

General-Purpose Vision Models to Dermatology-Specific Foundation Models: Comparative Evaluation for Facial Skin Disease Classification

Uchenna Esther Okpete

Inje University

Md Ariful Islam Mozumder

Inje University

Osita Sunday Nnyigide

Korea University of Technology and Education

Haewon Byeon

Korea University of Technology and Education

Hee-Cheol Kim

heeki@inje.ac.kr

Inje University

Article

Keywords: Facial skin disease, Transfer learning, Vision transformer, Foundational models, Convolutional neural network, Large language model

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11133435/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

General-Purpose Vision Models to Dermatology-Specific Foundation Models: Comparative Evaluation for Facial Skin Disease Classification

Uchenna Esther Okpete¹, Md Ariful Islam Mozumder¹, Osita Sunday Nnyigide²,
Haewon Byeon^{2*}, Hee-Cheol Kim^{1*}

¹Institute of Digital Anti-Aging HealthCare, Inje University, Gimhae, 50834, South Korea.

²Worker's Care & Digital Health Lab, Department of Future Technology, Korea University of
Technology and Education, Cheonan 31253, South Korea

*Corresponding author(s). E-mail(s): heeki@inje.ac.kr; bhwpuma@naver.com

Contributing authors: ucheokpete@gmail.com; arifulislamro@gmail.com;
ositasunday2@gmail.com

Abstract

Skin diseases represent a major global health burden, while access to dermatological expertise remains limited in many rural and resource-constrained settings. These disparities have increased interest in scalable digital tools for dermatological assessment and patient education. However, trustworthy clinical AI requires reproducible evaluation, standardized benchmarking, transparent reporting of failure patterns, and assessment of model robustness across different representations and classification strategies. This study presents a systematic transfer-learning framework for six-class facial skin disease classification, comparing seven pretrained visual representations comprising ResNet50, MobileNetV3-Large, ViT-B/16, Swin-T, DINOv2, DermLIP and Derm Foundation, under a common downstream classification framework. XGBoost was used as the primary downstream classifier, while an RBF-kernel support vector machine (SVM) was additionally evaluated as an ablation analysis to investigate the dependence of classification performance on downstream classifier selection. Performance was assessed using accuracy, macro-F1, macro-AUROC, per-class recall, confusion matrices, and bootstrap-derived 95% confidence intervals, with paired McNemar testing and Holm correction used to compare classifier. Derm Foundation achieved the highest performance among the evaluated feature representations when combined with XGBoost, obtaining 88.17% accuracy, 88.05% macro-F1, and 98.58% macro-AUROC on the testing set. At the class level,

recall was 94.5% for acne, 87.0% for atopic dermatitis, 100.0% for normal skin, 90.5% for psoriasis, 84.0% for rosacea, and 73.0% for seborrheic dermatitis. Classifier comparisons showed that the relative performance of XGBoost and RBF-SVM depended on the underlying feature representation. The principal residual classification difficulty involved seborrheic dermatitis, particularly its reciprocal confusion with rosacea. To demonstrate technical feasibility, the classification framework was integrated into a prototype Django-based web application architecture featuring an exploratory LLM module for automated patient-education content translation. Preliminary, descriptive specialist feedback was gathered to identify future engineering optimization constraints, such as inference latency and communication mechanics. These findings are limited to the evaluated testing dataset and do not establish external generalizability or clinical efficacy.

Keywords: Facial skin disease, Transfer learning, Vision transformer, Foundational models, Convolutional neural network, Large language model,

1. Introduction

Skin diseases are among the most common health conditions worldwide, affecting more than 4.7 billion people. They account for nearly 2% of global morbidity and rank as the fourth leading cause of disability [1,2]. In addition to their physical manifestations, visible dermatoses impose a considerable psychological burden, with affected individuals experiencing increased rates of depression, anxiety, social withdrawal, and diminished quality of life compared with the general population [3,4].

Facial dermatoses such as acne, atopic dermatitis, psoriasis, rosacea, and seborrheic dermatitis are common non-communicable conditions that may exhibit substantial visual overlap [5]. Differences in skin lesion morphology, and spatial distribution can be incredibly subtle, while the appearance of the same condition can vary according to individual characteristics, skin type, disease stage, and image-acquisition conditions. Consequently, automated image-based classification of facial dermatoses represents a challenging computer-vision problem in which clinically relevant distinctions may depend on relatively subtle visual characteristics.

Anatomical localization is an important consideration in the development and evaluation of image-based dermatological models. The appearance of a skin disorder can vary across anatomical sites due to regional differences in skin characteristics, surrounding tissue, and hair distribution. Consequently, machine learning models trained on heterogeneous body locations risk learning confounding, location-associated visual traits rather than true disease features. Indeed, recent evaluations of dermatology foundation models reveal site-dependent disparities in reliability, particularly underperforming on underrepresented facial and ear sites [6]. Restricting image acquisition to a consistent anatomical region reduces this confounding variability, aligning the training distribution with the clinical application to ensure the model learns representations strictly relevant to the diagnostic task.

Previous studies investigating automated classification and detection of facial skin conditions using clinical images have faced persistent barriers to clinical translation. Systemic bottlenecks include limited dataset sizes, diagnostic imbalances, limited demographic and skin-type diversity, and insufficient external validation [7,8-10]. Furthermore, high heterogeneity among published studies regarding model architectures, image-acquisition protocols, diagnostic definitions, and evaluation metrics makes direct performance comparisons difficult [11].

The suitability of a dataset depends greatly on the compatibility of its diagnostic categories, anatomical localization, image characteristics, and class distribution with the intended clinical task. As summarized in Table 1, widely used dermatological image datasets often fail to provide a suitable representation of the six facial categories considered in this study. For example, Fitzpatrick17k contains clinical images spanning

73 numerous conditions, but its diagnostic taxonomy includes broad dermatitis and eczematous categories that
74 do not directly correspond to specific atopic dermatitis category [12]. Similarly, HAM10000 and collections
75 within the ISIC Archive primarily feature dermoscopic images of neoplastic lesions [13]. Other large-scale
76 resources may provide broader diagnostic coverage but differ in anatomical distribution, acquisition setting,
77 accessibility, or diagnostic definitions.

78 As Table 1 demonstrates, dataset size alone does not guarantee suitability for a specific facial skin-disease
79 classification task. Controlled evaluation using a highly consistent target dataset provides an important
80 opportunity to isolate performance variations. By benchmarking architectures within a uniform
81 experimental pipeline, we can distinguish differences directly attributable to the learned visual
82 representations from those artifacts introduced by downstream experimental variability [12,24].

83 Deep learning models excel at learning hierarchical visual representations directly from images, minimizing
84 the need for manual feature engineering. However, training these networks from random initialization
85 requires substantial labeled data and computational resources. Transfer learning mitigates these constraints
86 by reusing representations learned from large-scale source datasets or related tasks and has therefore
87 become widely used in medical-image classification [25,26]. Pretrained representations provides robust
88 visual features while lowering the computational and data requirements associated with training large
89 models entirely from scratch.

90 Transfer learning is typically implemented via feature extraction, fine-tuning, or hybrid methods. Feature
91 extraction freezes the pretrained encoder and uses it to transform input images into static representations
92 for a downstream classifier (Figure 1). In fine-tuning, some or all pretrained parameters are updated using
93 target-domain data. Hybrid approaches combine pretrained feature extractors with external machine-
94 learning models. These strategies differ in their computational requirements and in the degree to which the
95 pretrained representation can adapt to the target domain. Notably, frozen encoders coupled with lightweight
96 downstream classifiers have demonstrated competitive performance in medical image classification while
97 dramatically reducing computational costs [27]. For datasets in which annotated clinical images are limited
98 relative to the complexity of modern deep-learning architectures, frozen feature extraction provides an
99 attractive framework because it allows pretrained visual representations to be evaluated without updating
100 millions of model parameters.

101 An important but comparatively under-examined consideration is the interaction between the pretrained
102 representation and the downstream classifier. A feature representation that performs well with one classifier
103 may not produce equivalent performance with another because different classifiers impose different
104 decision boundaries and assumptions on the underlying feature space. While recent studies have evaluated

dermatology foundation models using various downstream architectures, such as multilayer perceptrons, XGBoost, or lightweight adaptation layers, comparisons across such studies remain limited [28,29]. Confounding variations in datasets, preprocessing, training strategies, and validation procedures obscure whether performance gains stem from the representation itself or the downstream setup. Thus, an important methodological gap remains in the controlled comparison of pretrained visual representations for multiclass facial dermatoses classification under a common feature-extraction and downstream-classification framework.

A standardized frozen-feature framework resolves this problem by holding key experimental factors constant while varying the pretrained representation. Under such a framework, the same facial image cohort, preprocessing procedure, training and validation partitions, evaluation metrics, and downstream classification strategies can be applied across multiple pretrained representations. This enables a more direct assessment of the extent to which differences in performance are associated with the characteristics of the learned representation. Furthermore, evaluating multiple downstream classifiers on the same extracted features allows researchers to quantify the performance dependency on classifier selection alone.

To address these gaps, the present study systematically evaluates seven pretrained visual representations for the automated classification of six facial skin-image categories: acne, atopic dermatitis, normal skin, psoriasis, rosacea, and seborrheic dermatitis. The evaluated architectures encompass conventional convolutional neural networks (CNNs), vision transformers (ViTs), self-supervised models, and dermatology-specific representations, including a dermatology foundation model. All pretrained feature extractors are evaluated within a common frozen-feature transfer-learning framework. XGBoost serves as the primary downstream classifier, while a radial basis function (RBF) kernel support vector machine (SVM) is evaluated as an ablation analysis to determine whether classification performance depends on the choice of downstream classifier.

Specifically, the study aims to:

- Compare pretrained image representations from convolutional neural network (CNN), transformer, self-supervised, and dermatology-specific models for six-class facial skin-image classification.
- Evaluate the effect of downstream classifier selection by comparing XGBoost with an RBF-kernel SVM using the same extracted feature representations.
- Assess classification performance comprehensively using accuracy, balanced accuracy, macro-F1, macro-AUROC, per-class recall, confusion matrices, and bootstrap-derived 95% confidence intervals.

136 • Quantify classifier-dependent differences using paired statistical comparisons of image-level
137 predictions obtained from the same testing images.

138

2. Related Works

2.1 Transfer Learning in Medical Imaging

Transfer learning is highly relevant to medical image analysis because the development of large, expert-annotated clinical datasets is frequently constrained by the high image-acquisition costs, expert annotation labor, and strict quality control. Consequently, pretrained neural networks have been widely adopted across medical imaging tasks, including classification, detection, segmentation, and severity assessment, with the potential to reduce the amount of target-domain data and computational resources requirements relative to training from scratch [27,30,31]. In medical image analysis, transfer learning can be implemented through several configurations. Kim et al., categorized these approaches into feature extraction, feature-extractor hybrid configurations, fine-tuning, and fine-tuning from scratch [32]. Among studies using a single approach, feature extraction was the most frequently reported configuration ($n = 38$), followed by fine-tuning from scratch ($n = 27$), whereas pretrained fine-tuning alone was less common ($n = 3$) [32]. These findings highlight substantial methodological variation in how pretrained representations are adapted to medical imaging tasks. While feature extraction reduces computational demand and preserves predictive performance, fine-tuning offers deeper adaptation but demands extensive optimization. Because prior literature often relies on specialized, fine-tuned, or task-specific architecture rather than a shared, frozen-feature protocol, the existing evidence is insufficient to isolate downstream classifier effects or establish absolute performance rankings under identical conditions [32].

Regarding the transferability of deep visual representations, early layers of convolutional networks typically learn general visual patterns (e.g., edges and textures), whereas later layers increasingly encode representations specific to the source dataset and task. Yosinski et al. systematically investigated this general-to-specific transition and demonstrated that feature transferability decreases as the source and target tasks become increasingly dissimilar³⁴. This general-to-specific property provides an important conceptual foundation for feature-based transfer learning. General visual information learned during pretraining may remain useful when applied to a new domain, whereas highly specialized representations may be less transferable when the source and target tasks differ substantially. Thus, the usefulness of a pretrained model cannot necessarily be inferred solely from its performance on its original pretraining task. Instead, transferability must be evaluated in relation to the target domain and task.

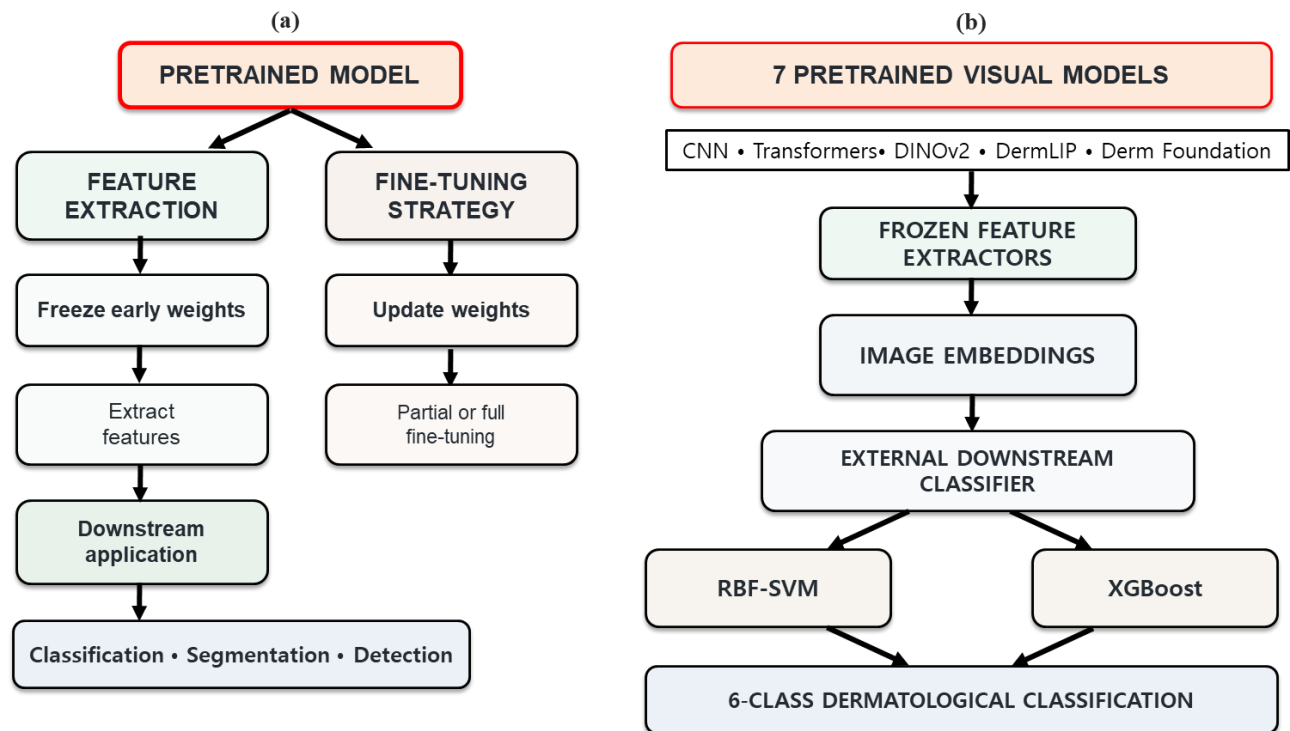


Fig. 1. Transfer Learning: General Strategy and Study-Specific Implementation (a) General transfer learning taxonomy (b) Frozen-feature extraction approach used in this study

2.2 Transfer Learning for Dermatological Image Classification

In dermatological image analysis, transfer learning has been widely adopted using pretrained CNNs including VGG, ResNet, DenseNet, EfficientNet, and MobileNet variants [35]. These approaches allow researchers to apply deep visual representations to skin-disease classification without requiring the development of large-scale networks from random initialization. Comparisons across VGG16, Inception-v3, ResNet, MobileNet, NasNet, and EfficientNet reported that EfficientNet variant achieved higher accuracies of 94.0%, 92.0%, and 93.8% for selected lesion classes [36]. MobileNet has been shown to outperform baseline CNNs on DermNet images across accuracy, precision, recall, and F1-score [37]. However, conventional ImageNet pretraining uses general, natural images and may inadvertently capture non-clinical features. For instance, a variational autoencoder (VAE) trained directly on dermatological data demonstrated a near-zero overfitting gap versus an increasing gap for ImageNet transfer learning³⁸. Although the resulting representations contain useful low-level and mid-level visual information, their ability to capture disease-specific characteristics may be limited by the difference between the source domain and the target dermatological domain [38]. Puls et al. evaluated 16 pretrained models across four datasets, finding that CLIP-ViT-B and ViT-H-14 performed best overall, while CLIP-ResNet50 exhibited

the lowest variability [39]. Conversely, Kornblith et al. reported weak transfer to fine-grained datasets, concluding that transferability is highly sensitive to pretraining choices [40].

Recent dermatological image-analysis research has increasingly investigated alternative pretrained representations and domain-specific transfer learning strategies, including transformer architectures, self-supervised representations, and dermatology foundation models [12, 41,33].

2.3 Transformer and Self-Supervised Representations

Transformers represent an image as a sequence of visual tokens and use self-attention mechanisms to model relationships between them. Pure transformer can operate directly on image patches without convolutional layers; when pretrained on sufficiently large datasets, these models transfer effectively to smaller downstream benchmarks [42].

The Vision Transformer (ViT): divides an image into fixed-size patches and processes them as a token sequence using global self-attention. Its main advantage is the ability to integrate information across the whole lesion, without relying on progressively stacked convolutional receptive fields. ViT can transfer effectively after large-scale pre-training, but its weaker local inductive bias may limit performance on small, imbalanced dermatology datasets⁴². In a controlled HAM10000 comparison, ViT-B/16 achieved an intermediate macro-F1 of 0.7631 [43].

The Swin Transformer addresses these limitations by utilizing hierarchical representations and shifted local-attention windows. Local windows reduce computational cost, while shifting the windows allows information to flow across neighbouring regions [44]. This combination is well suited to skin lesions, where fine texture and global structure must be analysed together. In the same HAM10000 benchmark, Swin-T performed best, achieving a macro-F1 of 0.8582 and outperforming ViT-B/16 by 9.5 percentage points [45].

DINOv2: provides a general-purpose self-supervised ViT representation. In the HAM10000 benchmark, it achieved competitive macro-AUC but a lower macro-F1 than Swin-T under a fixed, 25-epoch adaptation budget., underperforming with a fixed 25-epoch adaptation budget. This underperformance has been attributed to limited adaptation constraints and mismatches between pretraining and downstream fine-tuning resolutions [45]. Hence, self-supervised methods like DINOv2 warrant rigorous evaluation with longer fine-tuning, resolution matching, calibration, and domain-specific pre-training rather than a minimal linear head alone.

2.4 Dermatology-Specific Foundation Models

Derm Foundation is an image-only dermatology foundation model engineered to accelerate clinical AI development by generating 6,144-dimensional embeddings from skin images. Validation reports indicate that simple downstream classifiers trained on these embeddings improve diagnostic accuracy by 10–15% over standard computer-vision baselines when data is limited [45]. The model is designed explicitly to support downstream classification tasks but does not diagnose independently.

DermLIP is a contrastive vision-language model utilizing a CLIP-style architecture, combining a ViT-B/16 vision encoder with a GPT-2 text encoder. This multimodal model was trained on 403,563 dermoscopic and clinical image–text pairs from the Derm1M dataset [46]. Its framework is optimized for zero-shot and few-shot classification, image–text retrieval, and automated concept annotation rather than isolated, single-modality diagnostic tasks.

2.6 Research Gap and Motivation

While the existing literature establishes the clear value of transfer learning for dermatological image analysis, several distinct methodological gaps remain:

- **Heterogeneous Paradigms:** Pretrained representations are developed using increasingly diverse architectures and training paradigms—ranging from conventional CNNs to transformers, self-supervised networks, and domain-specific foundation models.
- **Variable Adaptation:** Prior transfer-learning studies vary wildly in the extent to which pretrained parameters are fine-tuned or adapted to target datasets.
- **Non-Standardized Downstream Architectures:** Evaluation pipelines and downstream classifiers are rarely standardized, making direct comparisons between underlying visual representations mathematically confounding.

Accordingly, this study systematically benchmarks seven pretrained visual representations spanning four distinct paradigms:

- **Conventional CNNs:** ResNet50 and MobileNetV3-Large
- **Vision Transformers:** ViT-B/16 and Swin-T
- **Self-Supervised Learning:** DINOv2
- **Dermatology-specific Models:** DermLIP and Derm Foundation

The resulting embeddings are evaluated using a unified classification framework driven by XGBoost as the primary classifier, alongside RBF-SVM for comprehensive ablation analysis. This allows us to answer two complementary questions: (1) how effectively do different frozen representations transfer to multiclass

247 facial dermatoses classification, and (2) how significantly does downstream classifier selection influence
248 their relative performance?

249

3. Methodology

3.1 Study design

The proposed framework, illustrated in Figure 2, consists of three primary computational stages: (1) Frozen feature extraction from facial skin images diverse pretrained architectures (ResNet50, MobileNetV3-Large, ViT-B/16, Swin-T, DINOv2, DermLIP, and Derm Foundation) (2) Evaluation of the extracted feature embeddings using XGBoost classifier and a RBF-SVM for multiclass diagnostic mapping; and (3) and the integration of the LLM module within a web-based application to translate point predictions into accessible, patient-centered educational guidance. By holding the feature extractors fixed, this unified framework provides a controlled experimental framework for comparing the evaluated representations. It eliminates downstream experimental artifacts and variations that typically arise from model-specific fine-tuning protocols, reducing variation attributable to differences in downstream training procedures.

3.2 Dataset description and preparation

The present study involves a secondary analysis of a publicly available, de-identified clinical dataset. The data were obtained from the AI-Hub Skin Disease Diagnostic Medical Images dataset, a large-scale clinical repository compiled across 10 academic university hospitals in South Korea under the auspices of the National Information Society Agency (NIA) [23]. All source data were strictly anonymized and de-identified by the generating institutions prior to public deployment to safeguard patient confidentiality. All personal identifying information, including direct patient identifiers, was entirely removed before public release, ensuring that individual identities cannot be traced back to specific participants. Because this investigation utilizes strictly de-identified, publicly accessible retrospective data, that does not involve direct patient recruitment, and features no direct contact with human subjects, this study was deemed exempt from formal Institutional Review Board (IRB) review under standard institutional guidelines for secondary data use.

The integrated dataset contains images from multiple diagnostic categories, including 19,984 images of normal skin, 10,616 of atopic dermatitis, 10,854 of psoriasis, 10,207 of acne, 4,357 of seborrheic dermatitis, and 5,502 of rosacea. For each of the six target disease classes, 800 frontal-view images and 800 side-view images were used for training, giving 1,600 images per class. While, 100 frontal-view images and 100 side-view images were reserved for the testing set. Thus, the training set comprised 9,600 images (1,600 images

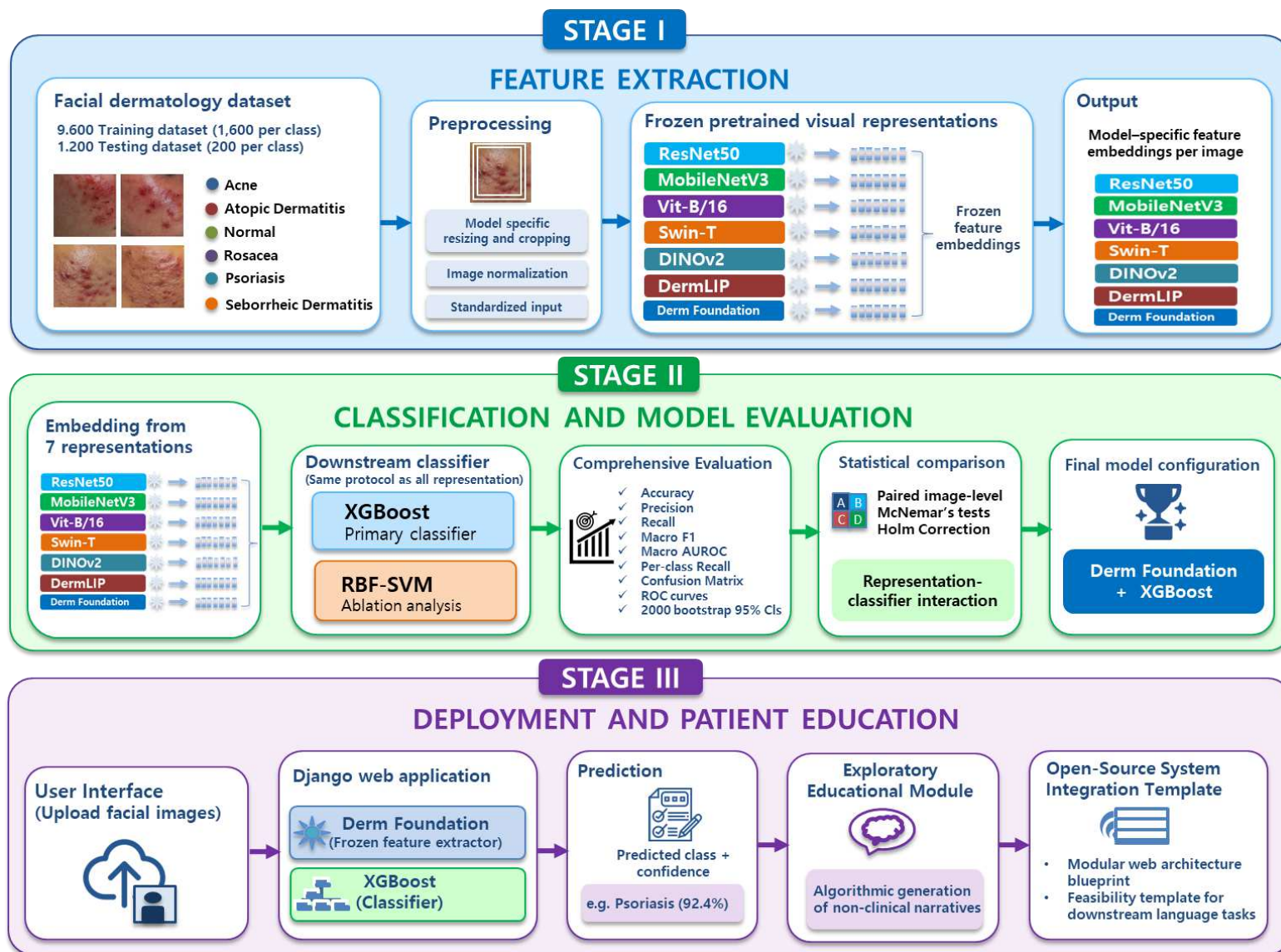


Fig. 2. The overall workflow of the proposed framework for multi-class facial skin disease classification and patient education

Table 1. Characteristics of publicly available clinical dermatology image datasets relevant to facial skin-disease classification.

Dataset	Approx. size	Total disease categories	Acne	Atopic dermatitis	Normal	Psoriasis	Rosacea	Seborrheic dermatitis	Availability	Ref.
Fitzpatrick17k	16,577	114	Yes	No ^a	No	Yes	Yes	Yes	Public	[14]
SCIN	10,000+	419	Limited	Limited	No	Yes	Yes	Yes	Public	[15]
SkinDisNet	1,710	6	No	Yes	Yes	No	No	Yes	Public	[16]
Xiangya-Derm	>150,000	571	No	Yes (Eczema)	No	Yes	No	No	Restricted	[17]
DDI	656	78	Limited	Yes	No	Yes	Limited	Limited	Stanford's research-use terms	[18]
DermNet	>23,000	23	Yes	Yes	No	Yes	Yes	Yes	Viewable; terms apply ^b	[12]
HAM10000	10,015	7	No	No	No	No	No	No	Public	[19]
ISIC	25,000–33,000+	Variable ^c	Limited	No	No	No	No	No	Public	[20]
PAD-UFES-20	2,298	6	No	No	No	No	No	No	Public	[21]
Derm7pt	>2,000	20	Limited	Limited	No	Limited	Limited	Limited	Public	[22]
AI hub	106,776 clinical images	32 diseases + normal skin	Yes	Yes	Yes	Yes	Yes	Yes	Platform access	[23]

^aDiagnostic terminology in Fitzpatrick17k includes broad dermatitis/eczema categories that do not necessarily correspond directly to the specific atopic dermatitis category used in the present study. ^bPublicly viewable; license dependent for research use..

^bISIC is an archive containing multiple datasets/challenge collections, not one fixed disease dataset. For the commonly compared 2016–2020 collections, some reviews summarize 9 diagnostic categories and should not be presented as the total number of diseases in the entire ISIC Archive.

3.3 Transfer learning-based feature extraction

This study employed a frozen-feature transfer learning framework to evaluate the quality of pretrained visual representations for six-class facial skin-disease classification. Rather than fine-tuning the pretrained networks on the target dataset, each pretrained model was used as a fixed feature extractor. This approach enabled heterogeneous visual representations to be compared under a common downstream classification framework while minimizing variation attributable to model-specific parameter optimization.

For all models, the pretrained parameters were frozen during feature extraction, and no target-domain gradient updates were performed.

Given an input facial skin image I , each pretrained model f_m generated a feature representation:

$$E_m = f_m(I) \quad (1)$$

Where I denotes the input image, f_m denotes the feature-extraction function associated with pretrained model m , and E_m denotes the resulting image embedding. The extracted embeddings represent the visual information encoded by the corresponding pretrained model and were subsequently used as inputs to the downstream classification stage.

The dimensionality of the extracted representation depended on the architecture (Table 2). ResNet50 generated a 2,048-dimensional representation, MobileNetV3 a 960-dimensional representation, ViT-B/16 a 768-dimensional representation, Swin-T a 768-dimensional representation, DINOv2 a 768-dimensional representation, DermLIP a 512-dimensional representation, and Derm Foundation a 6,144-dimensional representation. The structural properties, learning paradigms, baseline classification hyperparameters, and total training runtimes for each of these evaluated architectures are systematically detailed in Table 2.

For Derm Foundation, images were encoded as PNG images and provided to the model's inference signature according to the released implementation. For DermLIP, the model-specific preprocessing pipeline was applied and the resulting image embeddings were L2-normalized prior to downstream classification. Equivalent model-specific preprocessing was applied to the other pretrained representations according to their respective implementations.

3.4 Downstream Classification

The extracted feature embeddings were used as inputs to XGBoost as the primary downstream classifier. XGBoost was selected as the main classifier for evaluating the discriminative utility of the pretrained feature representations. The same XGBoost configuration was applied independently to the embeddings

Table 2. Structural properties, hyperparameters, and computational overhead of the evaluated feature extractors with XGBoost as the primary classifier.

Model	Architecture	Learning paradigm	Embedding dimension	Classifier hyperparameter	Total training time
ResNet50	CNN	supervised	2048	300 estimators; max depth = 6; learning rate = 0.05; objective = multi:softprob; mlogloss; seed = 42	0.31 h
MobileNetV3-Large	CNN	supervised	960	300 estimators; max depth = 6; learning rate = 0.05; objective = multi:softprob; mlogloss; seed = 42	0.11 h
ViT-B/16	Transformer	supervised pretraining	768	300 estimators; max depth = 6; learning rate = 0.05; objective = multi:softprob; mlogloss; seed = 42	7.88 h
Swin-T	Hierarchical Transformer	supervised pretraining	768	300 estimators; max depth = 6; learning rate = 0.05; objective = multi:softprob; mlogloss; seed = 42	0.22 h
DINOv2	Vision Transformer	self-supervised	768	300 estimators; max depth = 6; learning rate = 0.05; objective = multi:softprob; mlogloss; seed = 42	0.38 h
DermLIP	Vision-language model	dermatology-specific	512	300 estimators; max depth = 6; learning rate = 0.05; objective = multi:softprob; mlogloss; seed = 42	0.34 h
Derm Foundation	Foundation model	dermatology-specific	6144	300 estimators; max depth = 6; learning rate = 0.05; objective = multi:softprob; mlogloss; seed = 42	9.11 h

generated by each pretrained model to maintain a consistent downstream classification procedure across feature extractors (Table 2).

The overall classification process can be represented as:

$$I \rightarrow f_m \rightarrow E_m \rightarrow g_c \rightarrow \hat{y}_{m,c} \quad (2)$$

Where g_c represents downstream classifier c , and $\hat{y}_{m,c}$ denotes the predicted class produced by the combination of pretrained representation m and downstream classifier c .

Primary XGBoost classification: For the primary classification experiments, XGBoost was configured for six-class classification using the multi:softprob objective. The model was trained with 300 boosting estimators, a maximum tree depth of 6, and a learning rate of 0.05. The multiclass logarithmic loss

(mlogloss) was used as the evaluation metric. The random seed was set to 42, and parallel computation was enabled using `n_jobs = -1`. The classifier generated class-probability estimates for each of the six diagnostic categories, which were subsequently used for multiclass AUROC calculation.

RBF-SVM ablation analysis: As an ablation analysis, an RBF-kernel support vector machine (SVM) was additionally applied to the same extracted feature representations. The purpose of this analysis was to determine whether the observed classification performance was dependent on the choice of downstream classifier and to assess the robustness of the pretrained representations under an alternative classification strategy.

Before SVM classification, the extracted feature embeddings were standardized using `StandardScaler`, with the scaling parameters estimated from the training features. The standardized features were then classified using an RBF-kernel SVM with $C = 10$ and `gamma="scale"`. Class weighting was set to balanced. Probability estimation was enabled (`probability=True`). To obtain class-probability estimates for multiclass AUROC calculation. The decision function was configured using a one-versus-rest strategy.

The pretrained feature-extraction models were kept frozen throughout both the primary XGBoost experiments and the SVM ablation analysis. Thus, differences between XGBoost and SVM performance reflected the interaction between the fixed pretrained representation and the downstream classifier rather than differences in target-domain fine-tuning.

3.5 Model Evaluation

The performance of each pretrained representation–classifier combination was evaluated on the held-out testing set. Accuracy, balanced accuracy, macro-F1, and macro-AUROC were calculated, together with per-class recall, confusion matrices, and receiver operating characteristic curves. The testing set contained 1,200 images, with 200 images from each of the six diagnostic categories.

Accuracy measures how close the observed measurement is to the true value and was calculated as the proportion of correctly classified images among all testing images:

$$\text{Accuracy} = \frac{TP+TN}{P+N} \quad (3)$$

Per-class recall was calculated for each diagnostic category to identify differences in sensitivity across acne, atopic dermatitis, normal skin, psoriasis, rosacea, and seborrheic dermatitis.

$$\text{Per-class recall} = \frac{TP}{TP+FN} \quad (4)$$

Balanced accuracy was calculated as the arithmetic mean of the recall obtained for each of the six classes. Because the internal testing set contained an equal number of images from each class, balanced accuracy and overall accuracy were expected to be similar and, under the observed class-balanced evaluation, were identical.

$$\text{Balanced accuracy} = \frac{TPR + TNR}{2} \quad (5)$$

Macro-F1 was calculated as the unweighted mean of the F1 scores across the six classes, giving equal importance to each diagnostic category regardless of class frequency.

$$\text{Macro F1} = \frac{1}{N} \sum_{i=1}^N F1_i \quad (6)$$

Macro-AUROC was calculated using a one-versus-rest approach across the six classes and macro-averaged across class-specific AUROC values.

$$\text{Macro-AUROC} = \frac{1}{N} \sum_{i=1}^N AUROC_i \quad (7)$$

Absolute Confusion matrices were generated to characterize the distribution of correct and incorrect predictions among the six classes. While, multiclass receiver operating characteristic (ROC) curves were also generated using a one-versus-rest strategy to visualize the discriminative performance of the evaluated models across the six diagnostic categories.

3.6 Confidence Interval Estimation

To quantify the uncertainty associated with the observed test performance, nonparametric bootstrap resampling was used to estimate 95% confidence intervals for the primary performance metrics. For each model–classifier combination, 2,000 bootstrap samples were generated by resampling the validation observations with replacement. For each bootstrap sample, accuracy, balanced accuracy, macro-F1, and macro-AUROC were recalculated using the corresponding predictions and class-probability estimates.

The 95% confidence intervals were calculated using the percentile method, with the lower and upper bounds corresponding to the 2.5th and 97.5th percentiles, respectively, of the bootstrap distribution. For bootstrap samples in which the metric could not be validly calculated, particularly for multiclass AUROC when one or more classes were absent from a resampled dataset, the corresponding iteration was excluded from the confidence-interval calculation. A minimum number of valid bootstrap iterations was required before

reporting a confidence interval. This approach provided uncertainty estimates for model performance without assuming a specific parametric distribution for the evaluation metrics.

3.7 Statistical Comparison

To compare the performance of the SVM and XGBoost classifiers across the different feature extractors, a two-sided McNemar's exact test was conducted on the 1,200 test images. The exact p -value is calculated based on the discordant pairs using the cumulative binomial distribution under the null hypothesis ($p = 0.5$)

$$P_{exact} = 2 \cdot \sum_{k=0}^{\min(b,c)} \binom{b+c}{k} \left(\frac{1}{2}\right)^{b+c} \quad (8)$$

Where, b represents the count of images correctly classified by XGBoost but incorrectly by SVM, and c represents the count of images correctly classified by SVM but incorrectly by XGBoost.

To control the family-wise error rate across the $m = 7$ feature extractor comparisons, p -values, were adjusted using the Holm-Bonetti stepwise procedure. The unadjusted p -values were ranked in ascending order ($p_1 \leq p_2 \leq \dots \leq p_m$) and each adjusted p -values was calculated sequentially using the formula:

$$P_{i,adjusted} = \min (1, \max_{j \leq i} [(m - j + 1) \cdot p_j]) \quad (9)$$

Statistical significance was evaluated at the ($\alpha = 0.05$) level.

3.8 Web Application Implementation and LLM-Assisted Patient Education

To demonstrate the technical feasibility of integrating the classification framework into a user-facing application, a web-based application was developed using the Django framework. The application, hosted locally during development and testing, provides an intuitive interface that enables users to upload facial skin images for automated disease classification and receive educational information about the predicted condition. The application was used to demonstrate the integration of image preprocessing, feature extraction, XGBoost classification, and patient-oriented educational content generation. This implementation was not intended to establish clinical efficacy or diagnostic validity.

Following image submission, the uploaded image was received and validated by the Django server and subsequently preprocessed using the same image preprocessing procedures employed during model development. The processed image was then passed to the best performing frozen feature extractor to generate a high-dimensional dermatological image representation. The resulting feature embedding was

subsequently provided to the trained XGBoost classifier, which generated a prediction for one of the six target categories.

The predicted disease label and its associated confidence score are then forwarded to an LLM to generate patient-friendly educational information tailored to the predicted skin condition. The generated educational content is returned to the Django server together with the classification result. The results interface therefore displayed the predicted disease category, the associated confidence score, and the LLM-generated educational information within a unified web-based interface.

4. Results

4.1 Training Convergence of the Pretrained Models

Training and testing loss dynamics were examined over 300 boosting rounds to assess optimization behavior and convergence across the evaluated models (Figure 3). All models demonstrated a marked reduction in training loss during the early stages of training, followed by progressively slower improvement as training proceeded. By the later epochs, the training-loss curves had approached relatively low and stable values, indicating convergence of the optimization process.

Testing loss showed a similar initial decline across models, although the final levels differed substantially between representations. The Derm Foundation embeddings achieved the lowest loss throughout the later stages of training, followed sequentially by DermLIP and DINOv2, whereas the general-purpose representations exhibited higher test-loss plateaus. Crucially, testing loss trajectories remained stable during the later stages of training with no late-epoch increases, indicating the absence of severe overfitting.

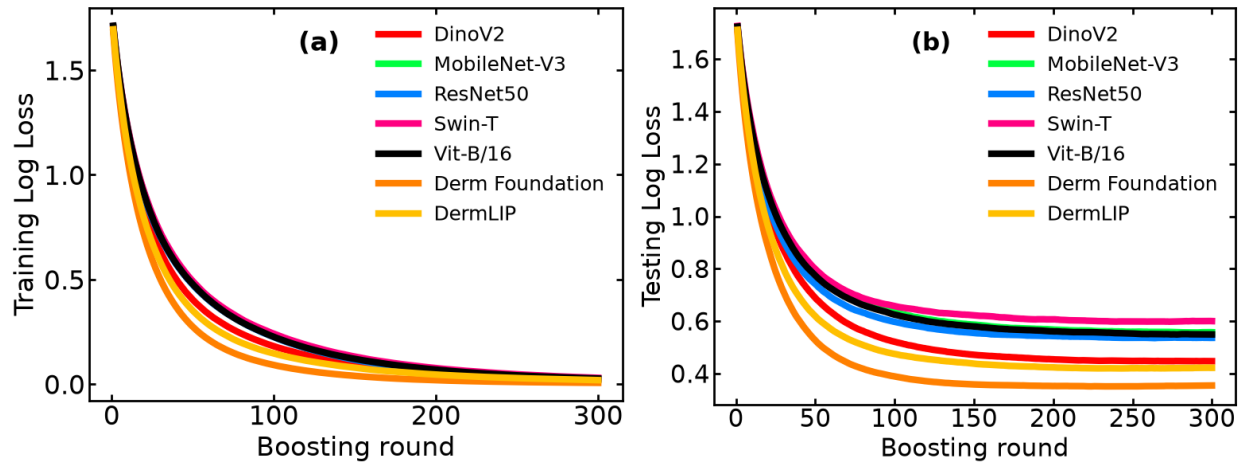


Fig. 3. Training and testing loss trajectories across 300 boosting rounds for the seven evaluated feature representations.

4.2 Overall Performance of Pretrained Feature Representations

Substantial differences in classification performance were observed among the seven pretrained feature representations when evaluated using XGBoost as the primary downstream classifier (Table 3). Derm Foundation achieved the highest overall performance, with an accuracy of 88.17%, macro-F1 of 88.05%, and macro-AUROC of 98.58%. DermLIP achieved the second-highest accuracy at 83.75%, followed by DINOv2 at 82.58%. Conversely, the ImageNet-pretrained CNN and transformer representations showed lower accuracy, ranging from 75.67% (Swin-T) to 78.75% (MobileNetV3-Large).

The relatively narrow 95% confidence intervals derived from bootstrapping demonstrate that these performance variations are statistically stable. The pronounced superiority of Derm Foundation and DermLIP confirms that domain-specific pretraining generates highly discriminative features for facial skin-disease classification compared to general-purpose ImageNet weights. Furthermore, the minimal discrepancy between accuracy and macro-F1 scores across the top performers implies that these models provide uniform classification power across all six categories, rather than having their metrics inflated by a few majority classes

Table 3. Multiclass classification performance of frozen feature representations optimized via the primary XGBoost downstream classifier.

Feature extractor	Accuracy (95% CI)	Macro-F1 (95% CI)	Macro-AUROC (95% CI)
ResNet50	78.58% (76.33–80.83)	78.46% (76.13–80.76)	97.07% (96.56–97.58)
MobileNetV3-Large	78.75% (76.50–81.08)	78.62% (76.37–80.81)	96.77% (96.21–97.33)
ViT-B/16	77.83% (75.42–80.17)	77.63% (75.36–79.88)	96.80% (96.24–97.35)
Swin-T	75.67% (73.17–78.00)	75.54% (73.13–77.69)	96.32% (95.67–96.92)
DINOv2	82.58% (80.33–84.67)	82.68% (80.53–84.72)	97.82% (97.37–98.23)
DermLIP	83.75% (81.67–85.83)	83.69% (81.63–85.71)	98.13% (97.75–98.49)
Derm Foundation	88.17% (86.33–89.92)	88.05% (86.23–89.80)	98.58% (98.21–98.92)

4.3 Ablation analysis: Comparative performance of feature representations

To explore the interaction between the extracted feature spaces and downstream decision boundaries, an ablation analysis was conducted by swapping the primary XGBoost classifier with an RBF-SVM (Figure 4 and Table 4). Figure 4 compares classification performance across the seven feature extractors using XGBoost and RBF-SVM as common downstream classifiers. Overall, the results demonstrate that the choice of feature representation had a substantial influence on classification performance, with the dermatology-specific representations achieving the strongest results. Derm Foundation maintained peak performance across both architectures, yielding 88.17% accuracy with XGBoost and 87.83% with RBF-SVM. DermLIP also demonstrated strong performance, achieving 83.75% and 84.50% accuracy with XGBoost and RBF-SVM, respectively, followed by DINOv2 with 82.58% and 82.83%, respectively. These findings indicate that representations developed specifically for dermatological imagery may provide more informative features for the present six-class classification task than general-purpose ImageNet-pretrained representations.

The close correspondence between accuracy and macro-F1 across the experiments indicates that the observed differences were not primarily attributable to a single aggregate metric. For Derm Foundation, for example, XGBoost achieved 88.17% accuracy and 88.05% macro-F1, while RBF-SVM achieved 87.83% accuracy and 87.73% macro-F1. Similarly, DINOv2 and DermLIP showed relatively small differences between accuracy and macro-F1 for both downstream classifiers. Because the testing set contained 200 images per class, accuracy and balanced accuracy were equivalent in this evaluation.

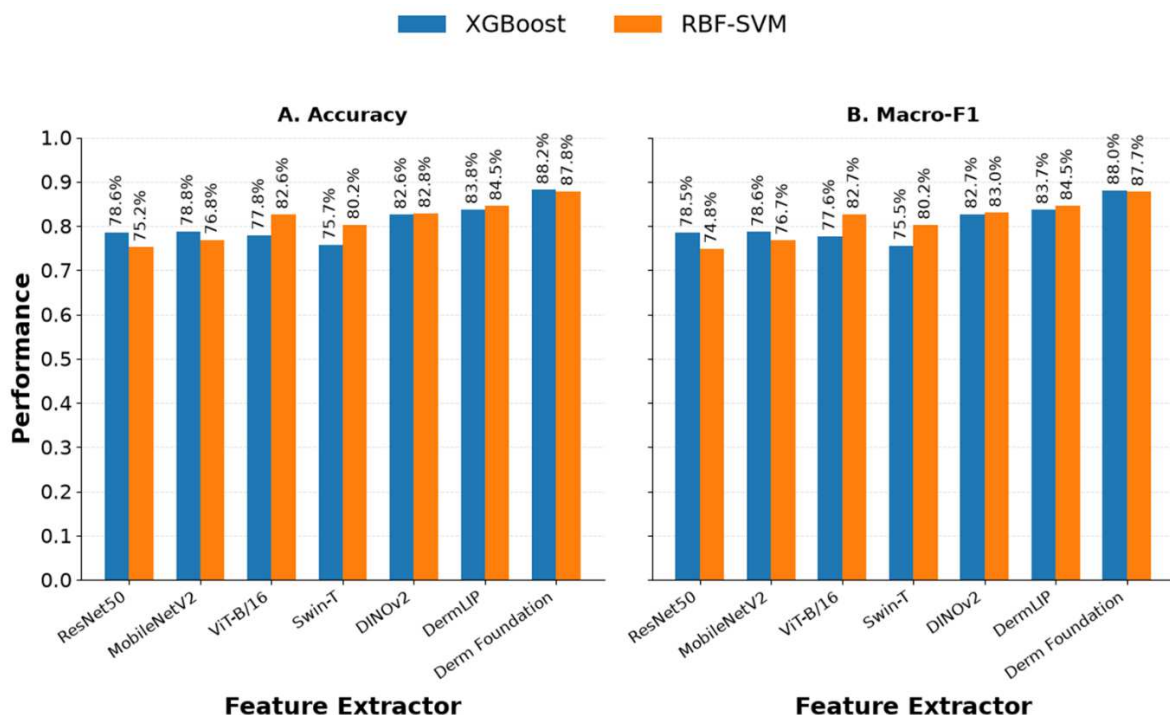


Fig. 4. Comparative validation accuracy across the seven frozen feature extractors using XGBoost vs. RBF-SVM downstream classifiers.

4.4 Ablation analysis: Comparison of downstream classifiers

The effect of the downstream classifier was highly dependent on the underlying model architecture rather than uniformly favoring one algorithm. The RBF-SVM substantially improved the performance of the pure transformer architectures, increasing the accuracy of ViT-B/16 (+4.75 percentage points) and Swin-T (+4.58 percentage points). Conversely, XGBoost proved superior for standard CNN architectures. XGBoost performed better for ResNet50 (−3.41 percentage points for SVM relative to XGBoost) and MobileNetV3-Large (−1.92 percentage points).

For the self-supervised and domain-specific representations, the performance differences between the two classifiers were minimal; for DINOv2 (+0.25 percentage points), Derm Foundation (−0.34 percentage points), and DermLIP (+0.75 percentage points). The extreme proximity of the XGBoost (88.17%) and RBF-SVM (87.83%) accuracy for Derm Foundation implies that a sufficiently mature, domain-specific feature space reduces sensitivity to downstream classifier selection. This pattern suggests a single classifier universally dominates; instead, optimal downstream selection depends entirely on the mathematical structure of the learned feature space.

Table 4. Ablation comparison of validation accuracy between XGBoost and RBF-SVM classifiers.

Feature extractor	XGBoost Accuracy (%)	RBF-SVM Accuracy (%)	Difference (XGB – SVM)
ResNet50	78.58%	75.17%	+3.41 pp
MobileNetV3-Large	78.75%	76.83%	+1.92 pp
ViT-B/16	77.83%	82.58%	−4.75 pp
Swin-T	75.67%	80.25%	−4.58 pp
DINOv2	82.58%	82.83%	−0.25 pp
DermLIP	83.75%	84.50%	−0.75 pp
Derm Foundation	88.17%	87.83%	+0.34 pp

4.5 Statistical Comparison of XGBoost and RBF-SVM

To evaluate whether observed performance gaps between XGBoost and RBF-SVM were statistically meaningful, paired image-level predictions were analyzed using an exact McNemar test with a Holm-Bonferroni correction (Table 5).

The paired analyses confirmed that the statistical significance of classifier differences depended on the feature representation (Table 5). For CNN architecture, ResNet50, XGBoost produced significantly more correct predictions than RBF-SVM (95 versus 54 discordant cases favoring XGBoost; exact McNemar test, Holm-adjusted $P=0.0049$). Conversely, RBF-SVM significantly outperformed XGBoost for the hierarchical Swin Transformer, Swin-T (92 versus 37 discordant cases favoring RBF-SVM; Holm-adjusted $P<0.001$).

Importantly, no statistically significant differences in prediction accuracy were observed for MobileNetV3-Large, DINOv2, DermLIP, or Derm Foundation after applying the Holm correction. These paired analyses therefore did not support a universal advantage of XGBoost or RBF-SVM. Instead, the effect of the downstream classifier was representation-dependent.

Table 5. Paired image-level statistical comparisons between XGBoost and RBF-SVM predictions across the testing cohort (n = 1,200).

Feature extractor	XGB correct	SVM correct	XGB only correct (b)	SVM only correct (c)	Exact McNemar (<i>p</i> -value)	Holm-adjusted (<i>p</i> -value)	Significant
ResNet50	943	902	95	54	0.00098	0.00492	Yes
MobileNet V3-Large	945	922	79	56	0.05789	0.23158	No
Swin-T	908	963	37	92	1.4×10^{-6}	8.4×10^{-6}	Yes
DINOv2	991	994	47	50	0.83921	1.00000	No
DermLIP	1005	1014	42	51	0.40692	1.00000	No
Derm Foundation	1058	1054	47	43	0.75203	1.00000	No

4.6 Per-Class Diagnostic Performance

Figure 5 demonstrates substantial heterogeneity in class-specific sensitivity across the evaluated feature extractor–classifier combinations. Normal skin consistently achieved high recall across the evaluated representations, whereas seborrheic dermatitis generally showed the lowest or among the lowest recall values, with rosacea also showing comparatively lower sensitivity for several representations.

Dermatology-specific representations demonstrated markedly stronger and more balanced recall profiles compared to general-purpose weights. When paired with XGBoost, Derm Foundation achieved high, well-distributed sensitivity across the board: 94.5% for Acne, 87.0% for Atopic Dermatitis, 100.0% for Normal Skin, 90.5% for Psoriasis, 84.0% for Rosacea, and 73.0% for Seborrheic Dermatitis. DermLIP also demonstrated a robust class-wise profile, particularly showing high sensitivity for Normal Skin and Rosacea.

Also, seborrheic dermatitis showed the lowest or among the lowest recall values across the evaluated feature representations, indicating that this class remained comparatively difficult to discriminate irrespective of the downstream classifier.

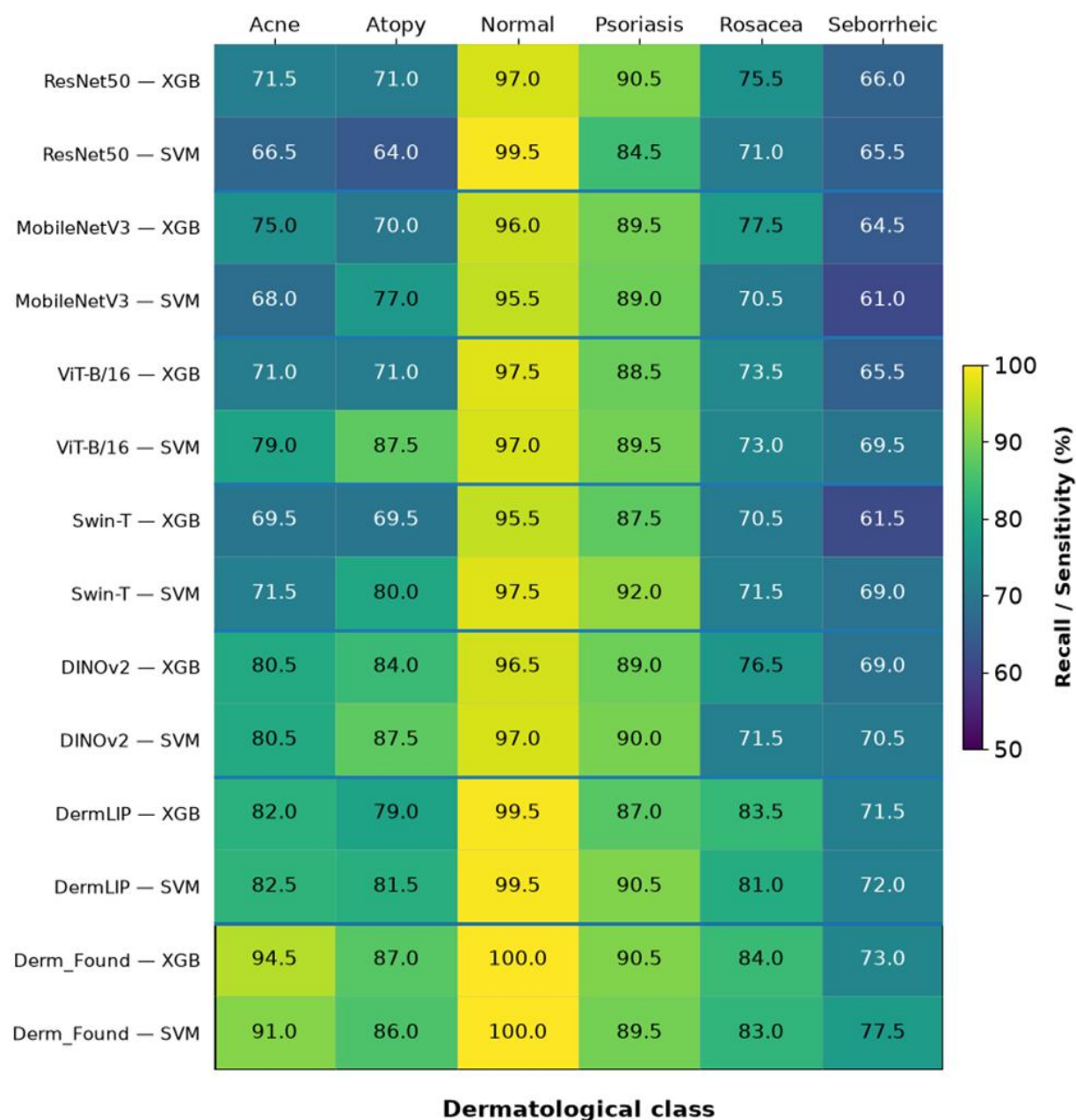


Fig. 5. Per-class recall (sensitivity) across feature extractor–classifier combinations. Recall was calculated for each of the six dermatological classes on the balanced testing set ($n = 1,200$; 200 images per class). Higher recall indicates a greater proportion of images belonging to the corresponding class that were correctly identified. XGBoost and RBF-SVM results are shown for each frozen pretrained feature extractor.

4.7 Confusion-Matrix Analysis

To gain deeper insight into the classification behavior of the seven feature extractors and uncover the specific error typologies driving these recall variations, a multi-class confusion matrix analysis was conducted on the testing set. Overall, the confusion matrices showed substantial differences in class-separation ability across feature representations. The dermatology-specific models, particularly Derm Foundation, demonstrated the strongest predictions performance along the main diagonal, indicating more consistent discrimination between the six diagnostic categories.

Derm Foundation produced the highest number of correct classifications among the evaluated representations in the testing set, correctly classifying 189/200 acne images, 174/200 atopic dermatitis images, all 200/200 normal skin images, 181/200 psoriasis images, 168/200 rosacea images, and 146/200 seborrheic dermatitis images (Figure 6). Misclassification was nevertheless observed for seborrheic dermatitis, which was most frequently confused with rosacea (27 cases), and rosacea was conversely misclassified as seborrheic dermatitis (24 cases). This distinct bidirectional pattern indicates substantial visual overlap between these two inflammatory facial conditions, challenging even the strongest domain-specific feature representation.

The corresponding RBF-SVM confusion matrix showed a similar overall pattern but with somewhat different class-specific error distributions. With Derm Foundation features, RBF-SVM correctly classified 182/200 acne images, 172/200 atopic dermatitis images, 200/200 normal skin images, 179/200 psoriasis images, 166/200 rosacea images, and 155/200 seborrheic dermatitis images (Figure 6). Seborrheic dermatitis remained the most difficult category, with 17 cases misclassified as rosacea, while 20 rosacea images were classified as seborrheic dermatitis. Thus, the reciprocal seborrheic dermatitis–rosacea confusion persisted with RBF-SVM, although the number of seborrheic dermatitis images assigned to rosacea was lower than that observed with XGBoost (17 versus 27 cases). Conversely, RBF-SVM produced 20 rosacea-to-seborrheic dermatitis errors compared with 24 for XGBoost. These findings demonstrate that the underlying Derm Foundation representation retained similar class-level error structure under both downstream classifiers, while the magnitude and distribution of individual errors varied with classifier choice.

In contrast, the general-purpose feature extractors suffered from much wider error patterns (Figure S2). Swin-T, for example, only correctly identified 123/200 seborrheic dermatitis images, misclassifying 35 cases as psoriasis and 28 as rosacea. Similarly, DINOv2 correctly classified 138/200 seborrheic dermatitis images while misinterpreting 34 as psoriasis and 15 as rosacea

A recurrent error pattern across several models involved confusion between seborrheic dermatitis, psoriasis, and rosacea. ResNet-50 misclassified 38 seborrheic dermatitis images as psoriasis and 36 rosacea images as seborrheic dermatitis. MobileNetV3-Large also showed substantial confusion between seborrheic dermatitis and psoriasis, with 33 seborrheic dermatitis images predicted as psoriasis, while 33 rosacea images were classified as seborrheic dermatitis. ViT-B/16 demonstrated a similar pattern, with 31 seborrheic dermatitis images predicted as psoriasis and 31 rosacea images predicted as seborrheic dermatitis. This recurring confusion pattern suggests that these classes remained difficult to distinguish within the evaluated image representations.

DINOv2 improved discrimination across several classes compared with the conventional ImageNet-pretrained representations, particularly for acne, atopic dermatitis, and psoriasis. DermLIP also showed strong diagonal dominance, correctly classifying 164 acne, 158 atopic dermatitis, 199 normal skin, 174 psoriasis, 167 rosacea, and 143 seborrheic dermatitis images. However, similar to the other models, residual confusion remained most evident between seborrheic dermatitis and psoriasis or rosacea.

4.8 Multiclass Discriminative Performance: ROC Analysis

The multiclass ROC analysis assessed the threshold-independent discriminative ability of the evaluated feature representations and downstream classifiers across the six facial skin condition. Class-specific ROC curves and AUC values were examined to determine how effectively each model distinguished each disease category from the remaining classes across different classification thresholds.

For the Derm Foundation representation with XGBoost, the ROC curves demonstrated strong discrimination across all six categories, with a macro-AUROC of 0.9858. Class-specific AUC values were 0.997 for acne, 0.983 for atopic dermatitis, 1.000 for normal skin, 0.988 for psoriasis, 0.984 for rosacea, and 0.962 for seborrheic dermatitis (Figure 7). The high AUC values indicate strong threshold-independent discrimination within the testing dataset. Seborrheic dermatitis showed the lowest AUC (0.962), although discrimination remained high relative to chance performance. This is consistent with the residual confusion involving this category observed in the per-class recall and confusion-matrix analyses. This pattern indicates that the Derm Foundation representation provided strong overall separation of the six classes while retaining comparatively greater overlap for seborrheic dermatitis.

The RBF-SVM models also demonstrated high multiclass discriminative performance across the evaluated feature representations. For Derm Foundation, the macro-AUROC was 0.9850, indicating discrimination comparable to that obtained with XGBoost. More broadly, the differences in macro-AUROC between XGBoost and RBF-SVM were relatively small for the strongest representations, including DINOv2, DermLIP, and Derm Foundation (Figure S3). These findings are consistent with the accuracy-based

analysis, in which classifier differences varied according to the underlying feature representation rather than showing a uniform advantage for either downstream classifier.

Across the ROC analyses, the consistently high AUC values indicate that the learned representations retained substantial class-discriminative information across a range of decision thresholds. However, high AUROC does not imply complete separation of individual disease categories at the selected classification threshold.

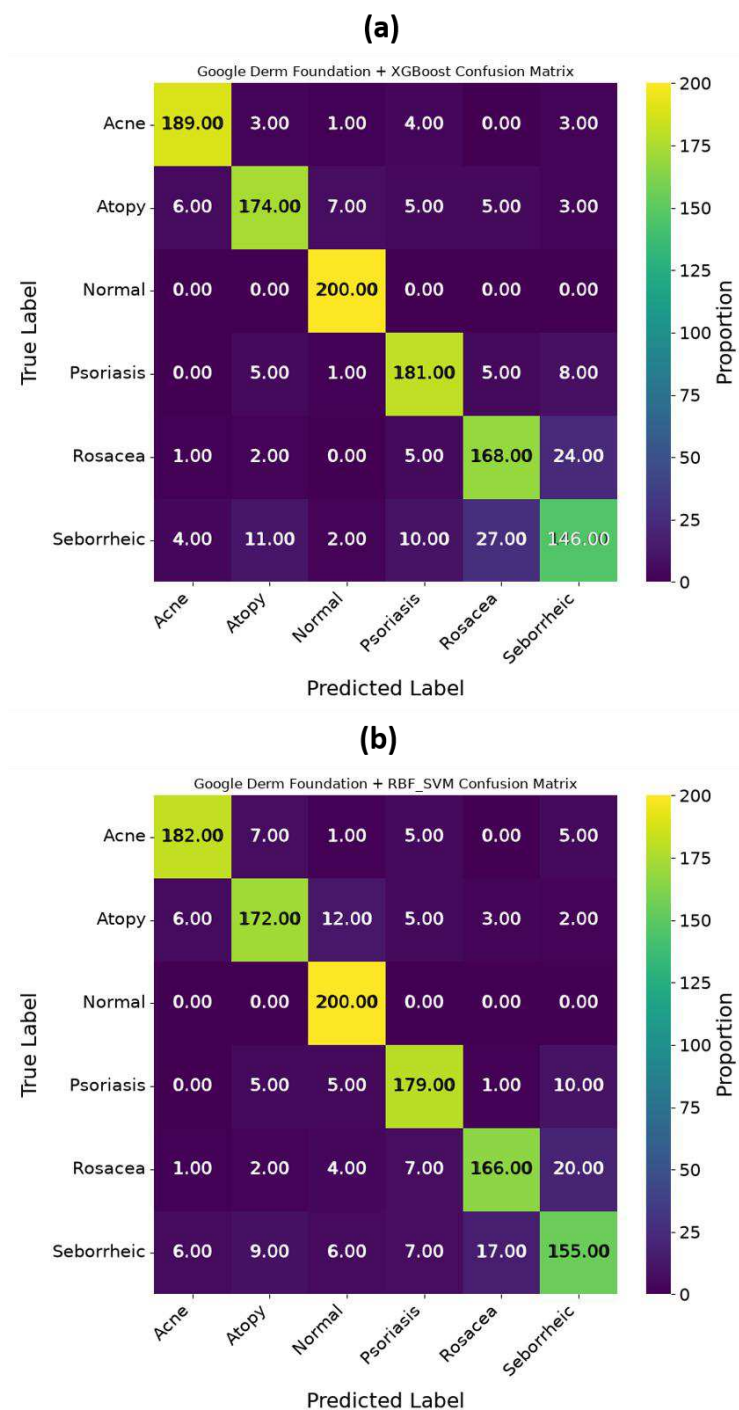


Fig. 6. Multi-class confusion matrices for the peak-performing Derm Foundation feature extractor with the two downstream classifiers. Confusion matrices show the classification performance of the Derm Foundation features combined with XGBoost (a) and RBF-SVM (b) on the balanced testing set ($n = 1,200$; 200 images per class). Multi-model confusion matrix grid for all other baseline models are available in Supplementary Figures S1 and S2.

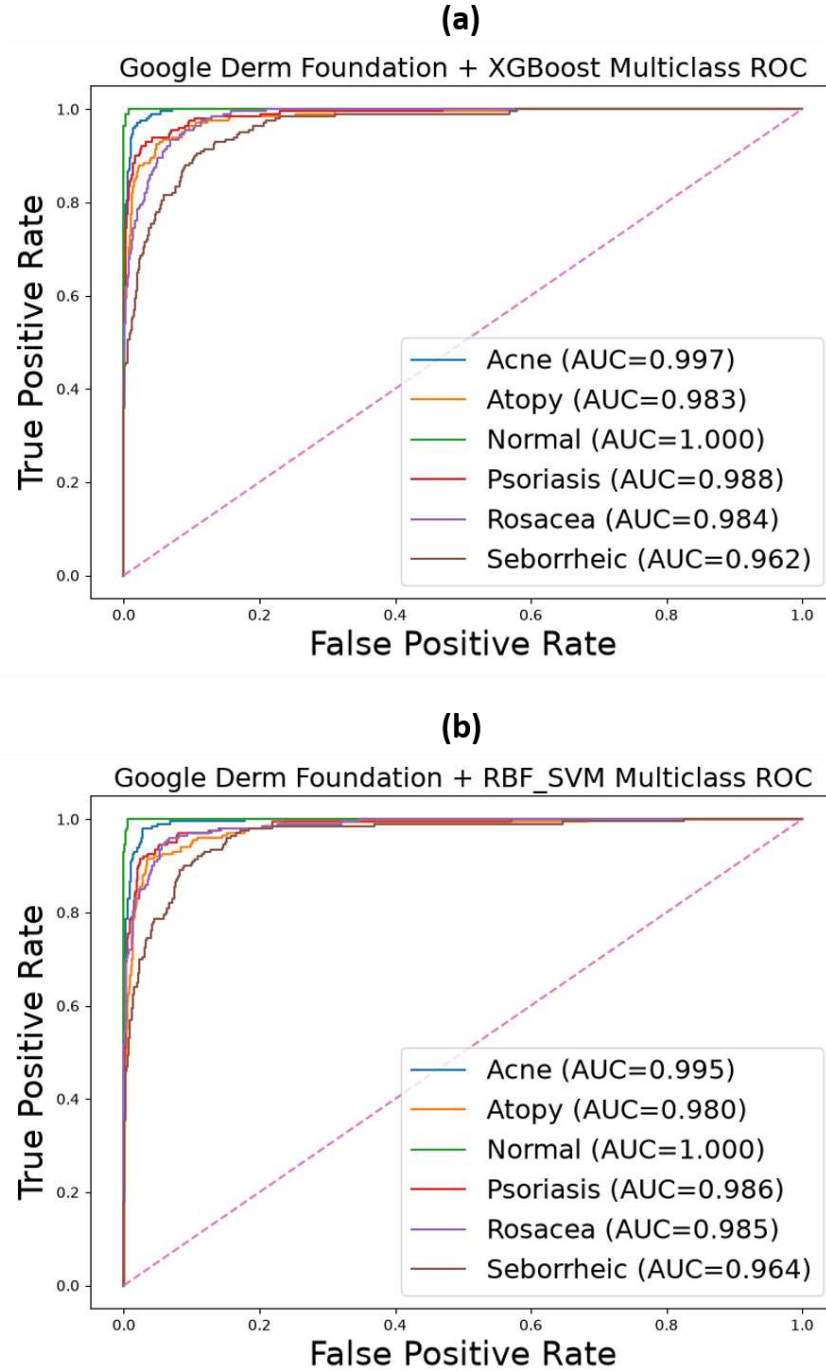


Fig. 7. Class-wise Receiver Operating Characteristic (ROC) curves for the Derm Foundation feature representation with the two downstream classifiers. Class-specific ROC curves are shown for the Derm Foundation features combined with (a) XGBoost (primary classifier) and (b) RBF-SVM (ablation classifier) on the test dataset. The curves and corresponding Area Under the ROC Curve (AUROC) metrics are presented individually for each of the six-target facial skin-conditions. Individual ROC analyses for the other feature extractors are provided in Supplementary Figures S3 and S4.

4.9 Architectural Framework for LLM-Assisted Patient Education

The exploratory LLM component was utilized to translate the static predicted disease categories into natural, patient-oriented educational narratives. Crucially, the LLM operated entirely downstream from the core diagnostic pipeline; it did not participate in image classification, feature extraction, or probability vector estimation, and was strictly isolated from modifying or evaluating the classification framework's predictions [47,48]. This architectural stage is presented strictly as a structural proof-of-concept for multi-modal pipeline integration and does not constitute a clinically validated deployment.

Qualitative inspection of generated outputs showed that the system produced educational text related to the predicted disease categories, including descriptions of common features and general information about the condition. The generated explanations were presented in clear, non-technical language intended for patients without specialized dermatological knowledge.

Unlike traditional rule-based educational architectures, the generative framework dynamically scales without requiring manually curated text templates for each distinct clinical condition. This algorithmic flexibility provides high architectural scalability, allowing the future integration of additional dermatological taxonomies with minimal modifications to the core system.

However, because generative language architectures are inherently susceptible to factual errors, omissions, and hallucinations, formal regulatory evaluations of semantic accuracy, completeness, hallucination rates, readability metrics, and clinical safety were not performed in this investigation. Consequently, the LLM component must be regarded purely as an experimental, unvalidated architectural prototype rather than a clinical information system.

5. Discussion

5.1 Principal Findings and Architectural Comparisons

In this study, we systematically evaluated seven pretrained visual representations for six-class facial skin-condition classification within a frozen-feature transfer learning framework. The principal finding was that dermatology-specific pretrained representations (DermLIP-PanDerm and Google derm foundation) provided stronger discriminative features than general-purpose ImageNet-pretrained convolutional and transformer models. When paired with the primary XGBoost classifier, Derm Foundation achieved the highest overall performance across all global metrics, yielding a multiclass accuracy of 88.17%, a macro-F1 score of 88.05%, and an outstanding macro-AUROC of 98.58%. DermLIP and the self-supervised DINOv2 demonstrated the next strongest performance, whereas ResNet50, MobileNetV3-Large, ViT-B/16, and Swin-T consistently achieved lower overall accuracy. The relative superior performance of DermLIP is consistent with its documented superiority over other foundation models across broader dermatological benchmarks⁴⁶. Similarly, our findings corroborate independent medical computer vision studies indicating that DINOv2 can outperform conventional CNN and transformer baselines on dermatologic image-classification tasks [49].

A second important finding was that the effect of the downstream classifier was representation-dependent. After applying a Holm-Bonferroni correction for multiple comparisons, the primary XGBoost classifier significantly outperformed the RBF-SVM when utilizing ResNet50 embeddings. Conversely, the RBF-SVM significantly reversed this dynamic when paired with the hierarchical Swin Transformer. Interestingly, no statistically significant performance variations were observed between the two classifiers when deploying DINOv2, DermLIP, or Derm Foundation embeddings. These findings suggest that downstream classifier selection cannot be considered independently of the underlying pretrained representation and that no single classifier is uniformly optimal across feature spaces. This structural phenomenon is consistent with multi-class medical imaging benchmarks showing that lightweight adapter performance fluctuates based on how well an encoder's native feature density maps to a given algorithm's decision boundaries and structural assumptions [50].

This specific error pattern is highly plausible from a clinical and physiological perspective. Both facial conditions frequently present in identical central facial distributions, specifically the malar region, the nasolabial folds, and the glabella, and display highly overlapping physical characteristics. Seborrheic dermatitis is characterized by erythematous plaques covered by greasy, yellowish scales, while rosacea presents with persistent centrofacial erythema, telangiectasia, papules, and pustules [51]. During early or mild disease stages, or when altered by lighting conditions in a photograph, the greasy scales of seborrheic dermatitis can visually mimic the shiny, erythematous surface of papulopustular rosacea. Consequently,

differentiating between them remains a notorious clinical hurdle that frequently results in misdiagnoses during manual primary care examinations [51]. Recent clinical guidance explicitly emphasizes that seborrheic dermatitis can closely mimic or concurrently overlap with rosacea and localized psoriasis, necessitating meticulous differential evaluation and occasionally even biopsy [52]. Thus, findings revealed dermatology-specific pretraining improved representation quality overall but did not eliminate completely the challenges posed by highly ambiguous, visually overlapping inflammatory dermatoses when relying solely on 2D pixel information.

The prototype web application implementation demonstrates the technical feasibility of structurally linking deep visual feature extractors with generative language models. Traditional deep learning diagnostics frequently act as isolated "black boxes" that yield single point predictions or raw probability vectors without actionable clinical context, limiting their utility for downstream patient translation. By transforming technical classifications into coherent, patient-oriented educational narratives, this modular prototype provides a technical framework to bridge the gap between raw machine output and meaningful patient literacy. This pipeline alignment is consistent with modern teledermatology engineering concepts that focus on developing scalable multi-modal interfaces to enhance the transparency and accessibility of automated classification systems [53].

The qualitative evaluation spotlighted critical areas for future engineering optimization. First, variability in network processing speeds across different image uploads was flagged as a threat to smooth user adoption. Restructuring the application backend to eliminate processing latency fluctuations and establish uniform execution speeds is an essential requirement prior to any broader clinical testing. Furthermore, designing frontend interface mechanics that translate raw numerical prediction percentages into clear, intuitive qualitative uncertainty categories (e.g., high, moderate, or low confidence boundaries) remains a vital future user-experience constraint [54].

5.2 Limitations and Future Work

An important limitation of the present study is the absence of independent external validation on separate image datasets. Although our models were evaluated on an independent internal testing cohort and demonstrated consistent performance tiers across multiple configurations, their classification accuracy remains unverified when exposed to images acquired from disparate institutions, unseen clinical protocols, varying imaging environments, and diverse skin-tone distributions. During dataset preparation, external public repositories such as DDI, Fitzpatrick17k, and SCIN were meticulously considered but ultimately excluded. This decision was driven by the fact that their baseline diagnostic taxonomies did not correspond sufficiently to the six specific target categories defined in our framework. Prioritizing strict clinical

classification over a forced, artificial label mapping preserved the medical validity of our study setup, though it temporarily prevented a broad assessment of external generalization.

Accordingly, future research should focus on determining whether the strong performance tiers demonstrated by these dermatology-specific representations persist under clinically meaningful distribution shifts. Validating this pipeline against multi-institutional facial-image cohorts containing highly diverse patient populations, comprehensive skin-tone representations across the entire Fitzpatrick scale, and varied consumer camera settings will be the most direct test of external robustness. Additionally, a targeted benchmarking comparison between full end-to-end fine-tuning and frozen-feature adaptation could clarify the exact inflection points where the heavy computational cost of updating a pretrained backbone yields a meaningful diagnostic advantage over the highly efficient frozen strategy evaluated here. Finally, conducting formal safety, and factual accuracy assessments of the downstream LLM-generated educational text will be a critical prerequisite to support the safe translation of this framework into real-world consumer workflows. These subsequent investigations will directly build upon our current findings, systematically verifying the robustness, edge-computing efficiency, and clinical usability of domain-specific representation learning.

6. Conclusion

In conclusion, this study demonstrated that pretrained dermatology-specific visual representations can provide effective frozen features for six-class facial skin-condition classification. Derm Foundation achieved the strongest overall performance with XGBoost, while classifier ablation demonstrated that downstream classifier effects were dependent on the underlying representation. The principal residual classification difficulty involved seborrheic dermatitis and rosacea, identifying a key limitation in relying solely on 2D pixel information for visually overlapping inflammatory facial conditions. Furthermore, the integration of the classifier with a Django-based web application and an LLM-assisted educational module successfully demonstrated the technical feasibility and architectural scalability of multi-modal system integration. Independent external dataset validation and prospective clinical evaluations remain necessary prerequisites before this framework can be considered for routine clinical use.

Contributions

U.E.O, O.S.N and M.A.I conceived and designed the experiment, U.E.O and O.S.N performed the experiment and data analysis, U.E.O and O.S.N wrote the manuscript and U.E.O, O.S.N, M.A.I, H.B and H-C.K reviewed the manuscript.

Funding

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government(MSIT)(IITP-2024-RS-2024-00436773).

Conflicts Of Interest

The authors declare no conflicts of interest.

Data availability statement

The clinical skin image datasets analyzed during the current study are available from the AI-Hub platform (National Information Society Agency, Republic of Korea) repository at https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=230&utm_source=chatgpt.com_

References

1. Huai, P. et al. Global burden of skin and subcutaneous diseases: an update from the Global Burden of Disease Study 2021. *Br. J. Dermatol.* **192**, 1136–1138 (2025).
2. Li, N. et al. Skin and subcutaneous disease burden, 1990–2021: a GBD review. *Int. J. Surg.* **112**, 7858–7868 (2026).
3. Germain, N. et al. Stigma in visible skin diseases: a literature review and development of a conceptual model. *J. Eur. Acad. Dermatol. Venereol.* **35**, 1493–1504 (2021).
4. Liu, L. et al. Acne and risk of mental disorders: a two-sample Mendelian randomization study based on large genome-wide association data. *Front. Public Health* **11**, 1156522 (2023).
5. Tuchinda, P. et al. An updated review for management of atopic dermatitis. *Asian Pac. J. Allergy Immunol.* **43**, 963–983 (2025).
6. Chu, Y., Oh, B. & Yang, S. Evaluating the trustworthiness of foundation models for skin lesion segmentation. In *Proc. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* 573–583 (2025).
7. Waqar, M. et al. A comparative study of convolutional neural networks (CNN) and Inception V3 for facial skin disease classification. *Kashf J. Multidiscip. Res.* **2**, 232–258 (2025).
8. Rahman, S. et al. From algorithm to clinic: a critical review of deep learning advances and challenges in dermatological diagnosis and care delivery. *ICT Express* **12**, 732–745 (2026).
9. Abdulqader, Z. & Abraham, A. Deep learning-based skin disease detection and classification: a systematic literature review. *Dasinya J. Eng. Inform.* **2**, 1 (2026).
10. Aljohani, B. H., Yafooz, W. M. & Alsaedi, A. AI-driven dermatological diagnostics: a systematic review of machine and deep learning methods for skin disease classification. *Array* **30**, 100797 (2026).
11. Safi, I., Safi, F. M. & Pashay, R. Systematic review of deep learning and machine learning models for eczema (atopic dermatitis) image detection. *Technology* **1**, 35–45 (2025).
12. Yan, S. et al. A multimodal vision foundation model for clinical dermatology. *Nat. Med.* **31**, 2691–2702 (2025).
13. Kozachok, A. et al. Searching for robust augmentations to improve out-of-domain generalization in dermoscopic skin cancer classification. Preprint at <https://arxiv.org/abs/2607.26765> (2026).
14. Groh, M. et al. Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 1820–1828 (2021).
15. Ward, A. et al. Creating an empirical dermatology dataset through crowdsourcing with web search advertisements. *JAMA Netw. Open* **7**, e2446615 (2024).
16. Sultana, M. et al. SkinDisNet: a comprehensive dataset of clinical images and metadata for skin disease detection and classification. *Data Brief.* **50**, 112239 (2025).
17. Huang, K. et al. The classification of six common skin diseases based on Xiangya-Derm: development of a Chinese database for artificial intelligence. *J. Med. Internet Res.* **23**, e26025 (2021).

18. Daneshjou, R. et al. Disparities in dermatology AI performance on a diverse, curated clinical image set. *Sci. Adv.* **8**, eabq6147 (2022).
19. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **5**, 180161 (2018).
20. Rotemberg, V. et al. A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Sci. Data* **8**, 34 (2021).
21. Pacheco, A. G. C. et al. PAD-UFES-20: a skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data Brief* **32**, 106221 (2020).
22. Kawahara, J., Daneshvar, S., Argenziano, G. & Hamarneh, G. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE J. Biomed. Health Inform.* **23**, 538–546 (2019).
23. AI Hub. AI Hub dataset: facial skin disease image dataset. <https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=230>.
24. Huang, K. et al. The classification of six common skin diseases based on Xiangya-Derm: development of a Chinese database for artificial intelligence. *J. Med. Internet Res.* **23**, e26025 (2021).
25. Banu, M. S., Dhavapandiammal, A. & Palanisamy, K. Exploring deep transfer learning for medical image processing and analysis: a comprehensive analysis across modalities. *Curr. Med. Imaging* **22**, e15734056456208 (2026).
26. Bisht, Y. S. Enhancing medical image classification using self-supervised transfer learning. *In Proc. 3rd International Conference for Innovation in Technology (INOCON) 2008–2013* (2024).
27. Taha, A. & Lukmanov, R. Where semantics meets space: a layer-wise analysis of frozen encoders for efficient polyp segmentation. *Dokl. Math.* **112**, 592–599 (2025).
28. Mahbod, A., Ecker, R. & Woitek, R. Fusion of foundation and vision transformer model features for dermatoscopic image classification. Preprint at <https://arxiv.org/abs/2505.16338> (2025).
29. Yuceyalcin, F., Yilmaz, A. & Temelkuran, B. A hierarchical benchmark of foundation models for dermatology. Preprint at <https://arxiv.org/abs/2601.12382> (2026).
30. Kora, P. et al. Transfer learning techniques for medical image analysis: a review. *Biocybern. Biomed. Eng.* **42**, 79–107 (2022).
31. Matsoukas, C., Haslum, J. F., Sorkhei, M., Söderberg, M. & Smith, K. What makes transfer learning work for medical images: feature reuse & other factors. Preprint at <https://arxiv.org/abs/2203.01825> (2022).
32. Kim, H. E. et al. Transfer learning for medical image classification: a literature review. *BMC Med. Imaging* **22**, 69 (2022).
33. Nasir, I. M. et al. Transformer-based foundation learning for robust and data-efficient skin disease imaging. *Diagnostics* **16**, 440 (2026).
34. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? *In Advances in Neural Information Processing Systems* **27**, 3320–3328 (2014).
35. Naqvi, M. et al. Skin cancer detection using deep learning—a review. *Diagnostics* **13**, 1911 (2023).

36. Abir, S. I. et al. Precision lesion analysis and classification in dermatological imaging through advanced convolutional architectures. *J. Comput. Sci. Technol. Stud.* **6**, 168–180 (2024).
37. Keerthika, S., Jayadharshini, P. & Santhiya, S. Transfer learning framework for early and accurate skin disease detection using DermNet images. *In Proc. 2025 IEEE 5th International Conference on ICT in Business Industry & Government (ICTBIG) 1–7* (2025).
38. Matas, I. et al. Mitigating overfitting in medical imaging: self-supervised pretraining vs. ImageNet transfer learning for dermatological diagnosis. Preprint at <https://arxiv.org/abs/2505.16773> (2025).
39. Puls, E. D., Todescato, M. V. & Carbonera, J. L. An evaluation of pre-trained models for feature extraction in image classification. Preprint at <https://arxiv.org/abs/2310.02037> (2023).
40. Kornblith, S., Shlens, J. & Le, Q. V. Do better ImageNet models transfer better? *In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2661–2671 (2019).
41. Azizi, S. et al. Big self-supervised models advance medical image classification. *In Proc. IEEE/CVF International Conference on Computer Vision* 3458–3468 (2021).
42. Dosovitskiy, A. et al. An image is worth 16×16 words: transformers for image recognition at scale. Preprint at <https://arxiv.org/abs/2010.11929> (2020).
43. Aydemir, S. B. A comparative benchmark of transformer-based architectures for dermoscopic skin lesion classification on the HAM10000 dataset. *J. Comput. Bioeng. Clin. Med.* **1**, 16–19 (2026).
44. Liu, Z. et al. Swin transformer: hierarchical vision transformer using shifted windows. *In Proc. IEEE/CVF International Conference on Computer Vision* 9992–10002 (2021).
45. Steiner, D. F. et al. Health-specific embedding tools for dermatology and pathology. Google Research (2024). <https://github.com/Google-Health/derm-foundation>.
46. Yan, S. et al. Derm1M: a million-scale vision-language dataset aligned with clinical ontology knowledge for dermatology. *In Proc. IEEE/CVF International Conference on Computer Vision* 12681–12690 (2025).
47. Bajwa, J. et al. Artificial intelligence in healthcare: transforming the practice of medicine. *Future Healthc. J.* **10**, 234–240 (2023).
48. Singhal, K. et al. Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
49. Taufik, E. A., Khondoker, A., Parsa, A. F. & Mostafa, S. A. Visual bias and interpretability in deep learning for dermatological image analysis. *In Proc. 2025 4th International Conference on Image Processing and Media Computing (ICIPMC) 45–49* (IEEE, 2025).
50. Li, X. et al. From embeddings to accuracy: comparing foundation models for radiographic classification. *J. Imaging Inform. Med.* **39**, 3196–3207 (2026).
51. Choi, J. W., Kim, B. R., Lee, H. S. & Youn, S. W. Characteristics of subjective recognition and computer-aided image analysis of facial erythematous skin diseases: a cornerstone of automated diagnosis. *Br. J. Dermatol.* **171**, 252–258 (2014).
52. Turchin, I. et al. Current understanding of seborrheic dermatitis: presentation, diagnosis, and special populations. *J. Cutaneous Med. Surg.* **29**, 16S–23S (2025).

53. Navarrete-Dechent, C. Teledermatology and artificial intelligence. *Iproceedings* **8**, e36894 (2022).
54. Malvey, J. Risks and benefits of artificial intelligence in teledermatology. *Iproceedings* **8**, e36891 (2022).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementary.docx](#)