Identifying Bias in Skin Lesion Classification and Segmentation

Armin Hadzic[1] $PhilipMathew$[1] $JohnAucott$[2]
Philippe Burlina[1]
[1]$TheJohnsHopkinsUniversityAppliedPhysicsLaboratory$
[2]$JohnsHopkinsUniversitySchoolofMedicine$
11100 Johns Hopkins Road Laurel, Maryland 20723
arminhadzic@outlook.com

Abstract

Skin diseases are of growing concern in the United States and correctly diagnosing the disease from a lesion can be difficult for clinicians. Introducing automated skin lesion segmentation and classification models to clinical diagnoses to help combat some of the uncertainty in a diagnosis. We propose an automated approach for semantic segmentation and classification of Erythema Migrans, Herpes Zoster, and Tinea Corporis. Our model demonstrates successful segmentation on a new skin disease segmentation dataset with a 0.7053 Jaccard score and 0.8225 Dice score and illustrates the existence of skin tone bias in segmentation models. We apply these lesion segmentation masks to the task of reducing skin tone bias in skin disease classification and show a 12.32% reduction in accuracy gap between dark and light skin tone examples on a benchmark dataset. Further, we explore bias reduction through adversarial debias resulting in a 3.84% improvement in performance for the underrepresented subpopulation.

1. Introduction

In 2008, an estimated 288,000 people in the United States were infected by Lyme disease, with 2.4 million people being tested for the disease [11]. The number of cases has grown as the CDC estimated 300,000 cases annually in 2013 [13] and then refined their estimates to nearly 476,000 cases per year during 2010-2018 [14]. With the expansion of Lyme disease in the United States, diagnosis and testing of Lyme disease is of growing importance. Lyme disease is caused by a tick bite which infects a person with the bacteria *Borrelia burgdorferi*. 70-80% of the time a red oval lesion, called Erythema Migrans (EM), forms on the skin at the site of the bite [18]. Diagnosis of EM can be challenging as only 20% of United States patients have the traditional bull's eye lesion [23] and can be mistaken for other diseases such as Herpes Zoster (HZ) [16].

Improving the accuracy and speed of skin disease diagnosis is essential to ensuring diseases are treated quickly and, in the case of EM, do not develop into later stages of the disease. The medical field has been turning to Artificial Intelligence (AI), and often deep convolutional neural networks (CNN), to assist clinicians in timely and accurate diagnosis [6, 8]. CNNs have shown success in medical image classification and segmentation due to their ability to learn image representations from data which, traditionally, were poorly captured by hand-crafted features. However, CNNs have also been shown to exhibit bias inherent in training data [10], which is of growing concern when CNNs are applied to medical applications. In response, the AI community has been investigating bias mitigation strategies like data generation for underrepresented subpopulations [19] and adversarial debias [26]. While applying CNNs to dermatology is of growing interest, little work has been done on identifying and reducing the prevalence of bias in CNN prediction for skin disease classification and segmentation.

In this work, we address the challenge of identifying and quantifying skin disease segmentation bias. We present a new skin disease segmentation dataset with labeled skin diseases (EM, HZ, and Tinea Corporis (TC)) and annotated skin tones (light skin and dark skin tones). We train and evaluate a segmentation network on this dataset and identify bias in semantic segmentation with respect to skin tones. We then examine the problem of bias in skin disease classification by augmenting the [3] skin disease dataset with Individual Topology Angle (ITA) as a proxy for skin tone labels. This skin disease classification dataset is then used to illustrate the bias reduction benefits of using skin lesion masks constructed by the segmentation network, in addition to adversarial debias approaches.

2. Related Work

We provide an overview of prior work in skin disease classification, medical image segmentation, and bias mitigation methods in the domain of medical imaging.

2.1. Skin Disease Classification and Segmentation

Deep CNNs have gained popularity for automated melenoma skin lesion segmentation due to their promising performance despite the prevalence of fuzzy borders, inconsistent lighting conditions, and image artifacts [2, 27]. Recently, ITA has been used as a proxy for skin tone labels in medical imagery for segmentation and classification tasks. Little bias was found in skin disease segmentation and classification models using the SD-136 [22] and ISIC2018 [5] datasets [12]. Using publicly accessible real-world imagery from [3] and our own dataset we were able to find that were these datasets biased towards lighter skin tones as reflected by ITA and, in the case of our segmentation dataset, confirmed by hand annotated skin tone labels. Our models were able to pick up on this bias for classification and segmentation tasks, resulting in a need for strategies to mitigate this bias.

2.2. Bias Mitigation

Addressing bias in deep learning models can be categorized as fitting into the following three categories [4]: (1) pre-processing such as sampling strategies and masking sensitive factors, (2) in-processing like adversarial debias, and (3) post processing such as thresholding. In this work we examine pre-processing via masking sensitive factors and in-processing with adversarial debias. Masking sensitive factors in imagery has shown to improve fairness in object detection and action recognition [24]. Adversarial debiasing operates on the principle of simultaneously training two networks with different objectives [7, 15, 21]. The competing two-player optimization paradigm was then applied to the task of improving model fairness by maximizing equality of opportunity [1], a less constrained variation on equality of odds. This debiasing technique has shown success in tabular data [26], word embeddings, and imagery [28].

3. Datasets

We present a skin disease segmentation dataset for the task of semantic segmentation. As detailed in Table 1, our dataset includes a training, validation, and test split composed of 113, 30, and 42 samples respectively. Each split includes images of diseased skin that was annotated by a medical technician and supervised by a clinician. The 185 images, sourced from publicly available sources, depict examples of the skin diseases: EM, HZ, and TC. No examples of normal non-diseased skin (NO) were included in this dataset. Segmentation masks identify three possible regions: background, skin, and lesions.

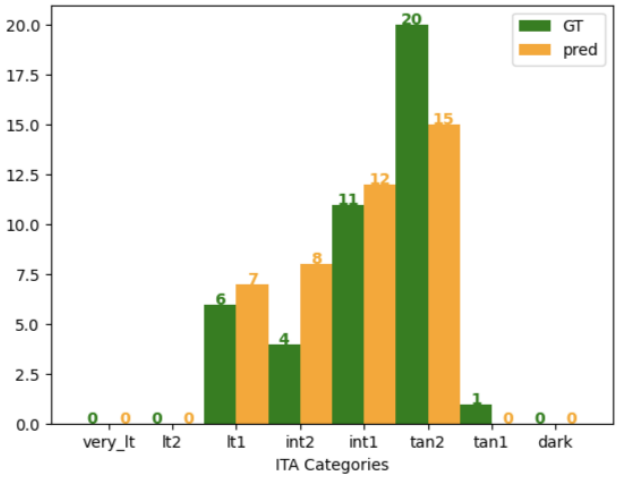

Figure 1: Segmentation dataset predicted (pred) and known (GT) test ITA distribution.

We incorporate the Individual Typology Angle (ITA)[25] as a proxy for skin tone color. ITA can be computed directly from an image and is correlated to the Fitzpatrick Skin Type, which is typically used to characterizing skin tone color in dermatology. ITA is calculated using CIELab color space and is calculated as follows:

$$ITA = \frac{180}{\pi} \arctan\left(\frac{L-50}{b}\right). \quad (1)$$

ITA uses the image luminance L (level of gray from black to white), and the blue to yellow balance b in relation to the intensity of pigmentation.

Additionally, for the task of classifying skin diseases we used the skin disease classification from [3]. This dataset includes 2574 labeled training examples for normal skin, EM, HZ, and TC. We also used ITA to further categorize disease skin samples as being ls or ds, as further detailed in Table 2. Similar to the segmentation dataset this classification dataset has a majority of ls examples for all three disease types across each dataset split.

4. Methods

We present two methods of reducing skin tone bias in skin disease classification artificial neural network models: (1) masking regions of images that do not depict lesions, and (2) utilizing adversarial training to discourage the classifier from learning biased representations.

4.1. Segmentation Masks

We train a U-Net[20] semantic segmentation network to segment images of diseased skin into three categories: background, skin, and lesion.

Test Split	NO		EM		HZ		TC		Total		Total Samples
	ls	ds	ls	ds	ls	ds	ls	ds	ls	ds	
Train	0	0	35	3	34	4	28	9	97	16	113
Validation	0	0	5	5	5	5	5	5	15	15	30
Test	0	0	7	7	7	7	7	7	21	21	42

Table 1: Disease segmentation dataset splits, annotated skin tones, and skin disease types. Skin tones is annotated as dark skin tone (ds) or light skin tone (ls). Each split is also characterized by the number of normal skin (NO), Erythema Migrans (EM), Herpes Zoster (HZ), and Tinea Corporis (TC) samples.

Test Split	NO		EM		HZ		TC		Total		Total Samples
	ls	ds	ls	ds	ls	ds	ls	ds	ls	ds	
Train	700	57	578	41	522	75	517	84	2317	257	2574
Validation	36	2	31	10	35	6	26	5	128	23	151
Test	86	4	73	7	51	9	66	6	276	26	302

Table 2: Disease classification dataset splits, ITA-based skin tones, and skin disease types. Skin tones are grouped together if the ITA categories are tan2, tan1, or dark into a ds class, and for all the other ITA categories were grouped into a ls class. Each split is also characterized by the number of NO, EM, HZ, and TC samples.

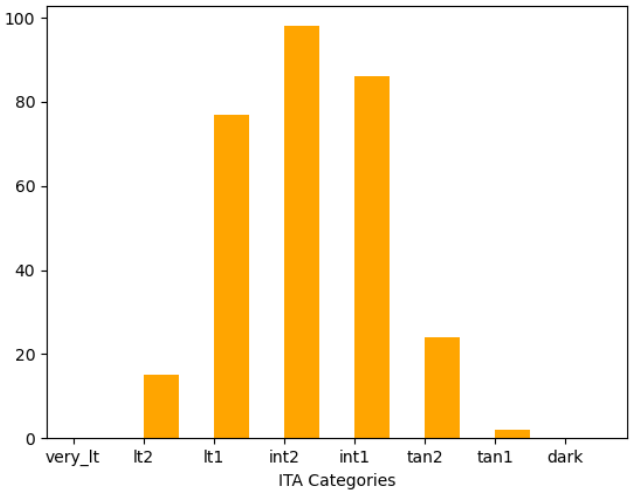

Figure 2: Classification dataset predicted test ITA distribution.

4.2. Adversarial Debias

While many prior works have used the formal definition of demographic parity from [17, 9] to express fairness, we favor the more strict equality of odds definition[17, 9], expressed below:

$$P(\hat{Y} = \hat{y}|S = s, Y = y) = P(\hat{Y} = \hat{y}|Y = y), \forall s, y, \hat{y}. \quad (2)$$

Equality of odds states that predictions $\hat{Y}$ are fair when the probability of producing said prediction given the true outcome Y, are conditionally independent of the protected factor S.

Following the work of [26], we use the adversarial debias technique to maximize fairness according to Equation 2. Adversarial debias uses a separate classifier (A with parameters Θ_A) to try to predict S using Y and the task prediction classifier's (F with parameters Θ_F) internal representation of a given input image X. The internal representation R is the task prediction classifier's output prior to applying *softmax*. The pair of classifiers are optimized consecutively using the total loss described as follows:

$$\min_{\Theta_F} \max_{\Theta_A} L(F(X;\Theta_F), Y) - \beta L(A(R, Y; \Theta_A), S). \quad (3)$$

The pair of categorical cross entropy losses ($L(\cdot;\cdot)$) are balanced using the hyperparameter $\beta \in 0, \cdots, 1.0$.

4.3. Metrics

The fairness of both classification and segmentation models are evaluated using a series of joint utility functions that balance the performance of said models and the gap between subpopulations of the protected factor. This joint utility function was first introduced by [19] for the case of classification accuracy, expressed as follows:

$$CAI_\alpha = \alpha(acc^b_{gap} - acc^d_{gap}) + (1-\alpha)(acc^d - acc^b). \quad (4)$$

The CAI_α uses α to balance the importance of a reduction in the accuracy gap and an improvement in accuracy between the baseline acc^b and debiased acc^d methods. Where the accuracy gap of the baseline acc^b_{gap} and debiased acc^d_{gap}

methods are computed as the difference between the accuracy of the subpopulations of S, which in this case is binary: light skin tone and dark skin tone. Moreover, [19] also introduced a variation of the fairness utility based on area under the curve AUC metric, as illustrated below:

$$CAUCI_\alpha = \alpha(AUC^b_{gap} - AUC^d_{gap}) + (1-\alpha)(AUC^d - AUC^b). \quad (5)$$

$CAUCI_\alpha$ operates in a similar manner to CAI_α where AUC gap (AUC^b_{gap} and AUC^d_{gap}) reduction importance is balanced with an increase in the change of AUC (AUC^d - AUC^b) for the baseline b and debiased d methods.

5. Results and Discussion

Developing methods for automated skin disease segmentation and classification models are essential components to improving the performance of medical diagnostics and monitoring treatment. We present and evaluate the performance of a skin disease segmentation network and then apply masks produced by that network to skin disease classification bias reduction. We then compare several methods of bias reduction with respect to skin tone, for skin disease classification models.

5.1. Skin Disease Segmentation

We evaluated the performance of our segmentation network on the skin disease segmentation dataset for the diseases EM, HZ, and TC. The segmentation network successfully performed semantic segmentation on background, skin, and lesions with a $Jaccard$ score of 0.7053 and $Dice$ score of 0.8225, as detailed in Table 3. We also evaluated the network on the light (ls) and dark (ds) skin tone subpopulations of the dataset, which revealed the network's bias towards ls subpopulations as performance was worse for the minimum performing (min subpop.) ds subpopulation both in terms of $Jaccard$ score (0.0809) and $Dice$ score (0.0743). These results have inspired us to investigate methods of reducing bias in segmentation networks in future work.

In Figure 3 we present test examples of our network's skin disease segmentations (Pred Mask) for ds in the left two columns and ls in the right two columns. When comparing each predicted mask with the medical technician annotated masks (GT Mask), the predicted mask matches well with the ground truth mask for both ls and ds samples. Overall, predicted segmentation masks for images of well-defined borders around lesions perform well. However, the quad of four images in the lower left corner of Figure 3 depict an interesting scenario where the lesion extends to the bottom center of the image, yet the segmentation network is unable to pick up on this subtlety. While this occurrence is a signal that further network improvements are needed to further improve performance, humans have also been known to have difficulties on subtle lesion skin tone changes. The lower right quad also illustrates unclear distinctions between healthy skin and a lesion where healthy skin is circumscribed within the outer lesion ring, which can result in the texture matching convolution layers of the segmentation network.

5.2. Disease Classification

We investigated the impact of skin segmentation masks or adversarial debias at reducing the bias in skin disease classification. Each method was evaluated on the skin disease classification dataset from [3], which contained labeled images of normal undiseased skin (NO) and skin with EM, HZ, and TC lesions. We first introduce our baseline method, then discuss masking and adversarial training debias approaches for reducing classification bias, and followed by dataset characteristics.

We initiated the experiments by training an off-the-shelf image classifier for skin disease classification. The baseline method, a ResNet30 image classifier, was shown to have a 3.52% accuracy and 0.0194 AUC improvement (see Table 4) over the [3] ResNet50 method. This performance improvement is likely attributed to the baseline's smaller network (ResNet30) overfitting less on the small training (2574 samples) set due to fewer trainable parameters. However, the high performing baseline also exhibited a 13.15% accuracy gap between ls and ds subpopulations. Next, we explored methods of reducing this accuracy gap using skin segmentation masks.

Next, we evaluated the benefits of automated skin disease segmentation masks for improving fairness, across skin tones, in skin disease classification. Skin segmentation masks have the potential to reduce bias in skin disease classification by masking out all regions of the image that do not contain lesions. Since the disease classification dataset did not include annotated skin segmentation masks we employed our trained segmentation network to generate skin and background masks for NO, EM, HZ, and TC skin images. Examples of predicted segmentation masks applied to images from the disease classification dataset are illustrated in Figure 3. While the segmentation network was never trained on NO skin, it was often able to correctly produce masks that covered all regions of the image, allowing a classifier to learn a correlation between fully masked image and NO skin. We applied these predicted background and skin masks to the imagery used to train the baseline image classifier and referred to this method as Image+Mask. According to the results in Table 4, the Image+Mask method successfully reduced the baseline method's accuracy gap down to 0.83% but did, unfortunately, also have a reduction in accuracy and AUC when compared to the baseline method. The lack of change in the minimum overall subpopulation accuracy between the Image+Mask and baseline

$Jaccard$	J_{gap}	$Dice$	D_{gap}	min subpop.
0.7053 (0.0035)	0.0809 (0.0001)	0.8225 (0.0029)	0.0743 (0.0002)	ds

Table 3: Segmentation Bias Test Results

Metrics	Baseline (Image Only)	Image+AD	Image+Mask	Image+Mask+AD	ResNet50 (Image Only) [3]
acc	**85.10 (4.02)**	81.79 (4.35)	73.84 (4.96)	72.52 (5.04)	81.58 (4.36)
acc_{gap}	13.15 (12.98)	5.33 (11.69)	**0.83 (11.87)**	4.82 (10.91)	-
acc_{min} (subpop.)	73.08 (ds)	**76.92 (ds)**	73.08 (ds)	72.10 (ls)	-
$CAI_{0.5}$	-	**2.2550**	0.5300	-2.1250	-
$CAI_{0.75}$	-	5.0375	**6.4250**	3.1025	-
AUC	**0.9725 (0.0185)**	0.9555 (0.0233)	0.9072 (0.0327)	0.9053 (0.0330)	0.9531 (0.0238)
AUC_{gap}	0.0331 (0.0714)	**0.0094 (0.0469)**	0.0275 (0.0618)	0.0414 (0.0536)	-
AUC_{min} (subpop.)	0.9420 (ds)	**0.9548 (ls)**	0.9050 (ls)	0.9023 (ls)	-
$CAUCI_{0.5}$	-	**0.0034**	-0.0299	-0.0378	-
$CAUCI_{0.75}$	-	**0.0135**	-0.0121	-0.0230	-

Table 4: Skin disease classification and associated bias. Sample types include normal skin (NO) and diseased skin: EM, HZ, and TC. Samples contained skin tones as a protected factor. Skin tones were classified as ls and ds based on ITA.

methods revealed that the Image+Mask method made no improvements to the ds subpopulation but merely degraded the performance of the ls subpopulation. Reducing the performance of the highest performing subpopulation without making any improvements to the worse performing subpopulation is not desirable since the model did not learn to perform better at any subpopulation. However, should an application warrant accuracy gap reduction as more important than accuracy or AUC improvement, then the Image+Mask method would be desirable as further illustrated by the best conjunctive accuracy improvement with $\alpha = 0.75$.

Investigating alternative methods for bias reduction, we transitioned from investigating input imagery modifications to alternative training procedures via adversarial debias (AD). AD training methods, introduced by [26], have been shown to reduce bias in classification problems [19]. We augmented the baseline image classifier with an adversarial module whose task is to try to predict the binary skin tone from the logits of the classifier network, we refer to this method as Image+AD. Table 4 shows that the Image+AD method was able to reduce the accuracy gap between the two skin tone subpopulations by 7.82% in addition to improving the classifier's accuracy and AUC on the subpopulation with the worst performance, ds, when compared with the baseline approach. Moreover, the fairness utility metrics $CAI_{0.5}$, $CAUCI_{0.5}$, and $CAUCI_{0.5}$ reflect that the AD method has the overall most fair performance when considering the trade off between performance improvement and gap reduction compared with the baseline method. We also explored masking non-lesion regions of images and training using AD, but found no benefit when compared to the Image+AD and Image+Mask methods.

We then examined the skin disease classification dataset characteristics and the best performing model's (Image+AD) performance on this in further detail. When constructing the skin segmentation dataset we found limited quantities of dark skin tone (ds) images with EM in the public domain. This disparity in ds EM images was also existed in disease classification dataset, outlined in Table 2, where only 1.59% training and 2.32% test images contained this combination. Given how few images were readily available, we were surprised to find that Image+AD model correctly classified all of the ds with EM test images, but did find misclassifications for light skin tone with EM test images. We suspect that an extended EM imagery dataset with more ds examples would further elucidate the capability of the Image+AD model at correctly classifying images of EM with ds. Qualitative examples of the Image+AD method's classification performance are depicted in Figure 4. We observed that most normal skin examples were classified correctly, regardless of skintone. Many of the correctly classified normal skin images were either digitally enhanced images or images of models with almost entirely clear skin, which resulted in smooth consistent textures that were more easily identifiable by the convolution layers of the convolutional neural network model. Incorrectly classified normal skin images tended to contain some discoloration on the skin as illustrated by the red line across the abdomen of the person in Figure 5 image (top row, second from the right). Overall, performance on some of these challenging scenarios could be further improved by increasing the volume of training imagery, in particular to ds examples.

6. Conclusion

We presented a study on identifying and quantifying the bias learned by a semantic segmentation model on a novel skin disease segmentation dataset. Our model successfully segmented EM, HZ, and TC lesions from skin and background on challenging real-world imagery with $Jaccard$ and $Dice$ scores of 0.7053 and 0.8225 respectively. We illustrated the skin tone bias reduction benefits of applying skin lesion masks, created by the segmentation model, for the task of skin lesion classification on a benchmark skin disease dataset [3] down to 0.83% acc_{gap} and improving fairness by 6.4250 $CAI_{0.75}$. Furthermore, we demonstrated the benefit of incorporating adversarial debias methods to improve performance of underrepresented populations while balancing overall classification accuracy. These advancements aid in moving CNNs closer towards addressing some of the challenges of operationalizing these models for clinical use.

In future work we intend to examine methods of reducing semantic segmentation bias with respect to a protected factor, such as skin tone. Existing methods of debiasing CNNs have been largely applied to the case of classification, but, as our work has illustrated, also need to be redesigned for the more challenging problem of segmentation.

References

[1] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017. 2

[2] Lei Bi, Jinman Kim, Euijoon Ahn, Ashnil Kumar, Michael Fulham, and Dagan Feng. Dermoscopic image segmentation via multistage fully convolutional networks. *IEEE Transactions on Biomedical Engineering*, 64(9):2065–2074, 2017. 2

[3] Philippe M Burlina, Neil J Joshi, Phil A Mathew, William Paul, Alison W Rebman, and John N Aucott. Ai-based detection of erythema migrans and disambiguation against other skin lesions. *Computers in Biology and Medicine*, 125:103977, 2020. 1, 2, 4, 5, 6

[4] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020. 2

[5] 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. 2

[6] Y Fujisawa, Y Otomo, Y Ogata, Y Nakamura, R Fujita, Y Ishitsuka, R Watanabe, N Okiyama, K Ohara, and M Fujimoto. Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumour diagnosis. *British Journal of Dermatology*, 180(2):373–381, 2019. 1

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2

[8] Yanyang Gu, Zongyuan Ge, C Paul Bonnington, and Jun Zhou. Progressive transfer learning and adversarial domain adaptation for cross-domain skin disease classification. *IEEE journal of biomedical and health informatics*, 24(5):1379–1393, 2019. 1

[9] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016. 3

[10] Katherine Hermann, Ting Chen, and Simon Kornblith. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33, 2020. 1

[11] Alison F. Hinckley, Neeta P. Connally, James I. Meek, Barbara J. Johnson, Melissa M. Kemperman, Katherine A. Feldman, Jennifer L. White, and Paul S. Mead. Lyme Disease Testing by Large Commercial Laboratories in the United States. *Clinical Infectious Diseases*, 59(5):676–681, 05 2014. 1

[12] Newton M Kinyanjui, Timothy Odonga, Celia Cintas, Noel CF Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R Varshney. Fairness of classifiers across skin tones in dermatology. In *International Conference on Medical Im-*

age Computing and Computer-Assisted Intervention, pages 320–329. Springer, 2020. 2

[13] Bridget M Kuehn. Cdc estimates 300 000 us cases of lyme disease annually. *Jama*, 310(11):1110–1110, 2013. 1

[14] Kiersten J Kugeler, Amy M Schwartz, Mark J Delorey, Paul S Mead, and Alison F Hinckley. Estimating the frequency of lyme disease diagnoses, united states, 2010–2018. *Emerging Infectious Diseases*, 27(2):616, 2021. 1

[15] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2

[16] Daniel R Mazori, Charisse M Orme, Adnan Mir, Shane A Meehan, and Andrea L Neimann. Vesicular erythema migrans: an atypical and easily misdiagnosed form of lyme disease. *Dermatology online journal*, 21(8), 2015. 1

[17] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019. 3

[18] Robert B. Nadelman. Erythema migrans. *Infectious Disease Clinics of North America*, 29(2):211–239, 2015. 1

[19] William Paul, Armin Hadzic, Neil Joshi, Fady Alajaji, and Phil Burlina. Tara: Training and representation alteration for ai fairness and domain generalization. *arXiv preprint arXiv:2012.06387*, 2020. 1, 3, 4, 5

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[21] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32:3358–3369, 2019. 2

[22] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A benchmark for automatic visual classification of clinical skin disease images. In *European Conference on Computer Vision*, pages 206–222. Springer, 2016. 2

[23] Carrie D. Tibbles and Jonathan A. Edlow. Does This Patient Have Erythema Migrans? *JAMA*, 297(23):2617–2627, 06 2007. 1

[24] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[25] Marcus Wilkes, Caradee Y Wright, Johan L du Plessis, and Anthony Reeder. Fitzpatrick skin type, individual typology angle, and melanin index in an african population: steps toward universally applicable skin photosensitivity assessments. *JAMA dermatology*, 151(8):902–903, 2015. 2

[26] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. 1, 2, 3, 5

[27] Lei Zhang, Guang Yang, and Xujiong Ye. Automatic skin lesion segmentation by coupling deep fully convolutional networks and shallow network with textons. *Journal of Medical Imaging*, 6(2):024001, 2019. 2

[28] Yi Zhang and Jitao Sang. Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4346–4354, 2020. 2

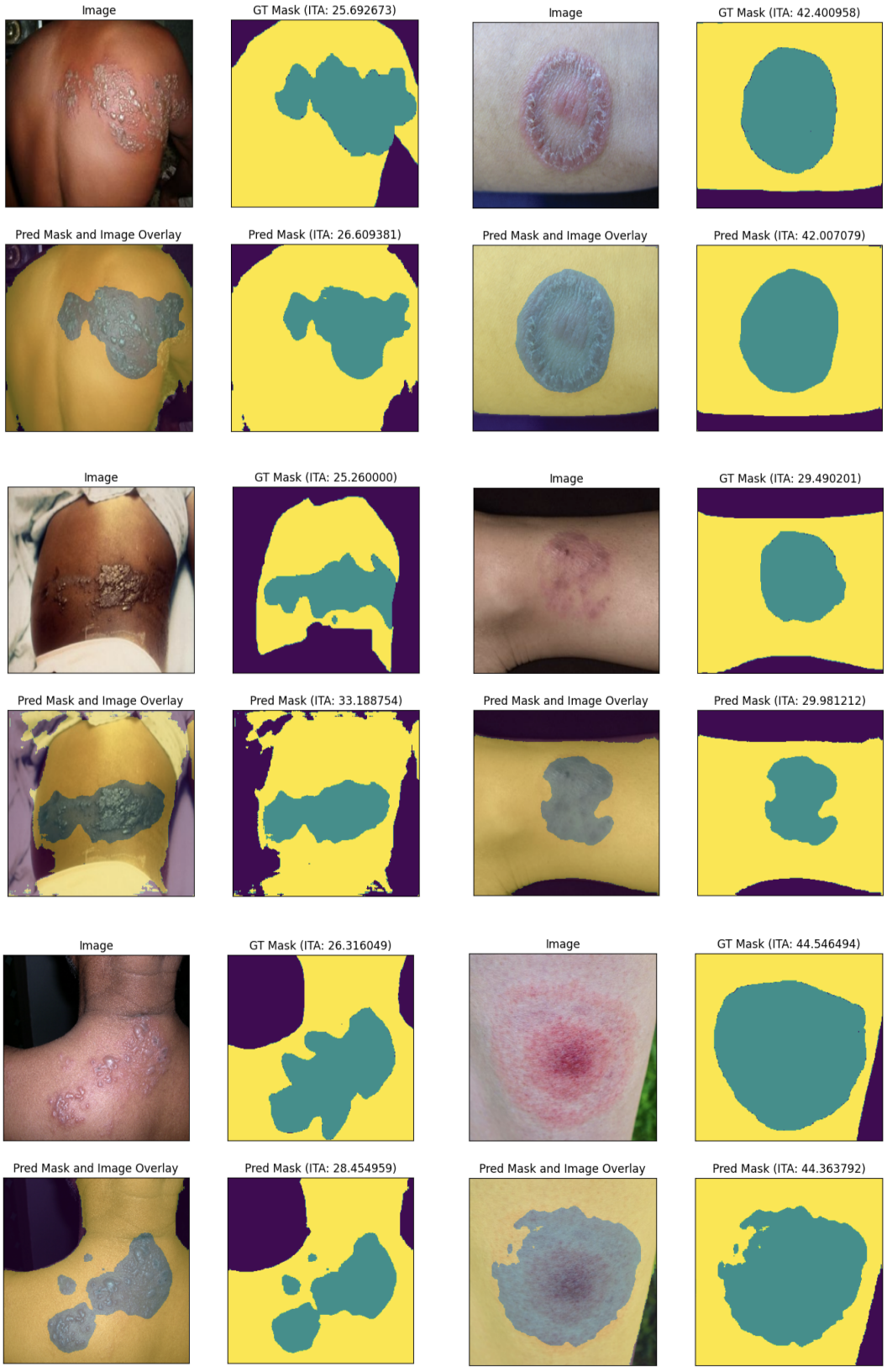

Figure 3: Segmentation examples and corresponding ITA. The segmentation masks show background (purple), skin (yellow), and lesions (blue).

Figure 4: Examples of Image+AD model's disease classification predictions (Pred) compared with the known disease (GT) for NO, EM, HZ, and TC.

Figure 5: Disease classification predictions for the Image+AD model where the model performed poorly. Predictions (Pred) are compared with the known disease (GT) for NO, EM, HZ, and TC.

Figure 6: Disease classification predictions and corresponding masks for the Image+Mask model. Predictions (Pred) are compared with the known disease (GT) NO, EM, HZ, and TC.