Beyond Accuracy: A Cost-Aware Approach to Skin Lesion Detection Across Skin Tone Imbalances

1st Md Mohit Hasan
Dept. of Computer Science and Engineering)
Northern University Bangladesh
Dhaka, Bangladesh
hasan_41220200217@nub.ac.bd

2nd Mahbuba Tasnime Suchi
Dept. of Computer Science and Engineering
Northern University Bangladesh
Dhaka, Bangladesh
suchi_41210301603@nub.ac.bd

3rd Md Hasibul Habib
Dept. of Computer Science and Engineering
Northern University Bangladesh
Dhaka, Bangladesh
habib_41210301601@nub.ac.bd

4th Sumya Akter
Dept. of Computer Science and Engineering
Northern University Bangladesh
Dhaka, Bangladesh
sumya.akter@nub.ac.bd

5th Zarin Tasnim Rothy
Dept. of Computer Science and Engineering
Northern University Bangladesh
Dhaka, Bangladesh
Rothi_NUBCSE@nub.ac.bd

6th A.M.Tayeful Islam
Dept. of Computer Science and Engineering
Northern University Bangladesh
Dhaka, Bangladesh
Tayef_NUBCSE@nub.ac.bd

7th Tanmoy Sarkar Pias
Dept. of Computer Science
Virginia Tech
Blacksburg, USA
tanmoysarkar@vt.edu

8th David Eisenberg
Dept. of Information Management and Business Analytics
Montclair State University
New Jersey, USA
eisenbergd@montclair.edu

9th Simon Bin Akter*
Dept. of Computer Science and Engineering
Northern University Bangladesh
Dhaka, Bangladesh
simon.akter@nub.ac.bd

***Abstract*—Skin lesion prediction using artificial intelligence (AI) models is highly dependent on skin tone, yet current approaches largely overlook this critical factor. The Fitzpatrick 17k dataset, which contains six skin tone categories: lighter to darker, is severely imbalanced, with most models biased toward lighter skin tones. Previous efforts to improve overall accuracy fall short: overall accuracy fails to reflect true performance across imbalances. This creates a significant gap, as effective skin lesion detection must work across all skin tones, not just a few. To address this, we introduce the Cost-Aware EfficientNet (CAEN) model, combining cost-sensitive learning (CSL) and attention mechanisms to tackle imbalanced data and ensure the model generalizes well across all skin tones with detailed interpretability. Rather than simply improving accuracy, our model enhances class-specific performance, achieving 79% recall for non-neoplastic, 88% for benign, and 80% for malignant lesions. This indicates an overall improvement in darker tones of approximately 44.86% compared to state-of-the-art results from prior studies. Furthermore, it remains robust across augmented test conditions, such as changes in brightness, contrast, blur, and zoom, providing balanced outcomes for diverse skin tones. This novel approach offers a significant leap toward fair and reliable skin lesion prediction for all skin tones with interpretability.**

***Index Terms*—Skin Disease, Fitzpatrick 17k, Imbalance, Cost-Sensitive Learning, Attention Mechanisms.**

I. INTRODUCTION

Skin lesions affect individuals across all skin tones, but diagnosing them can be difficult due to variations in pigmentation [1]. The Fitzpatrick 17k [2] dataset provides a valuable resource for classifying skin diseases using artificial intelligence (AI) [3]–[5] approaches, as it covers a diverse range of skin tones [6]–[9]. However, a major challenge in applying AI to this dataset is the imbalance among different skin tone classes [10]. Ensuring accurate diagnosis for all skin tones is critical, as biased AI models can lead to incorrect diagnoses, especially for underrepresented skin tones [1], [9], [10].

Previous studies using the Fitzpatrick 17k dataset have applied AI models for skin lesion classification, often using accuracy as the main performance measure [1], [6]–[9]. However, this metric can hide how much better these models perform on majority classes (more common skin tones) compared to minority classes (less represented skin

* Corresponding Author

tones), leading to biased results [11]–[13]. While augmentation techniques were applied, they focused on improving overall accuracy and did not clearly show whether the models were effective at identifying minority skin disease classes across different skin tones [1], [6]–[9]. Prior studies [7]–[10] did not explore how models trained on one skin tone performed against others, missing a clearer analysis of bias. Additionally, these studies [7]–[10] showed significantly lower performance on darker tones compared to lighter ones. It raises questions about the generalizability of these models across different skin tones, highlighting a potential gap in understanding their applicability and effectiveness across diverse populations. Furthermore, these studies [1], [6]–[10] did not address the impact of image quality during testing, such as variations in brightness, contrast, blur, and zoom, which could affect model performance.

The need for accurate diagnosis across all skin tones makes it essential to address the limitations of prior approaches [1], [6]–[10]. The reliance on overall performance metrics does not offer a reliable measure of fairness in class-wise predictions, especially when dealing with imbalanced datasets like Fitzpatrick 17k [12], [13]. These studies also fail to demonstrate the model's generalizability and robustness across different skin tones, as indicated by their poor performance on darker skin tones [7]–[10]. Therefore, it is essential to explore how imbalances can be more carefully addressed, improve fairness, and ensure more equitable predictions across all skin tones.

This research addresses and improves upon key questions, as outlined below.

1) Are AI algorithms biased when detecting skin lesions across different skin tones?
2) Can we rely on accuracy metrics when the dataset is imbalanced, like in the Fitzpatrick 17k skin tone dataset, despite prior research [1], [6]–[9] efforts to improve accuracy for each skin tone?
3) Can a model trained on lighter skin tones accurately predict skin lesions on darker skin tones and vice versa?

This research addresses the class imbalance in the Fitzpatrick 17k [2] dataset by exploring various modeling approaches combined with augmentation techniques. The primary contribution is the evaluation of class-wise performance and its improvement instead of just overall metrics, alongside the introduction of a novel Cost-Aware EfficientNet (CAEN) model to effectively handle class imbalance. CAEN, based on EfficientNet architecture, incorporates dynamic cost-sensitive learning (CSL) [14] and attention mechanism [15], fine-tuned to predict skin lesions, including **i. non-neoplastic**, **ii. benign**, and **iii. malignant**, across six skin tones from light to dark. The study discusses the issue of generalizability and robustness by training the proposed model on lighter skin tones and testing it on darker tones, and vice versa. It also ensures generalizability and robust, reliable performance across all skin tones through the proposed model. Additionally, we have tested our proposed model across different augmented samples from the test set, varying brightness, contrast, blur, and zoom, highlighting the importance of fairness in AI models to ensure accurate predictions for all skin tones. Our study also includes detailed interpretability [3], [11], [16] of its predictions across different skin tones, ensuring transparency in how the model makes decisions for diverse populations.

The findings of this research have significant implications for dermatology by enhancing class-wise accuracy in skin lesion classification across diverse skin tones. This work promotes more equitable diagnostic methods and serves as a valuable reference for future researchers aiming to select strategies that mitigate bias in predictions [1] across different skin tones.

A. Related Works and Their Limitations

Previous studies [17], [18] on skin disease prediction have often overlooked how well models perform across different skin tones. Although datasets like Fitzpatrick 17k aim to address skin tone representation, prior studies [1], [6]–[10] have not effectively demonstrated how model improvements vary by skin tone. Many studies [1], [6]–[9] rely on overall accuracy metrics, which can be misleading, reflecting high performance on majority classes while neglecting minority groups [12], [13].

In contrast, our study introduces a new CAEN modeling technique that shows clear class-wise recall improvements in skin disease prediction for all skin tones. For instance, state-of-the-art recalls from prior research [10] for benign and malignant classes across light skin tones improved from 0.52 and 0.73 to **0.86** and **0.75**. For moderate skin tones, these classes improved from 0.60 and 0.65 to **0.88** and **0.84**, and for dark skin tones, from 0.55 and 0.45 to **0.88** and **0.86**. These studies [7]–[10] consistently showed lower performance for minority classes, such as benign and malignant, on darker tones. However, we have significantly improved performance on darker tones for these minority classes. Further, we validate the generalizability and robustness of our model by testing it on a diverse set of augmented samples with varying image qualities, ensuring balanced performance, which has been lacking in previous research.

II. MATERIALS AND METHODS

We utilized the Fitzpatrick 17k [2] dataset, which includes six different skin tones, ranging from lighter to darker, to predict skin lesions using AI approaches as detailed in **Fig 1**. To address biases caused by class imbalances and limited samples from specific skin tones, we fine-tuned a novel CAEN modeling technique. We compared our proposed model with several existing techniques, analyzing class-wise recall performance to identify improvements. We also examined how performance varies by training the proposed model on lighter skin tones and testing it on darker skin tones, as well as the reverse scenario. Additionally, we tested the model's performance under varying conditions, such as changes in brightness, contrast, blur, and zoom in the samples from the test set. Lastly, this study also comprehensively interprets

its predictions across various skin tones. This comprehensive evaluation allowed us to assess the robustness and reliability of our model across different skin tones and image quality variations.

A. Data Description

The Fitzpatrick 17k [2] dataset comprises a diverse collection of skin disease images categorized by six distinct skin tones: **Type I (very fair)**, **Type II (fair)**, **Type III (medium)**, **Type IV (olive)**, **Type V (brown)**, and **Type VI (dark)**. In our study, we utilized a total of 16,012 images from this dataset to predict skin lesions, specifically focusing on three categories: **(i) non-neoplastic**, **(ii) benign**, and **(iii) malignant** lesions. However, it is important to note that the images for different skin lesions are heavily imbalanced, with very limited samples available for certain skin tones represented in **Table I**, posing challenges for accurate model training and evaluation.

TABLE I
DISTRIBUTION OF SKIN LESIONS ACROSS DIFFERENT FITZPATRICK SCALES (SKIN TONES). THE DATA DISTRIBUTION OF BENIGN AND MALIGNANT CASES APPEARS TO BE HIGHLY IMBALANCED COMPARED TO NON-NEOPLASTIC CASES, AND THERE ARE ALSO SIGNIFICANTLY FEWER IMAGES REPRESENTING DARK SHADES.

Fitzpatrick Scale (Skin Tone)	Non-Neoplastic	Benign	Malignant
Type I (very fair)	2050	444	453
Type II (fair)	3395	671	742
Type III (medium)	2377	475	456
Type IV (olive)	2113	367	301
Type V (brown)	1227	159	147
Type VI (dark)	530	44	61

B. Train/Test Formulation

The dataset is divided into a training set and a testing set in an 80:20 split ratio. To ensure balanced training and improve model performance, we implemented data augmentation strategies tailored to each skin tone. Specifically, we conducted augmentation separately for each skin tone to balance the benign and malignant lesion classes against the non-neoplastic class. This process involved applying various transformations, including rotation, brightness adjustment, contrast enhancement, zooming, and blurring, to create a more diverse set of training images. To effectively address the class imbalances within our dataset, we significantly increased the number of benign samples by a factor of 5.55, aligning it more closely with the non-neoplastic class. Similarly, we augmented the malignant samples by a factor of 4.86 to achieve a comparable balance represented in **Table II**. These augmentation ratios were determined through iterative testing of various sampling strategies within our predictive model, allowing us to optimize the balance for each class and improve the overall robustness of our model.

C. Proposed Cost-Aware EfficientNet

The proposed CAEN model represented in **Equation 1** builds upon the EfficientNet architecture, integrating CSL [14] and an attention mechanism [15] to tackle class imbalance and improve model performance. CSL [14] assigns higher weights to underrepresented classes, such as benign and malignant cases, ensuring the model focuses more on accurately detecting these categories that are often overlooked in imbalanced datasets. The attention mechanism [15] allows the model to focus dynamically on the most important areas within an image, improving its ability to capture subtle differences between classes. These improvements make CAEN better at mitigating bias, as it directly addresses the imbalance issue that standard EfficientNet [19] struggles with. Furthermore, CAEN is more effective at developing a model that generalizes well across different skin tones, as the attention mechanism enables the model to focus on key visual patterns rather than being biased toward skin tone variations, which can be limited in certain datasets. This ability to extract relevant features across diverse cases allows CAEN to perform robustly in a wider range of real-world scenarios compared to the standard EfficientNet [19], which tends to struggle with such diversity due to its reliance on the limited sample present for particular scenarios.

TABLE II
COMPARISON OF NON-NEOPLASTIC, BENIGN, AND MALIGNANT COUNTS PRE- AND POST-AUGMENTATION IN TRAIN SET. THE AUGMENTATION PROCESS INVOLVED APPLYING SEVERAL TRANSFORMATIONS TO THE IMAGES, SUCH AS VARYING ROTATION, ADJUSTING BRIGHTNESS, ENHANCING CONTRAST, ZOOMING, AND ADDING BLUR.

Fitzpatrick Scale (Skin Tone)	Non-Neoplastic	Benign (Pre → Post)	Malignant (Pre → Post)
Type I (very fair)	1654	347 → 1884	365 → 927
Type II (fair)	2729	535 → 3108	589 → 2905
Type III (medium)	1901	377 → 2049	366 → 1391
Type IV (olive)	1692	295 → 1604	237 → 1264
Type V (brown)	973	137 → 744	120 → 636
Type VI (dark)	409	35 → 194	48 → 252

$$\text{CAEN}(x) = \sigma\left(w \cdot \text{Att}(f(x)) + b\right) \cdot \text{CSL}(y) \tag{1}$$

The description of the **Equation 1**: Where: x represents the input image, while $f(x)$ denotes feature extraction from the EfficientNet base model. The variable w signifies the weights of the dense layer, and b indicates the bias term. The function σ represents the softmax activation function, which is applied to the output of the attention mechanism, $\text{Att}(f(x))$, that highlights relevant features in the extracted data. Additionally, y represents the true labels, and $\text{CSL}(y)$ denotes the cost-sensitive function that adjusts the loss based on class weights, enhancing the model's ability to address class imbalances effectively.

III. RESULTS AND ANALYSES

First, the proposed CAEN model was trained on images of various skin tones together and compared with existing models. Next, it was trained and tested separately for each skin type, and the model with the best accuracy was compared to the CAEN model. Finally, the proposed model was trained on light skin tones and tested on dark tones, and vice versa to

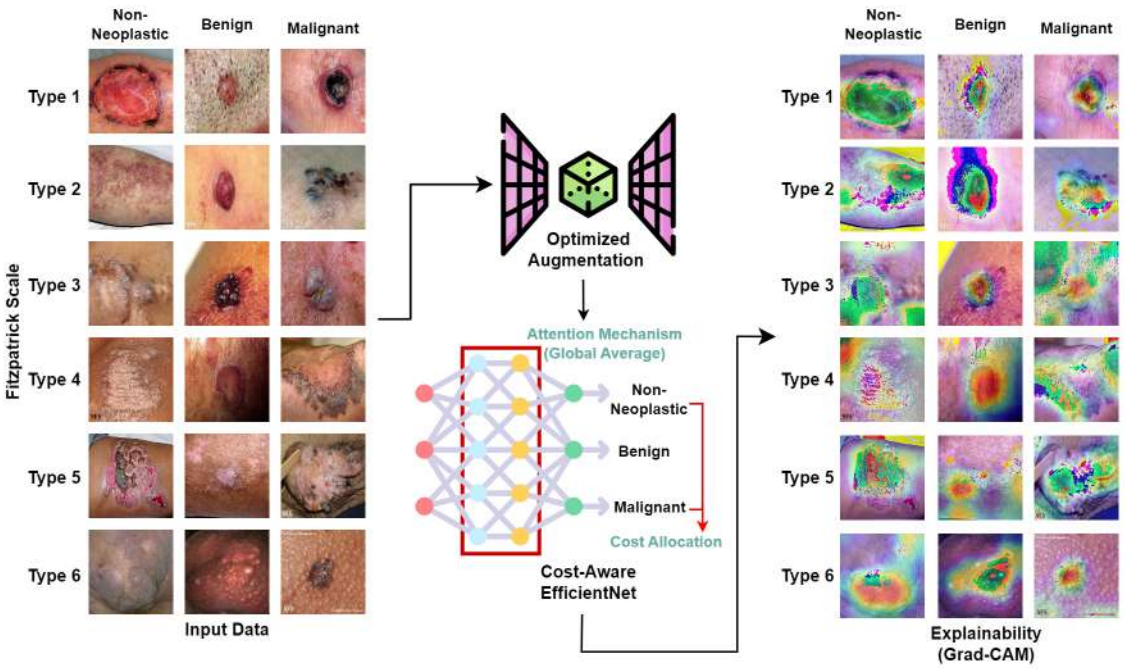

Fig. 1. **Complete workflow of improved skin lesion prediction across heavy skin tone imbalances.** This includes optimized data augmentation, a cost-aware model for imbalance correction, and detailed explainability.

examine how model performance varies. Results are computed ten times each by changing the model hyper-parameters and conventional probability threshold [20], and the standard deviation (SD) [21] for each is calculated to demonstrate variance in results. Finally, our study also offers an in-depth explanation of its predictions for different skin tones.

A. Experimental Results

TABLE III
CLASS-WISE RECALL AND ACCURACY COMPARISON OF THE PROPOSED MODEL WITH EXISTING APPROACHES. THE MODEL IS TRAINED USING DATA FROM ALL SKIN TONES TOGETHER. THE TEST SETS ARE DIVIDED INTO LIGHT (I-II), MODERATE (III-IV), AND DARK (V-VI) SKIN TONES, WITH CLASS-WISE RECALL AND ACCURACY PRESENTED FOR EACH AND OVERALL IN THE LAST SECTION. AUGMENTED INDICATES THAT THE MODEL WAS TESTED ON AN AUGMENTED VERSION OF THE TEST SET, WHICH VARIED BRIGHTNESS, CONTRAST, BLUR, AND ZOOM.

Fitzpatrick Scale (Skin Tone)	Non-Neoplastic	Benign	Malignant	Accuracy (%)
Inception ResNet V2	0.01 ± 0.9	0.01 ± 0.6	0.92 ± 0.4	17.28 ± 0.8
CLIP	0.14 ± 0.8	0.41 ± 0.7	0.07 ± 0.4	19.54 ± 0.9
NASNetLarge	0.74 ± 0.6	0.20 ± 0.7	0.36 ± 0.8	58.26 ± 0.7
Inception V3	0.69 ± 0.5	0.27 ± 0.8	0.38 ± 0.9	55.96 ± 0.8
Xception	0.69 ± 0.7	0.35 ± 0.5	0.37 ± 0.9	58.81 ± 0.7
MobileNet V2	0.57 ± 0.8	0.37 ± 0.7	0.53 ± 0.6	64.20 ± 0.9
ResNet 50	0.88 ± 0.5	0.41 ± 0.6	0.70 ± 0.4	80.64 ± 0.6
VGG 19	0.83 ± 0.6	0.50 ± 0.7	0.62 ± 0.8	76.02 ± 0.7
EfficientNet B0	0.84 ± 0.4	0.50 ± 0.9	0.70 ± 0.7	78.35 ± 0.8
CAEN (Augmented)	0.75 ± 0.3	0.82 ± 0.2	0.75 ± 0.3	75.25 ± 0.3
CAEN (Light)	0.75 ± 0.3	0.86 ± 0.1	0.75 ± 0.2	76.43 ± 0.3
CAEN (Moderate)	0.85 ± 0.2	0.88 ± 0.4	0.84 ± 0.3	84.53 ± 0.2
CAEN (Dark)	0.74 ± 0.4	0.88 ± 0.2	0.86 ± 0.1	84.98 ± 0.2
CAEN (Overall)	0.79 ± 0.4	0.88 ± 0.3	0.80 ± 0.4	80.67 ± 0.3

The **Table III** compares different models for skin lesion classification across three classes: non-neoplastic, benign, and malignant, with accuracy shown for each. Some models, like VGG 19, EfficientNet B0, and ResNet 50, achieved higher overall accuracy than the proposed CAEN model, but their class-wise performance, especially for the minority classes (benign and malignant), was significantly lower. The CAEN model, however, showed balanced and improved performance across all classes. It performed particularly well on dark skin tones, which had fewer samples, while also maintaining good accuracy on light skin tones. Even when tested on augmented data with varying image quality, CAEN continued to perform well, especially on darker tones, while balancing performance for lighter tones.

Previous studies [7]–[10] have conducted methods to improve accuracy for each skin tone outcome. Hence, in this comparison in **Table IV**, models were trained separately for each skin type, and the model with the highest overall accuracy is compared with CAEN. While ResNet 50 achieved higher overall accuracy, its performance for minority classes like benign and malignant was significantly lower. In contrast, CAEN maintained a balanced performance across all classes, indicating that focusing solely on accuracy in imbalanced data can be misleading, and class-wise performance provides a more reliable evaluation.

Previous studies [7]–[10] failed to investigate how models trained on one skin tone performed when applied to others, overlooking a more thorough analysis of potential skin tone bias. In **Table V**, we evaluated the CAEN model by training it on one skin tone and testing it on others to see how generaliza-

TABLE IV

CLASS-WISE RECALL PERFORMANCE COMPARISON OF MODELS TRAINED SEPARATELY BY SKIN TONE. MODELS WERE TRAINED ON DATASETS DIVIDED BY SKIN TONES. THE PROPOSED MODEL IS COMPARED WITH THE ONE ACHIEVING THE HIGHEST ACCURACY TO HIGHLIGHT THE UNRELIABILITY OF ACCURACY METRICS ACROSS IMBALANCED DATA.

Fitzpatrick Scale	ResNet 50				CAEN			
	Non-Neoplastic	Benign	Malignant	Accuracy (%)	Non-Neoplastic	Benign	Malignant	Accuracy (%)
Type 1	0.94 ± 0.7	0.14 ± 0.8	0.57 ± 0.9	76.27 ± 0.6	0.68 ± 0.3	0.66 ± 0.2	0.73 ± 0.4	68.50 ± 0.3
Type 2	0.94 ± 0.5	0.21 ± 0.6	0.58 ± 0.8	78.12 ± 0.7	0.68 ± 0.2	0.57 ± 0.1	0.75 ± 0.4	67.85 ± 0.3
Type 3	0.92 ± 0.8	0.30 ± 0.9	0.52 ± 0.7	77.34 ± 0.4	0.66 ± 0.2	0.70 ± 0.3	0.74 ± 0.4	67.92 ± 0.2
Type 4	0.97 ± 0.9	0.33 ± 0.7	0.33 ± 0.5	82.01 ± 0.8	0.75 ± 0.3	0.78 ± 0.4	0.70 ± 0.2	75.04 ± 0.4
Type 5	0.99 ± 0.4	0.25 ± 0.6	0.07 ± 0.8	83.01 ± 0.9	0.80 ± 0.2	0.68 ± 0.3	0.74 ± 0.4	78.55 ± 0.3
Type 6	1.00 ± 0.8	0.00 ± 0.4	0.33 ± 0.9	87.30 ± 0.5	0.78 ± 0.4	0.78 ± 0.3	0.69 ± 0.1	76.92 ± 0.2

TABLE V

COMPARISON OF RECALL GENERALIZABILITY WHEN TRAINED ON LIGHTER TONES AND TESTED ON DARKER TONES, AND VICE VERSA. THE SKIN TONE ON WHICH THE MODEL IS TRAINED IS MARKED IN BOLD, WHILE PERFORMANCE IS ALSO TESTED ON THE OTHER TWO SKIN TONES. LIGHTER TONES INCLUDE FITZPATRICK TYPES I-II, MODERATE TONES INCLUDE TYPES III-IV, AND DARKER TONES INCLUDE TYPES V-VI.

Skin Tone	Non-Neoplastic	Benign	Malignant	Accuracy
Lighter	0.72 ± 0.4	0.49 ± 0.3	0.70 ± 0.4	68.29% ± 0.3
Moderate	0.71 ± 0.4	0.52 ± 0.3	0.60 ± 0.2	66.67% ± 0.2
Darker	0.74 ± 0.4	0.35 ± 0.4	0.53 ± 0.2	69.73% ± 0.3
Moderate	0.73 ± 0.3	0.61 ± 0.4	0.58 ± 0.4	72.48% ± 0.3
Lighter	0.66 ± 0.4	0.49 ± 0.2	0.63 ± 0.3	65.62% ± 0.3
Darker	0.79 ± 0.4	0.53 ± 0.3	0.53 ± 0.3	78.92% ± 0.4
Darker	0.82 ± 0.4	0.45 ± 0.3	0.62 ± 0.4	79.82% ± 0.3
Lighter	0.61 ± 0.4	0.41 ± 0.3	0.49 ± 0.4	58.92% ± 0.2
Moderate	0.70 ± 0.4	0.32 ± 0.3	0.49 ± 0.2	65.36% ± 0.3

tion varies across different skin tones. From our results, when the model was trained on lighter tones, it performed better on darker tones for non-neoplastic cases, but its performance for benign and malignant cases was lower compared to when it was trained on moderate tones. For models trained on moderate tones, the results varied: they performed better on non-neoplastic cases for darker tones than for lighter tones, but both darker and lighter tones showed lower performance for benign cases. When the model was trained on darker tones, its performance on lighter and moderate tones dropped, especially for non-neoplastic cases on lighter tones and benign cases on moderate tones, which were significantly low.

B. Explainability

Gradient-Weighted Class Activation Mapping (Grad-CAM) [22] in the **Fig 1** highlights affected areas in skin lesions, distinguishing non-neoplastic, benign, and malignant types. While green is prominent, indicating significant regions, orange and yellow mark the most critical areas. Blue and pink/magenta shades primarily indicate less significant regions but still contribute to the lesion assessment. The color intensity reflects how the model identifies skin tone across the Fitzpatrick scale, as seen from the consistent focus on lesion areas across diverse skin tones (Type 1 to Type 6), capturing lesions accurately despite varying pigmentation.

C. Discussion

In this research, we aimed to address key questions regarding the performance of AI algorithms in detecting skin lesions across various skin tones. The following summarizes our findings based on the results presented in **Table III**, **Table IV**, and **Table V**, offering insights into the addressed questions.

Question 1: In our research, we first explored whether AI models are biased when detecting skin lesions on different skin tones. The results in **Table IV** highlight that models trained separately for each skin tone, like ResNet 50, achieved higher overall accuracy but performed poorly on minority classes, particularly for darker skin tones. This indicates a significant bias, as the model tended to favor lighter tones, confirming that AI models can indeed be biased based on skin tone. Additionally, **Table V** illustrates that the performance of the CAEN model varied significantly depending on the skin tone it was trained on. This variation suggests that models can be biased and can not be generalized well across different skin tones, as their effectiveness fluctuates depending on the training set.

Question 2: We questioned the reliability of accuracy metrics in imbalanced datasets, like the Fitzpatrick 17k skin tone dataset. From our comparisons in **Table IV**, it became evident that focusing solely on overall accuracy can be misleading. Although ResNet 50 in **Table IV** had a higher accuracy, its performance on benign and malignant cases was significantly lower, highlighting the limitations of using accuracy metrics alone. This aligns with findings in prior studies [7]–[10] that often fail to adequately address class imbalances, leading to potentially biased conclusions about model effectiveness.

Question 3: Finally, we examined whether a model trained on one skin tone can accurately predict lesions on others. Previous studies [7]–[10] failed to assess model performance when trained on one skin tone and tested on another, missing a critical opportunity to comprehensively evaluate skin tone

bias. Additionally, their improvements mainly benefited light skin tones, failing to generalize across other tones [7]–[10]. This emphasizes the need for a model that delivers consistent performance across all skin tones, as confirmed by the results in **Table V**, which show the presence of skin tone bias. Our study advocates for the CAEN model, as outlined in **Table III**, which trains on all skin tones simultaneously. By utilizing CSL and an attention mechanism, the CAEN model effectively generalizes its performance across diverse skin tones, ensuring better reliability and reducing bias.

IV. LIMITATIONS

The Fatzpartc 17k [2] dataset has very few images of malignant lesions, especially for dark skin tones, making it hard to train a model to accurately classify these lesions on darker skin. Besides, the area of skin disease is often underrepresented, with many images showing only small portions of the affected skin, making it difficult to distinguish between categories, which could be improved by reshaping the images under the guidance of a skin disease professional for better visibility and prediction. Additionally, besides the images, important clinical information, such as patient history and other findings, is essential for accurate diagnosis but is not provided in the dataset.

V. CONCLUSION

In this research, we addressed the bias of AI algorithms in detecting skin lesions across diverse skin tones. Previous research [1], [6]–[10] has mainly focused on increasing the overall accuracy of skin lesion detection across various skin tones. However, our experimentation, as shown in **Table IV**, indicates that AI models often perform better on lighter skin tones, leading to lower accuracy for darker skin tones, even when the overall accuracy appears high. Therefore, improving overall accuracy can not be a reliable strategy for imbalanced datasets like Fitzpatrick 17k. Previous studies [7]–[10] did not examine how models trained on one skin tone performed on others, missing a comprehensive analysis of potential skin tone bias. Furthermore, the enhancements they achieved primarily favored light skin tones and did not generalize well to other tones [7]–[10]. This underscores the necessity for a model that provides consistent performance across all skin tones, as evidenced by the results in **Table V**, which indicate the existence of skin tone bias.

Our findings highlight the unreliability of overall accuracy metrics in imbalanced datasets represented in **Table IV** and advocate for the CAEN model, which effectively generalizes class-wise recall performance across all skin tones by employing dynamic CSL [14] and attention mechanism [15]. Class-wise performance metrics were used to effectively indicate improvements, rather than focusing solely on overall accuracy. Even when tested on augmented data with varying image quality—such as differences in brightness, contrast, blur, and zoom—CAEN continued to perform well, especially on darker tones, while balancing performance for lighter tones, demonstrating its reliability across diverse conditions. Our study further presents an extensive analysis of its predictions across various skin tones, promoting transparency in the model's decision-making for diverse populations. These improvements not only enhance class-wise accuracy in skin lesion classification but also promote equitable diagnostic methods in dermatology, ensuring that all patients receive accurate and fair assessments, regardless of their skin tone.

REFERENCES

[1] E. Akuffo-Addo, L. Samman, L. Munawar, M. Akbik, N. Kokikian, R. Wescott, J. J. Wu, Assessing gpt-4's diagnostic accuracy with darker skin tones: underperformance and implications, Clinical and Experimental Dermatology 49 (10) (2024) 1244–1245.

[2] Kaggle, Fitzpatrick 17k dataset, accessed: 2024-02-04 (2023).

[3] S. Bin Akter, T. Sarkar Pias, S. Rahman Deeba, J. Hossain, H. Abdur Rahman, Ensemble learning based transmission line fault classification using phasor measurement unit (pmu) data with explainable ai (xai), Plos one 19 (2) (2024) e0295144.

[4] M. A. I. Siddique, A. Z. B. Aziz, A. Matin, An improved deep learning based classification of human white blood cell images, in: 2020 11th International Conference on Electrical and Computer Engineering (ICECE), IEEE, 2020, pp. 149–152.

[5] R. Rahman, A. F. Rakib, M. Rahman, T. Helaly, T. S. Pias, A real-time end-to-end bangladeshi license plate detection and recognition system for all situations including challenging environmental scenarios, in: 2021 5th International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), IEEE, 2021, pp. 1–6.

[6] J. Schneider, I. Tejani, T. Jarmain, R. Moy, et al., Diagnosis of skin disease in moderately to highly pigmented skin by artificial intelligence, Authorea Preprints (2023).

[7] S. Du, B. Hers, N. Bayasi, G. Hamarneh, R. Garbi, Fairdisco: Fairer ai in dermatology via disentanglement contrastive learning, in: European Conference on Computer Vision, Springer, 2022, pp. 185–202.

[8] M. Dominguez, J. T. Finnell, Unsupervised softotsunet augmentation for clinical dermatology image classifiers, in: AMIA Annual Symposium Proceedings, Vol. 2023, American Medical Informatics Association, 2023, p. 329.

[9] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, O. Badri, Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1820–1828.

[10] A. Pundhir, S. Verma, B. Raman, Towards ethical dermatology: Mitigating bias in skin condition classification, in: 2024 International Joint Conference on Neural Networks (IJCNN), IEEE, 2024, pp. 1–8.

[11] S. B. Akter, S. Akter, T. S. Pias, Stroke probability prediction from medical survey data: Ai-driven analysis with insightful feature importance using explainable ai (xai), in: 2023 26th International Conference on Computer and Information Technology (ICCIT), IEEE, 2023, pp. 1–6.

[12] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, N. Japkowicz, The class imbalance problem in deep learning, Machine Learning 113 (7) (2024) 4845–4901.

[13] B. Cao, Y. Liu, C. Hou, J. Fan, B. Zheng, J. Yin, Expediting the accuracy-improving process of svms for class imbalance learning, IEEE Transactions on Knowledge and Data Engineering 33 (11) (2020) 3550–3567.

[14] I. Araf, A. Idri, I. Chairi, Cost-sensitive learning for imbalanced medical data: a review, Artificial Intelligence Review 57 (4) (2024) 80.

[15] S. V. Moravvej, S. J. Mousavirad, M. H. Moghadam, M. Saadatmand, An lstm-based plagiarism detection via attention mechanism and a population-based approach for pre-training parameters with imbalanced classes, in: Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part III 28, Springer, 2021, pp. 690–701.

[16] S. B. Akter, S. Akter, M. D. Tuli, D. Eisenberg, A. Lotvola, H. Islam, J. F. Fernandez, M. Hüttemann, T. S. Pias, Fair and explainable myocardial infarction (mi) prediction: Novel strategies for feature selection and class imbalance correction, Computers in Biology and Medicine 184 (2025) 109413.

[17] A. K. Verma, S. Pal, S. Kumar, Comparison of skin disease prediction by feature selection using ensemble data mining techniques, Informatics in Medicine Unlocked 16 (2019) 100202.

[18] A. A. Elngar, R. Kumar, A. Hayat, P. Churi, Intelligent system for skin disease prediction using machine learning, in: Journal of Physics: Conference Series, Vol. 1998, IOP Publishing, 2021, p. 012037.

[19] V. Goutham, A. Sameerunnisa, S. Babu, T. B. Prakash, Brain tumor classification using efficientnet-b0 model, in: 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), IEEE, 2022, pp. 2503–2509.

[20] J. S. Aguilar-Ruiz, M. Michalak, Classification performance assessment for imbalanced multiclass data, Scientific Reports 14 (1) (2024) 10759.

[21] S. Prateek, R. Garg, K. Kumar Saxena, V. Srivastav, H. Vasudev, N. Kumar, Data-driven materials science: application of ml for predicting band gap, Advances in Materials and Processing Technologies 10 (2) (2024) 708–717.

[22] C. Van Zyl, X. Ye, R. Naidoo, Harnessing explainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of grad-cam and shap, Applied Energy 353 (2024) 122079.