DermaCon-IN: A Multi-concept Annotated Dermatological Image Dataset of Indian Skin Disorders for Clinical AI Research

Shanawaj S Madarkar[*†‡] **Mahajabeen Madarkar**[*§] **Madhumitha Venkatesh**[*†¶]
Teli Prakash[‖] Konda Reddy Mopuri[†] Vinaykumar MV[**] KVL Sathwika[††]
Adarsh Kasturi[**] Gandla Dilip Raj[**] PVN Supranitha[**] Harsh Udai[†]

Abstract

Artificial intelligence is poised to augment dermatological care by enabling scalable image-based diagnostics. Yet, the development of robust and equitable models remains hindered by datasets that fail to capture the clinical and demographic complexity of real-world practice. This complexity stems from region-specific disease distributions, wide variation in skin tones, and the underrepresentation of outpatient scenarios from non-Western populations. We introduce DermaCon-IN, a prospectively curated dermatology dataset comprising 5,450 clinical images from 2,993 patients across outpatient clinics in South India. Each image is annotated by board-certified dermatologists with 245 distinct diagnoses, structured under a hierarchical, aetiology-based taxonomy adapted from Rook's classification. The dataset captures a wide spectrum of dermatologic conditions and tonal variation commonly seen in Indian outpatient care. We benchmark a range of architectures, including convolutional models (ResNet, DenseNet, EfficientNet), transformer-based models (ViT, MaxViT, Swin), and Concept Bottleneck Models to establish baseline performance and explore how anatomical and concept-level cues may be integrated. These results are intended to guide future efforts toward interpretable and clinically realistic models. DermaCon-IN provides a scalable and representative foundation for advancing dermatology AI in real-world settings.

1 Introduction

Skin diseases pose a significant global health challenge, affecting billions of individuals and ranking among the leading causes of disease burden. The Global Burden of Disease 2019 [39] study identified dermatological conditions as the *fourth* leading cause of nonfatal morbidity worldwide. Common ailments such as fungal infections, acne, scabies, and eczema impact millions globally [21], underscoring the urgent need for improved diagnostic tools and equitable access to care.

Artificial intelligence (AI) has emerged as a promising solution for enhancing dermatological diagnosis and triage, particularly in resource-constrained regions with limited access to dermatologists. However, a critical bottleneck remains: the lack of representative training data that adequately

[*]Equal contribution
[†]Department of Artificial Intelligence, Indian Institute of Technology Hyderabad, India
[‡]Indian Navy
[§]Department of Dermatology, S R Patil Medical College, India
[¶]TCS Research Scholar
[‖]Department of Dermatology, S Nijalingappa Medical College, India
[**]Department of Dermatology, Sri Chamundeshwari Medical College, Hospital & Research, India
[††]Interns at Indian Institute of Technology Hyderabad, India

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

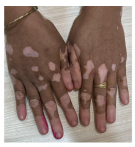
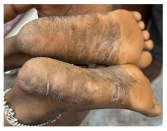
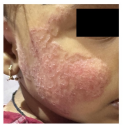
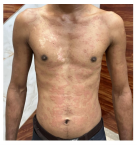
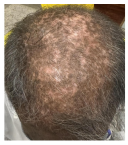
DermaCon-IN sample images					
Descriptors	Patch, White (Hypopigmentation)	Fissure, Hyperkeratotic plaques, Scale	Erythema, Plaque, Scale	Erythema, Wheal	Patch, Plaque, White (Hypopigmentation)
Body Parts	Upper Extremities Hands (Manus) Fingers (Digits)	Lower Extremities Soles (Plantar Region)	Head Cheeks	Trunk Abdomen, Thorax, Upper Extremities Shoulders	Head Scalp
Main Class	Pigmentary Disorders	Keratanisation Disorders	Infectious Disorders	Inflammatory Disorders	Inflammatory Disorders
Sub Class	Pigmentary Disorders	Keratanisation Disorders	Infectious-Fungal	Inflammatory (Other)	Eczema and Dermatitis
Disease Label	Vitiligo	Palmoplantar Keratoderma	Tinea Faciei	Urticaria	Seborrheic Dermatitis

Figure 1: Sample images from DermaCon-IN dataset with skin lesion descriptors, body parts, main class, sub class, and disease labels.

captures diversity in disease presentations and skin tones based on regional relevance [4]. Most AI models for dermatology to date have been developed and benchmarked using datasets predominantly sourced from North American, European, or Australasian populations [1, 26]. This geographic skew has introduced performance biases, especially for underrepresented groups such as individuals, patients presenting with diseases more prevalent outside Western contexts, and those with darker skin tones.

Recent work reveals the consequences of biased dermatology datasets [1, 28]. Public benchmarks focus on pigmented lesions and melanoma, mirroring Western priorities, while overlooking common conditions in tropical regions, like fungal infections, scabies, pigmentary, and nutritional disorders [37, 39, 21, 11, 20]. A "one-size-for-all" approach fails across populations, demanding region-specific resources. Under-representation of darker skin tones compounds this gap: [4] reports 30-40% accuracy drop on darker skin in DDI, while trained on Fitzpatrick17k's dataset (dominant lighter tones in the dataset). These biases lead to models that underperform on underrepresented phenotypes.

To address these limitations, we introduce a new dermatology image dataset curated from Indian outpatient clinics. To the best of our knowledge, it is the first densely annotated dataset centered around Indian skin phototype and is designed to improve diversity in both disease coverage and skin tone representation for dermatological AI research. It complements existing datasets by capturing the phenotypic and pathological landscape of a population historically underserved in global medical AI efforts. In addition, the dataset also aims to support explainable modeling by reflecting how dermatologists diagnose, through the combined use of anatomical location and visual descriptors. The Key contributions of this work are as follows:

- **South Asian Clinical and Phenotypic Representation.** DermaCon-IN developed in South Asia reflects regional disease patterns, such as the high prevalence of infectious etiologies (fungal, viral, parasitic) observed in tropical outpatient settings [42, 12, 35, 2, 19]. This contrasts with existing datasets dominated by inflammatory or neoplastic disorders common in Western contexts [45, 16]. The dataset also includes Fitzpatrick skin types IV–VI, which are typically underrepresented in existing resources [10], offering a path to reduce fairness gaps in clinically deployable AI models.

- **Multi-Concept Clinical Annotations.** Each image is annotated with two independent sets of clinically meaningful metadata: precise anatomical locations and lesion-level descriptors that capture surface and morphological features of skin lesions (seen in Figure 1). To the best of our knowledge, this is the first publicly available dataset to offer both annotation types at this scale and granularity, supporting structured supervision and interpretable modeling.

- **Clinically Aligned Hierarchical Labeling.** Disease labels are organized in a three-tier hierarchy: main diagnostic class, etiology-based subclass, and specific disease label. This structure is derived from Rook's *Textbook of Dermatology* (the clinical gold standard) [9] and adapted to Indian dermatology practice. It mirrors diagnostic workflows in real-world settings, enabling both coarse- and fine-grained modeling.

- **Benchmarking for Classification and Interpretability.** We provide baseline results for disease classification and for Concept Bottleneck Models (CBMs) [23] that leverage the concept annotations. These benchmarks demonstrate the dataset's relevance for both predictive accuracy and to evaluate whether models are learning medically meaningful concepts in alignment with expert reasoning.

2 Related Work

Table 1: Comparative Survey of existing Dermatology Datasets available for AI research [Columns: **A**: Neoplasm & Tumors Centric, **B**: Broad Skin Disease Spectrum, **C**: Dermoscopic Single Lesion Focus, **D**: Real-time Multi-lesion Multi-Focus, **E**: Body Part, **F**: Lesion Descriptor, **G**: Rook's classification labels]

Dataset	Disease Distribut.		Acquisit. Type		Dense Annotat.			Source of Images		Skin Tone	Classes	
	A	B	C	D	E	F	G	Web Scraped (Atlas)	Geographic Location	Present	#Images	Hierarchical #level[#count]
ISIC Archive [18]	✓	✗	✓	✗	✓	✗	✗	✗	Europe	✗	∼485,000	1[9]
HAM10000 [40]	✓	✗	✓	✗	✓	✗	✗	✗	Austria,Australia	✗	10,015	1 [7]
DERM12345 [46]	✓	✗	✓	✗	✗	✗	✗	✗	Türkiye	✗	12,345	3 [5,15,38]
BCN20000 [14]	✓	✗	✓	✗	✓	✗	✗	✗	Spain	✗	18,946	1 [8]
PH2 [27]	✓	✗	✓	✗	✗	✓	✗	✗	Portugal	✗	200	1 [3]
PAD-UFES-20 [29]	✓	✗	✗	✓	✓	✗	✗	✗	Brazil	✗	1,612	1 [6]
DDI [4]	✓	✗	✗	✗	✓	✗	✗	✗	USA	✓	656	1[78]
Derm7pt [22]	✓	✗	✓	✗	✗	✓	✗	✓	Italy	✗	1,011	2 [5, 20]
Fitzpatrick17k [10]	✗	✓	✗	✓	✗	✗	✗	✓	–	✓	16,577	3 [3,9,114]
SD-198 [36]	✗	✓	✗	✓	✗	✗	✗	✓	–	✗	6,584	1[198]
SkinCon [5]	✗	✓	✗	✓	✗	✓	✗	✓	–	✓	3,886	✗
PASSION [8]	✗	✓	✗	✓	✓	✗	✗	✗	Africa	✓	4,901	1[4]
SCIN [43]	✗	✓	✗	✓	✓	✗	✗	crowd-sourced	USA	✓	10,000+	1 [419]
DermaCon-IN	✗	✓	✗	✓	✓	✓	✓	✗	South India	✓	5,450	3 [8,19,254]

Neoplasm-Centric Benchmarks: Early dermatology AI models were trained on datasets focused on neoplasms and tumours, such as the ISIC [18], HAM10000 [40], DERM12345 [46], etc, as discussed in Table 1. These primarily contain dermoscopic images and omit common diseases like infectious and inflammatory disorders. Dermoscopic imaging focusing on a single lesion further abstracts clinical variability in lighting, context, and lesion complexity, limiting real-time applicability.

Atlas-Sourced Clinical Datasets: SD-198 [36] and Fitzpatrick17k [10] introduced clinical (non-dermoscopic) photographs to broaden the coverage of disease spectrum. However, both are derived from educational atlases (like DermNet), not clinical repositories, yielding limited annotations. Moreover, the Fitzpatrick17k [10] dataset excludes several prominent diseases, including Fungal and Viral infections, and has skewed tonal variation of over 75% belonging to Types I–III.

Fairness-Focused Collections: Datasets like DDI [4] and PASSION [8] emphasise tonal representation but trade off diagnostic breadth. DDI includes fewer than 80 disease labels, of which neoplastic or pigmentary offer a larger contribution. PASSION [30] has pediatric participants' images across only four conditions (eczema, fungal infections, scabies, impetigo), selected for regional prevalence. SCIN [43]dataset, on the other hand, introduces crowd-sourced images, expanding coverage to common non-neoplastic conditions but mirrors U.S. disease patterns [24], thus under-representing both high-burden infectious, pigmentary, etc, disorders seen in global contexts and darker skin tones.

Our Dataset in Context: In South Asia, outpatient dermatology is dominated by inflammatory, infectious, pigmentary, and appendageal diseases [42], yet remains underrepresented in existing datasets. We address this gap with a dataset of 5,450 high-resolution clinical images collected prospectively from 2,993 Indian patients. It covers the disease spectrum aligned with Indian and global burden data. Each image retains anatomical context and is annotated by board-certified dermatologists with standardised diagnosis, as well as Fitzpatrick and MST skin tone ratings, which align with patterns observed in the Indian context [32, 33, 15]. The distribution of which is shown in Figure 2. Unlike prior work such as SkinCon [5], which retrofitted a set of lesion descriptors onto existing datasets, we capture both lesion and anatomical concepts at source and leverage the full concept set for statistical validation and model benchmarking.

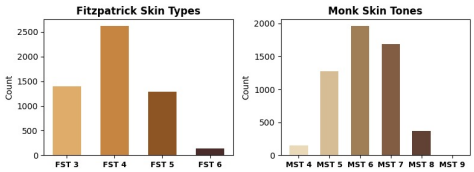


Figure 2: FST and MST distribution of skin tones of subjects in DermaCon-IN dataset

3 Data Collection Methodology

3.1 Clinical Setting and Data Sources

The DermaCon-IN dataset was developed via multi-institutional collaboration involving three tertiary-care hospitals and affiliated regional clinics across North Karnataka, South India. The cohort

represents a demographically and geographically diverse outpatient population from Karnataka, Maharashtra, Goa, and Andhra Pradesh. Data were collected between 2024 and 2025 under institutionally approved ethical protocols.

3.2 Image Acquisition Protocol

Image capture was designed to mirror real-world dermatology workflows, emphasizing both lesion detail and the broader anatomical region. Photographs were taken using high-resolution smartphone devices—108MP Android, 48MP iPhone Pro, and 12–36MP cameras—under ambient clinical lighting. Images include the affected body part with surrounding skin to preserve anatomical context and support spatial modeling. A standardized protocol guided acquisition across sites, allowing relevant variations in angle, distance, and lighting to reflect clinical realism, unlike prior datasets focused on dermoscopic or tightly cropped views.

3.3 Inclusion and Exclusion Criteria

Patients of all ages with clinically confirmed dermatologic conditions were included (with consent), contingent on diagnostic agreement by two board-certified dermatologists or follow-up validation. Only images meeting gradability standards and accompanied by complete metadata, including diagnosis, anatomical region, Fitzpatrick and Monk tone ratings, demographics, and diagnostic confidence, were retained. Exclusion criteria included poor image quality, visual obstructions (e.g., tattoos, accessories), metadata gaps, or ambiguous diagnoses, and patients unwilling to participate.

3.4 Annotation Process

The entire dataset was annotated by four board-certified Dermatologists with clinical experience of 11 years, 3 years, 3 years, and 1 year, respectively. The entire dataset was divided into four smaller subsets for labelling by these doctors based on the availability of the dermatologist. Labels followed a three-level disease taxonomy informed by Rook's Classification [9], which is considered a gold standard in Dermatology (Refer to Supplementary Sec. A). Annotations also included 47 lesion-level descriptors and 49 body part locations, along with patient metadata not linked to patients' privacy. Discrepancies were resolved via consensus or adjudication by a third expert.

3.5 Quality Control and Inter-Rater Agreement

To ensure consistency, 10–15% of each annotator's batch was randomly reviewed by another dermatologist. Inter-rater reliability, measured using Cohen's Kappa, achieved a score of 0.84 (Figure 3), aligning with accepted clinical annotation standards. Skin tone ratings were independently assigned by the set of trained experts, which was verified for consensus by the dermatologist. Anatomical site labels were validated using structured region maps.

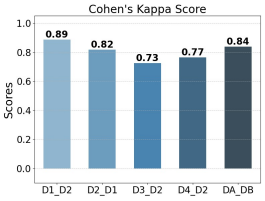


Figure 3: Cohen's Kappa scores for cross validation of annotations provided by 4 doctors (D1, D2, D3, D4)

3.6 Final Dataset Composition

The final composition contains $5{,}450$ high-resolution JPEG images across 8 top-level etiologic classes, 19 clinically meaningful subclasses, and 245 fine-grained disease labels. Each sample includes dense metadata: hierarchical disease labels, body parts, skin lesion descriptors, Fitzpatrick [7] and Monk skin tone [34] scores, diagnostic certainty, and image gradability. We consider 49 body parts and 47 lesion descriptors as concepts, which account for 96 unique concepts.

4 Dataset Overview and Statistics

Diagnostic structure and class granularity: DermaCon-IN is organized around an aetiologically informed, clinically validated taxonomy rooted in Rook's [9] classification and aligned with ICD-11 [44]. The dataset includes 8 high-level diagnostic categories ranging from Infectious to Neoplastic, including No Definite Diagnosis — reflecting the full spectrum of dermatological conditions prevalent in South Asian outpatient settings [42]. These main classes are further expanded into 19 subclasses to capture hybrid and co-occurring disease states. For example, subclass combinations such as Inflammatory + Infectious (Bacterial) reflect polymorphic real-world presentations of superimposed two diseases. Each top-level category is populated with sufficient training instances for stratified benchmarking; notably, high-burden groups like fungal infections and eczema dominate in volume, while rare but clinically significant categories—e.g., keratinisation disorders remain represented with enough density to enable few-shot generalisation. The Disease label reflects 245 distinct disease types, which follow a long-tailed distribution with a log-normal fit to leaf-node frequencies, yielding

an exponent of 1.8, as expected in real-world clinical settings. To better understand label diversity in the dataset, refer Figure 4. This visualization underscores the need for imbalance-aware learning strategies.

Lesion Descriptor annotations: The visual descriptors of skin lesions annotated in DermaCon-IN follow an established dermatological vocabulary commonly used by clinicians to describe skin lesions in line with extant clinical literature [3]. Inspired by frameworks [5] previously validated in dermatology AI literature, we consulted board-certified dermatologists to select clinically meaningful descriptors aligned with disease patterns frequently encountered in Indian outpatient settings. Rather than assigning fixed concepts per disease category, each image was independently annotated, reflecting the realistic variability in lesion appearance across patients, body regions, and skin types. Statistical analysis in Figure 5 of concept-disease correlations validated this flexible annotation strategy that supports concept-supervised learning approaches. Details of Descriptors utilized for annotation are provided in the Datasheet (Supplementary Material).

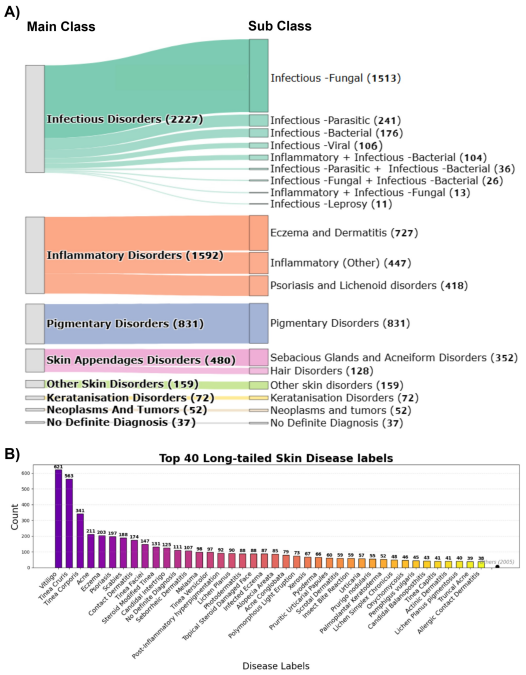


Figure 4: (A) Hierarchical structure of dermatological annotations (according to Rook's classification) showing main classes and their corresponding sub-classes with image counts (in bracket). (B) Distribution of the top 40 most frequent disease labels, illustrating the long-tailed nature of the dataset.

Anatomical Site annotations: We retain full views of affected anatomical regions wherever feasible. Partial cropping is applied in cases when identifiable facial features are present, in line with ethical guidelines to ensure patient privacy while preserving anatomical context reflecting how skin conditions are typically encountered in practice. This anatomical framing allows models to learn spatial priors: for instance, tinea capitis typically affects the scalp, while palmoplantar keratoderma appears on the palms and soles.

Such distributional cues are vital for accurate real-world diagnosis but are often lost in lesion-only images. Our dataset includes annotations for 49 distinct body sites, organised using both precise clinical terms and broader regional tags (e.g., trunk, extremities, scalp). A List of body parts annotated is provided in the the Datasheet (Supplementary Material).

Need for integration of Anatomical Context and Lesion Descriptors: Clinically, the same lesion type can indicate different diagnoses depending on its location; for example, a scaly plaque on the scalp might point to psoriasis, whereas a scaly plaque on the foot can suggest a fungal infection (tinea pedis). By jointly annotating lesion features (e.g., scales, plaques) and their anatomical context (e.g., scalp, foot), our dataset enables models to better reflect dermatologists' real-world diagnostic approach.

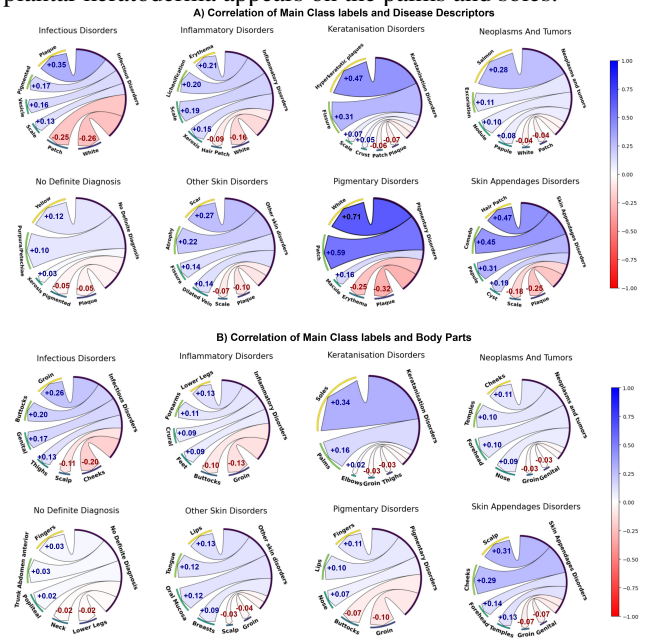


Figure 5: Chord diagrams of Pearson correlations (r) between eight disease classes and two concept types: **(A)** lesion descriptors and **(B)** body regions. Ribbon width represents within-class correlation strength, and colour encodes direction and magnitude (dark blue = positive; red = negative). Numbers on ribbons indicate r. Top-k positive and negative associations per class highlight distinctive clinical and anatomical patterns across disease groups.

Population structure and label density: The dataset reflects age and sex demographics typical of South Indian outpatient settings. The Dataset also adopts both Monk Skin Tone (MST) [34] and Fitzpatrick [7] scales, addressing the limitations of Fitzpatrick's UV-response bias. MST offers a perceptual alternative capturing wider tonal representation, especially for darker skin types, as the tonal scale has increased variation. The combined annotation aligns with Indian phenotypic distributions (MST 4–9, Fitzpatrick 3–4).

Statistical validation of concept and Anatomical coherence with Disease Labels: We examined Pearson correlation coefficients between (a) Disease descriptors and disease categories, and (b) anatomical body regions and disease categories, (refer Figure 5) to assess the biological plausibility and interpretability of our clinical annotations. Each chord diagram visualizes these relationships, where ribbon **width** indicates the strength of association between a concept and the class (within-class correlation) and ribbon **color** (dark blue $\rightarrow$ red) encodes the strength of correlation across classes (dark blue = strong positive; red = negative). Numeric labels on ribbons denote the actual correlation coefficients.

For instance, under *Pigmentary Disorders*, the descriptor *White* shows a moderately wide ribbon and dark-blue color ($r = +0.71$), reflecting a strong and distinctive association both within the class and relative to other classes. In contrast, *Hyperkeratotic plaques* under *Keratinization Disorders* display a wider but lighter-blue ribbon ($r = +0.47$), suggesting it is more class-specific but less distinctive across classes. Overall, the chord diagrams reveal statistically meaningful associations that align well with established dermatological knowledge. For instance, positive (high) correlations are observed between *erythema*, *vesicle*, and *scale* with inflammatory disorders, and between *hyperkeratotic plaques* and keratinisation disorders. Similarly, *white patches* and *pigmented lesions* show high positive associations with pigmentary and infectious disorders, respectively, underscoring the dermatologic specificity of the disease descriptors in general. Anatomical correlations further reinforce clinical fidelity. Keratinisation disorders predominantly localize to the *soles* and *palms*, consistent with plantar keratoderma patterns. Skin appendageal disorders, such as acne and seborrheic dermatitis, are strongly associated with sebaceous-rich zones like the *scalp* and *cheeks* [3, 31]. These findings validate the anatomical tropisms encoded in the dataset.

5 Challenges, Opportunities & Limitations

The dataset presents a range of challenges and opportunities that stem from the inherent complexity of real-world clinical data, offering both practical constraints and avenues for robust model development:

Resolution heterogeneity: Images were cropped post-acquisition to remove garments and background clutter where feasible, though incidental artifacts (e.g., jewelry) remain in some cases. The resulting variability in resolution, arising from diverse capture devices and aspect ratios, is a natural outcome, but advantageous. It reflects real-world conditions, where patients may crop or capture images themselves, and encourages model robustness to such variations. Image heights range from 296 to 4,608 pixels and widths from 346 to 4,608 pixels, with a mean resolution of 2,300x2,057 pixels. The average image area is 4.97M pixels ($\pm$3.01M), with an interquartile range of 2.62M–6.69M pixels (Figure 6), indicating consistently high-fidelity input for fine-grained modeling.

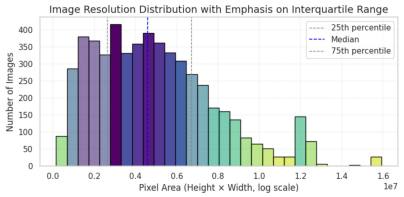


Figure 6: Resolution distribution of images in DermaCon-IN dataset

Hierarchical Labels and Clinical Distribution:

- **Class Imbalance:** Our hierarchical labelling across main and sub-classes reflects true outpatient frequencies, common conditions are well-represented, while others occur proportionally less. This mirrors real-world practice and enables models to learn from naturally occurring clinical distributions.
- **Long-tailed Disease Labels:** The dataset embraces the inherent long-tailed nature of dermatological diagnoses, where few diseases are prevalent and many are rare. This structure presents a valuable opportunity to train models that generalise across the full clinical spectrum, including rare disease categories.
- **Multi-disease co-occurrences:** A subset shows concurrent lesions from multiple disease types on the same anatomical site (e.g., *Inflammatory+Fungal*, *Fungal+Bacterial*). We

Table 2: Performance benchmarking of the proposed dataset on the 8-main class diagnosis classification task, conducted using both CNN- and ViT-based standard architectures. All metrics are averaged over 5 random seeds.

Model	Pre-trained	Accuracy	Balanced Acc.	Precision	Sensitivity	F1 Score
ResNet50 [13]	-	$47.45_{\pm 0.40}$	$23.93_{\pm 0.01}$	$46.59_{\pm 0.51}$	$47.44_{\pm 0.41}$	$46.43_{\pm 0.60}$
DenseNet121 [17]	-	$49.30_{\pm 0.30}$	$25.31_{\pm 0.01}$	$47.49_{\pm 0.86}$	$49.30_{\pm 0.29}$	$48.17_{\pm 0.74}$
ResNet50 [13]	ImageNet	$64.31_{\pm 0.22}$	$38.77_{\pm 1.40}$	$63.41_{\pm 0.29}$	$64.31_{\pm 0.23}$	$63.31_{\pm 0.25}$
DenseNet121 [17]	ImageNet	$65.20_{\pm 0.48}$	$37.31_{\pm 0.01}$	$64.62_{\pm 0.45}$	$65.20_{\pm 0.48}$	$64.37_{\pm 0.35}$
EffNet-B4 [38]	ImageNet	$64.28_{\pm 0.34}$	$35.58_{\pm 0.01}$	$63.53_{\pm 0.64}$	$64.27_{\pm 0.34}$	$63.38_{\pm 0.39}$
ViT-B/16-224 [6]	ImageNet	$64.09_{\pm 1.03}$	$34.56_{\pm 0.01}$	$62.59_{\pm 1.03}$	$62.88_{\pm 1.67}$	$62.98_{\pm 1.02}$
ViT-B/16-384 [6]	ImageNet	$66.95_{\pm 0.19}$	$35.78_{\pm 0.02}$	$65.39_{\pm 0.13}$	$66.95_{\pm 0.20}$	$65.78_{\pm 0.06}$
MaxViT-B/512 [41]	ImageNet	$66.92_{\pm 0.48}$	$36.07_{\pm 0.01}$	$66.30_{\pm 0.80}$	$66.92_{\pm 0.48}$	$65.90_{\pm 0.73}$
Swin-B/4W12-384 [25]	ImageNet	$\mathbf{70.41_{\pm 0.41}}$	$\mathbf{45.06_{\pm 0.02}}$	$\mathbf{69.83_{\pm 0.37}}$	$\mathbf{70.41_{\pm 0.42}}$	$\mathbf{69.69_{\pm 0.46}}$

Table 3: Performance comparison of model variants using the `Swin-B/4W12-384` backbone. The first two rows are baselines without a concept bottleneck (CB) layer, used for 8-main class (MC) and 19-subclass (SC) classification. The next four rows report CBMs trained with different concept sets: lesion descriptors (47), body parts (49), and both (96). The last merged rows present individual layer (SC & MC) performance of a Hierarchical CBM (Type 1 & 2) combining the CB layer with joint SC and MC classification as described in Fig 7. All metrics are results of end-to-end training and are averaged over 5 random seeds.

Classification head	Concepts	Accuracy	Precision	Sensitivity	F1 Score	Macro AUC
MC	-	$\mathbf{70.41_{\pm 0.41}}$	$\mathbf{69.83_{\pm 0.37}}$	$\mathbf{70.41_{\pm 0.42}}$	$\mathbf{69.69_{\pm 0.46}}$	$\mathbf{78.51_{\pm 0.59}}$
SC	-	$58.27_{\pm 0.22}$	$56.64_{\pm 0.45}$	$58.27_{\pm 0.22}$	$56.81_{\pm 0.43}$	$83.11_{\pm 2.55}$
(CBM-D) Concepts + MC	Descriptors	$68.57_{\pm 0.72}$	$67.63_{\pm 0.97}$	$68.55_{\pm 0.72}$	$67.69_{\pm 79.48}$	$85.18_{\pm 2.98}$
(CBM-B) Concepts + MC	Body parts	$68.38_{\pm 0.31}$	$67.69_{\pm 0.51}$	$68.31_{\pm 0.27}$	$67.90_{\pm 0.26}$	$84.96_{\pm 1.27}$
Concepts + MC	Descriptors & Body parts	$68.12_{\pm 0.43}$	$67.69_{\pm 0.71}$	$68.10_{\pm 0.48}$	$67.56_{\pm 0.48}$	$82.78_{\pm 2.70}$
(CBM-D) Concepts + SC	Descriptors	$56.42_{\pm 0.01}$	$55.72_{\pm 0.01}$	$56.51_{\pm 0.01}$	$55.63_{\pm 0.01}$	$80.16_{\pm 0.01}$
(CBM-B) Concepts + SC	Body parts	$55.88_{\pm 0.01}$	$55.47_{\pm 0.01}$	$55.90_{\pm 0.01}$	$54.87_{\pm 0.01}$	$78.94_{\pm 0.01}$
Concepts + SC	Descriptors & Body parts	$57.37_{\pm 0.01}$	$57.67_{\pm 0.01}$	$57.27_{\pm 0.01}$	$56.90_{\pm 0.01}$	$78.39_{\pm 0.02}$
(Type1) SC	Descriptors & Body parts	$53.98_{\pm 0.40}$	$56.12_{\pm 0.74}$	$53.98_{\pm 0.39}$	$54.49_{\pm 0.66}$	$76.13_{\pm 1.52}$
(Type1) MC	Descriptors & Body parts	$67.78_{\pm 0.47}$	$67.32_{\pm 0.67}$	$67.66_{\pm 0.48}$	$66.92_{\pm 0.37}$	$79.53_{\pm 0.62}$
(Type 2) SC	Descriptors & Body parts	$56.11_{\pm 0.67}$	$55.49_{\pm 0.0.62}$	$56.09_{\pm 0.9}$	$54.49_{\pm 0.66}$	$76.13_{\pm 1.52}$
(Type 2) MC	Descriptors & Body parts	$69.90_{\pm 0.20}$	$68.82_{\pm 0.36}$	$69.89_{\pm 0.19}$	$69.08_{\pm 0.31}$	$77.01_{\pm 2.24}$

represent these as dedicated subclasses to help models disentangle overlapping pathologies for the main-class label we follow dermatologists' treatment-priority logic—e.g., an infected eczema is assigned to the Infectious main class (not Inflammatory) because initial management targets the infection. Such cases, though less frequent, are observed in clinical settings, and our targeted misclassification analysis for these samples is provided in the Supplementary Sec. **B2**.

Instance-level concepts: These concepts describe what is visually observed in each image, such as scaling and erythema, and introduce two key characteristics that make DermaCon-IN a rich dataset for advancing clinical AI:

- **Class-agnostic semantics:** Descriptors (used as concepts) are shared across classes and are not rigidly tied to any single diagnostic category (e.g., *scaling* alone $\rightsquigarrow$ ichthyosis, whereas *scaling* + *erythema* $\rightsquigarrow$ psoriasis), with diagnostic meaning arising from their combinations. This invites the development of models capable of compositional and context-aware reasoning.
- **Long-tailed distribution:** The natural skew in concept frequencies mirrors real-world prevalence, where rare but critical findings coexist with common patterns. This creates opportunities to tackle challenges in multi-concept learning and rare concept detection—core problems in clinical AI.

Limitations. In accordance with ethical considerations, all images were anonymized by masking identifiable facial features, such as the eyes, and cropping facial regions where necessary. While essential for protecting patient identity, this may limit the model's ability to accurately learn or detect diseases that primarily manifest on the face.

6 Benchmarking with Models

6.1 Standard architectures

DermaCon-IN comprises high-resolution clinical photographs with multiple co-occurring lesions, varied anatomical regions, and hierarchically structured multi-label annotations. We selected archi-

tectures based on their complementary modelling strengths to benchmark model performance under these conditions. Convolutional neural networks such as ResNet50 [13], DenseNet121 [17], and EfficientNet (`EffNet-B4`) [38] are effective in capturing localised texture patterns and edge-level features, owing to their convolutional inductive biases and limited receptive fields. To complement this, we incorporated Vision Transformer (ViT) architectures [6], which leverage self-attention to relate spatially distant regions within an image. The ViT variants evaluated includes ViT-Base (`ViT-B/16-224` [6], `ViT-B/16-384`) [6], MaxViT-Base (`MaxViT-B/512`) [41], and Swin-Base (`Swin-B/4W12-384`) [25]. Among these, the Swin Transformer consistently achieved the best results for Main class prediction across evaluation metrics, demonstrating improved handling of both multi-class classification and class imbalance. Table 2 summarises the performance of all models considered. Swin Transformer's shifted window mechanism enables efficient modeling of non-contiguous regions, while its hierarchical representation captures both fine-grained lesion details and broader spatial patterns. These traits align closely with our dataset. We believe this alignment contributed to Swin's better performance.

Based on these observations, we adapted the `Swin-B/4W12-384` [25] variant of the Swin Transformer as the backbone for subsequent analysis and in concept bottleneck models (CBMs) [23] as shown in Table 3. The model was initialised with weights pretrained on ImageNet-22k and fine-tuned end-to-end on our dataset. Input images were resized and padded to 512×512 for further classification. We performed a stratified, subject-wise 80:20 split over Sub Class and reported all the results with the same split. We adapted weighted sampling strategies to handle class imbalance, and details of which are discussed in Supplementary Sec. B.

6.2 Concept Bottleneck Modeling for Interpretablity

Architecture. Given an input image x, the Swin-Transformer encoder E_θ yields a latent representation $z = E_\theta(x)$. A linear projection maps z to *concept logits* $c^\ell \in \mathbb{R}^{B+D}$, which are then passed through a sigmoid activation to obtain the interpretable *concept vector* $c = [c^{\text{bp}}, c^{\text{ld}}] = \sigma(c^\ell) \in [0,1]^{B+D}$, where c^{bp} denotes B **body-part** concepts and c^{ld} denotes D **lesion-descriptor** concepts. While both c and c^ℓ are used for interpretability and concept supervision, the downstream classifier f operates on the concept logits c^ℓ to produce task logits $y = f(c^\ell)$, predicting either a *Main-Class* (MC) or *Sub-Class* (SC) label. This architecture is presented in Figure 7(A).

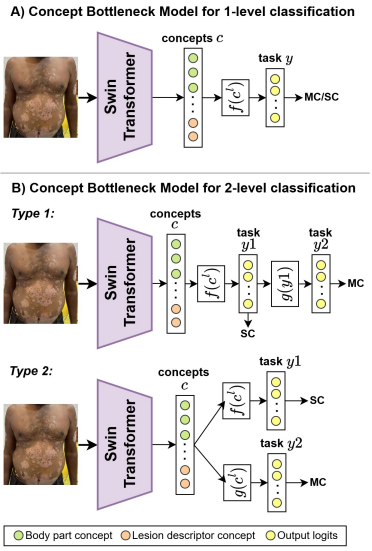


Figure 7: Architectural setup of Concept-bottleneck models for Main class (MC) and Sub class (SC) classification

Concept ablation and joint–stream effects. To quantify the contribution of each concept group, we trained two ablated models: **CBM-D**, which keeps only lesion-descriptor concepts c^{ld}, and **CBM-B**, which keeps only body-part concepts c^{bp}, both predicting MC labels. Each single–stream variant retained high accuracy (Table 2), demonstrating that the two concept families are independently learnable.

By contrast, the *full* CBM represents the true clinical diagnostic fidelity, in which both c^{bp} and c^{ld} co-exist in the bottleneck, achieved performance comparable to the individual concept streams, but revealed a systematic imbalance in activation: In many samples, only one concept group (typically descriptors) fired strongly, whereas the other (body parts) was under-activated (Fig. 8, bottom row). This led to a modest but consistent drop in overall accuracy, pointing to a *representational bottleneck* whereby competition for limited capacity biases the model toward a single semantic stream. These observations emphasize the need for improved multi-concept learning mechanisms that can balance several concept families simultaneously.

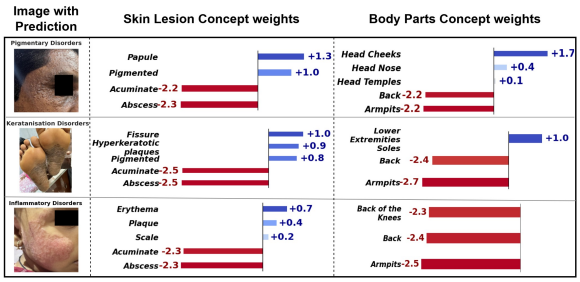


Figure 8: Bar plot of top and bottom-k contributing concepts (lesion descriptors and body parts) for the model's prediction. Contributions are shown as signed log-scaled weights derived from the CBM's intermediate logits.

Hierarchical CBMs. We explore two designs for joint SC–MC prediction (Figure 7(B)):

1. **Type 1 — cascade.** Concepts first predict sub-classes via $y1 = f(c)$. These logits are then mapped to the main classes through a second head $y2 = g(y1)$, ensuring taxonomy consistency by construction.
2. **Type 2 — parallel.** Both SC and MC are predicted from the shared concept vector through independent heads, $y1 = f(c)$ and $y2 = g(c)$, leveraging multi-task learning for implicit regularization.

Empirically, the *parallel* configuration surpassed the *cascade* alternative across all evaluation metrics (Table. 2), likely due to effective regularization and information sharing through the multi-task learning setup.

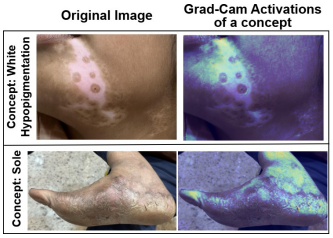


Figure 9: Examples of Grad-cam visualizations over Swin Transformer by choosing a specific concept with the best CBM model (Type-2).

Qualitative Analysis. To validate the spatial grounding of concepts, we employed Grad-CAM visualizations over Swin ViT on specific concept heads (Figure 9). For each selected concept (from both descriptor and body part categories), we backpropagated gradients from the concept prediction to the image space, producing activation heatmaps. These visualizations confirmed that the model's concept activations were often localized to semantically and anatomically appropriate regions, supporting the faithfulness of the learned representations.

To further assess whether the model's predictions are semantically grounded, we evaluate the alignment between its learned class-concept weights (MC branch of Type-2) and the statistical relevance (Pearson correlation, Section 4) of each concept to the class labels in the dataset (Figure 10). Alignment varies notably across diagnostic categories. *Pigmentary Disorders* and *Keratinisation Disorders* show strong Spearman correlations and statistically significant p-values ($p < 0.05$), suggesting that the model reliably prioritises clinically meaningful features for these classes. In contrast, *Neoplasms and Tumors* show weak or negative alignment, indicating reliance on non-semantic or latent cues, possibly due to low representation in the dataset. Whereas *No Definite Diagnosis* shows negative correlations as desired. Classes like *Skin Appendageal Disorders, etc.* exhibit partial alignment.

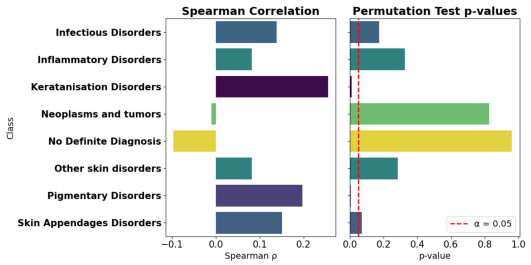


Figure 10: Class-wise assessment of alignment between the model's learned label weights over concepts and the dataset-derived correlation of concepts with class labels. The Spearman correlation measures the rank agreement between model-assigned concept importance and empirical concept–class correlations. Permutation test p-values (based on Pearson correlation) assess the statistical significance of this alignment.

Overall, these findings highlight that while the model is capable of learning semantically meaningful representations in some contexts, its reliance on concept supervision is uneven and class-dependent. This motivates future work in enforcing more robust concept–class alignment, particularly for clinically ambiguous or visually heterogeneous disease categories.

7 Conclusion and Future Directions.

DermaCon-IN captures real-world dermatological presentations from the South Indian population, offering concept-level annotations such as body parts and disease descriptors. It addresses gaps in skin tone diversity and regional disease patterns, serving both as a region-specific benchmark and a valuable addition to global dermatological datasets. By enabling evaluation beyond labels, it supports clinically grounded, interpretable learning—a step closer towards clinical deployment.

Targeted future work. Our findings point to several *method-driven* avenues for closing the identified gaps. These steps can help models to not only classify accurately but also reason in clinically meaningful ways. They are as follows:

- **Concept-weight normalisation:** Adding explicit regularisers that penalise disproportionate reliance on a single concept group, encouraging balanced gradient flow.
- **Hierarchy-consistent objectives:** coupling the SC and MC heads with cross-level consistency losses to discourage contradictory evidence pathways.

- **Curriculum and re-sampling strategies:** oversampling under-represented concept combinations (e.g. body-part signals within neoplasm classes) to equalize learning pressure across concept space.

Broader integration. Because DermaCon-IN offers strong coverage of South-Indian skin tones and disease spectra, its concept annotations complement existing global datasets. Merging these resources will enable training of larger, more broadly representative models—paving the way toward a unified *foundation-level* representation for dermatological imaging. Looking ahead, we will expand DermaCon-IN as versioned releases sourced from additional centers across India, thereby deepening geographic, phenotypic, and skin-tone coverage while preserving curation standards. In parallel, we plan to add pixel-level lesion masks for a subset of images, enabling segmentation and tighter alignment between concepts and spatial evidence. We hope this resource will move dermatology AI closer to responsible clinical use.

Ethical Clearance Statement: The dataset was collected in accordance with institutional ethical guidelines and has been approved by the Institute Ethics Committee of the Indian Institute of Technology Hyderabad under protocol number *IITH/IEC/2025/01/05*.

Implementation: The experiments were run on 4 GPUs of Nvidia-A6000, each of 42GB RAM. The dataset sizes around ~ 4 GB. The code used in this work is available at GitHub.

Availability and Licensing: The dataset can be downloaded at Harvard Dataverse. This work is licensed under CC BY-NC-SA 4.0. To view a copy of this license, visit creativecommons.org.

Funding and Acknowledgement: This work was not supported by any external funding. All contributors were involved out of self-motivation and shared interest in advancing dermatological AI research. We would also like to thank Dr. Deepanshu Bansal for his contributions to the dataset labelling process.

References

[1] N. Alipour et al. Skin type diversity in skin lesion datasets: A review. *International Journal of Dermatology*, 63:198–210, 2024. doi:10.1007/s13671-024-00440-0.

[2] P. Balasubramanian et al. Epidemiological study of skin disorders in andaman and nicobar islands. *Indian Journal of Dermatology*, 66(5):454–458, 2021. doi:10.4103/ijd.IJD_30_20.

[3] Jean L. Bolognia, Julie V. Schaffer, and Lorenzo Cerroni. *Dermatology*. Elsevier, 4 edition, 2017. ISBN 9780702062759.

[4] R. Daneshjou et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Scientific Reports*, 12:12565, 2022. doi:10.1126/sciadv.abq6147.

[5] R. Daneshjou et al. Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/hash/7318b51b52078e3af28197e725f5068a-Abstract-Datasets_and_Benchmarks.html`.

[6] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[7] Thomas B. Fitzpatrick. Soleil et peau. *Journal de Médecine Esthétique*, 2:33–34, 1975.

[8] P. Gottfrois et al. Passion for dermatology: Bridging the diversity gap with pigmented skin images from sub-saharan africa. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 703–712. Springer Nature Switzerland, 2024.

[9] Christopher E.M. Griffiths et al. *Rook's Textbook of Dermatology*. Wiley-Blackwell, 10 edition, 2024. ISBN 9781119709213. URL `https://www.wiley.com/en-us/Rook%27s%2BTextbook%2Bof%2BDermatology%2C%2B4%2BVolume%2BSet%2C%2B10th%2BEdition-p-00402062`. 4-volume set.

[10] M. Groh et al. 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 (CVPR)*, pages 1820–1828, 2021.

[11] Robert J Hay, Neil E Johns, Hannah C Williams, and et al. The global burden of skin disease in 2010: An analysis from the global burden of disease study 2010. *The Journal of Investigative Dermatology*, 134(6): 1527–1534, 2014. doi:10.1038/jid.2013.446.

[12] Roderick J. Hay et al. Skin disease in the tropics and the lessons that can be learned from leprosy and other neglected diseases. *Acta Dermato-Venereologica*, 100:adv00113, 2020. doi:10.2340/00015555-3469.

[13] Kaiming He et al. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[14] C. Hernández et al. Bcn20000: Dermoscopic lesions in the wild. *Scientific Data*, 11:641, 2024. doi:10.1038/s41597-024-03387-w.

[15] V. Hourblin et al. Skin complexion and pigmentary disorders in facial skin of 1204 women in 4 indian cities. *Indian Journal of Dermatology, Venereology and Leprology*, 80(5):395–401, 2014. doi:10.4103/0378-6323.140290. URL `https://doi.org/10.4103/0378-6323.140290`.

[16] Pengcheng Huai et al. Global burden of skin and subcutaneous diseases: an update from the global burden of disease study 2021. *British Journal of Dermatology*, ljaf071, 2025. doi:10.1093/bjd/ljaf071. Published: 11 April 2025.

[17] Gao Huang et al. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708, 2017.

[18] ISIC. Siim-isic 2020 challenge dataset, 2020. URL `https://doi.org/10.34970/2020-ds01`. Creative Commons Attribution-Non Commercial 4.0 International License.

[19] N. S. Jayanthi et al. Epidemiological pattern of skin diseases among patients attending dermatological outpatient department at a tertiary care centre, north chennai. *Indian Journal of Clinical and Experimental Dermatology*, 3(4):134–137, 2017. doi:10.18231/2455-6769.2017.0032.

[20] Emily Johnson and Robert Lee. Global burden of skin diseases in 2019: Updated estimates from the global burden of disease study 2019. *The Lancet Global Health*, 8(11):e1539–e1540, 2020. doi:10.1016/S2214-109X(20)30386-7.

[21] C. Karimkhani et al. Global skin disease morbidity and mortality: An update from the global burden of disease study 2013. *JAMA Dermatology*, 153(5):406–412, 2017. doi:10.1001/jamadermatol.2016.5538.

[22] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh. 7-point checklist and skin lesion classification using multi-task multi-modal neural nets. *IEEE Journal of Biomedical and Health Informatics*, Apr 2018. doi:10.1109/JBHI.2018.2824327. Epub ahead of print.

[23] Pang Wei Koh et al. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 5338–5348, 2020.

[24] Melissa R. Laughter, Mayra B. C. Maymone, Chante Karimkhani, Chandler Rundle, Sophia Hu, Sophia Wolfe, Katrina Abuabara, Parker Hollingsworth, Gil S. Weintraub, Cory A. Dunnick, Adnan Kisa, Giovanni Damiani, Aziz Sheikh, Jasvinder A. Singh, Takeshi Fukumoto, Rupak Desai, Ayman Grada, Irina Filip, Amir Radfar, Mohsen Naghavi, and Robert P. Dellavalle. The burden of skin and subcutaneous diseases in the united states from 1990 to 2017. *JAMA Dermatology*, 156(8): 874–881, 2020. doi:10.1001/jamadermatol.2020.1573. URL `https://jamanetwork.com/journals/jamadermatology/fullarticle/2767074`.

[25] Ze Liu et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[26] M. López-Pérez et al. Are generative models fair? a study of racial bias in dermatological image generation, 2025. URL `https://arxiv.org/abs/2501.11752`.

[27] T. Mendonça et al. Ph²: A dermoscopic image database for research and benchmarking. In *Proceedings of the 35th International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, Japan, July 2013.

[28] A. O'Malley et al. Ensuring appropriate representation in artificial intelligence–generated medical imagery: Protocol for a methodological approach to address skin tone bias. *JMIR AI*, 3:e58275, 2024. doi:10.2196/58275. URL `https://ai.jmir.org/2024/1/e58275`.

[29] A. G. C. Pacheco et al. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 32:106221, 2020. doi:10.1016/j.dib.2020.106221.

[30] Adam Paszke et al. Pytorch: An imperative style, high-performance deep learning library. https://github.com/pytorch/pytorch, 2019.

[31] Ronald P. Rapini. *Dermatology: 2-Volume Set*. Mosby, 1 edition, 2007. ISBN 9780721601573.

[32] S. Sachdeva. Fitzpatrick skin typing: Applications in dermatology. *Indian Journal of Dermatology, Venereology and Leprology*, 75(1):93–96, 2009. doi:10.4103/0378-6323.45238. URL https://doi.org/10.4103/0378-6323.45238.

[33] R. Sarkar et al. A randomised study to evaluate the efficacy and effectiveness of two sunscreen formulations on indian skin types iv and v with pigmentation irregularities. *Indian Journal of Dermatology, Venereology and Leprology*, 85(2):160–168, Mar-Apr 2019. doi:10.4103/ijdvl.IJDVL_932_17.

[34] Candice Schumann, Femi Olanubi, Auriel Wright, Ellis Monk, Courtney Heldreth, and Susanna Ricco. Consensus and subjectivity of skin tone annotation for ml fairness. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/60d25b3210c92f5ba2002a8e1f1adf1c-Abstract-Datasets_and_Benchmarks.html.

[35] B. Shah et al. Epidemiological study of skin diseases in himatnagar. *International Journal of Research in Dermatology*, 5(2):342–345, 2019. doi:10.18203/issn.2455-4529.IntJResDermatol20190453.

[36] X. Sun et al. A benchmark for automatic visual classification of clinical skin disease images. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, volume 9910 of *Lecture Notes in Computer Science*, pages 206–222. Springer, Cham, 2016. doi:10.1007/978-3-319-46466-4_13.

[37] M. D. Szeto et al. Dermatologic data from the global burden of disease study 2019 and the patientslikeme online support community: Comparative analysis. *JMIR Dermatology*, 7:e50449, 2024. doi:10.2196/50449. URL https://pubmed.ncbi.nlm.nih.gov/39661989/.

[38] Mingxing Tan et al. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 6105–6114. PMLR, 2020.

[39] K. J. S. Thakur et al. The burden of skin diseases in india: Global burden of disease study 2017. *Indian Journal of Dermatology, Venereology and Leprology*, 87(6):764–771, 2021. doi:10.25259/IJDVL_978_20.

[40] P. Tschandl et al. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5:180161, 2018. doi:10.1038/sdata.2018.161. URL https://doi.org/10.1038/sdata.2018.161.

[41] Zihang Tu et al. Maxvit: Multi-axis vision transformer. In *European Conference on Computer Vision (ECCV)*, 2022.

[42] Katelyn Urban et al. The global, regional, and national burden of fungal skin diseases in 195 countries and territories: A cross-sectional analysis from the global burden of disease study 2017. *JAAD International*, 2:22–27, 2021. doi:10.1016/j.jdin.2020.10.003.

[43] A. Ward et al. Creating an empirical dermatology dataset through crowdsourcing with web search advertisements. *JAMA Network Open*, 7(11):e2446615, November 2024. doi:10.1001/jamanetworkopen.2024.46615.

[44] World Health Organization. *International Classification of Diseases, 11th Revision (ICD-11)*, 2019. URL https://icd.who.int/. Adopted by the 72nd World Health Assembly in 2019; came into effect on 1 January 2022.

[45] Aobuliaximu Yakupu et al. The burden of skin and subcutaneous diseases: findings from the global burden of disease study 2019. *Frontiers in Public Health*, 11:1145513, 2023. doi:10.3389/fpubh.2023.1145513.

[46] A. Yilmaz et al. Derm12345: A large, multisource dermatoscopic skin lesion dataset with 40 subclasses. *Scientific Data*, 11(1):1302, November 2024. doi:10.1038/s41597-024-04104-3.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes. Our main contributions are clearly outlined in Section 1, following the motivation established in Section 1. The positioning of the dataset is appropriately framed in Section 2, where we provide a high-level overview of DermaCon-IN. A more granular description is presented in Section 4. Finally, benchmarking results demonstrating the dataset's utility are detailed in Section 6.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation, as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes. We explicitly discuss the limitations of our DermaCon-IN in Section 5. In addition to outlining dataset-specific constraints, we also highlight the broader challenges associated with modeling our data, which we view as opportunities for future research and innovation 5.

   Guidelines:

   - The answer NA means that the paper has no limitation, while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed not to penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes. We provide complete training and inference code on GitHub, along with detailed documentation, model checkpoints, and dataset splits to enable full reproduction of our results (refer to Implementation subsection 7 for the GitHub link).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Yes. We provide open access to the complete dataset, accompanying metadata, and all code necessary to reproduce the main experimental results. Detailed instructions and the access link are provided in Section 7.

   Guidelines:

   - The answer NA means that the paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes. All necessary implementation details, training configurations are provided in the Supplementary Sec B. The data splits 6.1 are made publicly available, and comprehensive model details are described in the benchmarking section (Section 6 and at 7), ensuring transparency and reproducibility of results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes. We report appropriate measures of statistical significance, such as error bars, in the (Section 6). These are computed over 5 different seed values to ensure robustness and reproducibility of the reported results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Yes. We provide detailed information on the computational resources used in the experiments (please refer to 7), including GPU types, memory capacity.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes. Our research fully complies with the NeurIPS Code of Ethics. All data used in this work were collected and processed with appropriate approvals, and our study adheres to principles of fairness, transparency, and respect for privacy.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: [Yes] . Our work is centered on addressing gaps in dermatological AI for underrepresented populations and low-resource settings. While the full discussion is woven throughout the paper, which culminates in a summary of the potential societal benefits and is provided in the conclusion (Section 7).

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Yes. All patient images are de-identified and stripped of personally identifiable information prior to release. The dataset is hosted on a controlled platform (Harvard Dataverse) with clear licensing and intended-use guidelines. We explicitly discourage any downstream use beyond academic or clinical research aligned with the goals of this work. While access is currently open in compliance with NeurIPS reproducibility requirements, we intend to adopt a lightweight verification process post-publication to ensure responsible use offering continued public access to researchers whose affiliations and research purposes are consistent with ethical and scholarly standards.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes. All external assets are properly acknowledged as detailed in the paper. Our released dataset and code are licensed under the CC BY-NC-SA 4.0 license. Full licensing details are provided in Subsection 7.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Yes. We provide datasheets and detailed documentation for all newly introduced assets as part of the supplementary material.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Yes. This was an observational study involving clinical images collected under IRB-approved protocols. No crowdsourcing or task-based participation occurred. The study involved no risk to participants, as all images were de-identified and cropped to exclude any identifiable features. Full ethical details are provided in the supplementary material.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Yes. The study was conducted with approval from the appropriate Institutional Review Board (IRB), in accordance with institutional and national ethical guidelines(refer 7). All data were de-identified prior to use and the research posed no foreseeable risk to participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: Not applicable. Any assistance from LLMs was limited to spell check, grammar correction, and minor language refinement, and did not influence the scientific content, methodology, or originality of the work.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.