Bias in Dermatological Datasets: A Critical Analysis of the Underrepresentation of Dark Skin Tones in Melanoma Classification Images

Tommaso Ruga[1,2], Ester Zumpano[1,2], Eugenio Vocaturo[2,1], Luciano Caroprese[3], and Caterina Arlia[4]

[1] DIMES - University of Calabria, Rende(CS), Italy
{tommaso.ruga,e.zumpano}@dimes.unical.it
[2] CNR-NANOTEC National Research Council, Rende(CS), Italy
{eugenio.vocaturo}@cnr.it
[3] InGeo - University G. D'Annunzio, Chieti-Pescara
{luciano.caroprese}@ingeo.it
[4] DIMEG - University of Calabria, Rende(CS), Italy
{rlacrn98e68d086d}@studenti.unical.it

Abstract. Cutaneous melanoma presents a profound healthcare challenge, particularly for individuals with darker skin tones, where late diagnosis significantly increases mortality rates. Despite remarkable advancements in artificial intelligence for medical diagnostics, current dermatological image classification systems suffer from a critical ethical and methodological limitation: severe underrepresentation of diverse skin tones in training datasets. This research uses MultiExCam, our novel multi-approach explainable architecture, to quantitatively demonstrate the systemic bias in melanoma detection across different skin tones. Our contributions are threefold: first, we comprehensively analyze major dermatological image repositories, documenting the severe underrepresentation of Fitzpatrick skin types V-VI across all datasets examined; second, we introduce Pipsqueak, a meticulously curated dataset of melanocytic lesions in darker skin tones, which demonstrates the profound scarcity of diverse representation in existing resources; and third, through empirical validation, we quantify performance disparities that emerge when models trained predominantly on light skin images are applied to darker skin tones, revealing accuracy drops that could translate to potentially fatal clinical consequences. This work provides crucial evidence for the urgent need to develop more inclusive diagnostic technologies that can effectively serve all populations, regardless of skin tone, and challenges the field to prioritize deliberate collection of diverse dermatological data.

Keywords: Melanoma Classification · Dataset Bias · Skin Tone Diversity

1 Introduction

The landscape of medical artificial intelligence has witnessed remarkable progress, particularly in dermatological image classification [1, 2]. However, a critical and

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_32

often overlooked ethical challenge persists: the profound underrepresentation of diverse skin tones in medical imaging datasets. The Fitzpatrick scale, which categorizes skin tones from type 1 (lightest) to type 6 (darkest), reveals a stark reality in current medical AI research—most existing datasets are heavily skewed towards lighter skin types. This imbalance is not merely a statistical anomaly but a significant ethical and clinical concern. Melanoma, a potentially deadly form of skin cancer, presents differently across various skin tones. Existing AI diagnostic tools, trained predominantly on images of light-skinned individuals, risk misclassifying or completely missing crucial diagnostic markers in darker skin tones. Systematic reviews have consistently demonstrated that less than 5% of dermatological images represent Fitzpatrick skin types IV-VI, creating a dangerous knowledge gap that could lead to delayed diagnoses and potentially fatal consequences for individuals with darker skin. The challenges of our research are multifaceted:

- **Dataset Bias**: Existing repositories originate primarily from European and North American populations, reflecting a narrow demographic representation. Addressing these disparities requires concerted efforts to develop more inclusive dermatological datasets with comprehensive representation across the full spectrum of human skin diversity.
- **Algorithmic Limitations**: AI models predominantly on unbalanced dataset demonstrate measurably inferior performance when applied to darker skin tones. Research has documented elevated rates of false negatives for potentially life-threatening conditions such as melanoma when these algorithms are applied to individuals with darker skin, raising significant concerns regarding diagnostic equity.
- **Clinical Implications**: The underrepresentation translates to real-world healthcare disparities, with individuals of color often receiving later-stage diagnoses of skin cancers and other dermatological conditions.

The specific contributions of this research are threefold. First, we provide a comprehensive critical analysis of major dermatological image datasets, systematically documenting the severe underrepresentation of darker skin tones and the resulting implications for AI diagnostic accuracy. Second, we introduce Pipsqueak—a meticulously curated, small but symbolically significant dataset representing melanocytic lesions in darker skin tones—which demonstrates the profound scarcity of diverse representation in existing resources. Finally, through empirical validation using our MultiExCam architecture [3], we quantitatively demonstrate the performance disparities that emerge when models trained predominantly on light skin images are applied to darker skin tones, revealing accuracy drops that could translate to potentially fatal clinical consequences. By highlighting these disparities and offering both methodological insights and a concrete resource, our work contributes to the broader discourse on ethical AI in healthcare, advocating for more inclusive approaches to dermatological dataset development that can serve all populations equitably.
The paper is organized as follows: after a discussion of the current literature, the main open access datasets are analyzed, highlighting their characteristics

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_32

and any biases they may contain. In the following section, Pipsquek is introduced — A new (painstaking, compact) dataset of melanoma for dark skin color. Subsequently, the performance of MultExCam is tested on the newly proposed dataset, in order to assess how the biases present in the datasets used to train even the most effective models can affect their predictive capabilities. Finally, the obtained results are discussed and possible future directions are explored.

1.1 Related Works

Several papers in the literature highlight the challenges related to bias and the scarcity of datasets representing people with skin of color. We report in this section the key works in the domain.

The research by Rezk et al., [4, 5] addresses a critical healthcare disparity in dermatological diagnosis through artificial intelligence. Their work spans two complementary papers that focus on improving the diversity of skin tones in dermatological resources to develop more inclusive AI solutions for skin cancer detection. In their 2022 protocol paper published in JMIR Research Protocols, Rezk et al. outlined a four-phase approach to tackle the underrepresentation of darker skin tones in dermatological images. The protocol established a framework for quantifying this diversity gap, generating images for underrepresented skin tones, evaluating these generated images, and developing inclusive skin cancer detection models. Their preliminary analysis confirmed that darker skin tones were significantly underrepresented in existing dermatological datasets, with over 60% of images in their main dataset showing light skin [4]. Building on this protocol, their implementation paper published in JMIR Dermatology presented the execution and results of their approach. The researchers collected 1,701 clinical images from publicly available repositories, including DermNet NZ, ISIC-2018, and Dermatology Atlas, confirming that 84.1% of their dataset consisted of light and intermediate skin tones. They then applied two deep learning methods—style transfer (ST) and deep blending (DB)—to generate realistic images of skin lesions on darker skin tones [5]. Quantitative evaluation showed that the style transfer method outperformed deep blending, achieving a lower loss of realism score (0.23 vs. 0.63) and higher disease presentation similarity (0.44 vs. 0.17). In qualitative assessment, 62.2% of the ST-generated images were classified as real by participants in a visual Turing test, demonstrating high realism. Eight dermatologists correctly diagnosed the lesions in the generated images at a rate of 75%, comparable to their 76% accuracy with real images [5].

The research by Kushimo et al., [6] focuses on improving melanoma detection in people with darker skin tones using deep learning techniques. The authors highlight a critical healthcare disparity: while melanoma is less common in people with darker skin, when it does occur, mortality rates are higher due to delayed diagnosis and detection. This work addresses the lack of diversity in skin cancer diagnosis technology, which historically has been developed primarily using images of lighter skin. The researchers developed a custom deep learning model based on DenseNet121 architecture pre-trained on ImageNet, applying transfer learning to fine-tune its final layers for melanoma classification. They collected

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_32

approximately 100 clinical images of dark skin with various conditions, including 37 melanoma cases, and combined these with images from the HAM10000 dataset. After preprocessing steps including cropping, denoising, illumination correction, and segmentation, they performed data augmentation to increase their dataset to 18,374 images. Their model achieved remarkable accuracy: 99% for melanoma detection in white skin and 98% for dark skin. Other performance metrics were similarly impressive, with precision, recall, specificity, and F1-score values ranging from 0.90 to 0.99 across different skin tones. When compared to existing approaches in the literature, the proposed model outperformed many previous studies in accuracy and precision. This research represents an important step toward developing more inclusive AI diagnostic tools for dermatology that work effectively across all skin tones. The authors suggest that future work could explore using generative adversarial networks (GANs) to create synthetic melanoma images to further address the lack of clinical data for cancer in people of color.

Han et al., [7] revealed that when algorithms trained predominantly on images of East Asian skin were tested on skin lesion images of White patients from the USA, they significantly underperformed. This demonstrates how algorithms can fail to generalize across different ethnic populations. The authors found that among the datasets where Fitzpatrick skin type information was available (only 2,436 images from three datasets), only ten images were from individuals with Fitzpatrick skin type V, and merely a single image was from an individual with Fitzpatrick skin type VI. This extreme underrepresentation highlights the critical gap in darker skin representation in training datasets.

Navarrete-Dechent et al., [8], emphasized that most algorithms submitted to the ISIC 2018 challenge performed worse on test images from external institutions independent from the training dataset, indicating a lack of generalizability across different population demographics.

Cormier et al., [9], highlighted that prevalence, presentation, and types of skin cancer vary significantly in skin of color populations, with associated poorer outcomes, making the underrepresentation particularly problematic for accurate diagnosis in these communities.

Royse et al., [10], noted that the bias is especially concerning when detecting skin cancers that preferentially affect patients of skin of color, such as Kaposi sarcoma, further emphasizing the clinical consequences of dataset bias.

Adamson and Smith, [11], authored an influential perspective in JAMA Dermatology in 2018 discussing how excluding skin of color in training datasets poses significant risks. They warned that algorithms trained on homogeneous datasets could lead to incorrect diagnosis or completely missing skin cancers in people with darker skin. The authors argued that such limitations could widen existing racial disparities in dermatology, with the potential for AI to exacerbate rather than reduce healthcare inequities.

Han et al., [12], demonstrated in their study published in the Journal of Investigative Dermatology that algorithms trained on homogeneous populations perform better on those populations. Their AI model trained using skin images

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_32

from exclusively Asian populations achieved greater accuracy for classifying basal cell carcinomas and melanomas in Asian populations (96%) than in Caucasian populations (88-90%). This work provides empirical support for concerns that lack of diversity in training data leads to disparate performance across different populations.

These studies collectively highlight major challenges in developing equitable AI systems for dermatology. These works emphasize the need for more diverse, well-documented datasets with standardized reporting of demographic variables and clearly demonstrate the real performance gaps that emerge when algorithms are trained on limited population samples.

2 How biased are Open Access Datasets?

2.1 The ISIC (International Skin Imaging Collaboration) Dataset

The ISIC archive represents a significant and comprehensive collection of dermoscopic images, serving as a pivotal resource for skin cancer research and machine learning development. In 2020, the archive compiled images from multiple international sources, including hospitals and research centers in Spain, Australia, Austria, the United States, and other countries. The dataset encompasses a substantial collection of 7,311 images from the Hospital Clinic Barcelona, 8,449 images from the University of Queensland, 4,374 images from Medical University Vienna, 11,108 images from Memorial Sloan Kettering Cancer Centre, and 1,884 images from the Sydney Melanoma Diagnosis Centre. Despite its impressive scale and international collaboration, the ISIC archive epitomizes the critical representational limitations identified in the systematic review. The dataset predominantly features images from predominantly white-populated countries, predominantly in the Global North. Although the collection spans multiple international sites, it fundamentally lacks comprehensive demographic metadata, particularly regarding participant ethnicity and Fitzpatrick skin type. This metadata deficiency critically undermines the potential generalizability of machine learning models developed using these images. The geographical distribution of ISIC images starkly reflects the significant global bias in medical imaging datasets. With 56.3% of images hosted in the ISIC archive, the collection predominantly represents populations from Europe, North America, and Oceania, systematically underrepresenting populations with darker skin tones.

2.2 The HAM10000 Dataset

HAM10000 contains 10000 images, some of which refer to the same lesion but with a different degree of zooming. It refers to the third task of the 2018 challenge and includes, among others, 1134 Melanoma, and 6705 Common Nevus. The dataset contains RGB images with a resolution of 600x450 pixels, which show high heterogeneity as they have been acquired from different populations. It consists of three datasets, respectively, training, validation, and test set. In

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_32

this work, we merged and reshuffled the training and validation sets, then split them into training, validation, and test sets according to the ratio 70-15-15. To ensure concreteness in the results, the two considered classes, Common Nevus and Melanoma were balanced downwards on the basis of the number of examples in the least populous class, i.e. Melanoma. This resulted in 1134 items per class, subsequently divided into the three sets as previously described. Then, the data belonging to the training set were augmented through the use of the ImageDataGenerator function provided within the Keras library, to train the DL model. It is important to highlight, as reported in [13] that the HAM10000 dataset has significant limitations in terms of demographic representation. According to the systematic review published in The Lancet Digital Health, this dataset originates exclusively from Austria and Australia, reflecting the concerning geographical trend where 79% of skin cancer image datasets come only from Europe, North America, and Oceania. HAM10000 lacks adequate metadata on patient ethnicity and Fitzpatrick skin type. The review found that across all analyzed datasets, only 1.3% of images had associated ethnicity data and merely 2.1% included Fitzpatrick skin type information. In the rare cases where such data was available, there was massive under-representation of darker skin types: only ten images across all datasets were from individuals with Fitzpatrick skin type V, and just a single image from someone with type VI.

2.3 The PAD-UFES-20 Dataset

The PAD-UFES-20, published in 2020, is a skin lesion dataset from Brazil containing 2,298 clinical images collected from 1,373 patients. The images of skin lesions are captured through macroscopic photographs in PNG format. What makes PAD-UFES-20 notable is the inclusion of relevant clinical metadata with each image, such as patient age, sex, and the anatomical location of the lesions. However, the dataset demonstrates very limited diversity in skin tones as it only contains data for 1415 images with reported ethnicity and 2236 images with Fitzpatrick skin type information. Despite originating from Brazil, which has a diverse population, the dataset contains very limited representation of individuals with darker skin types (Fitzpatrick V-VI). Most images are from individuals with lighter to medium skin tones, with only a small fraction representing darker-skinned individuals.

2.4 The DermNet Dataset

DermNet represents one of the most extensive and publicly accessible online repositories of dermatological images, originally maintained by DermNet New Zealand. This comprehensive database encompasses thousands of clinical images documenting a wide spectrum of dermatological conditions and has been extensively utilized for both educational purposes and artificial intelligence development in dermatology. Despite its considerable size and widespread usage, DermNet exhibits significant disparities in the representation of diverse skin tones, reflecting a broader issue prevalent across dermatological datasets.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_32

Systematic reviews have consistently identified a severe underrepresentation of darker skin types within the DermNet collection, with Fitzpatrick skin types V and VI constituting less than 5% of the total image repository. This imbalance creates a foundational bias that potentially compromises the development of equitable diagnostic technologies. The dataset further lacks comprehensive demographic metadata, with the majority of images omitting critical information regarding Fitzpatrick skin type classification or patient ethnicity. This absence of demographic documentation significantly impedes precise quantification of the diversity gap and complicates efforts to address representational imbalances. The geographical origin of the images demonstrates a pronounced bias, with a predominance of specimens originating from Oceania, Europe, and North America, thereby limiting the representation of dermatological characteristics common in African, South Asian, and Middle Eastern populations.

2.5 The Fitzpatrick17k Dataset

The Fitzpatrick17k dataset, is a comprehensive collection of 16,577 annotated skin images sourced from DermaAmin and Atlas Dermatological. This dataset uses the Fitzpatrick scale, which ranges from type 1 (lightest) to type 6 (darkest) to categorize skin tones. The dataset exhibits significant bias in its distribution of skin tones. Specifically, it contains 895 images for Fitzpatrick type one, 1,413 for type two, 931 for type three, 668 for type four, 305 for type five, and only 104 for type six. This severe imbalance skews heavily toward lighter skin tones, with types 1-3 comprising approximately 75% of the dataset, while the darkest skin tone (type 6) represents just 2.4% of the images. The researchers demonstrated this bias empirically by training an EfficientNet model on the original dataset, which achieved an overall median accuracy of 82.1%. When analyzed by skin tone, the model showed significant disparities in performance, with a notably lower accuracy for Fitzpatrick type 6 (darker skin) compared to other categories.

2.6 The MED-NODE Dataset

The MED-NODE dataset, published in 2015 and originating from the Netherlands, represents an early contribution to dermatological image classification using macroscopic imaging. Captured with a Nikon D3 or Nikon D1x camera equipped with a Nikkor 2.8/105 mm micro lens, the dataset contains 170 images focusing on two distinct skin lesion categories. While modest in size, the dataset attempted to provide a standardized approach to skin lesion image analysis using high-quality photographic equipment. However, the MED-NODE dataset exemplifies the profound representational limitations systematically documented in the comprehensive review. The dataset entirely omits crucial demographic metadata, including the number of participants, ethnicity, and Fitzpatrick skin type information. This metadata absence critically undermines the potential generalizability of any machine learning models developed using these images, rendering it impossible to assess the dataset's demographic diversity and potential inherent biases. Originating from a European country, MED-NODE contributes to

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_32

the problematic trend where the vast majority of dermatological image datasets are sourced from a narrow geographical region, predominantly representing populations with lighter skin tones.

2.7 The PH^2 Dataset

The PH^2 dataset, emerging from Portugal and published in 2013, offers a focused collection of dermoscopic images specifically curated for dermatological research. Comprising 200 images, the dataset utilizes the Tuebinger Mole Analyzer System for image acquisition, representing an early contribution to skin lesion classification research. The dataset encompasses three distinct skin lesion categories, providing a relatively narrow but targeted approach to image classification. Despite its potential scientific value, the PH^2 dataset manifests the systemic representational limitations extensively documented in the systematic review. The dataset provides no reported information about the number of participants, a critical omission that undermines the potential generalizability of any derived machine learning models. Moreover, the dataset completely lacks metadata regarding participant ethnicity and Fitzpatrick skin type, rendering it impossible to assess the demographic diversity of the underlying image collection. The absence of comprehensive demographic information in the PH^2 dataset reflects a broader methodological weakness in dermatological imaging research. By failing to capture and report essential contextual metadata, such datasets risk creating machine learning models that may perform inconsistently across different population groups.

3 Pipsqueak: A new (painstaking, compact) Dataset of melanoma for dark skin color

No public available dataset in the state of the art literature exposes a balanced set of images, no dataset exists with melanoma images from different nuances of skin tone, from light to medium and dark. An initial reconnaissance of major dermoscopic image archives - from the ISIC dataset to Fitzpatrick17k, from PAD-UFES-20 to PH^2 - revealed an alarming disparity: less than 5% of images depicted patients with medium to dark skin tones. This limitation is not merely a statistical nuance, but a potential clinical risk with potentially fatal consequences.

We therefore created, and here we present it, *Pipsqueak*, a dataset stemmed from a systematic evaluation of several established dermatological image repositories. This process began with a careful examination of multiple datasets including ISIC, HAM10000, PAD-UFES-20, DermNet, Fitzpatrick17k, Med-Node, and PH^2, each presenting significant limitations in terms of skin tone diversity. More in details, PH^2, DermNet and MED-NODE dataset were manually analyzed for skin phototype variety. This analysis confirmed the absence of images representing Fitzpatrick V and VI skin types, further substantiating the systematic underrepresentation issue. The ISIC archive, one of the largest and most

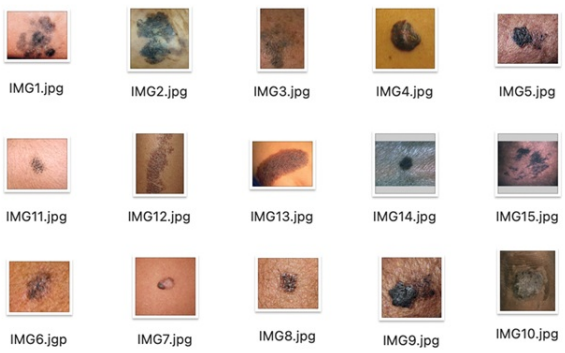
ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_32

utilized datasets for AI-driven skin lesion diagnosis, was filtered according to Fitzpatrick skin type classification. This revealed only 10 images corresponding to Fitzpatrick type V and a single image representing Fitzpatrick type VI. Crucially, when melanoma-specific filtering was applied, no samples from these darker skin types (Fitzpatrick V and VI) were found. The single image representing Fitzpatrick type VI depicted a benign lesion and was included in our nascent collection. The Fitzpatrick17k dataset proved to be more inclusive than other available resources, containing 16,577 clinical images labeled with skin conditions and skin types according to the Fitzpatrick classification system. Following consent-based access to the dataset, we implemented a focused filtering approach targeting melanoma and nevus images specifically from Fitzpatrick V and VI skin tones. This initial filtering was followed by manual analysis to ensure selection quality. Several challenges emerged during this process: many images depicted the same lesion from different angles, necessitating careful selection to avoid duplication; additionally, the clinical (non-dermoscopic) nature of these images required specialized consideration during the selection process. Despite Fitzpatrick17k's relative inclusivity, the final yield of usable images for our research totaled just 15 images—9 melanomas and 6 nevi—highlighting the profound scarcity of diverse representation even in the most inclusive available datasets. Our manual curation process was painstaking. Each image was a hard-won testament to the underrepresentation of skin of color in medical imaging. We sourced these rare images from various dermatological archives, carefully selecting instances that could provide meaningful insights into how skin lesions manifest across different skin tones. With unwavering determination, we undertook a meticulous manual selection process, sifting through multiple sources to carefully curate a mere 16 images representing patients with diverse dark skin tones. The resulting core dataset of 15 images from Fitzpatrick 17k, combined with the single Fitzpatrick VI image from ISIC, formed a nucleus of 16 images representing melanocytic lesions in darker skin tones 1. This extremely limited collection underscores the critical gap in representation that persists in dermatological imaging archives. Through sophisticated data augmentation techniques, we expanded this initial nucleus to generate a dataset of 80 images - a small but symbolically significant step towards more equitable medical AI representation. The augmentation process employed controlled transformations including subtle rotations, scaling, and controlled brightness adjustments that preserved the critical diagnostic features while expanding the training utility of these rare images. This approach maintained the integrity of the essential melanoma and nevus characteristics while creating sufficient variation to support machine learning model development. While modest in absolute terms, Pipsqueak represents a deliberate methodological commitment to addressing the profound equity gap in dermatological datasets. Its name—referring to something small yet significant—reflects its position as a crucial step toward developing more inclusive diagnostic technologies in dermatology.

It is important to acknowledge that we could have employed more extensive data augmentation techniques to further expand the dataset beyond 80 images.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_32

However, we deliberately chose to limit the augmentation process. Starting with a core dataset of merely 16 images and expanding it excessively would have introduced significant risks of overfitting and potentially created misleading patterns that don't accurately represent real-world melanoma presentations in darker skin tones. The fundamental issue is not simply the lack of augmented data but rather the scarcity of authentic source images representing diverse skin tones in dermatological archives. This limitation underscores a critical insight: the path forward cannot rely solely on computational techniques to manufacture diversity where original diversity is profoundly lacking. Instead, meaningful progress demands a fundamentally different approach—one that prioritizes comprehensive, deliberate collection of dermatological images across all skin types, with particular attention to currently underrepresented populations. Pipsqueak thus serves not only as a modest contribution to the existing landscape but also as a clear demonstration of the limitations of current approaches and the urgent need for systematic change in how dermatological imaging data is collected, categorized, and shared across the global research community.



(a) The 15 images extracted from Fitzpatrick 17k

(b) The unique image extracted from ISIC

Fig. 1: Pipsqueak: the core dataset of 16 images

4 MultiExCam's Experimentation as a measure of Bias Assessment

In the field of bias assessment, numerous techniques and methodological approaches exist. In this study, we opted for a comparative approach based on analyzing the performance of the MultiExCam architecture across two distinct datasets. The results reveal significant performance disparities that strongly indicate the presence of dataset implies a limited ability to generalize effectively.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_32

The section firstly briefly describes the MultiExCam architecture and then reports the results obtained by MultiExCam on the *HAM10000* [14] versus those obtained on Pipsqueak.

4.1 The MultiExCam architecture

In this section we briefly recall the MultiExCam Architecture, firstly introduced in [3] and under review at an international journal.

MultiExCam is a novel multi-approach and explainable architecture for skin lesion classification that integrates DL and ML techniques to achieve robust classification performance while maintaining interpretability. The proposed methodology consists of four stages: i) *Data Preprocessing*, ii) *Feature Extraction and Initial Classification*, iii) *ML Classification*, and iv) *Ensemble Classification*. In data preprocessing stage images are loaded, stored, and prepared for analysis. This crucial step involves image cleaning, resizing, and removal of artifacts, such as body hair. Additionally, data augmentation is applied to the training set to mitigate overfitting and enhance model generalization. In the second stage, a ResNet50 model, leveraging transfer learning, performs two crucial functions: i) *Initial classification*, providing a baseline classification of skin lesions, and ii) *Feature extraction*, extracting deep features from the penultimate layer for use in subsequent machine learning models. The third stage involves four distinct machine learning classifiers processing a combination of *CNN-extracted features* and *hand-crafted statistical features*. The diversity of these models - including an *Extra Trees Classifier*, a *K-Nearest Neighbors*, a *Support Vector Machine* (*NuSVM* variant), and a *XGBoost* model - contributes to the robustness of the final ensemble. The final stage performs an ensemble classification using a feed-forward neural network (FFNN) that acts as an expert classifier, producing a final classification by learning to optimally combine the inputs from previous stage. This approach allows for a more nuanced and accurate classification, leveraging the strengths of each individual model.

4.2 MultiExCam Experimental Results on HAM10000 Dataset

The MultiExCam architecture was initially validated on the HAM10000 dataset [14], with detailed results currently under review for publication in an international journal. The HAM10000 dataset contains 10,000 dermatoscopic images, including 1,134 melanoma and 6,705 common nevus samples. For experimental validation, the classes were balanced by downsampling to match the less populous melanoma class, resulting in 1,134 samples per class. The dataset was then split into training, validation, and test sets following a 70-15-15 ratio.

The architecture demonstrated promising performance across all evaluation metrics. After preprocessing steps including hair removal using the Black Hat filter technique, the ResNet50-based deep learning component achieved an accuracy of 91%, an F1-score of 91%, and an AUC of 0.97. The complete MultiExCam ensemble further improved these results, achieving an accuracy of 92%, an F1-score of 92%, while maintaining the AUC at 0.97. Notably, the ensemble

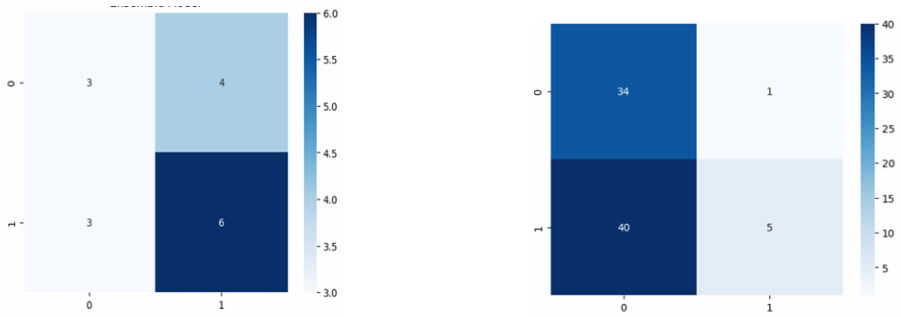
ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_32

approach enhanced both sensitivity and specificity compared to the individual models, with improvements especially visible in the confusion matrices showing a reduction in false negatives for melanoma cases.

4.3 MultiExCam Experimental Results on Pipsqueak Dataset

Performance Analysis: The experimental evaluation of the MultiExCam architecture on the Pipsqueak dataset provides critical insights into both the challenges and opportunities in melanoma detection across diverse skin tones. Our analysis encompasses two distinct experimental scenarios: first, utilizing the original 16-image core dataset, and subsequently, employing an augmented variant expanded to 80 images through controlled transformation techniques.

Results on Original Pipsqueak Dataset: The initial experiments conducted on the 16-image dataset exhibited several noteworthy characteristics. Figure 2a presents the confusion matrix for the ensemble model on the original dataset. The performance metrics reveal moderate classification capabilities. The model



(a) Confusion matrix for the ensemble model on the original Pipsqueak dataset (16 images).

(b) Confusion matrix for the ensemble model on the augmented Pipsqueak dataset (80 images).

Fig. 2: Confusion matrix for the ensemble model on the Pipsqueak dataset.

achieved an overall accuracy of 0.56, with precision values of 0.50 for nevus classification and 0.60 for melanoma identification. The recall metrics were 0.43 for nevus and 0.67 for melanoma cases, culminating in F1-scores of 0.46 and 0.63 respectively. Sensitivity was measured at 0.67, while specificity reached 0.43. These results, while demonstrating some discriminative ability, highlight the substantial challenges in achieving reliable classification with such a limited dataset. The disparity between nevus and melanoma metrics suggests an imbalanced performance profile that warrants careful consideration in clinical contexts.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_32

Results on Augmented Pipsqueak Dataset: To address the limitations imposed by the restricted dataset size, we implemented controlled augmentation techniques to expand the dataset to 80 images. Figure 2b illustrates the confusion matrix obtained from the ensemble model trained and evaluated on this augmented dataset. The augmentation process yielded significant performance improvements across all metrics. The ensemble model achieved an enhanced overall accuracy of 0.65, with precision values increasing to 0.46 for nevus classification and 0.83 for melanoma detection. Recall metrics showed substantial improvement at 0.97 for nevus cases and 0.31 for melanoma identification. The resulting F1-scores were 0.62 and 0.29 for nevus and melanoma classifications, respectively. Notably, sensitivity increased to 0.11, while specificity substantially improved to 0.97. The support metrics indicate a more robust evaluation framework, with 35 nevus and 45 melanoma instances comprising the augmented test set, compared to the original 7 nevus and 9 melanoma test cases.

4.4 Comparative Analysis and Discussion

The comparative analysis of performance metrics between the original and augmented datasets reveals several significant patterns. The augmentation process produced a substantial improvement in overall accuracy (from 0.56 to 0.65) and a marked enhancement in specificity (from 0.43 to 0.97), indicating the model's increased capability to correctly identify nevus cases. However, the augmentation appears to have introduced a trade-off in sensitivity, which decreased from 0.67 to 0.11, suggesting a reduced ability to correctly identify melanoma cases. This transformation in the performance profile indicates a potential shift in the decision boundary of the classifier, possibly influenced by the synthetic characteristics introduced during the augmentation process. The confusion matrices provide visual confirmation of this performance shift. In the original dataset, the model demonstrated a relatively balanced distribution of misclassifications, with 3 false positives and 3 false negatives. In contrast, the augmented dataset model exhibited a strong bias toward nevus classification, with only 1 false positive but 40 false negatives. This pattern suggests that while augmentation improved certain aspects of model performance, it potentially introduced or amplified biases in the classification algorithm. The substantial increase in false negatives in melanoma detection is particularly concerning from a clinical perspective, as missed melanoma diagnoses carry significant health implications. These findings underscore the complex challenges inherent in developing robust diagnostic algorithms for underrepresented populations in dermatological imaging. Furthermore, this comparative analysis reinforces our earlier assertion regarding the limitations of data augmentation as a comprehensive solution to the diversity gap in dermatological datasets. The transformation in performance characteristics following augmentation emphasizes that synthetic expansion cannot fully substitute for authentic diversity in source images, particularly for high-stakes medical diagnostics such as melanoma detection.

This research demonstrates that the underrepresentation of darker skin tones in dermatological image datasets creates profound challenges for AI-driven di-

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_32

agnostic technologies, particularly in melanoma detection. Our comprehensive analysis of leading public datasets reveals a systematic pattern of exclusion, with darker skin tones (Fitzpatrick types V and VI) comprising less than 5% of available images. This disparity is not merely a statistical anomaly but represents a significant barrier to equitable healthcare delivery. Through the painstaking curation of Pipsqueak, a compact yet symbolically important dataset of melanocytic lesions in darker skin tones, we have empirically demonstrated the limitations of existing computational approaches. The stark performance disparities observed when applying our MultiExCam architecture to Pipsqueak versus HAM10000 provide compelling evidence of the inherent biases perpetuated by current methodologies. Despite achieving enhanced overall accuracy through data augmentation (from 0.56 to 0.65), the concerning trade-off in sensitivity—with melanoma detection rates decreasing from 0.67 to 0.11—underscores the insufficient nature of synthetic data generation as a comprehensive solution. These findings highlight a critical insight: the path toward more inclusive dermatological AI cannot be achieved solely through computational techniques applied to fundamentally biased datasets. Instead, meaningful progress requires a paradigm shift in how dermatological imaging data is collected, categorized, and shared across the global research community.

5 Conclusion and Future Works

The profound scarcity of authentic source images representing diverse skin tones necessitates dedicated efforts to develop more comprehensive, demographically transparent datasets that accurately represent the full spectrum of human skin diversity. Beyond the technical implications, our work emphasizes the ethical imperative of addressing algorithmic bias in medical AI. The potential clinical consequences of misdiagnosis or delayed diagnosis due to skin tone-based performance disparities represent a tangible manifestation of healthcare inequity. By quantifying these disparities and demonstrating their clinical implications, we contribute to the growing discourse on ethical AI development in medicine. In conclusion, while Pipsqueak represents a modest contribution to the dermatological imaging landscape, its significance extends beyond its size. It serves as both a practical demonstration of existing limitations and a call to action for systematic change in research methodology. Future research should focus on examining additional open access datasets and establishing international collaborative initiatives specifically designed to collect dermatological images from diverse global populations, with particular emphasis on skin tones that are severely underrepresented in current datasets. These efforts must implement standardized protocols for comprehensive demographic data collection, including the systematic documentation of Fitzpatrick skin types, ethnicity, age, gender, and geographical origin. Future work must prioritize deliberate, inclusive data collection practices, transparent demographic reporting, and rigorous evaluation across diverse populations. Only through such comprehensive approaches can we develop truly equitable diagnostic technologies that serve all populations, regardless of skin

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_32

tone, and fulfill the promise of AI as a tool for reducing rather than reinforcing healthcare disparities.

References

1. E. Vocaturo, D. Perna, E. Zumpano, Machine learning techniques for automated melanoma detection, in: IEEE BIBM), 2019, pp. 2310–2317.
2. E. Vocaturo, E. Zumpano, Multiple instance learning approaches for melanoma and dysplastic nevi images classification, in: IEEE ICMLA, 2020, pp. 1396–1401.
3. T. Ruga, G. Musacchio, D. Maurmo, An ensemble architecture for melanoma classification, in: pHealth 2024, IOS Press, 2024, pp. 183–184.
4. E. Rezk, M. Eltorki, W. El-Dakhakhni, Leveraging artificial intelligence to improve diversity of dermatological skin color pathology: Protocol for an algorithm development and validation study (preprint), JMIR Research Protocols 11 (11 2021). doi:10.2196/34896.
5. E. Rezk, M. Eltorki, W. El-Dakhakhni, Improving skin color diversity in cancer detection: Deep learning approach, JMIR Dermatology 5 (2022) e39143. doi:10.2196/39143.
6. O. Kushimo, A. Salau, O. Adeleke, D. Olaoye, Deep learning model to improve melanoma detection in people of color, Arab Journal of Basic and Applied Sciences 30 (2023) 92–102. doi:10.1080/25765299.2023.2170066.
7. S. Han, I. Park, S. Chang, W. Lim, M. Kim, G. Park, J. Chae, C.-H. Huh, J.-I. Na, Augment intelligence dermatology : Deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders, Journal of Investigative Dermatology 140 (03 2020). doi:10.1016/j.jid.2020.01.019.
8. C. Navarrete-Dechent, S. Dusza, K. Liopyris, A. Marghoob, A. Halpern, M. Marchetti, Automated dermatological diagnosis: Hype or reality?, Journal of Investigative Dermatology 138 (06 2018). doi:10.1016/j.jid.2018.04.040.
9. J. Cormier, Y. Xing, M. Ding, J. Lee, P. Mansfield, J. Gershenwald, M. Ross, X. Du, Ethnic differences among patients with cutaneous melanoma, Archives of internal medicine 166 (2006) 1907–14. doi:10.1001/archinte.166.17.1907.
10. K. Royse, F. El Chaer, E. Amirian, C. Hartman, S. Krown, T. Uldrick, J. Lee, Z. Shepard, E. Chiao, Disparities in kaposi sarcoma incidence and survival in the united states: 2000-2013, PLOS ONE 12 (2017) e0182750. doi:10.1371/journal.pone.0182750.
11. A. Adamson, A. Smith, Machine learning and health care disparities in dermatology, JAMA Dermatology 154 (08 2018). doi:10.1001/jamadermatol.2018.2348.
12. S. Han, M. Kim, W. Lim, G. Park, I. Park, S. Chang, Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm, Journal of Investigative Dermatology 138 (02 2018). doi:10.1016/j.jid.2018.01.028.
13. D. Wen, S. M. Khan, A. J. Xu, H. Ibrahim, L. C. Smith, J. Caballero, L. Zepeda, C. de Blas Pérez, A. K. O. Denniston, X. Liu, R. N. Matin, Characteristics of publicly available skin cancer image datasets: a systematic review., The Lancet. Digital health (2021).
14. P. Tschandl, C. Rosendahl, H. Kittler, The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, Scientific data 5 (1) (2018) 180161.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97554-7_32