IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 May 2023, accepted 16 June 2023, date of publication 26 June 2023, date of current version 1 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3289320

RESEARCH ARTICLE

Assessing Bias in Skin Lesion Classifiers With Contemporary Deep Learning and Post-Hoc Explainability Techniques

ADAM CORBIN AND OGE MARQUES, (Senior Member, IEEE)
Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, FL 33431, USA
Corresponding author: Adam Corbin (acorbin3@fau.edu)

ABSTRACT As Artificial Intelligence (AI) is increasingly utilized in dermatology, ensuring fairness in the development of Machine Learning models is crucial, particularly in skin lesion classification, where decisions can significantly impact people's lives. This study investigates the presence of biases between different Fitzpatrick Skin Types in baseline pretrained models and evaluates various training techniques to mitigate these disparities. An unsupervised skin transformer is developed to adjust an image's Fitzpatrick Skin Type (FST), and joint regularization and synthetic image blending methods are employed to address bias concerns. Additionally, eXplainable Artificial Intelligence (XAI) techniques, such as Gradient-weighted Class Activation Mapping (Grad-CAM), are utilized to identify any underlying reasons for bias in the models. The results indicate that joint regularization and synthetic blending methods enhance the area under the curve performance and fairness. Meanwhile, XAI was found to be a valuable tool for fine-tuning Deep Learning models and uncovering problems. These findings can aid in developing accurate and unbiased skin lesion classification models, promoting equitable healthcare, and improving patient outcomes.

INDEX TERMS Bias, deep learning, explainable AI, fairness, skin lesion classification.

I. INTRODUCTION AND OVERVIEW

Classifying skin lesions is vital in dermatology, as it facilitates early diagnosis of skin diseases, leading to better patient outcomes. Deep learning (DL) techniques have improved the accuracy and efficiency of skin lesion classification. However, applying DL models in medical domains raises concerns about biases in the models that may cause incorrect predictions and discriminate against specific groups of people, particularly those with darker skin tones.

The need for fair and accurate skin lesion classification models has gained widespread recognition, and many attempts have been made to eliminate bias in machine learning models [1], [2], [3], [4], [5]. However, the field of research in this area is still developing, and a consensus on the most effective methods for addressing bias in skin lesion classification models needs to be reached.

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Wang.

This study uses contemporary DL techniques to assess the usefulness of popular post-hoc XAI techniques in detecting signs of bias against darker skin in skin lesion classifiers. We have developed a skin transformer to alter skin color and assess its impact on DL model results, as well as XAI techniques like Grad-CAM to uncover the underlying reasons for any biases discovered in the models.

In the experimental research study, FST metadata was generated, and this FST information was deployed to assess fairness across baseline transfer learning models, the CIRCLe model (join regularization), and models that utilized synthetic data. The findings indicate that both the join regularization and the use of synthetic data outperformed the baseline, with synthetic data emerging as the most effective in fairness, quantified in terms of Area Under the Curve (AUC). We also evaluate the utility of Post-Hoc XAI as an essential instrument for discerning crucial areas within images. XAI is particularly beneficial in revealing model limitations, such as focusing on image regions not pertinent to the skin lesion. The study advocates for incorporating XAI during model

development for optimization and refinement. The insights and outcomes of this research hold substantial implications for machine learning, particularly in contexts where fairness is important.

This work investigates the bias problem in a baseline binary classifier to distinguish between benign and malignant dermoscopic images. Our analysis indicates that the baseline classifier may be biased towards Fitzpatrick Skin Type (FST) 1-3 compared to FST 4-6. To address this issue, we propose several solutions, including modifications to the CIRCLe method and the use of synthetic data to train models. We hypothesize that these solutions improve fairness across FST.

II. LITERATURE REVIEW

A. GENERATING METADATA

One of the commonly recognized issues with public dermatology datasets is the lack of diversity in skin tone representation. Most available datasets predominantly feature images of lighter skin tones [2], [6], [7]. Furthermore, it has been observed that some AI diagnostic tools demonstrate better performance on lighter skin tones, potentially due to the absence of adequate training data for darker skin tones [8]. The FST scale is a widely-used classification system to determine an individual's skin tone. The scale consists of six classes and was first introduced in a study on sun exposure and its impact on different skin tones [9].

Type 1 and 2 consist of individuals with very fair to fair skin, often accompanied by red or blond hair and freckles. People within this category tend to burn easily when exposed to sunlight. While Type 1 individuals almost always burn and never tan, Type 2 individuals usually burn but may achieve a slight tan with repeated exposure. Type 3 includes those with a cream-white skin complexion, who may experience sunburn but can gradually tan over time. Type 4 comprises individuals with moderate brown skin, often of Mediterranean or Asian descent, who have the propensity to tan easily and seldom burn. Type 5 is characterized by people with dark brown skin, typically of Middle Eastern, Hispanic, or African descent, who tan very easily and rarely experience sunburn. Type 6 encompasses individuals with deeply pigmented dark brown to black skin, primarily of African ancestry, who are naturally protected against sunburn due to the high melanin content in their skin and never burn.

While there are other scales available, such as the Glogau scale [10], Roberts scale [11], and Baumann scale [12], the Fitzpatrick scale remains the primary scale used in dermatology research. However, metadata for dermatology images, including skin type and ethnicity data, are not frequently collected. For instance, in a review of various skin cancer datasets, only 2.1% of the 106,950 images in 21 open-access databases included FST data, and only 1.3% included ethnicity data [13]. A graphical representation of these findings is presented in Fig. 1.

The findings highlight the limited availability of public datasets containing FST data, which poses significant challenges to researchers seeking to utilize this information in their studies. Manual annotation of datasets by experts remains the most reliable approach to obtaining such data, but it is often prohibitively expensive and time-consuming. Consequently, researchers have developed alternative methods for obtaining skin-type classification data, including the use of mathematical techniques to extract this information from images. Specifically, one such approach involves computing the Individual Typology Angle (ITA) of an image using Equation (1). The ITA can then be mapped to an FST category, as shown in Table 1.

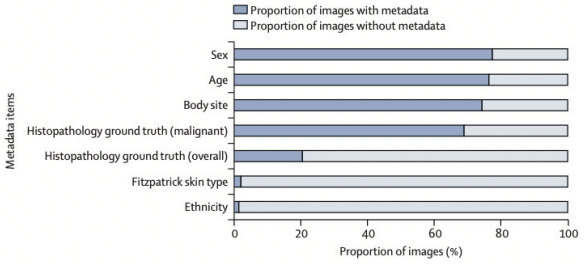

FIGURE 1. Provided metadata in publically available datasets.

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

TABLE 1. Fitzpatrick ITA values to skin type for six classes [14].

ITA Range	FST
$55^\circ < \text{ITA}$	Type1
$41^\circ < \text{ITA} \leq 55^\circ$	Type2
$28^\circ < \text{ITA} \leq 41^\circ$	Type3
$19^\circ < \text{ITA} \leq 28^\circ$	Type4
$10^\circ < \text{ITA} \leq 19^\circ$	Type5
$\text{ITA} \leq 10^\circ$	Type6

The two main components in the ITA Equation (1) are the b and L* values. These values come from the LAB color space which the images are converted from RGB to LAB color space. L* value represents lightness, ranging from 0 (representing black) to 100 (representing white). The b axis ranges from blue to yellow, where negative values signify blue and positive values signify yellow.

Fig. 2 shows a graph of how each component adjusted can change the ITA result.

In a study by Kinyanjui et al., [2], the International Skin Imaging Collaboration (ISIC) 2018 dataset was evaluated using different approaches to capture the FST by calculating the ITA on whole images and images with the skin lesion removed. The ISIC dataset comprises challenges hosted by Kaggle since 2016. Each year, a different dataset and challenges were created, starting with simple benign and malignant classifiers working up to multi-class classifiers and skin lesion segmentation. The ISIC 2018 dataset provides skin lesion pixel mask annotations that can remove skin lesions from the images before the ITA value is computed. To improve the estimation of FST, the team built a segmentation model to remove skin lesions and found that this

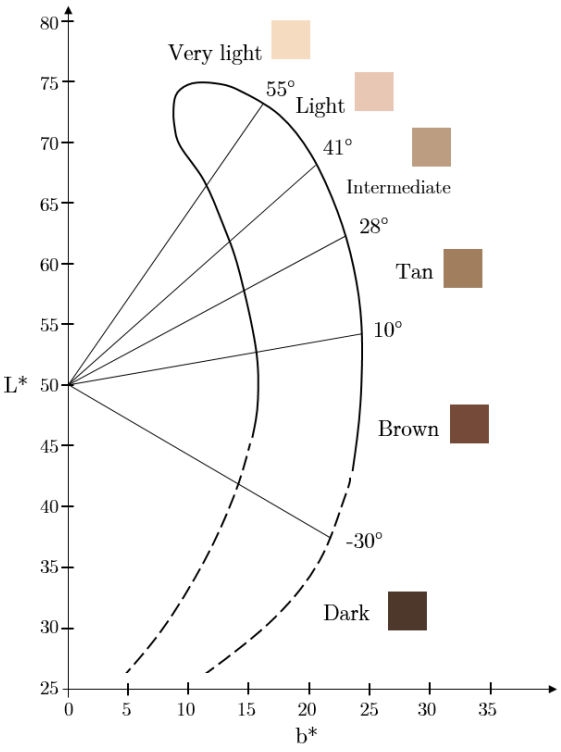

FIGURE 2. In this figure shows how the b and L* values contribute to the ITA value. There are also skin samples and a table of ITA values for FSTs.

approach was effective. Furthermore, they observed that their Machine Learning (ML) classification model did not show any correlation between performance and ITA value.

Groh et al. [3] conducted research on a dataset of 17,000 images that includes FST classification metadata. They also developed a new algorithm to classify skin tones using a customized YCbCr mask to remove non-skin pixels and backgrounds. YCbCr is a color space that separates image luminance from chrominance. The Y component represents the luminance or brightness of a color. In other words, it captures the amount of light in the color. The Cb component represents the chrominance relative to the blue color channel. Negative values in the Cb channel represent colors toward blue, while positive values represent colors away from blue. The Cr component represents the chrominance relative to the red color channel. Negative values in the Cr channel represent colors toward red, while positive values represent colors away from red.

To calculate the ITA value, they only used skin pixels identified by the mask. The customized YCbCr mask was designed to focus exclusively on skin pixels, as shown in Figure4. Equations 2-6 define the ranges of pixels that represent skin.

$$Cr > 135 \tag{2}$$
$$Cr \geq (0.3448 \cdot Cb) + 76.2069 \tag{3}$$
$$Cr \geq (-4.5652 \cdot Cb) + 234.5652 \tag{4}$$
$$Cr \geq (-1.15 \cdot Cb) + 301.75 \tag{5}$$
$$Cr \geq (-2.2857 \cdot Cb) + 432.85 \tag{6}$$

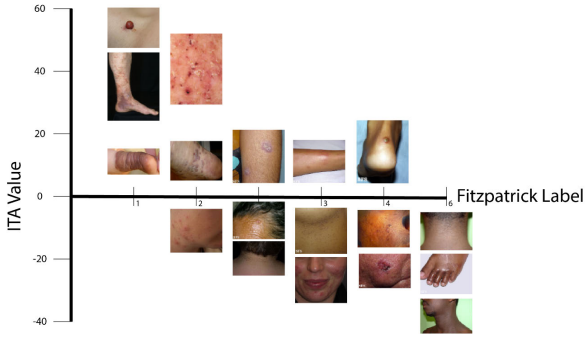

FIGURE 3. Sampled images plotted on a graph representing their respected ITA value and FST.

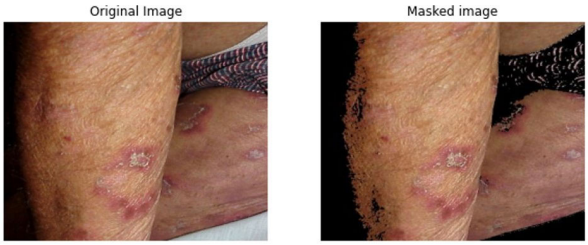

FIGURE 4. Example of an image with the applied YCbCr masked removing the background.

This mask would allow the Fitzpatrick classification to work on images beyond dermoscopic types. Fig. 3 shows a set of images plotted on a graph with the result of their ITA and FST.

In their work, Groh et al. [3] observed a high variance of ITA values, which resulted in misclassifications of FSTs for many images. To address this, they proposed an updated table of ITA for FST ranges, as shown in Table2. This new table helped to improve their performance in estimating FST, as evidenced by the results in Fig. 3, which demonstrated better agreement with expert results.

TABLE 2. Alternative mapping of ITA ranges to the six class FSTs [3].

ITA Range	FST
$40° < \text{ITA}$	Type 1
$23° < \text{ITA} \leq 40°$	Type 2
$12° < \text{ITA} \leq 23°$	Type 3
$0° < \text{ITA} \leq 12°$	Type 4
$-25° < \text{ITA} \leq 0°$	Type 5
$\text{ITA} \leq -25°$	Type 6

The methods available for generating FSTs provide opportunities for investigating dataset fairness and optimizing the data split between training and test sets. However, these methods also underscore the need for specific skin type metadata in datasets, which could facilitate data generation using Generative Adversarial Networks (GANs). Moreover, it is crucial to evaluate AI/ML solutions from a fairness perspective to ensure that all skin types perform similarly. These evaluations are essential in promoting equity in dermatological research and preventing the exacerbation of health disparities.

TABLE 3. Results of using the YCbCr mask versus full image ITA calculation with using both Table 1 (Kinyanjui) [2] and Table 2 (Empirical) [3] to convert to the FST. These results represent the ITA to FST plus or minus 1 point to the annotated label.

	Full Image		YCbCr Mask	
	Kinyanjui	Empirical	Kinyanjui	Empirical
Overall	45.87%	60.34%	53.30%	70.38%
Type 1	50.97%	65.35%	52.22%	66.00%
Type 2	42.60%	59.57%	49.15%	69.47%
Type 3	35.43%	55.20%	45.13%	66.41%
Type 4	34.09%	58.54%	40.24%	72.10%
Type 5	78.21%	65.49%	93.41%	82.26%
Type 6	74.80%	64.04%	90.71%	79.69%

B. USING METADATA TO EVALUATE FAIRNESS

It is important to leverage different data sources when developing AI products, including datasets that provide valuable metadata such as pixel mask annotation, text-based metadata such as sex, age, and skin lesion location, and in-depth electronic health records in some private datasets. However, the FST data, crucial for evaluating datasets and ML models, is not always readily available in these sources. Therefore, as discussed in previous sections, it is necessary to explore different methods for obtaining this data, such as expert annotation and mathematical extraction of the ITA value. By leveraging diverse data sources and obtaining accurate FST data, we can develop more effective and equitable AI products. For instance, in the absence of FST data, pixel mask annotation can be leveraged to calculate the ITA value. However, this method may lead to inaccurate results as skin lesions have a different color than the patient's skin color, resulting in the inclusion of non-skin pixels in the calculation. In such cases, the pixel mask can be inverted to obtain a skin mask that can be used to calculate the ITA value exclusively on skin pixels.

In their study, Yap et al. [15] have incorporated additional metadata, such as FST, into their image analysis using deep learning. They have adopted a multimodal approach that utilizes a late fusion technique [16] to integrate different image types and text-based metadata into a single solution. To this end, they developed and executed a matrix of 13 different experiments that involved combinations of both image types and text-based classification, isolating some features and exploring different approaches to using the Convolutional Neural Network (CNN) embedding network.

Groh et al. [3] developed a deep learning classifier using the Fitzpatrick 17k dataset, which includes data on 114 skin conditions. With the FST data now available, they could evaluate the model's performance across each of the different FSTs and modify and hyper-tune the model to focus on improving performance across all skin types. Through experimentation, they discovered that modifying the holdout selection for different training methods either improved or harmed the models, providing them with valuable insights for optimal model training.

A systematic review conducted by Hohn et al. [17] identified 11 studies that utilized multimodal approaches to image analysis in dermatology. The studies analyzed a variety of datasets, including the ISIC, International Symposium on Biomedical Imaging (ISBI), HAM10000, ImageNet Large Scale Visual Recognition Challenge (ILSVRC), and private datasets with additional metadata, such as age, sex, lesion location, lesion size, bleeding, pattern, and elevation. The review paper provides valuable insights for developing our proposed approaches to improve fairness in dermatology image analysis.

Pakzad et al. [5] have proposed a novel model pipeline in their recent study that aims to improve bias and fairness on the Fitzpatrick 17k dataset. In their proposed pipeline, they have used Skin Color Transformer by extending StarGAN architecture, which transforms an input image to represent a different FST. This transformation is done by passing both original and transformed images through a feature extractor, and a regularization loss is captured. These features are then passed through a classifier to compute the classification and the classification loss. The model is trained to minimize both regularization and classification losses. Additionally, the model is punished when the original and transformed images result in different classes.

The performance of the proposed model was evaluated across all FSTs using recall, F1 score, and accuracy metrics. To assess fairness, the authors used Equal Opportunity Difference (EOD) and Normalized Accuracy Range (NAR) as part of fairness computations. The authors tested their model with different "backbone" architectures and found that DenseNet-121 performed the best. The authors also conducted holdout experiments, similar to those conducted by Groh et al. [3], and found similar but improved results. Overall, the proposed pipeline by Pakzad et al. [5] demonstrates promising results for improving fairness and reducing bias in skin type classification.

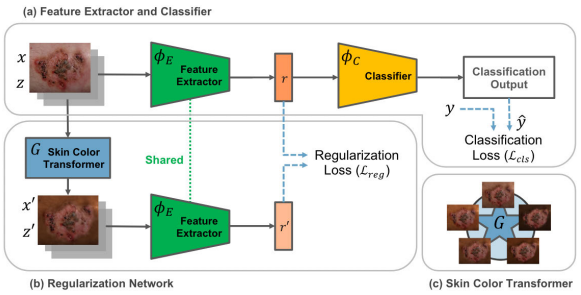

FIGURE 5. Overview of the CIRCLe model pipeline architecture [5].

C. MORE STANDARDIZATION

Standardization is crucial in dermatology AI, ensuring a consistent and reproducible evaluation of model performance, including fairness. Despite the rapid growth of AI in dermatology, there is a lack of standardization in the field, with researchers often experimenting with different techniques and methods. However, as AI becomes increasingly integrated into our lives, it is vital to establish standardized approaches to evaluate fairness at each process step.

A review paper by Young et al. [18] highlights the need for a more standardized metric approach for evaluating CNNs used to classify skin lesions. The paper notes that many studies use their unique way of computing metrics, making comparing results across different approaches difficult. The lack of standardization poses a challenge when evaluating the fairness of other AI models, and it is essential to develop standardized approaches to enable fair comparison of different approaches.

Establishing a standardized approach to evaluating fairness in dermatology AI is an ongoing challenge. However, it is essential to address this issue to ensure that AI models are transparent, unbiased, and equitable in their predictions. By developing standardized metrics and methods for evaluating fairness, we can help promote greater consistency, comparability, and transparency in dermatology AI research.

D. DEALING WITH IMBALANCED DATASETS

One of the most significant challenges in developing AI models for skin disease classification is dealing with imbalanced datasets. In the medical field, imbalanced datasets are common; some classes have significantly more samples than others. In dermatology, this is especially true for rare diseases, which can have limited samples, making it difficult to train a robust model.

Imbalanced datasets can result in models that perform well on majority classes but poorly on minority classes, leading to biased and unfair models. One approach to dealing with imbalanced datasets is oversampling, which involves replicating the minority class samples to balance the dataset. However, this approach can lead to overfitting and poor generalization performance.

Another approach is under-sampling, where the majority class samples are randomly removed to balance the dataset. However, this approach can result in a loss of information and poor performance on the majority class samples.

A more effective approach is to use techniques such as Synthetic Minority Oversampling Technique (SMOTE) [19] and Adaptive Synthetic Sampling (ADASYN) [20] that generate synthetic samples for the minority classes. SMOTE creates synthetic samples by interpolating between neighboring minority class samples in feature space, while ADASYN adapts the density distribution of the minority class samples to generate synthetic samples. These techniques preserve the original data distribution while balancing the dataset and can improve model performance in minority classes.

Another approach is to use class weighting during model training, where the loss function is weighted to give more importance to the minority class samples. This approach can help the model learn the minority class features better and improve its performance on the minority classes.

In addition to these techniques, it is also essential to evaluate the model's performance in all classes, including the minority classes. Metrics such as precision, recall, F1 score, and AUC can provide insights into the model's performance in individual classes. In particular, AUC can be useful for evaluating model performance on imbalanced datasets as it is less sensitive to class imbalance.

However, it is essential to note that no single technique can entirely solve the problem of imbalanced datasets. The choice of technique may depend on the dataset's characteristics and the specific issue being addressed. It is also essential to evaluate the impact of these techniques on fairness, as they may introduce new biases in the model.

In summary, dealing with imbalanced datasets is a critical challenge in developing AI models for skin disease classification. Techniques such as SMOTE, ADASYN, and class weighting, and proper evaluation metrics, can help improve model performance on minority classes and mitigate bias and unfairness. However, it is important to carefully evaluate the impact of these techniques on model performance and fairness.

III. MATERIALS AND METHODS

This section provides a detailed account of the experimental procedures and techniques used in the study, which is outlined in Fig. 6. The present study utilized the Kaggle melanoma 2020 dataset, comprising 37,648 images. In the data curation phase, skin lesion masks were generated, and Fitzpatrick skin type metadata and augmented images for certain training methods. Subsequently, the dataset was partitioned into specific training, validation, and test subsets. Thereafter, an ablation study was performed, constituting a widely adopted method [21] of evaluating different components of deep learning models, involving the addition or removal of specific model parts coupled with modifications in various hyperparameters. The study culminated with a comprehensive analysis of the results, encompassing fairness metrics and the application of the XAI technique using the Grad-CAM method.

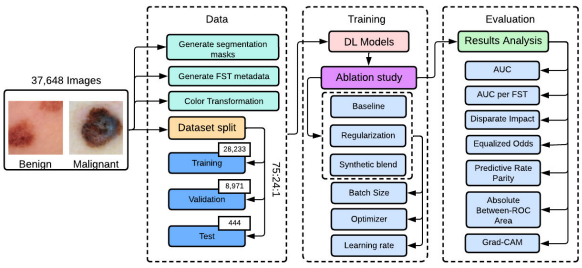

FIGURE 6. This figure provides a comprehensive overview of the dataset and the analysis pipeline. It begins by presenting the dataset and its partitioning, then training steps and an ablation study. The figure concludes by showcasing the analysis of the results, which includes multiple measurements.

Our approach is primarily inspired by the work of Pakzad et al. [5] on the CIRCLe pipeline architecture. They proposed using taking an image and augmenting it to a different FST and passing both the original and augmented images through the DL model. The learned features were then regularized to mitigate bias towards different skin tones.

Here is an outline of our approach:

1) **Data**
   a) **Generate skin lesion masks** - Using DoubleU-Net trained on ISIC 2018 Lesion Boundary segmentation.
   b) **Generate FST for all images** - The dataset does not provide this data, and it is required to evaluate fairness.
   c) **Color Transformation** - Transform each image to a different FST. This is used for the synthetic training method
   d) **Split dataset** - Split the dataset into training, validation, and testing.
2) **Ablation study** - Test each of the training methods by modifying the batch size, optimizer, and learning rate.
   a) **Baseline training**: Our baseline model is trained using the Melanoma 2020 challenge Kaggle dataset outlined in Section III-A1. Our baseline model is evaluated with the evenly distributed FST dataset.
   b) **Joint regularization training**: an image is subjected to a series of transformations using different FSTs. These transformed images and the original image are then passed through the model to compute their respective features. These features calculate and apply a regularization term to the models loss function. This regularization encourages the transformed images to produce similar results as the original image, while a greater deviation between them would harm the models performance. This method is outlined in Section III-B3.
   c) **Synthetic blend training**: The baseline model is retrained with additional synthetic data generated from our skin transformer outlined in Section III-A2. The training data is a blend of synthetic data and real images. Our training method uses 50% synthetic images and 50% real images.
3) **Evaluation** - Analyze each model's AUC across each FST. Also, evaluate the fairness metrics. Finally, evaluate the usefulness of XAI.

We hypothesize that each experiment should become a more fair model between the baseline, joint regularization, and synthetic blend training methods.

A. DATA

1) DATASET

This work explores melanoma classification using the Kaggle 2020 Melanoma Competition dataset [22]. The dataset comprises 37,648 dermoscopic images of either benign or malignant skin lesions selected from over 2,000 patients. Each image has a specific patient's unique identifier. This is important when building a training and test dataset to ensure they are stratified between the two. The malignant diagnoses in the dataset have been verified through medical experts, but the dataset at the current time has not had a thorough peer review.

This dataset provides the dermoscopic images and a set of useful metadata, which includes image name, patient id, sex, approximate age, anatomical location of image site, detailed diagnosis, and an indication of benign or malignant.

This study utilized a dataset that lacked the annotation of FST. Thus, the skin type was estimated using the methodology described in our prior research [23]. Because we also calculated the skin lesion masks, as described in Section III-A3, the FST was estimated with the skin lesions removed. This information is used in Section IV-C to analyze the performance of the models concerning each FST. In Fig. 7, we can see an overwhelming number of FST 1 relative to 2-6.

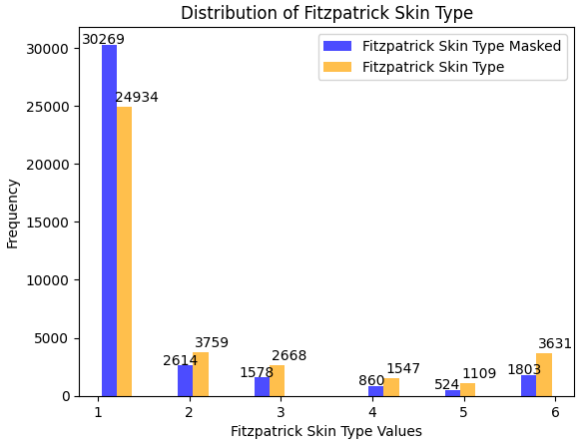

FIGURE 7. Distribution of the FST in the Kaggle 2020 Melanoma classification competition. The FST uses the approach proposed in [23] whereas the approach labeled as *FST Masked* removes the skin lesion before the FST is estimated.

Fig. 8 presents a set of samples of each estimated FST. Upon careful examination of the sample set, it is apparent that some images were misclassified due to artifacts, and excessive hair in row 3 column 5. Furthermore, images containing relatively larger lesions as observed in row 3 column 6, might have potentially undermined the accuracy of the FST estimation.

In the Kaggle 2020 Melanoma Competition, Chris Deotte provided a valuable resource to the community by presenting a triple-stratified dataset [24]. This dataset was split into test and validation sets, with a balanced distribution of patients, patient count, and diagnosis, ensuring fairness and accuracy in the evaluation process. Deotte was trying to avoid any data leakage between the test and validation sets, which could significantly impact the results and interpretation of the study.

In our experiments, we use Deotte's approach as a baseline for training but create a special validation dataset. We divided the data in a 75:24:1 between training, testing, and validation. The validation dataset selects 74 images of each FST class to create our benchmarking validation dataset. Of the 74 images, 37 are benign, and 37 are malignant for each FST. This brings

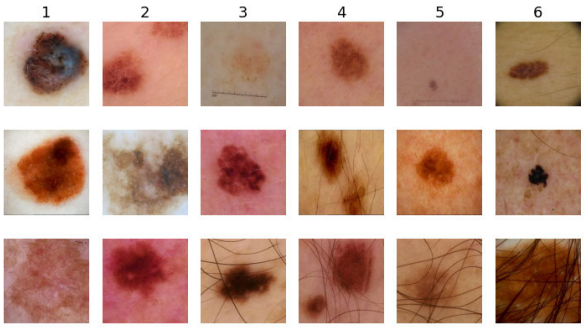

FIGURE 8. Sample set of images in each of the estimated FST. The columns represent the estimated FST going from 1 to 6 from left to right. The top two rows show examples where the estimation appears to be accurate whereas the last row examples of cases that might raise concerns about the estimation method.

a total of 444 total images in the validation dataset. This dataset is used to test our different experiments.

The *Synthetic blend* training method utilized in our study involves the generation of a new dataset by transforming every training image to be darker from a FST perspective. Due to limitations in the color transformer, we decided to restrict the FST shift to a maximum of 2 types. This restriction ensured that an image with an initial FST value of 3 would only transform to a maximum value of 5. This decision was made based on the observation that results beyond this maximum value did not yield satisfactory visual outcomes.

During the training process, we randomly generated a number between 0 and 1. If the generated number was greater than 0.5, we utilized the transformed image. Conversely, if the number was less than or equal to 0.5, we used the original image. This approach ensured that the training model was exposed to a mixture of transformed and original images to enhance the robustness of the trained model.

The incorporation of the Synthetic blend training method in our study aimed to improve the training process by expanding the diversity of the training dataset. By introducing transformed images, we sought to increase the variability of the dataset and ensure that the model could effectively generalize to a broad range of real-world scenarios. However, we acknowledge that this approach is not without limitations, particularly regarding the extent of image transformation allowed and the choice of transformation method.

2) COLOR TRANSFORMATION

Because many datasets lack darker skin type samples, there have been attempts to augment the datasets to generate darker skin examples. There are many different approaches to balancing datasets: Undersampling, oversampling, Synthetic Minority Oversampling Technique (SMOTE), data augmentation techniques, and the use of GANs to generate more images [5], [25], [26], [27]. For our experiments, we would like to augment the datasets for test and validation purposes. To evaluate the effects of skin color on the performance of the DL model, we use a color transformation algorithm that adjusts the skin color of the lesion images. The algorithm darkens the skin lesion images by converting from one FST to another. The color transformation is performed using color space manipulations and is inspired by Pakzad et al. [5], which uses StarGANs, although it is simplified for this study.

One of the reasons we decided to use a custom color transformer is that the StarGAN model needs to be trained and might need to be re-trained based on the dataset being used for classification. Also the StarGAN model adjusts the color of the full image including the skin lesion and we wanted to leave the skin lesion unchanged for our color transformer. Because of this it is important to have masks for the skin lesion which is discussed in section III-A3. Therefore, we propose a method that does not need to be trained and generates images that transform into different FSTs.

In this approach, the focus is on converting dermoscopic images from one FST to another. To accomplish this, only a few parameters need to be adjusted. The FST can be calculated using the ITA, represented by Equation (1). In the ITA Equation (1), the two main components are the b and L* values. Fig. 2 in part (a) shows a graph of how each component adjusted can change the ITA result. Using the ITA value, the FST can be determined by using Table 1 as a lookup reference. The accurate computation of the ITA value can be challenging in the presence of artifacts such as stickers or pen markings with colors distinct from the patient's skin tone. These artifacts introduce inaccuracies to the computation and can lead to incorrect ITA values. Additionally, the color of the skin lesion itself can differ from the patient's typical skin color, which can also cause incorrect ITA computations. To address this issue, we propose leveraging skin lesion masks to remove the lesion from the image before computing the ITA value. This approach ensures that the ITA value is calculated accurately, even in the presence of skin lesions or artifacts, and improves the overall robustness and accuracy of our proposed method.

Fig. 9 shows the pipeline of how the image is transformed. This utility expects an image, the skin lesion pixel mask annotation, the current FST, and desired, which is FST' as inputs. This is needed because we use the ISIC datasets, which do not provide FST data. The skin lesions are important to the downstream machine learning classifier, and we want to preserve them when doing the skin transformations. The skin lesion mask is used as a filter out all of the skin lesions before the color transformation. That leaves only the skin to be adjusted with the skin transformation.

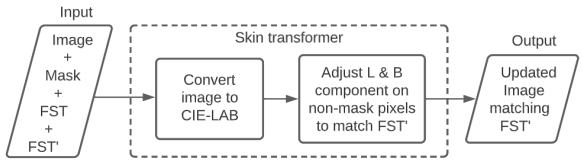

FIGURE 9. The skin transformer that uses in an image, the skin lesion mask, the FST, and desired FST' is returned a new image modified to match the desired FST.

To calculate the FST, the image must be converted from the RGB color space to the LAB color space. The LAB color space has three components: L, A, and B. The L component represents the lightness intensity that the color reflects or emits. The a component represents the color spectrum between green and red. The b component represents the color spectrum between blue and yellow.

Our approach involves adjusting the L and B components to convert the skin type of an image. The L component affects the intensity of the image, while the B component affects the yellow-ness of the image.

A random ITA value is selected in the range for the desired FST from 1 to make these adjustments. The difference is selected between the original ITA value, and this randomly selected ITA value in Equation (7). Based on this difference, the b and L values are then adjusted with a scaling factor in Equations 8 and 9.

$$ita_{\Delta} = ITA - ITA' \tag{7}$$

$$b = b + (ita_{\Delta} \cdot 0.5) \tag{8}$$

$$L = L - (ita_{\Delta} \cdot 0.12) \tag{9}$$

By adjusting these parameters, we aim to achieve a conversion of the image to a different FST while preserving the relevant information in the dermoscopic image.

We pre-processed all the original images to compute the ITA values and converted FST to be saved off so they did not need to be recomputed during the training and testing phases.

3) SEGMENTATION MASKS

To perform the skin color transformation, we aimed to separate the lesion from the surrounding skin and only modify the skin. Separating the lesion and the skin can be achieved using segmentation masks. In this study, we considered three scenarios for the availability of segmentation masks: no masks, automatically generated masks, and manually generated masks (available as ground truth). The Kaggle Melanoma 2020 dataset does not provide skin lesion masks, so we generate the masks by using a finely-tuned segmentation for skin lesions called DoubleU-Net [28]. The DoubleU-Net model was trained and tested on the ISIC 2018 Lesion Boundary segmentation challenge which they provided pre-trained weights. In generating the masks using the DoubleU-Net, we found the results needed to be rotated by 270 degrees and then flipped vertically to achieve the correct mask. Figure 10 shows the sample results of the skin lesion segmentation masks.

Masks were generated for each image in the Kaggle Melanoma 2020 dataset. These masks were used in the regularization training method and to develop the synthetic transformed images.

B. TRAINING

1) MODELS

For this study, we used the EfficientNet model family and ResNet. EfficientNet is a family of image classification models developed to increase image classification accuracy while reducing computational costs. It is based on a model scaling method that balances the network's width, depth, and image resolution against available resources to improve overall performance. The B2 model has around 10M parameters, whereas popular alternatives such as ResNet-50 or ResNeXt-101 are between 25M-84M parameters.

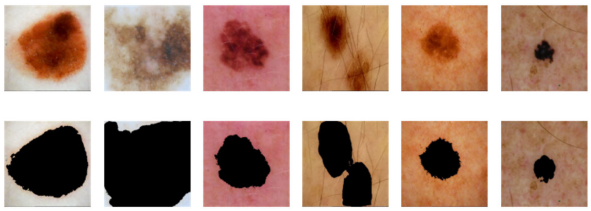

FIGURE 10. Sample set of images where the first row is the original image and the second row is the generated masks over the skin lesion.

We use the PyTorch implementation of the pre-trained EfficientNet B2, EfficientNet B4, and ResNet-50 models. The pre-trained weights come from this model being trained on the imagenet-1k [29], a dataset of 1,000 classes and over 1M training images and 100k test images. The pre-trained model was fine-tuned on the Kaggle Melanoma 2020 dataset of dermoscopic images of skin lesions using transfer learning. Transfer learning is a process in which a pre-trained model is fine-tuned on a smaller dataset to improve its performance for a specific task.

Because this model has a smaller set of parameters, the training time is improved over the larger models. Tan and Le in their release of EfficientNet [30] found that EfficientNet-B1 was 5.7x faster tan ResNet-152 on the inference latency and EffcientNet-B7 was 6.1X quicker than GPipe model. Users who require a more substantial model may experiment with the B0 to B7 models with minimal modifications to evaluate their suitability.

2) ABLATION STUDY

To evaluate the robustness of each training method, we conducted a comprehensive ablation study, focusing on modifying hyperparameters such as batch size, optimizer, and learning rate. We first selected the EfficientNet-B2 architecture as our reference model to streamline the ablation study. This decision allowed us to efficiently explore the impact of different hyperparameters on the model's performance.

1) **Batch Size Evaluation** - We commenced the study by examining the influence of varying batch sizes on the performance of the EfficientNet-B2 model. The chosen batch sizes for this investigation were 8, 16, 24, and 32. Upon identifying the optimal batch size, we investigated the most suitable optimizer for the model.
2) **Optimizer Selection** - We evaluated the following optimizers: Adam, AdamW, Adamax, and NAdam. This selection encompasses a diverse range of optimizers, highlighting the characteristics of stochastic optimization, from the adaptive moment estimation of Adamax to the decoupled weight decay regularization

implemented in the AdamW optimizer. All optimizers were readily available within the PyTorch framework.

3) **Learning Rate Exploration** - After selecting the most suitable optimizer, we explored the impact of various learning rates on the model's performance. The learning rates tested were 0.0005, 0.001, 0.0015, and 0.002.

Upon determining the optimal learning rate, we employed this combination of hyperparameters for all selected models to validate their efficacy in medical image analysis.

3) REGULARIZATION

The regularization is inspired by the Pakzad et al. [5] where they color transform an image and run it through the same feature extractor. Fig. 11 outlines our adaptation of the CIRCLe model. The regularization loss is added to the overall loss function to enforce the invariant condition between the original and synthetic images. The loss function includes a prediction loss, which calculates the cross entropy between the true labels and predicted probabilities, and a regularization loss, which calculates the squared error distance between the latent representations of the original image and the synthetic image. The final predicted class is the one with the highest predicted probability. A hyperparameter controls the trade-off between the prediction and regularization losses.

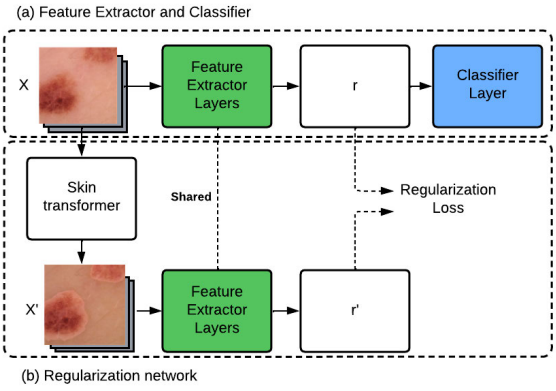

FIGURE 11. This figure depicts an updated classification architecture that extends CIRCLe. Figure (a) shows the process of passing a dermoscopic image X through the feature extractor layers of the model, resulting in a learned representation r. This representation is passed through the classification layer, producing the final classification output. Figure (b) illustrates the custom skin transformer responsible for processing the input image to produce a modified version X'. More details on this transformer can be found in Fig. 9. After X' is obtained, it is passed through the same feature extractor layers as in (a) to generate a newly learned representation r'. The regularization loss is then calculated using Mean Squared Error (MSE) between r and r', added to the classification loss. These losses are then used during backpropagation to train the model for better generalization across all FSTs. The ultimate goal of this architecture is to improve the accuracy and robustness of the classification model.

C. EVALUATION

1) METRICS

There are many reasons why evaluating fairness in AI systems is essential. One reason is that AI systems are increasingly used to make decisions that can significantly impact people's lives. Another reason is the potential for these systems to amplify existing societal biases. For example, if an AI system is trained on biased data, it may learn to perpetuate these biases.

An approach to evaluate fairness is to use statistical measures, such as comparing the proportion of individuals from different groups treated the same by the AI system.

Haas [31] and Gardner et al. [32] have created a standard set of metrics that can be used to evaluate fairness in Table 4

TABLE 4. List of proposed fairness metrics provided by the Haas [31] and Gardner et al. [32].

Fairness Metrics	Definition
Disparate impact	The ratio of statistical parity between classes is close to 1.
Equalized odds	When both true positive and false positive rates are the same between different classes.
Predictive rate parity	Classes have the same positive predicted value.
Absolute Between-ROC Area	The difference between the positive and negative Receiver Operating Characteristic (ROC) curves.

2) POST-HOC XAI TECHNIQUES

Evaluating the performance of a model using metrics such as the area under the Receiver Operating Characteristic Curve (AUROC) only tells part of the story. These quantitative measures must fully capture the complex and nuanced ways bias can manifest in a models predictions.

To address this gap, recent studies [33] have turned to post-hoc explainability techniques such as Grad-CAM to gain deeper insights into the decision-making processes of deep learning models. Grad-CAM provides a visual explanation of the models predictions by highlighting the regions of an image that the model used to make its decision. By examining the explanations generated by Grad-CAM, researchers can better understand how biases in the data or the models architecture may affect its predictions.

To understand the underlying reasons behind any biases found in the DL models, we use post-hoc XAI techniques. In this study, we use the off-the-shelf library implementations of Grad-CAM. This popular post-hoc XAI technique generates heatmaps that highlight the regions of the image that contribute most to the models prediction. Although it provides some good benefits, such as debugging modes, there are still doubts that XAI should be involved with the highly impactful decision-making of a critical system such as health care.

IV. RESULTS

To evaluate the performance and fairness of the models, all results are based on binary classification between benign and malignant. We use the AUC metric to measure the model's performance as it provides a reliable measure of classification performance when the data is imbalanced [34]. To ensure that the model's fairness is evaluated across different FSTs,

we will aggregate the binary classification results into buckets based on the FST. This approach allows us to analyze the model's performance and fairness across different FST groups, providing insights into how well the model generalizes across different skin types.

The performance of the deep learning models was assessed and divided into five distinct sections, which are outlined below:

1) **Ablation Evaluation** - The ablation study was conducted on the EfficientNet-B2 model, focusing on optimizing hyperparameters, including batch size, optimizer, and learning rate. The results from this analysis are presented in this section.
2) **Model Performance Evaluation** - We evaluated the overall performance of the deep learning models by examining the AUC. This metric provides insight into the models' discriminative ability and overall effectiveness in medical image analysis.
3) **Fitzpatrick Skin Type Model Performance** - In this section, we assessed the performance of the models using AUC, aggregated by the FST. This analysis aimed to determine whether the models' performance varied significantly across different skin types.
4) **Fairness Metrics** - We applied specific fairness metrics to evaluate the overall fairness of the models. The FSTs were divided into two groups, enabling us to investigate the degree to which model performance was equitable across different skin types.
5) **XAI Evaluation** - XAI techniques were assessed to determine whether the models' predictions were based on appropriate regions of the input images. This evaluation provided insights into the transparency and interpretability of the models, which are crucial factors in the context of medical image analysis.

By aggregating the results in this manner and conducting multiple evaluations, we can ensure the reliability and robustness of our proposed approach and promote the development of more accurate and fair models for skin lesion diagnosis.

A. ABLATION EVALUATION

The ablation study was divided into 3 different case studies, batch selection, optimizer selection, and learning rate selection. The evaluation criteria used the highest AUC to move on to the subsequent case study. The AUC results are from the testing dataset. Once a case study was completed, that hyperparameter was not changed for the remainder of the case studies.

1) ABLATION STUDY 1: BATCH SIZE SELECTION

In this study, we evaluated the performance of the EfficientNet-B2 model using different batch sizes with a fixed Adam optimizer with a fixed learning rate of 0.0005. Specifically, we tested batch sizes of 16, 24, and 32, and found that the model achieved the highest AUC of 0.938 when trained with a batch size of 32, as shown in Table 5.

TABLE 5. Case Study 1: Batch size Model performance comparison.

Model	Batch size	Optimizer	Learning Rate	AUC
EfficientNet B2	16	Adam	0.0005	0.908
EfficientNet B2	24	Adam	0.0005	0.933
EfficientNet B2	32	Adam	0.0005	**0.938**

2) ABLATION STUDY 2: OPTIMIZER SELECTION

In this study, we evaluated the performance of the EfficientNet-B2 model using different optimizers with a batch size of 32 with a fixed learning rate of 0.0005. Specifically, we tested the Adam, NAdam, Adamax, and AdamW optimizes, and found that the model achieved the highest AUC of 0.958 when trained with the NAdam optimizer, as shown in Table 6.

TABLE 6. Case Study 2: Optimizer Model performance comparison.

Model	Batch size	Optimizer	Learning Rate	AUC
EfficientNet B2	32	Adam	0.0005	0.938
EfficientNet B2	32	AdamW	0.0005	0.952
EfficientNet B2	32	Adamax	0.0005	0.951
EfficientNet B2	32	NAdam	0.0005	**0.958**

3) ABLATION STUDY 3: LEARNING RATE SELECTION

In this study, we evaluated the performance of the EfficientNet-B2 model using learning rates with a fixed NAdam optimizer with a fixed batch size of 32. Specifically, we tested learning rates of 0.0001, 0.0005, 0.001, 0.0015, and 0.002, and found that the model achieved the highest AUC of 0.958 when trained with a learning rate 0.0005, as shown in Table 7.

TABLE 7. Case Study 3: Learning Rate Model performance comparison.

Model	Batch size	Optimizer	Learning Rate	AUC
EfficientNet B2	32	NAdam	0.0001	0.942
EfficientNet B2	32	NAdam	0.0005	**0.958**
EfficientNet B2	32	NAdam	0.001	0.936
EfficientNet B2	32	NAdam	0.0015	0.931
EfficientNet B2	32	NAdam	0.002	0.913

Following the completion of all three ablation studies, we selected a batch size of 32, the NAdam optimizer, and a fixed learning rate of 0.0005 for subsequent model testing across different training types and architectures. These parameters were found to yield the best performance across the range of experiments conducted.

B. MODEL PERFORMANCE EVALUATION

The performance of the models was evaluated using AUC and tested based on the baseline transfer learning, joint regularization, and synthetic blend methods. Table 8 and Fig. 12 show all the results for each model for each training type.

The results of the experiments demonstrate that all of the models exhibited improved performance upon the

TABLE 8. Performance metrics for all models.

Training type	Model	AUC
Baseline	ResNet50	0.800
	EfficientNet - B2	0.845
	EfficientNet - B4	**0.858**
Regularization	ResNet50	0.840
	EfficientNet - B2	**0.881**
	EfficientNet - B4	0.873
Synthetic blend	ResNet50	0.896
	EfficientNet - B2	**0.901**
	EfficientNet - B4	0.900

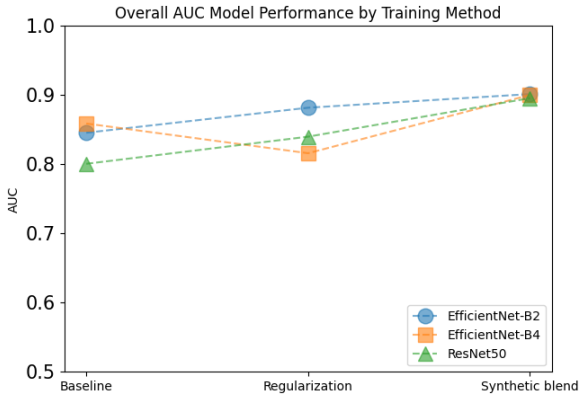

FIGURE 12. The figure depicts a line graph that illustrates each models Area Under the Curve (AUC) results, based on the employed training method.

implementation of joint regularization and synthetic blending methodologies, with the exception of a minor decline in performance observed in the case of EfficientNet-B4 concerning the synthetic blending technique. Among the models examined, EfficientNet-B2 emerged as the highest-performing model with an AUC score of 0.845 at baseline, 0.881 with joint regularization, and 0.901 with synthetic blending. In contrast, ResNet50 exhibited comparatively poorer performance across all training methods. These findings hold significant implications for the development and optimization of machine learning models for AUC assessment.

C. FITZPATRICK SKIN TYPE MODEL PERFORMANCE

Table 9 displays the results of the AUC metrics for each type of FST under the different training types. The AUC metric measures the model's ability to distinguish between the positive and negative samples. Based on the table, it is evident that the AUC scores differ significantly between the different types of FSTs.

For the Baseline training type, the EfficientNet-B2 model performed best in Type 1. The model also performed relatively well in Type 3 and Type 4. However, the model's performance declined significantly in Type 6. This decline in performance can be attributed to the inherent differences in the types of FSTs, which present unique challenges for machine learning models.

For the Regularization training type, the EfficientNet-B2 model showed significant improvements in performance across all types of FSTs, except Type 2. Specifically, the model showed the best performance in Type 1, representing a substantial improvement over the Baseline training type. The model's performance in Type 5 and Type 6 also improved. This improvement can be attributed to the regularization method, which provides a way to penalize the model's performance when it is not performing well on transformed images.

The EfficientNet-B2 model showed the best overall performance for the Synthetic blend training type in Type 1, Type 2, and Type 3. The model also showed a significant improvement in performance in Type 4. However, the model performance in Type 5 and Type 6 declined compared to the Regularization training type. The performance improvement can be attributed to the synthetic blend, which provided more samples of darker images, thereby improving the model's ability to learn diverse images.

The results indicate that the EfficientNet-B2 model is the best-performing model across all types of FSTs under the three different training types. However, the model's performance can be further improved with more training data. Moreover, the results indicate that the regularization and synthetic blend training types can significantly improve the models performance, thereby providing better AUC scores for almost all FSTs.

TABLE 9. AUC metrics for each FST comparing different training types for the EfficientNet - B2 model.

Training type	Type 1	Type 2	Type 3	Type 4	Type 5	Type 6
Baseline	0.959	0.818	0.920	0.907	0.897	0.650
Regularization	0.991	0.886	0.930	0.915	0.936	**0.668**
Synthetic blend	**0.997**	**0.910**	**0.959**	**0.942**	**0.938**	0.665

D. FAIRNESS METRICS

In this study, the fairness of the deep learning models was assessed using a set of commonly used fairness metrics, as shown in Table 10. The fairness metrics were computed for the top-performing models, which were ResNet50, EfficientNet B2, EfficientNet B4.

Table 10 presents the results of an empirical evaluation of different models trained with different methods. The metrics used to evaluate the models are disparate impact, equalized odds, predictive rate parity, and the Absolute Between-ROC Area (ABROCA). The training types evaluated are baseline, regularization, and synthetic blend.

Regarding Disparate Impact, the models showed varying levels of fairness, with values ranging from 0.711 for ResNet50 baseline to 1.000 for the EfficientNet B2 training with the synthetic blend. A value of 1 represents perfect fairness, so values above and below 1 indicate a degree of disparate impact. For Equalized Odds, the values ranged from 0.009 for EfficientNet B2 synthetic blend trained to .189 for EfficientNet B2 trained with Regularization. A value of 0 represents perfect fairness, so higher values indicate that

TABLE 10. Fairness metrics.

Training type	Model	Disparate Impact	Equalized Odds	Predictive Rate Parity	Absolute Between-ROC Area
Baseline	ResNet50	0.711	0.153	0.203	0.156
	B2	0.734	0.117	**0.000**	0.097
	B4	0.927	0.036	0.018	0.103
Regularization	ResNet50	0.714	0.162	0.025	**0.038**
	B2	0.949	0.036	0.051	0.111
	B4	0.740	0.189	0.040	0.052
Synthetic blend	ResNet50	0.871	0.090	0.023	0.115
	B2	**1.000**	**0.009**	0.023	0.115
	B4	0.987	0.090	0.018	0.104

the model is less fair. Predictive Rate Parity was calculated for all models, ranging from 0 for EfficientNet B2 baseline trained to 0.051 for EfficientNet B2 regularization trained. A value of 0 represents perfect fairness, so higher values indicate a less fair model. Finally, the ABROCA metric was computed, ranging from 0.038 for ResNet50 regularization to 0.156 for ResNet50 baseline. A zero value for ABROCA indicates an equally fair treatment of all groups, while a larger value corresponds to a reduction in fairness.

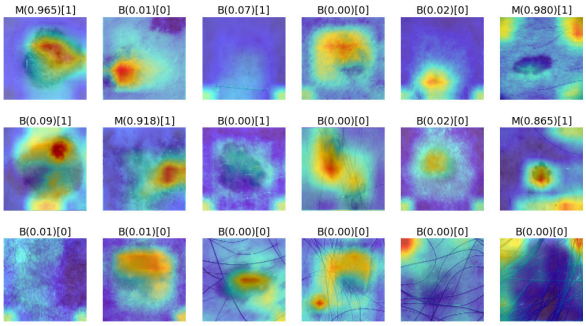

FIGURE 13. Sample of Grad-CAM on the EfficientNet-B2 model trained with synthetic images. Each image has an M and B representing model classification where M is malignant and B is benign. The number in the parenthesis represents the probability and the number in the square bracket represents the label where 1 is malignant, and 0 is benign.

E. XAI EVALUATION

The XAI approach of using Grad-CAM provides a valuable means of assessing model feature importance, as exemplified in the present study through the analysis of sample images depicted in Fig. 13. The results of this analysis demonstrate that the model assigns significant importance to the presence of hair, as evidenced by the last row of columns 3, 4, 5, and 6. Additionally, the model correctly identifies the skin lesion as a crucial feature in the images depicted in row 1, column 1, and row 2, column 2, accurately classifying them as malignant. In contrast, the model recognizes the presence of a dark corner in row 1, column 6, despite its lack of relevance to the skin. Although the model correctly classifies the image in row 1, column 3, it displays a degree of importance towards the ruler in the image, which is not a relevant feature. Similarly, in row 2, column 1, the model inaccurately classifies the lesion as benign. At the same time, the activation map highlights the lesion presented as a significant feature, displaying a deep red area on the lesion.

V. DISCUSSION

In dermatology, specifically skin lesion classification, the issue of bias is a critical concern because human health is involved. We must ensure that ML models are trained and tested on diverse and representative data to prevent potential patient harm. Awareness of bias in these models is crucial in ensuring they are fair and accurate. As depicted in Fig. 7, a substantial proportion of the FST 1 class was observed compared to the other FST classes, particularly FST 3, 4, and 5, which showed the lowest count. Given that the FST classification is not a flawless method, it is hypothesized that some misclassifications may have occurred, such as instances where hair or other artifacts negatively impacted the accuracy of the classification. Still, we are focused on melanoma classification.

From the results in Section IV, we discuss the same structure as follows.

- The choice of EfficientNet-B2 as the best model can be attributed to its ability to strike a good balance between model complexity and size. It is possible that with more training data, other models could perform better as well. Notably, the performance of the models varied across the different training methods, with the synthetic blend method yielding the best results, followed by the regularization method and then the baseline method. Compared to the baseline, the marginal improvement in performance observed with the regularization method can be attributed to its ability to penalize the model when it performs poorly against transformed images. This encourages the model to learn more robust features and reduces the impact of noise or artifacts in the training data. In contrast, the synthetic blend method performed the best among the three training methods. This can be attributed to the fact that it provided more samples of darker images, thereby offering a better balance between light and dark images and presenting more opportunities for the model to learn from a diverse range of images. Therefore, the synthetic blend approach helps reduce bias in the training data and promotes greater generalization in the models predictions.
- The top models from Section IV-C were selected for further analysis, where the data was aggregated based on the FST. The results indicated that a combination of joint regularization and synthetic blend training yielded the best overall performance. While the regularization training method performed best for the Type 6 skin type, the synthetic blend training method proved more effective for Type 1-5 skin types. Notably, the Synthetic Blend training method produced the greatest improvement in performance for Type 1 and Type 4 skin types, with a difference of 0.038 and 0.035, respectively, when compared to the baseline training method. These findings underscore the importance of considering joint regularization and synthetic blend training when evaluating the performance of skin type classification models. Moreover, the results suggest that the observed distribution of

skin types, as depicted in Fig. 7, in the dataset may significantly impact the performance of the classification models. Specifically, the misclassification of Type 6 skin types may be attributed to the unique characteristics of this skin type, which poses a challenge for all models.

- The fairness metric results from Table 10 suggest that different training methods and models can have a significant impact on the fairness and performance of machine learning models. While the baseline models showed relatively poor fairness performance, the EfficientNet-B2 model trained with baseline methods achieved the best predictive rate parity. The regularization models improved fairness, with the EfficientNet-B2 model achieving the best balance among different metrics. The synthetic blend models resulted in the best fairness performance, with the EfficientNet-B2 achieving the best balance among different metrics.
- The Grad-CAM approach allows for the selective evaluation of specific layers, providing ML researchers with the opportunity to focus on retraining or removing layers that may contribute to model-related issues. The highlights observed in the sample images depicted in Fig. 13, such as the negative impact of hair, rulers, and dark corners on the models' performance, could guide researchers in preprocessing the data before testing. Modifying the model to avoid these types of artifacts in the image could enhance its performance. Thus, Grad-CAM capability of selective layer evaluation could facilitate the identification of critical features for model performance and provide a roadmap for future improvements.

The utilization of regularization techniques in the training of machine learning models is an effective method for mitigating overfitting and improving model performance. In the case of the current study, regularization was applied to the images used for model training to diversify the sample and provide a more comprehensive representation of the underlying distribution. This diversity is crucial in skin condition classification, as skin types can vary widely, and the model must generalize effectively to make accurate predictions. The use of regularization in this context is beneficial as it helps the model learn from a more diverse range of skin types, enabling it to generalize its understanding of the underlying patterns and relationships.

VI. CONCLUSION

This study explored the integration of post-hoc explainability techniques, specifically Grad-CAM, into the assessment of skin lesion classifiers. The results demonstrate that incorporating such techniques can provide a more comprehensive understanding of the biases and limitations of these models, thus ensuring fairness and accuracy in their predictions.

This work is essential in developing fair and accurate skin lesion classification models. Our results offer valuable insights into the current biases present in DL models and suggest ways for improvement. Additionally, the study is relevant to the broader field of machine learning and its applications in medical domains, where fairness and accuracy are important.

In summary, the key contributions of this work reported are as follows:

- *Color skin transformer* - Develop an algorithm to convert an image between different FSTs
- *Apply regularization & data augmentation to improve fairness* - Used regularization and data augmentation through skin transformation to improve fairness in the disease classifier by reducing the variance in predictions based on skin type.
- *Evaluation of different models* - Evaluated the performance of several deep learning models on a dataset of images with different FSTs, and determined the best performing model.
- *Improved AUC* - Achieved improved AUC in disease classification by using the best performing model and incorporating the color skin transformer and regularization.

For future work, several potential improvements to these models have been identified. Firstly, augmenting the training data with additional synthetic data generated using GANs may enhance the models ability to learn and identify underrepresented classes. Secondly, exploring alternative loss functions may improve fairness in the model predictions. Additionally, pre-processing steps, such as removing hair from images before computing the FST may yield more accurate results. The color transformer could be fine-tuned for better visual performance for adjusting FST greater than 2 steps. Also, we could consider evaluating different layers using Grad-CAM to understand if the layers need to be modified or removed to help improve fairness performance. If desired, adjusting threshold values based on the AUROC graphs may improve these models' overall accuracy.

REFERENCES

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

[2] N. M. Kinyanjui, T. Odonga, C. Cintas, N. C. F. Codella, R. Panda, and P. Sattigeri, "Fairness of classifiers across skin tones in dermatology," in *Medical Image Computing and Computer Assisted Intervention—MICCAI* (Lecture Notes in Computer Science), A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, Eds. Cham, Switzerland: Springer, Sep. 2020, pp. 320–329.

[3] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. Badri, "Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset," 2021, *arXiv:2104.09957*.

[4] H. Kim, G. A. Tadesse, C. Cintas, S. Speakman, and K. Varshney, "Out-of-distribution detection in dermatology using input perturbation and subset scanning," 2021, *arXiv:2105.11160*.

[5] A. Pakzad, K. Abhishek, and G. Hamarneh, "CIRCLe: Color invariant representation learning for unbiased classification of skin lesions," 2022, *arXiv:2208.13528*.

[6] R. Daneshjou, M. P. Smith, M. D. Sun, V. Rotemberg, and J. Zou, "Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review," *JAMA Dermatol.*, vol. 157, no. 11, pp. 1362–1369, Nov. 2021. [Online]. Available: https://jamanetwork.com/journals/jamadermatology/fullarticle/2784295

[7] P. Tschandl, "Risk of bias and error from data sets used for dermatologic artificial intelligence," *JAMA Dermatol.*, vol. 157, no. 11, pp. 1271–1273, Nov. 2021, doi: 10.1001/jamadermatol.2021.3128.
[8] R. Daneshjou, K. Vodrahalli, W. Liang, R. A. Novoa, M. Jenkins, V. Rotemberg, J. Ko, S. M. Swetter, E. E Bailey, O. Gevaert, P. Mukherjee, M. Phung, K. Yekrang, B. Fong, R. Sahasrabudhe, J. Zou, and A. Chiou, "Disparities in dermatology AI: Assessments using diverse clinical images," 2021, *arXiv:2111.08006*.
[9] T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types I through VI," *Arch. Dermatol.*, vol. 124, no. 6, pp. 869–871, Jun. 1988.
[10] R. Glogau and S. Matarasso, "Chemical face peeling: Patient and peeling agent selection," *Facial Plastic Surg.*, vol. 11, no. 1, pp. 1–8, Jan. 1995, doi: 10.1055/s-2008-1064510.
[11] W. E. Roberts, "Skin type classification systems old and new," *Dermatol. Clinics*, vol. 27, no. 4, pp. 529–533, Oct. 2009. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0733863509000540
[12] L. Baumann, "Understanding and treating various skin types: The Baumann Skin Type Indicator," *Dermatol. Clinics*, vol. 26, no. 3, pp. 359–373, Jul. 2008.
[13] D. Wen, S. M. Khan, A. J. Xu, H. Ibrahim, L. Smith, J. Caballero, L. Zepeda, C. de B. Perez, A. K. Denniston, X. Liu, and R. N. Matin, "Characteristics of publicly available skin cancer image datasets: A systematic review," *Lancet Digit. Health*, vol. 4, no. 1, pp. E64–E74, Jan. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2589750021002521
[14] A. Chardon, I. Cretois, and C. Hourseau, "Skin colour typology and suntanning pathways," *Int. J. Cosmetic Sci.*, vol. 13, no. 4, pp. 191–208, Aug. 1991, doi: 10.1111/j.1467-2494.1991.tb00561.x.
[15] J. Yap, W. Yolland, and P. Tschandl, "Multimodal skin lesion classification using deep learning," *Exp. Dermatol.*, vol. 27, no. 11, pp. 1261–1267, Nov. 2018, doi: 10.1111/exd.13777. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/exd.13777
[16] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 96–108, Nov. 2017. [Online]. Available: http://ieeexplore.ieee.org/document/8103116/
[17] J. Höhn et al., "Integrating patient data into skin cancer classification using convolutional neural networks: Systematic review," *J. Med. Internet Res.*, vol. 23, no. 7, Jul. 2021, Art. no. e20708. [Online]. Available: https://www.jmir.org/2021/7/e20708
[18] A. T. Young, M. Xiong, J. Pfau, M. J. Keiser, and M. L. Wei, "Artificial intelligence in dermatology: A primer," *J. Investigative Dermatol.*, vol. 140, no. 8, pp. 1504–1512, Aug. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0022202X2031201X
[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," 2011, *arXiv:1106.1813*.
[20] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw., IEEE World Congr. Comput. Intell.*, Jun. 2008, pp. 1322–1328. [Online]. Available: http://ieeexplore.ieee.org/document/4633969/
[21] V. Parthipan, "Image down-scaler using the box filter algorithm," M.S. theses, Rochester Inst. Technol., Dept. Elect. Microelectron. Eng., Rochester, NY, USA, 2017. [Online]. Available: https://scholarworks.rit.edu/theses/9704
[22] V. Rotemberg et al., "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Sci Data*, vol. 8, p. 34, Jan. 2021, doi: 10.1038/s41597-021-00815-z.
[23] A. Corbin and O. Marques, "Exploring strategies to generate Fitzpatrick skin type metadata for dermoscopic images using individual typology angle techniques," *Multimedia Tools Appl.*, vol. 82, pp. 23771–23795, Nov. 2022.
[24] C. Deotte. (2020). *SIIM-ISIC Melanoma Classification—Triple Stratified KFold*. [Online]. Available: https://kaggle.com/competitions/siim-isic-melanoma-classification
[25] W. Badr. (Dec. 2020). *Having an Imbalanced Dataset? Here Is How You Can Fix It.* [Online]. Available: https://towardsdatascience.com/having-an-imbalanced-dataset-here-is-how-you-can-solve-it-1640568947eb
[26] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," *J. Big Data*, vol. 5, no. 1, p. 42, Dec. 2018. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-018-0151-6
[27] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, p. 60, Dec. 2019. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0197-0
[28] D. Jha, M. A. Riegler, D. Johansen, P. Halvorsen, and H. D. Johansen, "DoubleU-Net: A deep convolutional neural network for medical image segmentation," 2020, *arXiv:2006.04868*.
[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[30] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html
[31] C. Haas, "The price of fairness—A framework to explore trade-offs in algorithmic fairness," in *Proc. 40th Int. Conf. Inf. Syst. (ICIS)*. Atlanta, GA, USA: Association for Information Systems, 2019, pp. 1–18.
[32] J. Gardner, C. Brooks, and R. Baker, "Evaluating the fairness of predictive student models through slicing analysis," in *Proc. 9th ACM Int. Conf. Learn. Anal. Knowl.*, 2019, pp. 225–234, doi: 10.1145/3303772.3303791.
[33] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
[34] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016786550500303X

ADAM CORBIN received the bachelor's, master's, and Ph.D. degrees in computer engineering from Florida Atlantic University (FAU), Boca Raton, FL, USA, in 2011, 2012, and 2023, respectively.

From 2012 to 2022, he was a Software Engineer with GE Aerospace, where he primarily focused on developing flight management systems. Since 2013, he has been an Adjunct Professor with the Saint Petersburg State College, teaching introduction to software courses. In 2022, he transitioned to a role as a Software Manager with Fivetran. His research interests include examining the fairness and bias of machine learning models, particularly within dermatology applications.

OGE MARQUES (Senior Member, IEEE) received the master's degree in electronic engineering from the Philips International Institute, The Netherlands, in 1989, and the Ph.D. degree in computer engineering from Florida Atlantic University (FAU), Boca Raton, FL, USA, in 2001.

He has more than 35 years of teaching experience in different countries (USA, Austria, Brazil, The Netherlands, Spain, France, and India). He is currently a Professor of computer science and engineering with the College of Engineering and Computer Science, a Professor of biomedical science (secondary) with the Charles E. Schmidt College of Medicine, and a Professor of information technology (by courtesy) with the College of Business, FAU. His research interests include image processing, medical image analysis, computer vision, human vision, data science, artificial intelligence, and machine learning. He is the author of 11 technical books, one patent, and more than 100 refereed scientific articles in his fields of expertise.

Dr. Marques is a Senior Member of the Association for Computing Machinery (ACM). He is also a Tau Beta Pi Eminent Engineer and a member of the Honor Societies of Sigma Xi, Phi Kappa Phi, and Upsilon Pi Epsilon. He is also a fellow of the NIH AIM-AHEAD Consortium and a fellow of the Leshner Leadership Institute, American Association for the Advancement of Science (AAAS).