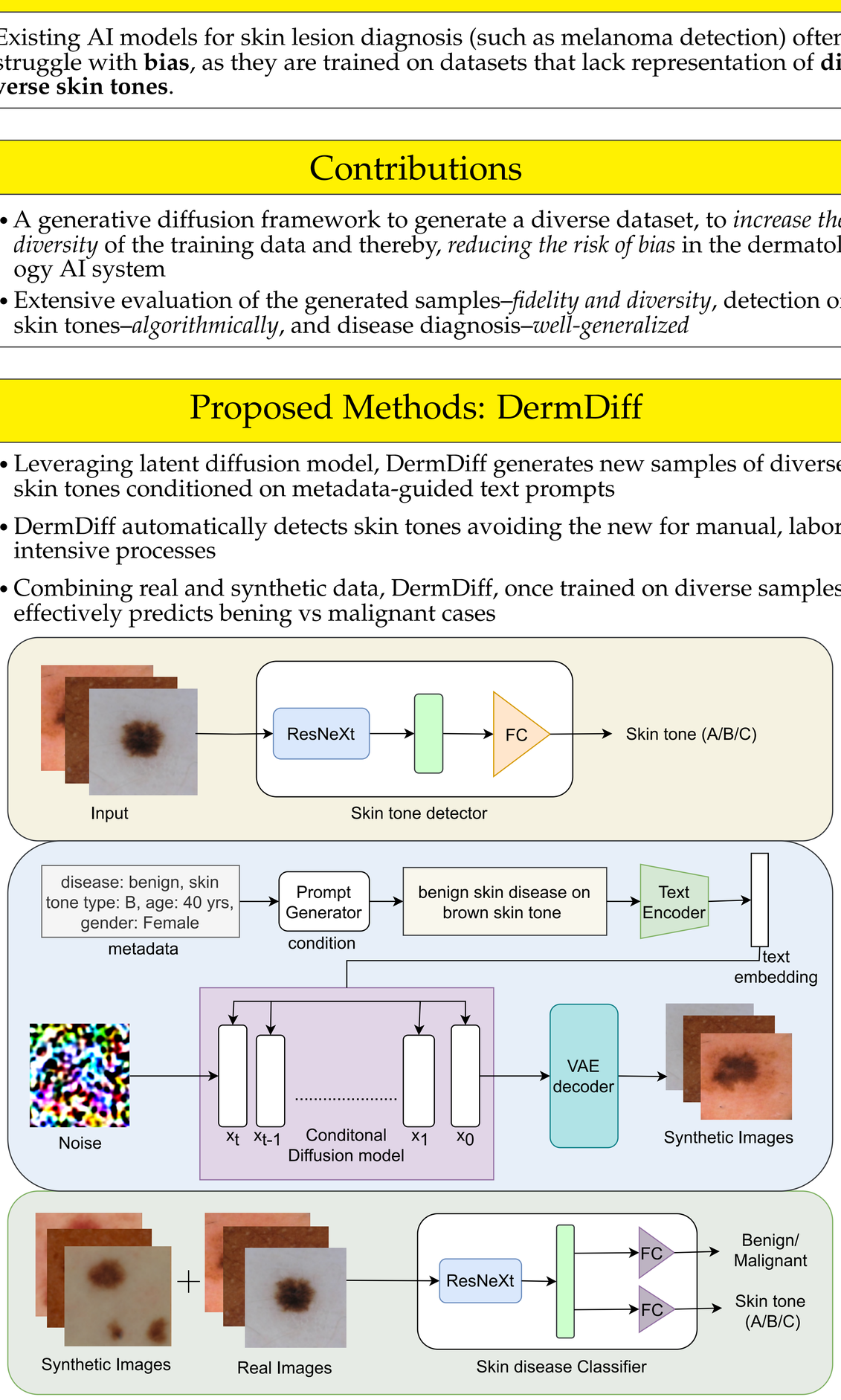

Figure 1: Proposed DermDiff framework to mitigate racial bias in dermatology diagnosis by generating new images of diverse skin tones.

Datasets

Table 1: Distribution of dermatology datasets used in our experiments.

Dataset	Size	Benign	Malignant	#Skin tones	Mode
Fitzpatrick	16,577	2,234	2,263	6	Train
ISIC	57,964	52,874	5,090	–	Train
DDI	656	485	171	3	Test

Table 2: Dermatology Skin types based on Fitzpatrick Skin Type (FST).

Acknowledgements

This work was funded by the United in True Racial Equity (UNITE) Research Priority Area at the University of Kentucky.

Results & Discussion

- DermDiff-generated samples are found to be of high quality and well-representative, both qualitatively and quantitatively (FID and MS-SSIM scores)
- Evaluation of skin tone detection and malignancy detection reveals the improved performances of DermDiff, even on a separate dataset (DDI)

Table 3: Quantitative comparison against the traditional Individual Typology Angle (ITA)-based skin tone detection on the DDI dataset.

Approach	Accuracy			F1-score		
	A	B	C	A	B	C
ITA	**0.89**	0.14	0.18	0.51	0.22	0.27
Fitzpatrick	0.84	**0.62**	0.40	**0.82**	**0.63**	0.44
Synthetic	0.58	0.51	**0.57**	0.55	0.50	0.61
Fitzpatrick+Synthetic	0.67	0.60	0.50	0.62	0.54	**0.64**

Table 4: Evaluation of the generated samples in terms of fidelity (FID) and diversity (MS-SSIM) across benign and malignant cases.

Metric	Compared Dataset	All (60,000)	Benign (30,000)	Malignant (30,000)
FID	ISIC + Fitzpatrick	86.53	25.77	117.01
	Fitzpatrick	20.34	79.14	15.35
MS-SSIM	–	0.35	0.46	0.27

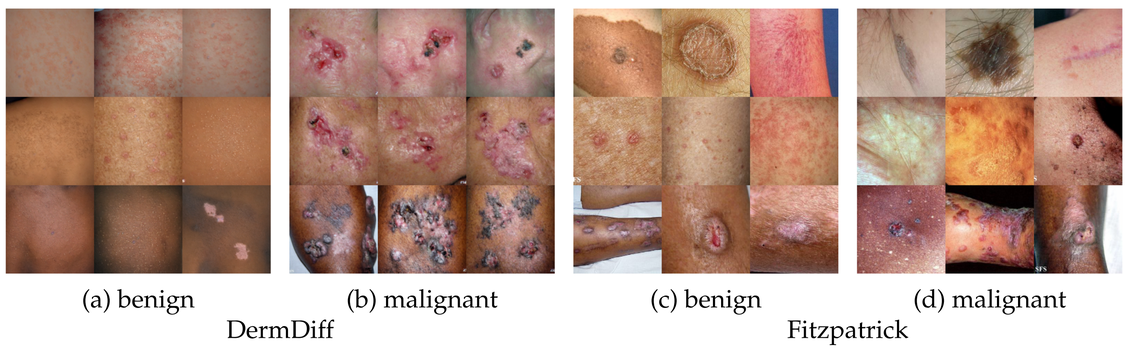

Figure 2: Visual comparison of the dermoscopic images in the Fitzpatrick17k dataset with the DermDiff-generated samples.

Table 5: Diagnostic performance on DDI dataset when the model was trained on (a) Fitzpatrick, (b) ISIC, and (c) combined Fitz+ISIC and synthetic image samples.

Training data	Accuracy			F1-score			AUC		
	A	B	C	A	B	C	A	B	C
Fitzpatrick	0.63	0.61	0.56	0.59	0.60	0.49	0.70	0.71	0.51
ISIC	**0.76**	0.69	**0.76**	0.43	0.41	0.43	0.58	0.67	0.52
Real+Synthetic	0.71	**0.77**	0.68	**0.63**	**0.72**	**0.52**	**0.72**	**0.78**	**0.58**

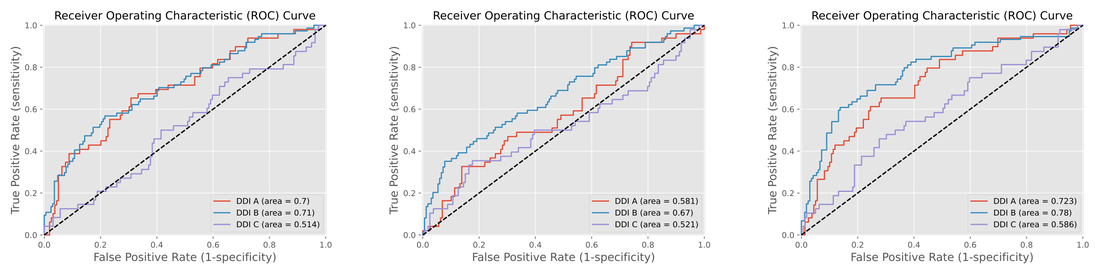

Figure 3: DDI ROC comparisons for training on Fitz (left), ISIC (middle), and Fitz+ISIC+Synthetic images.

Conclusion and Future Work

- DermDiff can generate diverse dermoscopic images based on skin tones and disease statuses
- Newly generated images, once combined with real ones, improve performances of downstream tasks such as disease classification and skin tone detection
- Our ongoing work focuses on better prompting the diffusion model with additional attribute information
- We also plan to evaluate the generated samples by validating on other downstream tasks (e.g., lesion segmentation)

References

[1] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on CVPR. (2022) 10684–10695

[2] Fitzpatrick, T.B.: The validity and practicality of sun-reactive skin types i through vi. Archives of dermatology **124**(6) (1988) 869–871

[3] Kinyanjui, N.M., Odonga, T., Cintas, C., Codella, N.C., Panda, R., Sattigeri, P., Varshney, K.R.: Estimating skin tone and effects on classification performance in dermatology datasets. preprint arXiv:1910.13268 (2019)