EdgeMixup: Embarrassingly Simple Data Alteration to Improve Lyme Disease Lesion Segmentation and Diagnosis Fairness

Haolin Yuan[§], John Aucott[‡], Armin Hadzic[†], William Paul[†], Marcia Villegas de Flores[§], Philip Mathew[§†], Philippe Burlina[†], Yinzhi Cao[§]

[§]Johns Hopkins University
[†]Johns Hopkins Applied Physics Lab, [‡]Johns Hopkins University School of Medicine
{hyuan4, mvilleg5, yinzhi.cao}@jhu.edu,
{william.paul, philip.mathew, Philippe.Burlina}@jhuapl.edu
armin.hadzic.scholar@gmail.com, jaucott2@jhmi.edu

Abstract. Lyme disease is a severe skin disease caused by tick bites, which affects hundreds of thousands of people. One task in diagnosing Lyme disease is lesion segmentation, i.e., separating benign skin from lesions, which can not only help clinicians to focus on lesions but also improve downstream tasks such as disease classification. However, it is challenging to segment Lyme disease lesions due to the lack of well-segmented, labeled Lyme datasets and the nature of Lyme, e.g., the typical bull's eye lesion and its closeness to normal skin. In this paper, we design a simple yet novel data preprocessing and alteration method, called EdgeMixup, to help segment Lyme lesions on imbalanced training datasets. The key insight is to deploy a linear combination of lesion edge, either detected or computed, and the source image highlights the affected lesion area so that a learning model focuses more on the preserved lesion structure instead of skin tone, thus iteratively improving segmentation performance. Additionally, the improved edge from lesion segmentation can be further used for Lyme disease classification—e.g., in differentiating Lyme from other similar lesions including tinea corporis and herpes zoster—with improved model fairness on different subpopulations.

1 Introduction

Medical Image Analysis has greatly benefited from advances in AI [1] yet some improvements still remain to be addressed, importantly in areas that allow both algorithmic performance and fairness [2], and in certain medical applications that promise to significantly lessen morbidity and mortality. Early detection of skin lesions is such an endeavor as it can aid in identifying infectious diseases with cutaneous manifestations. Lyme disease is an example of that with a potentially diagnostic skin lesion [3]—which is caused by the bacterium *Borrelia burgdorferi* and leads to nearly 476,000 cases per annum during 2010–2018 [4]. The earliest and most treatable phase of Lyme disease is manifested via a red concentric lesion at the site of a tick bite, called erythema migrans (EM) [5]. While the

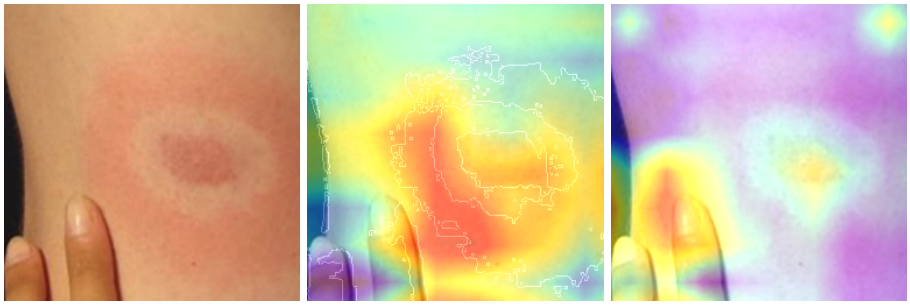
EM pattern may appear simple to recognize, its diagnosis can be challenging for those with or without a medical background alike, as only 20% of United States patients have the stereotypical bull's eye lesion [6]. When skin lesions are atypical they can be mistaken for other diseases such as tinea corporis (TC) or herpes zoster (HZ), two other diseases acting as confusers for Lyme, considered herein. This has increased interest in medical applications of deep learning (DL), and using deep convolutional neural networks (CNNs), to assist clinicians in timely and accurate diagnosis of conditions including Lyme disease, TC and HZ [7–9].

One important diagnosis task is to segment Lyme lesion, particularly the EM pattern, from benign skins. Such DL-assisted segmentation not only helps clinicians in pre-screening patients but also improves downstream tasks such as lesion classification. However, while Lyme disease lesion segmentation is intuitively simple, it is challenging due to the following reasons. First, there lacks of a well-segmented dataset with manual labels on Lyme disease. On one hand, some datasets—such as HAM10000 [10] and ISBI Challenges [11]—have manual annotated segmentations for diseases like melanoma, but they do not have Lyme disease lesions. On the other hand, some datasets—such as Groh et al. [12]—have Lyme disease and skin tone and classification labels, but not segmentation.

Second, the segmentation of Lyme lesion is itself challenging due to the nature of EM pattern. Specifically, a typical Lyme lesion exhibits a bull's eye pattern with one central redness and one outer circle, which is different from darkness lesion in cancer-related skin disease like melanoma. Furthermore, clinical data collected for training is usually imbalanced in some properties, e.g., more samples with light skins compared with dark skins. Therefore, existing skin disease segmentation [13] as well as existing general segmentation works, such as U-Net [14], polar training [15], ViT-Adapter [16], and MFSNet [17], usually suffer from relatively low performance and reduced fairness [2, 18, 19].

In this paper, we present the first Lyme disease dataset that contains labeled segmentation and skin tones. Our Lyme disease dataset contains two parts: (i) a classification dataset, composed of more than 3,000 diseased skin images that are either obtained from public resources or clinicians with patient-informed consent, and (ii) a segmentation dataset containing 185 samples that are manually annotated for three regions—i.e., background, skin (light vs. dark), and lesion—conducted under clinician supervision and Institutional Review Boards (IRB) approval. Our dataset with manual labels is available at this URL [20].

Secondly, we design a simple yet novel data preprocessing and alternation method, called EdgeMixup, to improve Lyme disease segmentation and diagnosis fairness on samples with different skin-tones. The key insight is to alter a skin image with a linear combination of the source image and a detected lesion boundary so that the lesion structure is preserved while minimizing skin tone information. Such an improvement is an *iterative* process that gradually improves lesion edge detection and segmentation fairness until convergence. Then, the detected, converged edge in the first step also helps classification of Lyme diseases via mixup with improved fairness. Our source code is available at this URL [20].



(a) Original image (b) Heat map for EdgeMixup (c) Heat map for legacy analysis

Fig. 1: A motivating example to illustrate why EdgeMixup improves model performance and reduces biases via mixing up lesion boundary with original image (Heatmap is generated via Grad-CAM).

We evaluate EdgeMixup for skin disease segmentation and classification tasks. Our results show that EdgeMixup is able to increase segmentation utility and improve fairness. We also show that the improved segmentation further improves classification fairness as well as joint fairness-utility metrics compared to existing debiasing methods, e.g., AD [21] and ST-Debias [22].

2 Motivation

In this section, we motivate the design of EdgeMixup by showing that added lesion boundary helps a DL model focus more on the lesion part instead of other features such as skin or background. Note that not all skin disease datasets are carefully processed either due to the large amount of work required or the scarcity of data samples collected, e.g., SD-198 [23] contains samples that are taken under variant environments. Specifically, we train two ResNet-34 models using the same dataset with and without EdgeMixup for a classification task of skin disease. We keep all hyper-parameters exactly the same for two models, and only augment the same image with and without mixing lesion boundary up with the original image. We generate initial lesion edges using EdgeMixup, which we will elaborate in following sections. Figure 1 shows the original image (Figure 1a) as well as two models' attention as heat-maps where red color represents the highest attention, yellow a higher attention, and purple the least attention. EdgeMixup helps the model to focus more on the lesion area comparing Figure 1b and 1c. The reason is that a legacy diagnosis has no information about lesion and does not know where to locate its focus, thus easily gets distracted by fingers instead of the lesion pattern.

3 Method

In this section, we first give the definition for model fairness, and we then describe the design of EdgeMixup for the purpose of de-biasing in Figure 2 and

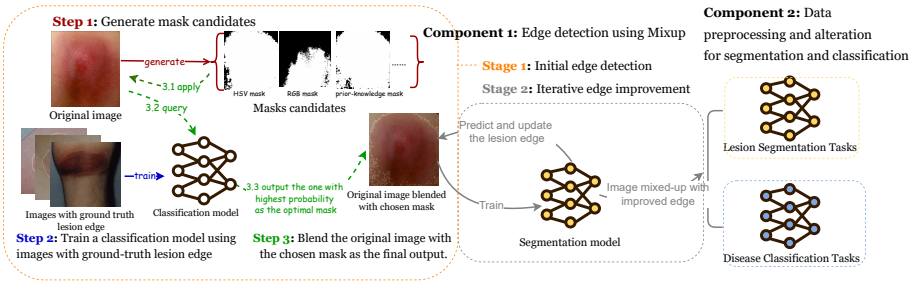


Fig. 2: EdgeMixup Process

Algorithm 1. We consider any model f, either a classification model f_{class} or a segmentation model f_{seg}, to be biased against certain skin-tone st_2 if given metrics M and samples x_{st_1} and x_{st_2} from class y, where st_1 and st_2 are different skin-tones according to their ITA scores, $M(f(x_{st_1}), y) > M(f(x_{st_2}), y)$. If there exists a model f such that $M(f(x_{st_1}), y) = M(f(x_{st_2}, y))$, we consider it perfectly fair for st_1 and st_2 skin-tone samples.

EdgeMixup improves model fairness on light and dark skin samples in both segmentation and classification tasks, and it has two major components: (i) edge detection using mixup, and (ii) data preprocessing and alteration for downstream tasks. More specifically, our proposed edge detection has two parts: initial edge detection and iterative improvement.

Initial Edge Detection: The purpose of initial detection, which is documented in the `Initial_edge_detection` function of Algorithm 1, is to provide a starting point, i.e., a rough boundary, for the next step of iterative improvement. The high-level idea is that EdgeMixup detects several edge candidates using the color range of ground-truth lesions in both Red-Green-Blue (RGB) and Hue-Saturation-Value (HSV) color space and then selects the target edge using a learning model based on the output confidence score. First, EdgeMixup trains a classification model based on a mixup of the ground-truth segmentation under clinician supervision and the original image (Line 7). Second, EdgeMixup generates many edge candidates. For example, EdgeMixup collects the mean range of lesion color from the training set and use the range as threshold to filter out any given sample for a candidate mask (Line 9). Lastly, EdgeMixup selects an edge candidate with the highest confidence score output by the learning model (Line 11) and returns it as the edge for this given sample. Note that the initial edge detection is irrelevant to the sample size of a particular subpopulation, thus improving the fairness. That is, even if the original dataset is imbalanced, as long as one sample from a subpopulation exists, the color range of the sample's lesion is considered in the initial detection.

Iterative Edge Improvement: EdgeMixup includes iterative edge improvement in the training phase of our segmentation model to further improve model utility. The intuitive reason of utilizing such algorithm is that by applying the

Algorithm 1 Pseudo-code of EdgeMixup

Input: A labelled sample $(x, y) \in D$, mixup weights α, ground-truth edged training set $D^{train}_{edge\_gt}$
Output: dataset $D_{\texttt{final\_edge}}$ in which each sample has it lesion edge highlighted $(x_{\texttt{edge}}, y)$

function `main`( )
  $D_{\texttt{initial\_edge}}$ = `Initial_edge_detection`(D, α)
  $D_{\texttt{final\_edge}}$ = `Iterative_edge_improvement`$(D_{\texttt{initial\_edge}}, \alpha)$
  return $D_{\texttt{final\_edge}}$
end function
function `Initial_edge_detection`(D, α)
  Train classification model $m_{\texttt{class}}$ using $D^{\texttt{train}}_{\texttt{edge\_gt}}$
  for each sample $x \in D$ **do**
    Get all edge candidates $\{\texttt{edge}_1, \texttt{edge}_2, .., \texttt{edge}_n\}$ for each sample x
    Mixup each edge candidate with x
    Query $m_{\texttt{class}}$ using all mixed-up $\{x_{\texttt{edge}_1}, ...x_{\texttt{edge}_n}\}$ and choose the optimal edge $\texttt{edge}_{\texttt{opt}}$
    Generate edged sample $x_{\texttt{edge}}$ = `Mixup`$(x, \texttt{edge}_{\texttt{opt}}, \alpha)$
  end for
  return $D_{\texttt{edge}}$
end function
function `Iterative_edge_improvement`$(D_{\texttt{edge}}, \alpha)$
  Train the first model $m_{\texttt{iter}}$ using edged dataset $D^{\texttt{train}}_{\texttt{edge}}$
  Evaluate $m_{\texttt{iter}}$ using $D^{\texttt{test}}_{\texttt{edge}}$ and get `current_Jaccard`
  `best_Jaccard` = 0
  `iter` = 1
  while `current_Jaccard` > `best_Jaccard` **do**
    `best_Jaccard` = `current_Jaccard`
    Predict lesion masks using $m_{\texttt{iter}}$, convert them to lesion edge *edge*
    Generate new training set for next model `Mixup`$(D_{\texttt{train}}, \texttt{edge}, \alpha)$
    Train a model for next iteration $m_{\texttt{iter}+1}$
    Evaluate $m_{\texttt{iter}+1}$ using edged $D^{\texttt{test}}_{\texttt{edge}}$ and get `current_Jaccard`
    `iter` += 1
  end while
end function
function `Mixup`$(x, edge, \alpha)$
  return $(\alpha \cdot x + (1 - \alpha) \cdot edge)$
end function

(a) iteration 1 (loss= 0.15) (b) iteration 2 (loss = 0.13) (c) iteration 3 (loss = 0.08)

Fig. 3: Illustration of iterative edge improvement on different iterations with train loss

mixup of detected edge and original image, given the lesion boundary feature detected in the previous iteration, the next-iteration segmentation model can converge better and the lesion boundary predicted by it is fine-grained. Specifically, EdgeMixup iteratively trains segmentation models from scratch, and we let the model trained in the previous iteration to predict lesion edge, which is then mixed-up with original training samples as the new training set for next-iteration model. The high-level idea is that when the lesion is restricted in a small

Table 1: Annotated segmentation and classification dataset characteristics, broken down by ITA-based skin tones (light skin/ dark skin) and disease types.

Split	Skin					SD-sub					
	NO	EM	HZ	TC	Total	DF	KA	PG	TC	TF	Total
seg	-	62	62	61	185	30	30	30	30	30	150
	-	47/15	46/16	40/21	133/52	23/7	27/3	27/3	24/6	29/1	130/20
class	885	740	698	704	3027	40	40	40	40	40	200
	822/63	682/58	608/90	609/95	2721/306	36/4	36/4	29/1	33/7	30/0	164/16

affected area, further detection will refine and constrain the detected boundary. Besides, EdgeMixup calculates a linear combination of original image and lesion boundary, i.e., by assigning the weight of original image as α and lesion boundary as $1 - \alpha$. Figure 3 shows the edge-mixed-up images for different iterations. EdgeMixup removes more skin areas after each iteration and gradually gets close to the real lesion at the third iteration.

4 Datasets

We present two datasets: (i) a dataset collected and annotated by us (called Skin), and (ii) a subset of SD-198 [23] with our annotation (called SD-sub). First, We collect and annotate a dataset with 3,027 images containing three types of disease/lesions, i.e., Tinea Corporis (TC), Herpes Zoster (HZ), and Erythema Migrans (EM). All skin images are either collected from publicly available sources or from clinicians with patient informed consent. Then, a medical technician and a clinician in our team manually annotate each image. For the segmentation task, we annotate skin images into three classes: background, skin, and lesion; then, for the classification task, we annotate skin images by classifying them into four classes: No Disease (NO), TC, HZ, and EM. We name it as Skin-class for later reference. Second, we select five classes from SD-198 [23], a benchmark dataset for skin disease classification, as another dataset for both segmentation and classification tasks. Note that due to the amount of manual work involved in annotation, we select those classes based on the number of samples in each class. The selected classes are Dermatofibroma (DF), Keratoacanthoma (KA), Pyogenic Granuloma (PG), Tinea Corporis (TC), and Tinea Faciale (TF). We choose 30 samples in each class for segmentation task, and we split them into 0.7, 0.1, and 0.2 ratio for training, validation, and testing, respectively.

Table 1 show the characteristics of these two datasets for both classification and segmentation tasks broken down by the disease type and skin tone, as calculated by the Individual Typology Angle (ITA) [24]. Specifically, we consider tan2, tan1, and dark as dark skin (ds) and others as light skin (ls). Compared to other skin tone classification schemas such as Fitzpartick scale [25], we divide ITA scores into more detailed categories (eight). One prominent observation is

Table 2: Segmentation: Performance and Fairness (margin of error reported in parenthesis)

	Method	Unet	Polar	MFSNet	ViT-Adapter	**EdgeMixup**
Skin	`Jaccard`	0.7053(0.0035)	0.7126(0.0033)	0.5877(0.0080)	0.7027(0.0057)	**0.7807(0.0031)**
	J_{gap}	0.0809(0.0001)	0.0813(0.0001)	0.1291(0.0076)	0.2346(0.0035)	**0.0379(0.0001)**
SD-seg	`Jaccard`	0.7134(0.0031)	0.6527(0.0036)	0.6170(0.0052)	0.5088(0.0042)	**0.7799(0.0031)**
	J_{gap}	0.0753(0.0001)	0.1210(0.0003)	0.0636(0.0033)	0.2530(0.0021)	**0.0528(0.0001)**

that ls images are more abundant than ds images due to a disparity in the availability of ds imagery found from either public sources or from clinicians with patient consent.

5 Evaluation

We implement EdgeMixup using python 3.8 and Pytorch , and all experiments are performed using one GeForce RTX 3090 graphics card (NVIDIA).

Segmentation Evaluation. Our segmentation evaluation adopts four baselines, (i) a U-Net trained to segment skin lesions, (ii) a polar training [15] transforming images from Cartesian coordinates to polar coordinates, (iii) ViT-Adapter [16], a state-of-the-art semantic segmentation using a fine-tuned ViT model, (iv) MFSNet [17], a segmentation model with differently scaled feature maps to compute the final segmentation mask. We follow the default setting from each paper for evaluation. Our evaluation metrics include (i) `Jaccard` index (IoU score), which measures the similarity between a predicted mask and the manually annotated ground truth, and (ii) the gap between `Jaccard` values (J_{gap}) to measure fairness.

Table 2 shows the performance and fairness of EdgeMixup and different baselines. We compare predicted masks with the manually-annotated ground truth by calculating the `Jaccard` index, and computing the gap for subpopulations with ls and ds (based on ITA). EdgeMixup, a data preprocessing method, improves the utility of lesion segmentation in terms of `Jaccard` index compared with all existing baselines. One reason is that EdgeMixup preserves skin lesion information, thus improving the segmentation quality, while attenuating markers for protected factors. Note that EdgeMixup iteratively improves the segmentation results. Take our Skin-seg dataset for example. We trained our baseline Unet model for three iterations, and the model utility is increased by 0.0468 on `Jaccard` index while the J_{gap} between subpopulations is reduced by 0.0193.

Classification Evaluation. Our classification evaluation involves: (i) Adversarial Debiasing (AD) [26], (ii) DexiNed-avg, the average version of DexiNed [27] as an boundary detector used by EdgeMixup, and (iii) ST-Debias [22], a debiasing method augmenting data with conflicting shape and texture information. Our evaluation metrics include accuracy gap, the (Rawlsian) minimum accuracy

Table 3: Skin disease classification and associated bias. Samples contain skin tones as a protected factor. (margin of error reported in parentheses, subpopulation reported in brackets)

	Metrics	ResNet34	Baselines			**EdgeMixup (ours)**	
			AD	DexiNed-avg	ST-Debias	U-Net	Mask-based
Skin	acc	**88.08(3.66)**	81.79(4.35)	69.87(5.17)	76.52(5.23)	86.75(3.82)	86.09(3.90)
	$\texttt{acc}_{\texttt{gap}}$	16.38(12.21)	5.33(11.69)	19.79(13.52)	2.64(8.05)	8.280(9.66)	**1.923(8.49)**
	$\texttt{acc}_{\texttt{min}}$	73.33[ds]	76.92[ds]	51.85[ds]	71.12[ds]	79.41[ds]	**84.38[ds]**
	$\texttt{CAI}_{0.5}$	-	2.380	-10.81	1.090	3.385	**6.233**
	$\texttt{CAI}_{0.75}$	-	6.715	-7.110	7.415	5.743	**10.35**
	AUC	**0.977(0.02)**	0.956(0.02)	0.889(0.04)	0.933(0.03)	0.974(0.02)	0.973(0.02)
	$\texttt{AUC}_{\texttt{gap}}$	0.039(0.07)	0.009(0.05)	0.090(0.11)	0.035(0.04)	0.011(0.02)	**0.01 (0.05)**
	$\texttt{AUC}_{\texttt{min}}$	0.942[ds]	0.955[ds]	0.807[ds]	0.910[ds]	0.973[ds]	**0.964[ds]**
	$\texttt{CAUCI}_{0.5}$	-	0.004	-0.069	-0.024	0.012	**0.013**
	$\texttt{CAUCI}_{0.75}$	-	0.017	-0.060	-0.014	0.020	**0.022**
SD-sub	acc	75.60(14.26)	73.53(14.83)	63.13(16.08)	71.73(13.01)	74.17(13.30)	**76.47(11.26)**
	$\texttt{acc}_{\texttt{gap}}$	28.12(54.98)	25.00(54.30)	25.21(51.20)	18.66(13.21)	18.51(11.50)	**15.00(9.62)**
	$\texttt{acc}_{\texttt{min}}$	50.00[ds]	50.00[ds]	43.75 [ds]	70.59[ls]	72.11[ls]	**75.00[ls]**
	$\texttt{CAI}_{0.5}$	-	0.525	-4.780	2.795	4.090	**6.995**
	$\texttt{CAI}_{0.75}$	-	1.822	-0.934	6.127	6.850	**10.06**
	AUC	0.922(0.10)	0.962(0.06)	0.824(0.13)	0.941(0.08)	0.953(0.11)	**0.970(0.06)**
	$\texttt{AUC}_{\texttt{gap}}$	0.429(0.35)	0.319(0.36)	0.175(0.29)	0.255(0.31)	0.178(0.29)	**0.170(0.29)**
	$\texttt{AUC}_{\texttt{min}}$	0.500[ds]	0.650[ds]	0.650[ds]	0.711[ds]	0.784[ds]	**0.800[ds]**
	$\texttt{CAUCI}_{0.5}$	-	0.075	0.078	0.097	0.140	**0.153**
	$\texttt{CAUCI}_{0.75}$	-	0.092	0.166	0.135	0.196	**0.206**

across subpopulations, area under the receiver operating characteristic curve (AUC), and joint metrics ($\texttt{CAI}_\alpha$ and $\texttt{CAUCI}_\alpha$).

Table 3 shows utility performance (acc and AUC) and fairness results (gaps of acc and AUC between ls and ds subpopulations). We here list two variants of EdgeMixup, and one of which, "Unet", uses the lesion edge generated by the baseline Unet model while "mask-based" implements deep-learning model involved methodology introduced in Section 3. By adding the "Unet" variant, we demonstrate here that simply applying lesion edge predicetd by the baseline Unet model, while not optimal, efficiently reduces model bias on different skin-tone samples. EdgeMixup outperforms SOTA approaches in balancing the model's performance and fairness, i.e., the $\texttt{CAI}_\alpha$ and $\texttt{CAUCI}_\alpha$ values of EdgeMixup are the highest compared with the vanilla ResNet34 and other baselines.

6 Related Work

Skin Disease Classification and Segmentation: Previous researches mainly work on improving model utility for both medical image [28] and skin lesion [29] classification. As for skin lesion segmentation tasks, few works has been proposed due to the lack of datasets with ground-truth segmentation masks. International Skin Imaging Collaboration (ISIC) hosts challenges of International Symposium

on Biomedical Imaging (ISBI) [11] to encourage researches studying lesion segmentation, feature detection, and image classification. However, official datasets released, e.g., HAM10000 [10] only contains melanoma samples and all of the samples are with light skins according to our inspection using ITA scores.

Bias Mitigation: Researchers have addressed bias and heterogeneity in deep learning models [18, 30]. First, masking sensitive factors in imagery is shown to improve fairness in object detection and action recognition [31]. Second, adversarial debiasing operates on the principle of simultaneously training two networks with different objectives [32]. The competing two-player optimization paradigm is applied to maximizing equality of opportunity [33]. As a comparison, EdgeMixup is an effective preprocessing approach to debiasing when applied to skin disease particularly for Lyme-focused classification and segmentation tasks.

7 Conclusion

We present a simple yet novel approach to segment Lyme disease lesion, which can be further used for disease classification. The key insight is a novel data preprocessing method that utilizes edge detection and mixup to isolate and highlight skin lesions and reduce bias. EdgeMixup outperforms SOTAs in terms of Jaccord index for segmentation and $\texttt{CAI}_\alpha$ and $\texttt{CAUCI}_\alpha$ for disease classification.

Acknowledgement. This work was supported in part by Johns Hopkins University Institute for Assured Autonomy (IAA) with grants 80052272 and 80052273, and National Science Foundation (NSF) under grants CNS18-54000. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or JHU-IAA.

References

1. Daniel SW Ting, Yong Liu, Philippe Burlina, Xinxing Xu, Neil M Bressler, and Tien Y Wong, "Ai for medical imaging goes deep," *Nature medicine*, vol. 24, no. 5, pp. 539–540, 2018.
2. Philippe Burlina, Neil Joshi, William Paul, Katia D Pacheco, and Neil M Bressler, "Addressing artificial intelligence bias in retinal disease diagnostics," *Translational Vision Science and Technology*, 2020.
3. 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*, 2014.
4. 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*, 2021.
5. Robert B. Nadelman, "Erythema migrans," *Infectious Disease Clinics of North America*, 2015.

6. Carrie D. Tibbles and Jonathan A. Edlow, "Does This Patient Have Erythema Migrans?," *JAMA*, 2007.
7. Philippe M Burlina, Neil J Joshi, Elise Ng, Seth D Billings, Alison W Rebman, and John N Aucott, "Automated detection of erythema migrans and other confounding skin lesions via deep learning," *Computers in biology and medicine*, vol. 105, pp. 151–156, 2019.
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*, 2019.
9. 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*, 2020.
10. 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*, 2018.
11. Noel Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, Harald Kittler, and Allan Halpern, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," 2019.
12. Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri, "Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset," in *Proceedings of the IEEE/CVF CVPR*, 2021.
13. Muhammad Attique Khan, Muhammad Sharif, Tallha Akram, Robertas Damaševičius, and Rytis Maskeliūnas, "Skin lesion segmentation and multiclass classification using deep learning features and improved moth flame optimization," *Diagnostics*, 2021.
14. 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*, 2015.
15. Marin Benčević, Irena Galić, Marija Habijan, and Danilo Babin, "Training on polar image transformations improves biomedical image segmentation," *IEEE Access*, 2021.
16. Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao, "Vision transformer adapter for dense predictions," in *ICLR*, 2023.
17. Hritam Basak, Rohit Kundu, and Ram Sarkar, "Mfsnet: A multi focus segmentation network for skin lesion segmentation," *Pattern Recognition*, 2022.
18. Simon Caton and Christian Haas, "Fairness in machine learning: A survey," *arXiv preprint arXiv:2010.04053*, 2020.
19. Philippe Burlina, William Paul, Philip Mathew, Neil Joshi, Katia D Pacheco, and Neil M Bressler, "Low-shot deep learning of diabetic retinopathy with potential applications to address artificial intelligence bias in retinal diagnostics and rare ophthalmic diseases," *JAMA ophthalmology*, 2020.
20. "Edgemixup repository," `https://github.com/Haolin-Yuan/EdgeMixup`.
21. 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*, 2018.
22. Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, and Cihang Xie, "Shape-texture debiased neural network training," in *International Conference on Learning Representations*, 2021.

23. 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*, 2016.
24. 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*, 2015.
25. Thomas B. Fitzpatrick, "Soleil et peau," *Journal de Médecine Esthétique (in French)*, 1975.
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*, 2018.
27. X. Soria, E. Riba, and A. Sappa, "Dense extreme inception network: Towards a robust cnn model for edge detection," in *WACV*, Mar 2020.
28. Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang, "Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification," in *ICCV*, 2021.
29. Bhuvaneshwari Shetty, Roshan Fernandes, Anisha P Rodrigues, Rajeswari Chengoden, Sweta Bhattacharya, and Kuruva Lakshmanna, "Skin lesion classification of dermoscopic images using machine learning and convolutional neural network," *Scientific Reports*, 2022.
30. Haolin Yuan, Bo Hui, Yuchen Yang, Philippe Burlina, Neil Zhenqiang Gong, and Yinzhi Cao, "Addressing heterogeneity in federated learning via distributional transformation," in *ECCV*, 2022.
31. 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 *ICCV*, 2019.
32. 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*, 2019.
33. 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.