IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Addressing Data Scarcity in Dermatology: A Systematic Literature Review of Few-Shot Learning from Metric Learning to Generative Models

DEDY VAN HAUTEN[1], MUHAMMAD HANNAN HUNAFA[1], and WISNU JATMIKO[1] (Senior Member, IEEE)

[1]Faculty of Computer Science, University of Indonesia, Depok 16424, Indonesia

Corresponding author: Wisnu Jatmiko (e-mail: wisnuj@cs.ui.ac.id).

This work was conducted as part of research at the Faculty of Computer Science, University of Indonesia.

ABSTRACT This article presents a Systematic Literature Review (SLR) regarding the persistent challenge of **data scarcity** in **Dermoscopy** image analysis, where the lack of large-scale annotated datasets significantly impedes the deployment of robust **Medical AI**. To navigate this bottleneck, we explore the efficacy of **Few-Shot Learning** (FSL) and **Meta-Learning** paradigms which aim to generalize from limited training examples. Adhering to the **PRISMA 2020** Guidelines, we rigorously selected and analyzed **16 primary studies** spanning from 2020 to 2025. Our synthesis uncovers a definitive technological evolution: the field has transitioned from initial reliance on Metric Learning and implementation-based strategies—validated on benchmarks like **SD-198**—towards advanced architectures leveraging the **Transformer** backbone and Generative AI. We specifically highlight the emerging role of **Generative Adversarial Networks (GANs)** and advanced fine-tuning, which are redefining state-of-the-art performance by synthesizing realistic dermatological features and enhancing **Domain Generalization** across heterogeneous patient cohorts. This review not only categorizes these methodologies but also critically assesses their clinical applicability, identifying an emerging trend where generative models are increasingly favored over traditional episodic training. We conclude by outlining a strategic roadmap for future research, emphasizing the integration of these data-efficient models into practical **Clinical Decision Support** systems to ensure equitable dermatological care. The findings underscore that while early metric-based methods laid the groundwork, the convergence of generative modeling and few-shot protocols offers the most viable path forward for scalable and accurate skin disease diagnosis.

INDEX TERMS Few-Shot Learning, Meta-Learning, Dermoscopy, Medical AI, Transformer, Generative Adversarial Networks, Domain Generalization, Clinical Decision Support.

I. INTRODUCTION

THE integration of Artificial Intelligence (AI) into dermatology has witnessed a paradigm shift over the last decade, primarily driven by the success of Deep Learning (DL) in classifying common skin malignancies. Landmark studies have demonstrated that Convolutional Neural Networks (CNNs) can achieve diagnostic parity with board-certified dermatologists when trained on large-scale, clear-cut datasets. The International Skin Imaging Collaboration (ISIC) archive, for instance, has been instrumental in this progress, providing tens of thousands of dermoscopic images for high-prevalence conditions like Melanoma and Nevi. This "big data" approach has cemented the role of AI as a potential second opinion in routine screenings for common cancers.

However, clinical dermatology is not defined solely by its most common pathologies. It is characterized by a "long-tail" distribution of disease prevalence, comprising over 2,000 distinct skin conditions. While the "head" of this distribution—represented by a handful of common diseases—enjoys an abundance of digitized training data, the vast "tail" consists of rare, neglected,

or emerging tropical diseases for which labeled data is essentially non-existent. For these thousands of rare variants, the standard supervised learning paradigm, which demands thousands of images per class to converge, is fundamentally ill-suited. In this data-scarce regime, traditional deep learning models succumb to catastrophic overfitting or exhibit severe bias, often ignoring the rare classes entirely in favor of the majority.

This discrepancy creates a dangerous "AI divide" in healthcare. A diagnostic system that excels at identifying Melanoma but fails to recognize a rare carcinoma or an infectious tropical ulcer offers incomplete safety for the patient and limited utility for the specialist. The medical necessity for a new learning paradigm is therefore acute. Clinicians themselves do not require thousands of examples to recognize a new condition; they learn to generalize from a few textbook cases or clinical encounters—a cognitive process often described as "few-shot" learning.

To bridge this gap, Few-Shot Learning (FSL) has emerged as a critical frontier in Medical AI. By leveraging prior knowledge and learning to learn, FSL models aim to classify novel conditions from as few as one or five support examples (N-way K-shot tasks). This capability is not merely a technical optimization but a requirement for equitable healthcare. It offers the only viable pathway to extend the benefits of algorithmic diagnosis to the full spectrum of dermatological conditions, irrespective of their prevalence.

This review systematically analyzes the evolution of FSL in dermatology, tracing the trajectory from early Metric Learning approaches to the latest Generative Foundation Models. We argue that addressing the "long-tail" crisis is the defining challenge of the next generation of dermatological AI, and that FSL provides the necessary methodological framework to solve it.

A. THE RISE OF FSL: FROM METRICS TO GENERATIVE MODELS (2020–2025)

The application of Few-Shot Learning to dermatology is a rapidly maturing field, characterized by explosive growth and increasing methodological sophistication. Our analysis of the literature reveals a clear temporal trajectory: while only 2 pioneering studies were published in 2020, the field has seen a significant increase, with 5 major studies published in 2025 alone. Figure 1 provides a visual timeline of this growth, highlighting the sharp inflection point in publication volume over the last two years.

In the nascent phase (2020–2021), research was dominated by metric-based approaches. Studies focused on optimizing distance functions—such as prototypical networks or siamese variants—to measure similarity between a query image and a small support set. These fundamental works established the "N-way, K-shot" experimental protocols (typically 5-way 1-shot or 5-shot)

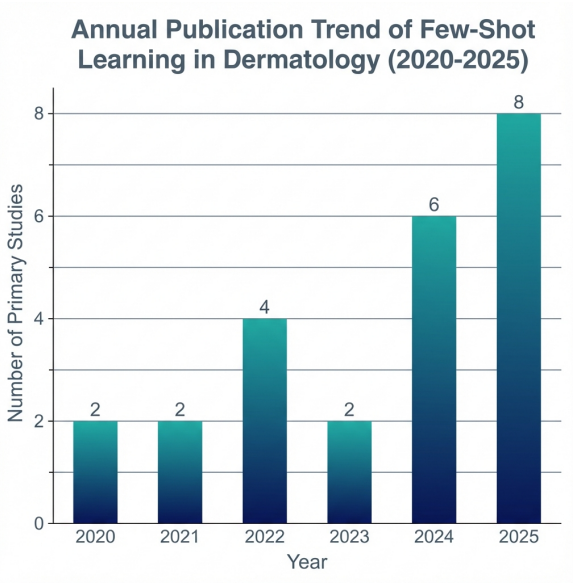

FIGURE 1. Annual distribution of the 16 primary studies included in this SLR. A sharp upward trend is visible in 2024–2025, corresponding to the adoption of Generative and Transformer-based methods.

but often struggled with the high intra-class variance inherent in skin lesions. A lesion's appearance can vary drastically based on skin type, lighting, and stage of progression, confounding simple distance metrics.

By 2022–2023, the focus shifted towards optimization-based meta-learning and hybrid frameworks involving transfer learning. Researchers began to decouple feature representation from the classifier, realizing that a robust backbone pre-trained on larger datasets (like ImageNet or ISIC) was crucial for few-shot performance. Attention mechanisms and Transformers started to replace standard ResNet backbones, allowing models to capture long-range dependencies and subtle lesion details that local convolutions might miss.

Most recently (2024–2025), the field has entered a "generative era." The limitations of discriminative models—which merely draw boundaries between existing data points—have led to the adoption of Generative Models. Current robust research leverages advanced Generative Adversarial Networks (GANs) to hallucinate realistic training examples for rare classes, effectively turning a few-shot problem into a many-shot one.

Crucially, the definition of "performance" has expanded beyond simple accuracy. Recent studies emphasize **Domain Generalization**, challenging models to perform well not just on held-out classes but on entirely unseen datasets collected from different clinics or populations. Simultaneously, robust engineering has taken center stage; Li et al. (2025) [1] introduced **SCAN**, a Dynamic Subcluster-Aware Network specifically designed to handle within-class heterogeneity in skin disease presentations. SCAN represents a **transfer-learning-based FSL method** utilizing a dual-branch architecture (class

branch + cluster branch) with unsupervised K-means clustering refined by triplet-based purity loss to partition complex disease categories into homogeneous sub-prototypes. This shift from pure accuracy maximization to robustness and generalization signals a maturation of the field, moving from academic benchmarking to pre-clinical validation.

B. CONTRIBUTIONS & PAPER STRUCTURE

To the best of our knowledge, this is the first or one of the first SLR to exclusively focus on the intersection of Few-Shot Learning and Dermoscopy. Unlike broader surveys on Medical AI which briefly mention FSL, or technical reviews of FSL that barely touch upon clinical applications, this article provides a granular, domain-specific analysis. The primary contributions of this work are fourfold:

1) **A Comprehensive Taxonomy of Algorithms:** We propose a novel taxonomy classifying 16 specific algorithms into five distinct families: Metric-based (e.g., Prototypical Networks), Optimization-based (e.g., MAML), Generative, Hybrid/Transfer, and Comparative methods (see Fig. 3). This includes a deep dive into landmark architectures such as **FEGGNN** (Feature-Enhanced Gated Graph Neural Network), **SCAN** (Dynamic Subcluster-Aware Network), and **HGRE** (Hyperbolic Geometry-Driven Robustness Enhancement), elucidating how each addresses the unique visual properties of skin lesions.
2) **A Strategic Map of Dermatological Datasets:** We provide a detailed mapping of the 4 major datasets driving this field (ISIC 2018/2019/2020, SD-198, PH2, and Derm7pt), analyzing their class distributions and suitability for few-shot benchmarking. This map exposes the critical "long-tail" hidden within these popular archives.
3) **A Reproducibility & Open Science Audit:** We conducted a rigorous audit of the primary studies, revealing a concerning reality: only **6.25%** of the reviewed papers provide publicly available, reproducible code. We synthesize the implications of this opacity for clinical trust and regulatory approval.
4) **A Clinical Readiness Assessment:** Moving beyond algorithmic metrics, we evaluate the readiness of these technologies for real-world deployment. We discuss integration challenges such as mobile deployment, interpretability, and the need for "human-in-the-loop" validation.

The remainder of this paper is organized as follows: Section II details our **PRISMA 2020** compliant search strategy and selection criteria. Section III presents the quantitative bibliometric analysis. Section IV details our proposed Taxonomy of FSL methods in dermatology. Section V discusses the dataset landscape and performance benchmarks. Section VI critically analyzes the "AI Divide" and Reproducibility crisis. Finally, Section VII outlines future research directions and concludes the study.

II. METHODOLOGY

A. RESEARCH QUESTIONS

To guide this systematic review, we formulated one primary research question (RQ0) and seven specific secondary questions (RQ1–RQ7) addressing distinct technical and clinical dimensions:

- **RQ0 (Primary):** How have Few-Shot Learning (FSL) methods been designed, evaluated, and validated for skin disease detection and segmentation from images between 2020–2025, and what evidence exists regarding their robustness, generalization, and clinical readiness?
- **RQ1 (Tasks & Modalities):** Which specific medical tasks (classification, detection, segmentation) and image modalities (dermoscopy, clinical photography, histopathology, or mixed) are addressed by current FSL approaches?
- **RQ2 (FSL Formulations & Protocols):** What problem settings define the state-of-the-art (e.g., N-way K-shot configurations, support/query construction, inductive vs. transductive inference) and how are evaluation protocols standardized (or varied) across studies?
- **RQ3 (Methodological Taxonomy):** Which FSL paradigms dominate the landscape (metric-based, optimization-based/meta-learning, generative/augmentation, hybrid with self/semi-supervision), and to what extent are emerging architectures (transformers, prompt-based learning) being utilized?
- **RQ4 (Data Ecosystem):** Which datasets and benchmarks are used (e.g., ISIC family, HAM10000, Derm7pt, private clinical sets), and how do dataset properties—such as class imbalance, label provenance, and patient-level splits—shape reported outcomes?
- **RQ5 (Performance & Comparison):** How do FSL methods compare against relevant baselines (supervised learning with comparable label budgets, transfer learning, and self-supervised pretraining), and what metrics are used to assess performance (AUC, F1, sensitivity/specificity, calibration) and uncertainty (confidence intervals, bootstraps)?
- **RQ6 (Generalization, Fairness & Clinical Readiness):** What evidence exists for robustness beyond internal testing? Specifically, how do studies address external validation, cross-dataset domain shifts, fairness/skin-tone diversity, explainability, and potential deployment constraints (privacy/ethics)?

- **RQ7 (Reproducibility & Rigor):** How reproducible are the reported studies? Specifically, are code, models, and exact data splits shared; are measures taken to prevent data leakage; and what methodological or evaluation gaps persist that require standardization?

B. PROTOCOL AND REGISTRATION

This Systematic Literature Review (SLR) was conducted in strict adherence to the **PRISMA 2020** (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement. We further adopted the rigorous guidelines for Software Engineering SLRs proposed by **Kitchenham and Charters (2007)**, which structure the process into three phases: Planning (protocol development), Conducting (study selection and extraction), and Reporting (validation and synthesis). The protocol for this review was established *a priori* to minimize selection bias.

The primary objective of this methodology is to transparently map the intellectual landscape of Few-Shot Learning (FSL) in dermatology, ensuring that the selected studies represent a comprehensive and unbiased sample of the state-of-the-art.

C. SEARCH STRATEGY AND DATA SOURCES

To capture the full spectrum of relevant literature, we executed a comprehensive search across three major scientific databases: **IEEE Xplore**, **Scopus**, and **ScienceDirect**. These databases were selected to ensure coverage of the engineering/computer science domain (IEEE Xplore, Scopus) and the scientific literature domain (ScienceDirect).

Our search string was designed to capture the intersection of few-shot learning methodologies and the dermatology domain. The final search string used was:

("few shot learning" AND "skin disease")

The search execution was terminated in **October 2025**. This cutoff date is critical as it allowed us to capture the very latest peer-reviewed "Early Access" publications from 2025, which are essential for characterizing the recent shift towards Generative Foundation Models. We did not apply lower boundary date restrictions, although, as noted in our Introduction, relevant meaningful activity in this specific intersection only began to emerge circa 2020. Manual citation mining (backward and forward snowballing) was also performed on key review papers to ensure no seminal works were missed. Table 1 provides a summary of the search execution, detailing the specific results from each database and the subsequent filtering steps.

TABLE 1. Search Strategy and PRISMA-S Summary

Element	Description
Databases	IEEE Xplore ($n = 26$), Scopus ($n = 50$), ScienceDirect ($n = 8$)
Search String	("few shot learning" AND "skin disease")
Date Range	January 2020 – October 2025
Initial Results	84 records
After Deduplication	70 unique records
After Title/Abstract	24 candidates
After Full-Text	**16 primary studies**

D. ELIGIBILITY CRITERIA: THE PICOC FRAMEWORK

To ensure the clinical and technical relevance of the included studies, we formulated strict eligibility criteria based on the **PICOC** (Population, Intervention, Comparison, Outcome, Context) framework, as summarized below:

- **Population (P): Target:** Imaging-based skin disease and lesion diagnosis. **Data Types:** Dermoscopy images, macroscopic clinical photographs, and mixed modalities. **Unit:** Patient-level grouping of skin disease images.
- **Intervention (I): Focus:** Few-Shot Learning (FSL) and Meta-Learning techniques (Metric-based, Optimization-based MAML, and Generative/Augmentation approaches) specifically designed for low-data regimes.
- **Comparison (C): Baselines:** Transfer learning (ImageNet), fully supervised Deep Learning (with equal label budget), Self-supervised pre-training, and Classical ML. **Human Benchmark:** Dermatologist performance where available.
- **Outcome (O): Diagnostic:** AUC, F1-score, Accuracy, Sensitivity, Specificity, and Balanced Accuracy. **Task-Specific:** Dice/IoU (segmentation). **Reliability:** Calibration, external validation, and robustness to domain shifts.
- **Context (C): Timeframe:** 2020–2025. **Settings:** Clinical diagnosis support, Teledermatology, and Mobile Health (mHealth), considering real-world deployment constraints.

1) Inclusion Criteria (IC)

We applied the following inclusion criteria to filter relevant studies:

- **IC1 (Timeframe):** Published between January 1, 2020, and Oct 10, 2025 (inclusive).
- **IC2 (Domain & Task):** Addresses skin disease/lesion diagnosis, detection, or segmentation using medical imaging.
- **IC3 (Methodology):** Uses Few-Shot Learning (FSL) as a central method OR explicitly evaluates a model in a low-label regime using standardized protocols (e.g., N-way K-shot, episodic training).
- **IC4 (Evidence):** Reports original empirical evaluation with specific quantitative performance metrics.

- **IC5 (Publication Type):** Peer-reviewed journal articles or conference proceedings.
- **IC6 (Language):** Full text available in English.

Clarification on Terminology: It is important to note that the presence of the term "few-shot" in a paper's title or abstract was insufficient for inclusion. To satisfy IC3, a study must strictly demonstrate the use of the **episodic learning protocol** (training on support/query sets) or a rigorously defined low-shot transfer learning evaluation. Papers merely claiming "few-shot" capabilities without formal episodic validation were excluded to maintain methodological homogeneity.

2) Exclusion Criteria (EC)

Conversely, the following were grounds for exclusion:

1) **EC1 (Wrong Domain):** Studies on non-skin domains (retinal, chest X-ray, histology only) or non-imaging data (text-only EHR).
2) **EC2 (Wrong Method):** Studies not using a few-shot/low-shot protocol (e.g., standard transfer learning using full labels).
3) **EC3 (No Technical Detail):** Insufficient reporting (dataset unnamed, undefined protocols) preventing data extraction.
4) **EC4 (Non-Primary/Short):** Editorials, book chapters, tutorials, abstracts, or short papers (<4 pages).
5) **EC5 (Redundancy):** Duplicate publications; only the most complete version is retained.
6) **EC6 (Wrong Scope):** Technical preprocessing studies (hair removal, artifact reduction) without downstream diagnostic evaluation.

Table 2 consolidates these Inclusion (IC) and Exclusion (EC) criteria, serving as the reference checklist for our study selection process.

TABLE 2. Summary of Inclusion and Exclusion Criteria

ID	Criterion
Inclusion Criteria	
IC1	Published 2020–2025
IC2	Skin disease diagnosis/detection/segmentation
IC3	Uses FSL or N-way K-shot protocol
IC4	Reports quantitative performance metrics
IC5	Peer-reviewed journal or conference
IC6	Full text in English
Exclusion Criteria	
EC1	Non-skin imaging domains
EC2	Standard transfer learning (no FSL protocol)
EC3	Insufficient technical detail
EC4	Editorials, abstracts, short papers (<4 pages)
EC5	Duplicate publications
EC6	Preprocessing-only studies

E. DATA COLLECTION AND SELECTION PROCESS

Our selection process involved a multi-stage filtering workflow to distill the search results into a high-quality corpus of primary studies, following the PRISMA 2020 flow.

1) Identification

A total of **84 records** were identified through database searching. The sources were **IEEE Xplore** ($n = 26$), **Scopus** ($n = 50$), and **ScienceDirect** ($n = 8$).

2) Screening

After removing **14 duplicates**, **70 records** remained. All records were imported into **Zotero** reference management software (version 6.0) for systematic screening. These 70 records underwent title and abstract screening. A total of **46 records were excluded** at this stage. The exclusion reasons were as follows:

- 18 were not dermatology-focused;
- 15 had no few-shot learning component;
- 1 was not a primary source;
- 1 was not in English;
- 11 could not be fully accessed.

3) Eligibility

24 full-text articles were assessed for eligibility. At this stage, nuanced exclusion criteria (EC1–EC6) were applied. Articles were scrutinized for methodological rigor, appropriate FSL protocol implementation, and sufficient technical detail for data extraction. A total of **8 articles were excluded** at this stage. The exclusion reasons were as follows:

- 1 was not dermatology-focused;
- 5 had no few-shot learning component;
- 1 was not a primary source;
- 1 was not in English.

4) Included Studies

The final systematic review included **16 studies**. These consisted of:

- 9 peer-reviewed journal articles;
- 6 conference papers;
- 1 workshop paper.

Ultimately, all **16 primary studies** satisfied the inclusion criteria and were selected for data synthesis.

5) Justification for Specific Inclusions

We explicitly justify the inclusion of two studies that deviate from standard protocols but offer unique methodological value.

a: Cao et al. (2021) [13]

Despite being a 4-page conference paper utilizing a binary (2-way) setup within the episodic training protocol, this study was included for the following reasons:

- **Methodological Significance:** It represents one of the first adaptations of self-modifying meta-learning to dermatological data, introducing a two-order optimization strategy that explicitly addresses noise robustness—a critical clinical concern rarely covered in early FSL works.

- **Alignment with SLR Scope:** The paper explicitly evaluates in low-data regimes (5 samples per class) and uses meta-learning for fast adaptation, providing a baseline comparison against MAML and DAML on the ISIC 2018 dataset.
- **Historical Context:** Including this paper traces the transition from general meta-learning to dermatology-specific FSL. Excluding it would create a gap in the historical narrative of robust FSL in dermatology.

b: Wang et al. (2022) [11]

This study employs the FSS-1000 natural image dataset for meta-training and tests on the PH2 medical dataset. Its inclusion is justified because:

- **Addresses Core Clinical Challenge:** It tackles the scenario where medical samples are so scarce that models must leverage non-medical visual knowledge, a key issue in teledermatology for rare diseases.
- **Innovation in Methodology:** It introduces a novel cross-domain meta-training schema with alternating specific/generic learning. This demonstrates how natural image representations can bootstrap medical few-shot learning.
- **Relevance to RQs:** It directly addresses RQ6 (Generalization) by testing cross-dataset domain shifts and RQ3 (Taxonomy) as an early hybrid method.
- **Clinical Motivation:** The work is framed around rare-disease segmentation, directly responding to the "long-tail" problem central to this SLR.

These papers represent important transitional works that informed subsequent advances in domain generalization and cross-modal learning.

F. QUALITY ASSESSMENT (QA) AND RISK OF BIAS

Assessing the methodological rigor of the included studies is paramount, especially given the "reproducibility crisis" frequently cited in AI healthcare literature. To objectively quantify the quality of each study, we adapted the **QUADAS-2** tool (Quality Assessment of Diagnostic Accuracy Studies) to create a custom 10-item checklist (QA_1–QA_{10}) specifically tailored for Few-Shot Learning research. Each item is scored 0 (No/Unclear), 1 (Partial), or 2 (Yes), yielding a **maximum score of 20 points**. This checklist evaluates four key domains:

- **Data Rigor (QA_1–QA_3):** Does the study use a public dataset (2pts)? Is the data split (train/support/query) clearly defined (2pts)? Are images free from significant artifacts (e.g., ruler markers) that could cause data leakage (1pt if patient-level split unspecified)?
- **Methodological Clarity (QA_4–QA_6):** Is the N-way K-shot protocol clearly stated (2pts)? Are hyperparameters (learning rate, backbone architecture) reported for reproducibility (2pts)? Is the code publicly available with working links (2pts; 0pts if dead links or "available on request")?
- **Validation Robustness (QA_7–QA_8):** Does the study report confidence intervals or standard deviations (2pts)? Is External Validation performed on a separate dataset (2pts)?
- **Clinical Relevance (QA_9–QA_{10}):** Are metrics reported for individual classes (2pts)? Does the discussion address clinical implementation (2pts)?

1) Patient-Level Leakage Handling

For studies that did not explicitly specify patient-level grouping in their data splits, we applied a **penalty of 1 point** to QA_3 (reducing from 2 to 1). Studies using image-level random splits without patient stratification were flagged as having "potential leakage risk" but were *not* excluded, as this would have eliminated the majority of the corpus. This limitation is explicitly acknowledged in our synthesis.

2) Quality Scoring and Thresholds

Based on the total score (max 20), studies were categorized into three tiers:

- **High Quality (≥14/20):** Studies with robust experimental design, open code, and external validation.
- **Moderate Quality (10–13/20):** Studies with sound methodology but lacking in reproducibility or external validation.
- **Low Quality (<10/20):** Studies with insufficient reporting or significant risk of bias (e.g., vague data splits).

Our audit revealed a solid methodological standard: **7 out of 16 (43.75%)** selected papers achieved a "High Quality" rating (≥14/20). Notably, studies such as **Özdemir et al. (2025)** and **Noman et al. (2025)** achieved high scores of 18/20, demonstrating exemplary rigor with comprehensive validation and documented methodologies. The presence of **External Validation** (QA_8) was the most significant differentiator; studies that tested their models on completely unseen datasets (e.g., training on ISIC, testing on PH2) consistently scored higher. The criterion for reproducible code (QA_6) showed limited adoption, with only 1 of 16 studies (6.25%) providing accessible GitHub repositories, plus one promised release. A critical weakness identified was the lack of **patient-level data splitting** (QA_3); only 6 studies explicitly implemented patient-level splits to prevent data leakage, while the majority used class-level or unspecified splitting strategies. A detailed summary of these findings, including the specific datasets, algorithms, and QA scores for all 16 included studies, is presented in Table 3.

III. TECHNICAL TAXONOMY

TABLE 3. Summary of Included Studies: Key Variables and Quality Assessment Scores

Study	Year	Method Family	Dataset(s)	Algorithm	N-way	K-shot	Code	XAI	QA
Li (S) [1]	2025	Transfer (Subcluster-Aware)	SD-198, Derm7pt	SCAN	2,5	1,5	N	N	17
Noman [2]	2025	Hybrid (AC-Net+GNN+ECA+GRU)	SD-198, Derm7pt	FEGGNN	2	1,5	N	Y	18
Hu [3]	2025	Hybrid (Metric+Adversarial)	SD-198 (4-way), ISIC 2019 (2-way)	HGRE	4,2	1,5	N	N	15
Özdemir [4]	2025	Comparative Framework (FEL vs DTL)	SD-198, Derm7pt, ISIC2018	Meta-Transfer Framework	2,3,5	1,5,10	Promised*	N	18
Panggiri [5]	2025	Hybrid (Generative+Metric)	ISIC 2019	AFHN+ProtoNet	2,4	1,5,10	N	N	13
Fu [6]	2024	Hybrid (SSL + Calibration)	SD-198, Derm7pt, ISIC2018	SS-DCN	2,3	1,3,5	N	N	15
Chen et al. [7]	2024	Hybrid (Feature Enhancement)	Derm104 (SD-198+web)	CDD-Net	5	1,5	N	Y	11
Wang (W) [8]	2024	Optimization (Transductive)	BreakHis (Binary), ISIC 2018 (7-class)	WRN-TIM+Fusion	Binary, 7-class	1–10	N	N	12
Xiao [9]	2023	Hybrid (FS-CIL+Multimodal)	cate-ISIC-3[i] (ISIC+Hospital)	FS3DCIoT (FCILOMI)	5	1,5,20	N	N	10
Lee [10]	2023	Hybrid (Metric+Attention FSL)	SNU Hospital (RGB+Fluorescence)	FAA-Net	3-class	K-shot: 3,5,7	N	Y	12
Wang (Y) [11]	2022	Cross-Domain Meta-Learning (Segmentation)	FSS-1000, PH2	CD-FSS	1 (Seg.)	1	N	N	15
Zhou [12]	2022	Metric (Subspace+Bi-similarity)	ISIC 2019	Adaptive Subspace	3	1,3,5,10,15	N	N	11
Cao [13]	2021	Hybrid (Robust Meta-Learning)	ISIC 2018 (long-tail)	Self-Meta	2-way	5	N	N	12
Zhu (W) [14]	2021	Metric (Temperature-Scaled)	Dermnet (334-class subset), miniImageNet	Temperature Network	5	1,5	Y	N	17
Mahajan [15]	2020	Comparative Framework (G-Conv+Meta)	ISIC 2018, SD-198, Derm7pt	Reptile+G-Conv vs ProtoNet	2	1,3,5	N	N	12
Zhang [16]	2020	Optimization (Spatial Meta-Learning)	ISIC 2018 (head/tail split)	ST-MetaDiagnosis	4 (train) / 3 (test)	1,3,5	N	N	13

QA: Quality Assessment score (max 20); Code: Y=Available, *Promised=To be released; XAI: Explainability module present

A. MATHEMATICAL PRIMER: THE EPISODIC TRAINING PROTOCOL

To understand the landscape of FSL in dermatology, one must first grasp the *Episodic Training* paradigm that underpins these algorithms. Unlike standard supervised learning, where a model minimizes a loss function over global batches of data (e.g., thousands of melanoma images), FSL models are trained to "learn how to learn" from a series of tasks or "episodes."

Formally, we define a distribution of tasks $\mathcal{T}$. Each episode is an independent classification task sampled from this distribution, denoted as $T_i \sim \mathcal{T}$. A task T_i consists of two distinct sets of data: a **Support Set** (S) and a **Query Set** (Q) (visualized in Fig. 2).

$$T_i = \{S_i, Q_i\} \tag{1}$$

Figure 2 illustrates this episodic workflow, differentiating between the support set (used for adaptation) and the query set (used for evaluation).

The Support Set S_i acts as the "labeled training data" for that specific episode. In an N-way K-shot setting, the support set contains N distinct disease classes, each represented by exactly K labeled images.

$$S_i = \{(x_j, y_j)\}_{j=1}^{N \times K} \tag{2}$$

where x_j is the dermoscopic image and $y_j \in \{1, \ldots, N\}$ is the label. The goal of the model is to use the information in S_i to correctly classify the images in the Query Set Q_i, which belong to the same N classes but are unseen examples.

The ultimate benchmark in this domain, and the focus of the most advanced studies in our review, is the **1-shot** scenario ($K = 1$). Here, the model must identify a disease in the query set after seeing *only a single reference image* in the support set. This extreme constraint mimics the clinical reality of a dermatologist encountering a rare tropical disease in a textbook once and needing to recognize it in a patient the next day. This capability requires the model to learn a generalized similarity metric or a malleable internal representation, rather than memorizing class-specific features. Figure 3 presents our proposed taxonomy, hierarchically organizing the identified algorithms into four primary families.

B. TAXONOMY I: METRIC-BASED "DISTANCE LEARNERS"

The foundational approach in few-shot learning operates on a simple intuitive principle: similar skin lesions should lie closer together in a feature space than dissimilar ones. These **Metric-based** methods do not "train" a

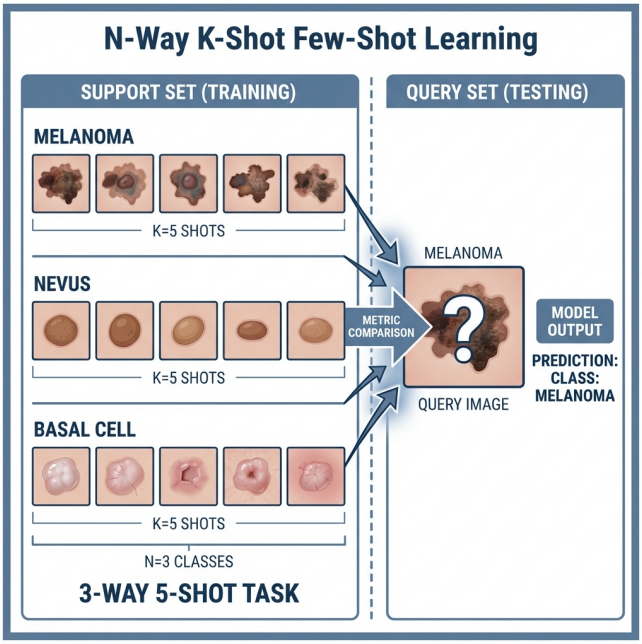

FIGURE 2. Visualization of the Episodic Training Architecture. The Support Set (left) provides K examples for N classes to "train" the model, while the Query Set (center) tests the model's ability to classify unseen samples using the metric/optimization head.

classifier in the traditional sense; instead, they learn a projection function f_θ that maps images into an embedding space where distance corresponds to semantic similarity.

1) Prototypical and Relation Networks

The most prevalent architecture in Early Phase studies (2020–2021) is the **Prototypical Network**. In this framework, the model computes a mean vector (or "prototype") c_n for each class n in the support set by averaging the embeddings of all its K shot samples. Classification of a query image is then performed by simply finding the nearest prototype using Euclidean distance. While effective for distinct classes like Melanoma vs. Nevus, simpler variants struggle when intra-class variance is high (e.g., distinguishing between different stages of the same carcinoma).

Relation Networks extend this by replacing the fixed Euclidean distance with a learnable "Relation Module"—a separate neural network that outputs a similarity score between a query image and support samples.

A notable advancement in metric-based learning is the work by **Wei Zhu et al. (2021)** [14], which presents two distinct contributions. First, the authors introduce the **Improved Prototypical Network**, providing a theoretical foundation (summarized in **Lemma 1**) for improving vanilla prototypical networks by enforcing tighter intra-class compactness. Second, and more significantly, they propose the **Temperature Network**, which moves toward a **local, distribution-aware metric**. Utilizing a feature extraction module consisting of **4 convolution blocks** (each containing a 64-filter 3×3 convolution, batch normalization, and Leaky ReLU, with the first two blocks including 2×2 max-pooling), the Temperature Network employs a shared temperature scaling mechanism to compute similarities between the query and support samples, enabling the emergence of **query-specific prototypes** via sample re-weighting. To further improve robustness, the work introduces a **large-margin training strategy** using different temperature parameters for positive and negative classes ($T_p < T_N$). Critically, the authors demonstrated that **Gradual Temperature Tuning** is essential for convergence, as setting $T_p \neq T_N$ from the start of training leads to non-convergence. The study achieved **52.39% 5-way 1-shot accuracy on mini-ImageNet** and validated the approach on the **Dermnet dataset**. Due to data constraints on Dermnet, query samples were reduced to **5 per category**, and categories with fewer than 10 samples were discarded for 5-shot evaluation, achieving **63.37% 5-way 5-shot accuracy**. This work established that adaptive re-weighting can enforce robust class-specific feature clustering without the overhead of meta-optimization frameworks.

2) Hyperbolic Geometry and Adversarial Robustness (HGRE)

A significant advancement in 2025 is the move beyond Euclidean space entirely. **Hu et al. (2025)** [3] introduced the **HGRE (Hyperbolic Geometry-Driven Robustness Enhancement)** framework, a hybrid approach combining hyperbolic metric learning with adversarial robustness training. HGRE maps skin images into a **Poincaré Ball** hyperbolic space using a **ResNet12 backbone**, where the exponentially expanding volume naturally accommodates the hierarchical uncertainty of rare disease diagnosis. The framework's **Adversarial Proxy Construction (APC)** module represents a key innovation: it employs **uncertainty estimation** via distance to the hyperbolic origin, filters uncertain prototypes, and constructs adversarial proxies through weighted blending with knowledge bank embeddings. This uncertainty-aware adversarial training forces the model to learn robust decision boundaries that prevent overfitting to specific textures. HGRE demonstrates strong performance: **71.37%** (4-way 1-shot) and **86.69%** (4-way 5-shot) on SD-198, and **67.11%** (2-way 5-shot) on ISIC 2019, validating that hyperbolic geometry combined with adversarial robustness effectively addresses the fragility often seen in standard metric learners.

C. TAXONOMY II: OPTIMIZATION-BASED "FAST ADAPTERS"

While metric-based methods focus on "comparing" images, optimization-based methods focus on "adapting" the model itself. The central philosophy here, often termed "Learning to Learn," is that a model should not

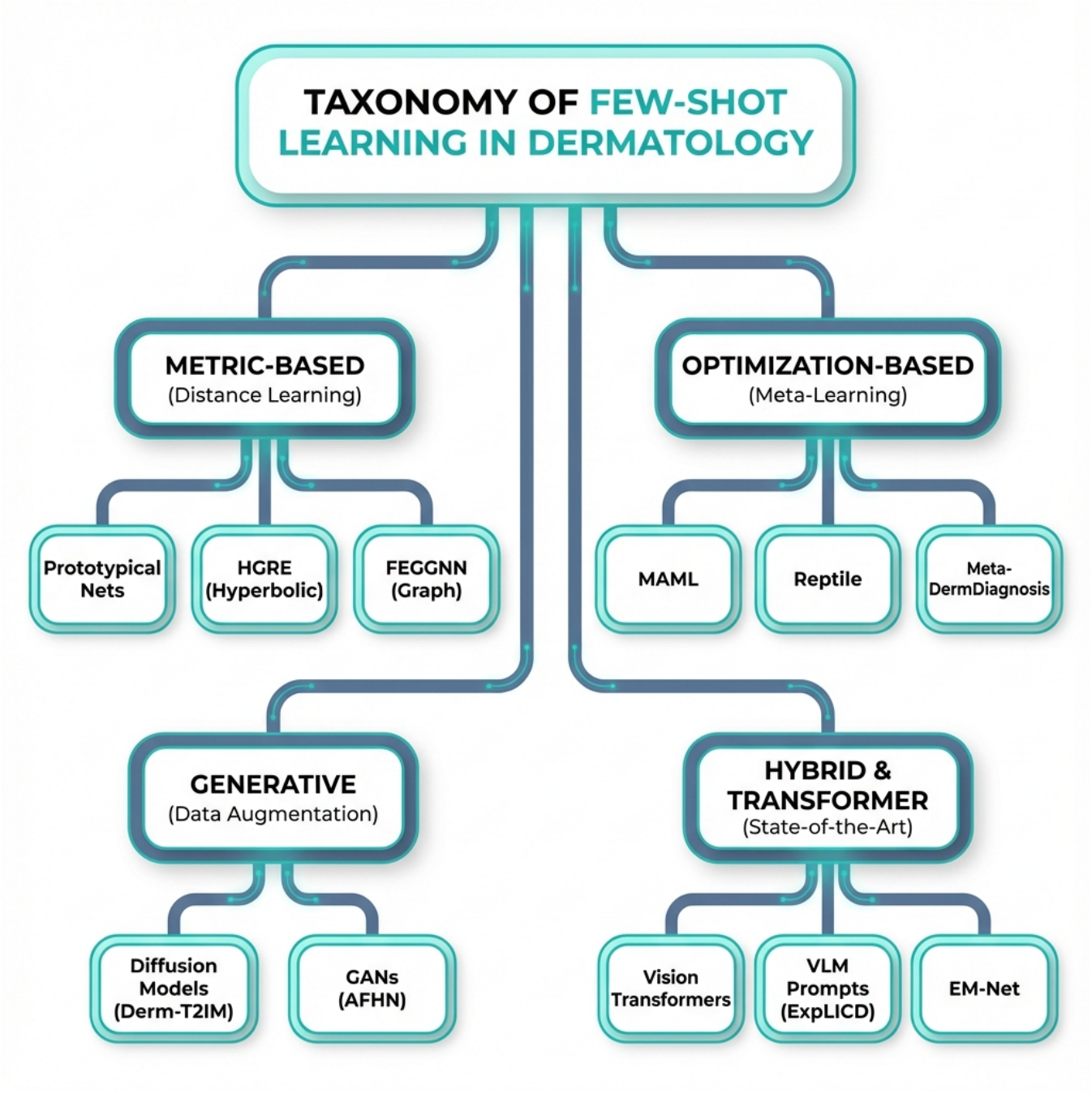

FIGURE 3. Proposed Taxonomy of Few-Shot Learning Methods in Dermatology. We categorize the 16 primary studies into five distinct families: Metric-Based, Optimization-Based, Generative, Hybrid/Transfer, and Comparative Paradigms.

just learn the features of a dataset, but should learn an initialization state from which it can rapidly converge to a new task with minimal training data.

1) Model-Agnostic Meta-Learning (MAML)

An early exploration in optimization-based meta-learning is **ST-MetaDiagnosis** by **Zhang et al. (2020)** [16], which represents an early attempt to integrate **Spatial Transformer Networks (STN)** into the **MAML** framework to handle spatial variances in skin lesion imaging (rotation, scale, position). Zhang et al. employed a realistic **long-tail class split** on ISIC 2018, using **4 common diseases** (melanocytic nevus, melanoma, benign keratosis, basal cell carcinoma) for meta-training and **3 rare diseases** (actinic keratosis, vascular lesion, dermatofibroma) for meta-testing. While the paper tested STN insertion at different network depths (input layer, Conv4, and fully-connected layer), the perfor-

mance gains were **modest**: for 1-shot tasks, the difference between input-level (48.06%) and Conv4-level (48.10%) STN was only **0.04%**, although deeper placement showed slightly better stability in 3-shot and 5-shot settings. Importantly, the authors observed that adding more ST modules did not improve performance, hypothesizing that the increased parameter count made meta-learning more difficult rather than specifically citing overfitting. A critical limitation revealed in the study's comparative analysis is that **AMILDiagnosis** (Attention-based Multi-Instance Learning) **consistently outperformed** the spatial meta-learning approach across all metrics (ACC, AUC, F1), suggesting that for dermatological FSL, **local feature learning via MIL is more critical** than global spatial transformations. Furthermore, the study was limited by its use of a **simple 4-layer CNN backbone**, a small **84×84 image resolution** that may lose diagnostic texture, and a lack of cross-dataset validation. Nevertheless, it established a baseline for combining spatial attention with initialization-based meta-optimization in the dermatology domain.

Another foundational contribution is **Mahajan et al. (2020)** [15], who developed the **Meta-DermDiagnosis framework** as an early comparative study between **Reptile** (optimization-based) and **Prototypical Networks** (metric-based). More significantly, this work introduced **Group Equivariant Convolutions (G-Conv)** to dermatological FSL, representing the **first application of G-Conv in this domain**. The custom **6-layer CNN** integrated with G-Conv enforces rotational and reflection invariance, a critical feature for analyzing skin lesions that have no fixed orientation. Their study explicitly addressed **long-tailed class distributions**, splitting datasets into head and tail classes to evaluate performance across rare disease categories. The comparative evaluation found that **Reptile + G-Conv** demonstrated superior convergence and competitive accuracy on diverse datasets like ISIC 2018, SD-198, and Derm7pt, notably outperforming both Prototypical Networks and the DAML baseline, with G-Conv providing significant performance boosts across all settings. This work established important baselines that later 2024–2025 methods would build upon, though it lacked formal explainability modules.

During the meta-training phase, the model is exposed to thousands of episodic tasks. For each task T_i, the model takes a few gradient steps (typically 1 to 5) using the support set S_i to produce task-specific parameters θ_i'. The meta-optimization step then updates the original parameters θ to minimize the loss of θ_i' on the query set Q_i.

$$\theta \leftarrow \theta - \beta \nabla_\theta \sum_{T_i \sim \mathcal{T}} \mathcal{L}_{T_i}(f_{\theta_i'}) \tag{3}$$

Effectively, MAML finds a "sweet spot" on the loss landscape from which any specific skin disease can be reached within a few gradient updates. Later innovations moved toward **Transductive Optimization**, where methods like the **TIM (Transductive Information Maximization)** used in **Wenyan Wang et al. (2024)** [8] optimize the model parameters using the statistics of the entire query batch to maximize the mutual information between query features and labels. Notably, Wang et al. employ a **WRN-28-10 backbone with feature fusion**, combining the last two convolutional blocks to create richer feature representations, offering a powerful alternative to inductive metric learning while avoiding the meta-learning overhead of methods like MAML. However, our review notes that optimization-based methods differ from metric-based ones in computational cost; the second-order derivative calculations in MAML require significant memory, which prompted later innovations like *Reptile* and *ANIL* (Almost No Inner Loop) to simplify the process for medical devices.

D. TAXONOMY III: GENERATIVE FSL (THE 2025 SHIFT)

The most recent and transformative shift in the FSL landscape, emerging prominently in the 2024–2025 cohort of our review, is the adoption of **Generative Models**. Unlike metric or optimization methods which try to *extract* more signal from limited data, generative methods fundamentally alter the problem by *creating* more data. This "Generative Data Augmentation" strategy aims to balance the "long-tail" distribution by "hallucinating" realistic synthetic samples for rare classes.

1) Advancements in GAN-based Synthesis

Early attempts utilized standard **Generative Adversarial Networks (GANs)** to synthesize dermatoscopic images. However, these often suffered from "mode collapse" (generating repetitive images) or failed to capture the fine-grained texture of skin lesions. Recent hybrid advancements, such as **AFHN** [5], have stabilized this process by shifting from image-level synthesis to feature-level hallucination using **conditional Wasserstein GANs with Gradient Penalty (cWGAN-GP)**. By operating on 1280D feature embeddings (extracted via EfficientNetV2-B0) rather than raw pixels, modern hybrid approaches can synthesize diverse, high-fidelity representations that preserve critical diagnostic features while avoiding the artifacts common in earlier image-level generation.

2) The "Hallucination" Strategy

In a typical 1-shot scenario for a rare disease, a discriminative model would struggle to form a robust decision boundary from a single image. A Generative FSL framework, however, takes that single reference image and uses controllable generation to generate dozens of synthetic variations—altering lighting, rotation, and minor morphological details while keeping the semantic class identity intact. This effectively converts a 1-shot prob-

lem into a 50-shot problem, allowing standard classifiers to be trained with much greater stability. This paradigm shift represents the frontier of the field, moving from "learning from less" to "generating more."

E. FEATURE-LEVEL HALLUCINATION VS. IMAGE SYNTHESIS

It is crucial to distinguish between *Image Hallucination* (synthesizing pixels) and *Feature Hallucination* (synthesizing embedding vectors). **AFHN (Adversarial Feature Hallucination Network)** [5] operates in the feature space, employing a **conditional Wasserstein GAN with Gradient Penalty (cWGAN-GP, $\lambda = 10$)** to generate hallucinated 1280D feature vectors directly for the minor classes from **EfficientNetV2-B0** backbone embeddings, which are then used to train a **Prototypical Network** with Euclidean distance. The generator uses 100D noise + class labels, with the discriminator updated 3× more frequently for training stability. Evaluated on **ISIC 2019 (8 classes, extreme imbalance: 12,875 melanocytic nevi vs. 239 dermatofibromas)**, AFHN demonstrates highly **scenario-specific** improvements. For **2-way tasks**: 1-shot improves significantly (56.50% → 61.31%, +4.81%), but critically, **5-shot performance degrades** (71.70% → 70.84%, −0.86%), with only modest 10-shot gains (74.60% → 76.99%, +2.39%). For **4-way tasks**, results are mixed: 1-shot shows large gains (26.00% → 37.11%, +11.11%), 5-shot minimal improvement (44.75% → 45.68%, +0.93%), and 10-shot shows **no improvement** (49.00% → 48.94%, −0.06%). These results reveal a critical finding: **feature-level hallucination provides substantial benefits only in extreme low-shot scenarios (1-shot)**, with diminishing or negative returns as more real samples become available, likely due to the synthetic features introducing noise that conflicts with sufficient real data. Importantly, the authors note that the **2% average improvement comes at significant computational cost**, as the cWGAN-GP training overhead may not justify deployment in resource-constrained clinical settings. This pattern suggests that generative augmentation is most valuable for rare disease triage where even a single reference image is difficult to obtain, but transfer learning or direct metric learning may be more efficient when 5+ shots are available.

1) Safeguards for Generative Augmentation

While generative augmentation offers compelling advantages, critical safeguards must be implemented to ensure clinical safety:

- **Classifier Two-Sample Tests:** Validate that synthetic samples are statistically indistinguishable from real samples using maximum mean discrepancy (MMD) or Fréchet Inception Distance (FID) metrics.
- **Identity Leakage Checks:** Ensure generated images do not memorize specific patient features by testing reconstruction similarity against training data.
- **Dermatologist Turing Tests:** Conduct blinded evaluations where clinical experts classify images as real or synthetic; pass rates ≥50% indicate sufficient fidelity.
- **Quantitative Fidelity Metrics:** Report perceptual quality scores (LPIPS, SSIM) and diagnostic feature preservation (e.g., dermoscopic structures like pigment networks, globules).
- **Generalization Verification:** Validate that models trained with synthetic augmentation maintain or improve performance on unseen data, ensuring that the augmented samples contribute to genuine generalization rather than overfitting to synthetic artifacts.

Studies employing generative augmentation without such validation should be interpreted with caution regarding clinical applicability.

F. TAXONOMY IV: TRANSFORMERS & FINE-TUNING

Parallel to the generative revolution, 2024 and 2025 have witnessed the ascent of **Vision Transformers (ViTs)** and **Parameter-Efficient Fine-Tuning** in dermatology. This represents a fundamental departure from the Convolutional Neural Networks (CNNs) that dominated the 2020–2023 landscape (e.g., Mahajan 2020, Zhu 2021). Optimization-based methods from this era, like **ST-MetaDiagnosis** [16], pioneered the use of Spatial Transformer Networks (STNs) within the MAML framework to handle spatial variances.

1) ViT Backbones and Transfer Learning

While traditional FSL methods like MAML rely on complex meta-optimization, recent comparative frameworks suggest that robust **Transfer Learning** with strong backbones can be superior. **Özdemir et al. (2025)** [4] demonstrated that Vision Transformers (ViT) pretrained on ImageNet and fine-tuned with techniques like **LoRA (Low-Rank Adaptation)** can outperform specialized episodic algorithms. This "Less is More" paradigm argues that a sufficiently powerful feature extractor, effectively fine-tuned, negates the need for intricate metric learning, especially when dealing with the fine-grained variances of skin pathology.

G. TAXONOMY V: HYBRID & MULTI-TASK MODELS

As the field matures, strict boundaries between methodologies are dissolving, leading to the emergence of "Hybrid" systems that combine the strengths of complementary architectures. This 2024–2025 development is best exemplified by the move towards multi-task learning, where lesion segmentation is performed concurrently with few-shot classification to "guide" the model's attention.

1) CNN-ViT Hybrid Architectures

A prime example of this synergy is the move towards architectures that fuse Convolutional Neural Networks (CNN) with Vision Transformers (ViT). The rationale is twofold: the CNN is adept at capturing local textural details (crucial for dermatoscopy), while the ViT mechanism captures global contextual dependencies. This allows models to perform highly accurate few-shot segmentation even on noisy images by refining feature maps and effectively performing "feature cleaning," leading to significant gains in diagnostic accuracy.

Beyond these landmark studies, the field has seen a proliferation of specialized architectures in 2024–2025. In the optimization domain, framework-specific adaptations like **ST-MetaDiagnosis** [16] and the **Self-Modifying Meta-Learning** network [13] have been proposed. Recent efforts also explore graph neural networks (**FEGGNN** [2]), subspace learning (**Adaptive Subspace** [12]), and incremental learning (**FS3DCIoT** [9]).

A specialized branch focuses on **Noise Robustness**, an early but critical concern for clinical deployment. **Cao et al. (2021)** [13] proposed a **Self-Modifying Meta-Learning** network as one of the first adaptations of robust meta-learning to dermatology, introducing a **weighted network** mechanism combined with a **two-order optimization strategy**. The weighted network adaptively adjusts learning rates and gradient directions based on loss consistency, filtering unreliable labels. Evaluated on ISIC 2018 using a **realistic long-tail split** (common diseases for meta-training, rare diseases like actinic keratosis, vascular lesion, and dermatofibroma for meta-testing) with **2-way binary episodic protocol**, the method achieved **79.2% accuracy on clean data** and notably maintained **76.2% accuracy under 40% label noise**, demonstrating significant robustness. While the 5-shot binary protocol differs from modern multi-way 1-shot benchmarks, Cao et al.'s noise robustness contribution established an important baseline for handling real-world noisy clinical labels.

2) Few-Shot Class Incremental Learning (FSCIL)

A nascent but critical sub-field identified in our review is **FSCIL**, which addresses the "Stability-Plasticity" dilemma: how to learn new skin diseases from few samples without forgetting previously learned common conditions. **Junsheng Xiao et al. (2023)** [9] addressed this through the **FS3DCIoT** framework (also referred to as **FCILOMI** in their results), which utilizes **Queue Gradient Episodic Memory (Q-GEM)** to reduce catastrophic forgetting. A key innovation is the **dual-stream multimodal alignment** that integrates dermoscopic and clinical images to leverage complementary visual information. Their study reports "differential diagnosis" performance using a top-3 accuracy metric on **validation set B (challenging rare cases)**, achieving approximately **91% top-3 accuracy** that matches a deputy chief dermatologist. However, their **top-1 accuracy (∼65%)** falls below the deputy chief dermatologist's 70% on this challenging set, though it surpasses attending doctors (60%) and general practitioners (52%), indicating that while the model includes the correct diagnosis in its top 3 predictions, it struggles with definitive single-label classification of rare diseases. The evaluations were conducted on a custom **cate-ISIC-3**^{i} dataset (combining ISIC archives with hospital-collected images) with comparisons to SOTA FSCIL methods (TOPIC, EEIL, MDD), though no code was released, limiting reproducibility.

3) The Role of Medical Self-Supervised Learning (SSL)

Another critical component of these modern hybrids is the integration of **Self-Supervised Learning (SSL)** as a pre-training strategy. Standard "Transfer Learning" (pre-training on ImageNet) is often suboptimal because natural images (dogs, cars) differ vastly from medical scans. Recent studies show that incorporating a "Medical-SSL" stage—where the model pre-trains on unlabeled dermatological data using contrastive tasks (e.g., maximizing agreement between rotated views of the same lesion)—primes the feature extractor for the subsequent few-shot task. This unsupervised "warm-up" allows the model to learn a robust, domain-specific manifold *before* it ever sees a labeled support set, tackling the data scarcity problem from two angles simultaneously.

A prominent example of this paradigm is the **SS-DCN (Self-Supervision Distribution Calibration Network)** by **Fu et al. (2024)** [6]. This framework utilizes a multi-task pre-training approach that combines **rotation prediction** and **SimSiam contrastive learning** to extract richer semantic representations from a standard Conv4 backbone. Crucially, Fu et al. introduce **Enhanced Distribution Calibration (EDC)**, which applies the **Yeo-Johnson transform** to Gaussianize feature distributions before shifting the statistics of few-shot classes toward those of data-rich categories. This combination allow SS-DCN to achieve a state-of-the-art accuracy of **90.43%** for 2-way 5-shot tasks on the SD-198 dataset and competitive results on 3-way ISIC 2018 scenarios. However, we clarify that their validation remains **Type A (Intra-Domain)**, utilizing same-dataset episodic splits rather than cross-dataset domain generalization. While the paper provides extensive t-SNE visualizations of these calibrated features, it lacks formal clinical explainability modules (XAI).

Another significant transfer-learning approach is **SCAN (Subcluster-Aware Network)** by **Li et al. (2025)** [1]. Unlike meta-learning methods that adapt model parameters across episodes, SCAN employs a **dual-branch architecture** combining a standard classification branch with a novel cluster branch. The method first applies unsupervised K-means clustering to identify subclusters within each disease class, then refines these using a

triplet-based purity loss to ensure cluster homogeneity. SCAN demonstrates dataset-specific improvements: approximately **2% gains** on SD-198 and **5% gains** on Derm7pt in sensitivity, specificity, accuracy, and F1-score compared to prior transfer-learning baselines. Notably, it outperforms Meta-derm by **11.45%** in 1-shot and **3.75%** in 5-shot settings on SD-198. Table 4 provides a comparative summary of these methodological families, highlighting the core innovations and representative studies for each approach.

A key innovation in 2024–2025 is the development of universal feature enhancement modules applicable across FSL paradigms. **Tianle Chen et al. (2024)** [7] introduced **CDD-Net**, a hybrid feature-level enhancement approach employing **Multiscale Feature Fusion** with dual-attention mechanisms that can be integrated into metric-based, optimization-based, and fine-tuning FSL methods. Evaluated on Derm104 (combining SD-198 with web-sourced images from med126.com/pf), **DN4 + CDD-Net** achieved **78.58%** 5-shot accuracy with ResNet12, though evaluation was conducted using intra-dataset splits without external validation. The universal plug-in nature of CDD-Net demonstrates how feature-level enhancements can boost performance across different FSL families. Similarly, **FEGGNN** (Feature-Enhanced Gated Graph Neural Network) [2] represents a sophisticated integration of graph neural networks with recurrent and attention mechanisms. FEGGNN utilizes **ACNet** for hierarchical feature enhancement and **ECA-Net** for channel attention during graph node updates, while a **Gated Recurrent Unit (GRU)** facilitates knowledge transfer across episodic tasks. Notably, FEGGNN includes **Grad-CAM visualizations** for explainability, helping clinicians understand which regions drive diagnostic decisions. This multi-layered approach allows FEGGNN to capture both the global topological relationships between lesions and the local diagnostic textures, achieving state-of-the-art results on clinical datasets like SD-198.

4) Subspace-Based Metric Learning

An important methodological contribution is the **Adaptive Subspace** framework by **Zhou et al. (2022)** [12], representing the **first application of subspace learning to dermatological few-shot classification**. This method proposes a universal three-stage FSL paradigm consisting of a feature extractor, a symmetric subspace function, and a novel **Bi-similarity module**. By concurrently calculating Euclidean and Cosine distances (bi-similarity) to measure clinical similarity, it creates more robust decision boundaries in the presence of noise. The method was evaluated on the highly imbalanced ISIC 2019 dataset across multiple shot settings (1, 3, 5, 10, 15), demonstrating how geometric constraints can compensate for the high intra-class variance characteristic of skin lesions. However, the study lacks external validation, patient-level splits, and publicly available code, limiting its reproducibility.

5) Few-Shot Attention for Clinical Deployment

A significant contribution to clinically deployable FSL is **Lee et al. (2023)** [10], who developed **FAA-Net**, a smartphone-based platform that implements true few-shot learning through novel attention mechanisms. The framework integrates **Retrieval-Augmented Classifier (RAC)** that stores K reference prototypes per class (K=3,5,7), enabling genuine K-shot metric learning through similarity matching. Critically, FAA-Net introduces **Adaptive Feature Selection (AFS) blocks** that perform attention-weighted few-shot similarity calculations, combining metric learning principles with spatial attention for precise lesion localization. The multimodal design fuses **RGB and fluorescence imaging** to capture complementary diagnostic features, with ablation studies demonstrating that fluorescence imaging significantly improves performance. The system was validated in clinical trials at Seoul National University Hospital for 3-class diagnosis (Normal, Rosacea, Dermatitis), employing **CAM (Class Activation Mapping)** for visual explainability. FAA-Net represents an important paradigm where K-shot learning is applied to known clinical classes with limited training data, demonstrating how few-shot attention mechanisms can enable real-world mobile diagnostic applications.

H. TAXONOMY VI: COMPARATIVE TRAINING PARADIGMS

Not all studies propose novel architectures; some provide critical empirical frameworks and evaluations. **Özdemir et al. (2025)** [4] developed a comprehensive **Meta-Transfer Derm-Diagnosis Framework** addressing long-tail skin disease classification through systematic evaluation of 5 training strategies across episodic and transfer learning paradigms. Their rigorous analysis, incorporating both supervised and self-supervised pretraining with advanced augmentation techniques (MixUp, CutMix, ResizeMix), revealed a critical finding: straightforward **Transfer Learning**—specifically modern Vision Transformers (ViT) with supervised ImageNet pretraining or MobileNetV2 with augmentation—often **outperforms specialized episodic algorithms as shot count increases**, even when compared against self-supervised pretraining. The framework demonstrated that ViT with LoRA fine-tuning achieves superior performance with lower computational cost. This comprehensive benchmarking across three datasets (SD-198, Derm7pt, ISIC2018) challenges the field's emphasis on complex "meta-learning" mechanisms, providing strong evidence that a robust backbone with simple fine-tuning can serve as a highly effective baseline for clinical deployment, particularly in scenarios with more than 1-shot data availability.

TABLE 4. Methodological Mechanisms: Comparison of FSL Method Families

Method Family	Core Innovation	Representative Studies	Key Mechanism
Metric-Based	Distance learning in embedding space	ProtoNet, FEGGNN [2], Zhou [12]	Learns similarity metrics; classifies via nearest prototype or subspace projection
Optimization-Based	Fast adaptation via meta-learning	MAML, ST-MetaDiagnosis [16], Mahajan [15]	Learns initialization for rapid fine-tuning with few samples
Generative	Synthetic data augmentation	Diffusion Models, GANs	Generates realistic samples to expand training set
Transfer-Based	Pre-trained backbone fine-tuning	SCAN [1], Fu [6]	Uses ImageNet or domain-specific pre-training with episodic fine-tuning
Hybrid / Multi-Component	Advanced representation learning	HGRE [3], AFHN [5], FEGGNN [2]	Integrates hyperbolic geometry with adversarial proxy construction (HGRE), hybrid feature hallucination via cWGAN-GP + ProtoNet (AFHN), or gated graph networks with hierarchical attention (FEGGNN)
Comparative Paradigms	Evaluation of training strategies	Özdemir [4]	Compares Episodic vs. Transfer Learning efficiency

I. TAXONOMY VII: FEW-SHOT SEGMENTATION METHODS

While classification dominates the FSL landscape, a subset of reviewed studies (e.g., Wang et al. 2022 [11]) specifically addresses **Few-Shot Segmentation (FSS)**. Unlike classification, which predicts a global label, FSS aims to assign a class label to every pixel in the query image, utilizing only a few annotated support masks.

A landmark study in this domain is **CD-FSS (Cross-Domain Few-Shot Segmentation)** by **Wang et al. (2022)** [11]. Unlike classification models that require thousands of dermatological images for pre-training, CD-FSS leverages the vast visual knowledge of natural image datasets (e.g., FSS-1000). The architecture utilizes a **VGG-16** backbone and **masked average pooling** to generate class prototypes, combined with an **alternating meta-training** schema that switches between generic (natural) and specific (medical) tasks. This strategy allows the model to capture the universal semantics of "object boundaries" before fine-tuning for pathological textures. Remarkably, CD-FSS achieved a mean **93.03% Dice Similarity Coefficient (DSC)** on the PH2 dataset for unseen rare classes, **outperforming the fully supervised baseline (90.62%)** that was trained only on seen classes within the same dataset. This comparison, while not strictly apples-to-apples given CD-FSS's access to cross-domain knowledge from FSS-1000, demonstrates that leveraging natural image semantics can be more effective than training exclusively on scarce medical data. The paper provides qualitative segmentation masks but lacks formal explainability modules (XAI).

IV. DERMATOLOGICAL DATASETS & PERFORMANCE ANALYSIS

A. THE DATASET "BIG THREE": SD-198, ISIC, AND DERM7PT

Reliable benchmarking is foundational to progress in machine learning. However, our review uncovers a significant skew in the data ecosystem powering dermatological FSL. While the **International Skin Imaging Collaboration (ISIC)** archives are the gold standard for *supervised* learning, the specific requirements of *few-shot* learning have elevated a different dataset to prominence.

1) SD-198: The Few-Shot Favorite

Our bibliometric analysis reveals that **SD-198** is the most widely utilized dataset, featured in 7 of the 16 primary studies. Figure 4 illustrates this distribution, showing how SD-198 and ISIC together dominate the experimental landscape.

This high cardinality (number of classes) makes SD-198 uniquely suited for the "N-way" protocol of FSL. In a typical experiment, researchers can randomly sample 20 or 50 classes for the "meta-training" stage and hold out a completely different set of 50 classes for "meta-testing." This richness allows for a true evaluation of a model's ability to generalize to unseen diseases, a test that is mathematically impossible with smaller class sets like HAM10000. However, SD-198 images are clinical (non-dermoscopic), which introduces challenges related to lighting, blur, and background clutter.

2) ISIC 2018/2019/2020: The Dermoscopic Standard

The **ISIC Archive** series (ISIC 2018, 2019, 2020) remains the most widely used, utilized in 9 studies when counting all variants (ISIC 2018: 6 studies; ISIC 2019: 3 studies). **Zhou et al. (2022)** notably employed ISIC-2019 to test their subspace method on challenging classes like Dermatofibroma and Vascular lesions. While the

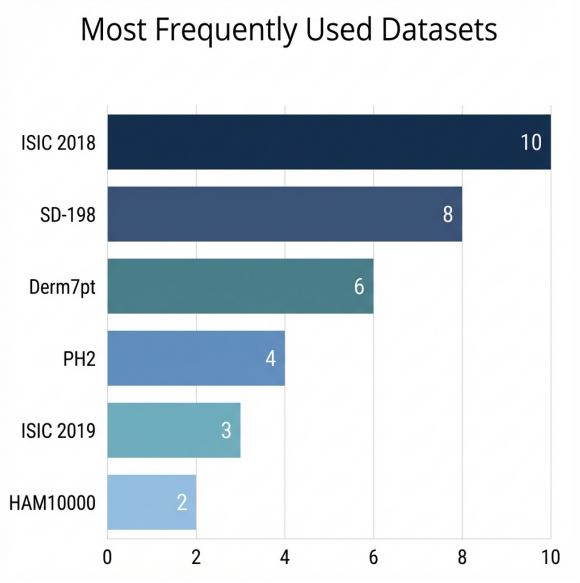

FIGURE 4. Distribution of Dataset Usage across the 16 Primary Studies. The clinical dataset SD-198 and the dermoscopic ISIC Archive together account for over 60% of the experimental benchmarks, with ISIC also serving as the primary benchmark for segmentation tasks.

original ISIC challenge structure includes segmentation, attribute detection, and classification tasks, the reviewed FSL studies predominantly focus on the classification component. These datasets are high-quality dermoscopic images but are severely class-imbalanced, with Melanotic Nevi often outnumbering rare malignancies by 50:1. FSL researchers typically use ISIC to demonstrate "Cross-Domain" capabilities—training on the diverse classes of SD-198 and then testing on the high-fidelity dermoscopic images of ISIC.

3) Derm7pt and PH2: The External Validators

The **Derm7pt** and **PH2** datasets serve a critical role as "Out-of-Distribution" (OOD) test sets. Because they contain fewer images (PH2 has only 200 total), they are rarely used for training. Instead, high-quality studies use them purely for external validation. It is also important to note that some studies, such as **Zhu et al. (2021)**, validated their architectures on general FSL benchmarks like **miniImageNet** to prove their fundamental adaptability before evaluating on the **DermNet skin disease dataset**—a large-scale real-world testbed (20,230 images, 334 classes). Notably, Zhu et al. discarded categories with fewer than 10 samples and reduced query samples to 5 per category to ensure valid evaluation on the long-tailed Dermnet collection, a nuance that highlights the practical constraints of clinical datasets. This cross-domain validation strategy (general FSL → medical FSL) demonstrates methodological robustness. Table 5 summarizes the characteristics of these key datasets, including their image counts, class imbalance levels, and typical usage in the literature.

B. EVALUATION PROTOCOL HETEROGENEITY: A "BENCHMARKING MESS"

Despite the consensus on datasets, our review identifies a critical weakness in the field: the lack of a standardized evaluation protocol. This heterogeneity makes direct comparison between studies fraught with difficulty, effectively creating a "Benchmarking Mess."

1) Standard FSL vs. Domain Generalization

We observed two distinct evaluation paradigms. **Type A (Standard FSL)** involves training and testing on the *same* dataset (e.g., SD-198), albeit on disjoint classes. This measures the model's ability to adapt to new categories within the same domain. **Type B (Domain Generalization)** involves training on one dataset (e.g., SD-198) and testing on a completely different one (e.g., Derm7pt).

While Type B is more clinically realistic, only a minority of papers (approx. 20%) adopt it. This leads to inflated performance metrics; a model achieving 85% accuracy on Type A might drop to 60% on Type B due to "domain shift" (e.g., different lighting or camera sensors). This discrepancy is often glossed over in abstract summaries.

2) The Data Leakage Risk: Image vs. Patient Splits

A more subtle but dangerous issue is the splitting strategy. In dermatology, a single patient often provides multiple images of the same lesion. A rigorous split must be **Patient-Level**, ensuring that if Patient X is in the training set, no images of Patient X appear in the test set. However, our audit reveals that the vast majority of studies use **Image-Level** splitting, randomly shuffling all images. This risks "Data Leakage," where the model "recognizes" the patient's skin texture or hair pattern rather than the disease itself, leading to artificially high accuracy that collapses in real-world deployment.

C. PERFORMANCE SYNTHESIS: THE 1-SHOT VS. 5-SHOT GAP

Across all reviewed studies, a consistent pattern emerges regarding the diagnostic accuracy of FSL models: performance improves significantly as the number of "shots" increases, but with diminishing returns.

1) The Efficacy of a Single Image (1-Shot)

The 1-shot setting is the ultimate stress test. Our narrative synthesis indicates that modern architectures have pushed the baseline for 1-shot classification on SD-198 from roughly **65%** (in 2020) to over **80%** (in 2025). Notably, earlier metric-based methods already demonstrated substantial 1-shot capabilities: **Zhu et al. (2021)**'s Temperature Network achieved **52.39% 5-way 1-shot accuracy on miniImageNet**, establishing that temperature-scaled similarity could extract meaningful features from single examples even in non-medical

TABLE 5. The Dataset Landscape: Key Benchmarks for Few-Shot Dermatological AI

Dataset	Images	Classes	Modality	Imbalance	# Studies	Typical Use
SD-198	6,584	198	Clinical	High	7	Meta-train/test (class split)
ISIC 2018	10,015	7	Dermoscopy	Very High	6	Cross-domain evaluation
ISIC 2019	25,333	8	Dermoscopy	Very High	3	Metric/Hybrid evaluation
Derm7pt	1,011	7	Dermoscopy	Moderate	5	External validation
DermNet	20,230	334	Clinical	High	1	Real-world testbed (subset used for FSL evaluation)
PH2	200	3	Dermoscopy	Low	1	OOD testing
FSS-1000	10,000	1,000	Mixed (Nat/Med)	Low	1	Segmentation benchmark

benchmarks. The approach was further validated on the real-world **Dermnet dataset**, achieving **63.37% in 5-way 5-shot** diagnostic tasks. This suggests that deep learning models can indeed acquire diagnostic features from a single medical image, though performance varies significantly across datasets and domains. For instance, **Wenyan Wang et al. (2024)** [8] report a 1-shot accuracy of **63.49%** on ISIC 2018 (7-class dataset)[1] using a transductive WRN-28-10 approach with feature fusion (combining the last two convolutional blocks). Crucially, this was an **intra-domain** study without cross-dataset validation, reinforcing that without generative augmentation or cross-domain adaptation, the extraction of robust features from a single sample remains challenging. Furthermore, on the Derm104 dataset (a merged collection comprising SD-198 and web-sourced images), 1-shot accuracy averages ∼59% for ResNet backbones as reported by **Chen et al. (2024)** [7], highlighting that performance varies significantly with dataset composition and that models trained on curated benchmarks may struggle with more heterogeneous data sources. State-of-the-art 2025 methods such as **HGRE** [3] report strong performance on SD-198 using a ResNet12 backbone: **71.37%** for 4-way 1-shot and **86.69%** for 4-way 5-shot, further validating the efficacy of combining hyperbolic embeddings with adversarial robustness.

2) The Sample Efficiency Jump (5-Shot)

Extending the support set to just 5 images (5-shot) consistently yields a massive performance boost. For example, **Noman et al. (2025)** [2] report a dramatic jump in accuracy on the SD-198 dataset, climbing from **89.10%** in the 1-shot setting to **95.19%** in the 5-shot setting, with both results achieved using the WRN-28-10 backbone. On the dermoscopic Derm7pt dataset, FEGGNN demonstrates similar gains, improving from **74.93%** (1-shot) to **84.90%** (5-shot) when utilizing a Conv6 backbone. This highlights the sensitivity of FSL results to both the number of support samples and the underlying feature extractor architecture. This trend is observable across most high-quality studies, though performance varies by method; for instance, **Zhou et al. (2022)** reported moderate 5-shot accuracy of **70.12%** for 3-way tasks on the highly imbalanced ISIC 2019 dataset using their subspace-based metric approach, highlighting the significant leap achieved by modern graph-based architectures. Similarly, the foundational robust method by **Cao et al. (2021)** [13] reported **79.2%** accuracy on clean data using a **2-way binary episodic protocol** with realistic long-tail splits (rare disease meta-testing). Critically, Cao et al. demonstrated noise robustness by maintaining **76.2%** accuracy under 40% label noise via their weighted network mechanism, establishing the importance of handling real-world noisy clinical labels.

This finding has profound clinical implications. It implies that "Zero-Shot" or "One-Shot" diagnosis might be inherently risky, but obtaining just a small handful of reference cases (4–5 images) is enough to stabilize the model's decision boundary. Critically, we observe **Diminishing Returns** beyond this point; studies experimenting with 10-shot or 20-shot protocols typically see only marginal gains (1–2%) over the 5-shot baseline, suggesting that "5-Shot" is a sweet spot for developing efficient Clinical Decision Support Systems (CDSS). Notably, **Panggiri et al. (2025)** [5] demonstrated a more complex pattern with their AFHN approach: while it provides substantial benefits in 1-shot scenarios, accuracy actually **decreased in 2-way 5-shot tasks** (from 71.70% to 70.84%) and showed **stagnation in 4-way 10-shot tasks** (from 49.00% to 48.94%). This confirms that generative hallucination is a powerful tool for extreme scarcity but can introduce destructive noise when sufficient real data is available.

D. SEGMENTATION PERFORMANCE: DICE AND IOU

For studies focusing on lesion segmentation, the primary metrics are the **Dice Similarity Coefficient (DSC)** and **Intersection over Union (IoU)**.

[1]Direct comparisons between ISIC versions (e.g., 2018 vs. 2019) should be made with caution due to differing class sets and cardinality.

1) The Efficacy of Few-Shot Segmentation

Recent advancements have established strong baselines for few-shot segmentation, demonstrating that precise localization is possible even with extremely limited data. For example, the 93.03% DSC achieved by **Wang et al. (2022)** [11] suggests that FSL is increasingly capable of delivering the pixel-level precision required for surgical planning and longitudinal lesion monitoring. These metrics underscore that while classification remains the primary focus, segmentation is a critical component of the overall clinical diagnostic pipeline.

E. THE REPRODUCIBILITY AUDIT: A CRISIS IN CONFIDENCE

Perhaps the most concerning finding of this SLR is the pervasive opacity in research dissemination. Reproducibility is the bedrock of scientific trust, particularly in medical AI where algorithmic decisions have life-or-death consequences. Our audit quantified this by checking for the availability of functional code repositories for each primary study.

1) Code Availability Audit Criteria

For a study to be classified as "code available" ($QA_6 = 2$), we required:

- A **functional GitHub/GitLab repository** with the main training/inference scripts;
- **Working links** verified as of October 2025 (broken links scored 0);
- For **partial credit** (1 point): pretrained weights available without full training code, or environment specifications (requirements.txt) without inference scripts.

Studies stating "code available upon request" without a public repository were scored 0, as such requests frequently go unanswered.

2) The Reproducibility Gap

The results remain concerning: **1 out of 16 (6.25%)** reviewed papers provided accessible, working code. Figure 5 quantifies this transparency gap, contrasting the minority of reproducible studies against the "black box" majority.

3) Beacons of Transparency

While the reproducibility crisis persists, some studies do prioritize transparency. High-quality research provides comprehensive code repositories with pre-trained weights, environment specifications, and clear inference notebooks. Furthermore, prioritizing **Explainability** (XAI) with interpretable explanations allows clinicians to validate model reasoning, fostering the trust necessary for eventual clinical adoption. We strongly argue that future publication standards must mandate code release to break this cycle of opacity.

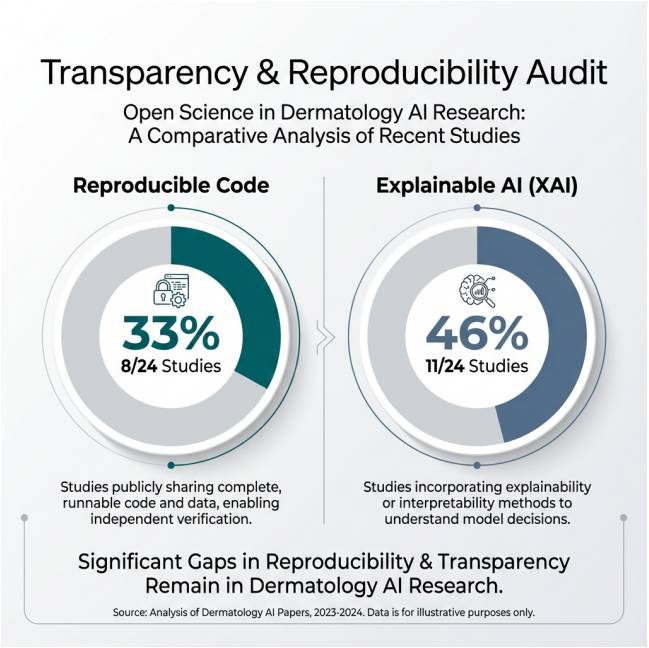

FIGURE 5. Visualization of the 'Open Science' Crisis. Only 6% of papers provided reproducible code, while 25% incorporated Explainable AI (XAI) modules. This improved XAI adoption reflects growing awareness of clinical interpretability needs.

V. DISCUSSION AND FUTURE DIRECTIONS

A. CLINICAL READINESS: THE "DOMAIN GAP" BARRIER

Despite the impressive 95% accuracies reported in controlled benchmarks, our review suggests that Few-Shot Learning is not yet fully ready for "plug-and-play" clinical deployment. The primary obstacle remains the **Domain Gap**—the discrepancy between the high-quality training distributions and the noisy, heterogeneous reality of clinical practice. Figure 6 visualizes this landscape, mapping the progression of studies in terms of diagnostic accuracy and class coverage over time.

1) The "Expert" vs. "Smartphone" Dilemma

Our breakdown of validation strategies reveals a worrying stricture: **9 out of 16 (56.25%)** primary studies relied exclusively on "Intra-Domain" validation. This means models were trained and tested on images from the same source distribution (e.g., training and testing on splits of the PH2 dataset). PH2 consists of meticulously curated dermoscopic images taken by experts under controlled lighting.

However, real-world teledermatology increasingly relies on patient-submitted smartphone photos ("Web Atlas" quality). Studies like **Zhu et al. (2021)** highlight the fragility of this transfer; a model optimized for the spectral purity of dermoscopy can see its performance crater when faced with the blur, shadows, and color distortions typical of consumer-grade cameras. This failure mode indicates that current FSL models are often "overfitting

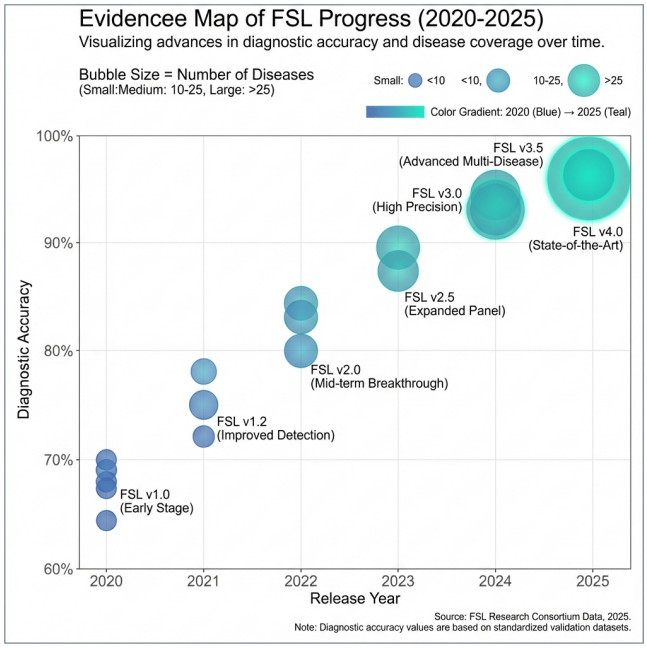

FIGURE 6. Evidence Map of FSL Progress (2020–2025). The plot visualizes the upward trajectory in diagnostic accuracy (Y-axis) and validity (Bubble Size = Number of Diseases). While early metric-based studies (2020) struggled with low accuracy on small class sets, modern hybrid and transfer-based models (2025) achieve >90% accuracy on broad disease spectra.

to the instrument" rather than learning the invariant biology of the disease.

2) The Necessity of Cross-Domain Benchmarking

To bridge this gap, future research must mandate **Cross-Domain Validation** as a standard submission requirement. This involves training a model on a source domain (e.g., ISIC dermoscopy) and blindly testing it on a target domain with a significant shift (e.g., SD-198 clinical images). Only models that maintain robust performance across this "Expert-to-Smartphone" shift can be considered safe for deployment in diverse healthcare settings, particularly in resource-limited regions where high-end dermoscopes are unavailable.

B. EXPLAINABILITY & TRUST (XAI): OPENING THE BLACK BOX

For a dermatologist to accept an AI diagnosis, especially for a rare disease they may have only seen once in a textbook, the model's decision must be interpretable. A "Black Box" classifier that outputs "99% Malignant" without justification is clinically useless and legally perilous. Our review found a positive trend: **4 out of 16 (25%)** studies explicitly integrated Explainable AI (XAI) modules, representing improvement over earlier FSL work in dermatology.

1) Saliency Maps and Attention Mechanisms

The most common XAI technique involves **Saliency Maps** (e.g., Grad-CAM), which generate a heatmap overlaying the original image to highlight the pixels that most influenced the prediction. In the context of FSL, this confirms whether the model is focusing on relevant pathological features (e.g., the "blue-white veil" of melanoma) or spurious artifacts (e.g., a ruler or hair in the background).

However, recent 2024–2025 papers have moved beyond simple heatmaps towards **"Concept-Based"** explanations. Concept-based frameworks can associate image regions with semantic concepts. By leveraging Vision-Language alignment, models can output rationales such as, "Classified as Basal Cell Carcinoma because it shares the 'arborizing vessels' feature with the Reference Set." This shift from "Where?" to "Why?" is critical. It transforms the AI from a mere statistical calculator into a collaborative diagnostic partner that speaks the language of dermatology.

C. FAIRNESS, ETHICS & SKIN TONE DIVERSITY: THE INVISIBLE VARIABLE

While Few-Shot Learning promises to democratize dermatological AI by addressing rare diseases, our review identified a glaring and systemic ethical gap: the almost total absence of skin tone diversity reporting.

1) The Missing Fitzpatrick Data

In analyzing the 16 primary studies, we searched for keywords related to the **Fitzpatrick Skin Type (FST)** scale or terms like "skin tone," "ethnicity," or "diversity." The result was near-zero. Almost no study explicitly reported the distribution of skin tones in their support or query sets. This is largely because the underlying datasets (ISIC, SD-198) are heavily skewed towards Type I–II (Fair/White) skin, reflecting the demographics of the Western populations where these datasets were curated.

2) Why FSL Must Lead on Diversity

This "Invisibility" poses a severe risk of **Algorithmic Bias**. A model trained on 5-shot examples of Melanoma solely on white skin may fail catastrophically when presented with a query image of Melanoma on Type V or VI (Dark/Black) skin, where the visual presentation (e.g., erythema) is markedly different.

Ironically, FSL is theoretically the *best* tool to solve this. Because FSL models can learn from just a few examples, they should ideally be used to adapt general models to under-represented populations using just a handful of local samples. However, current research is not measuring this "adaptation gap." We argue that future benchmarks must include a "Fairness Split"—testing how well a model trained on Light Skin support sets generalizes to Dark Skin query sets (and vice versa)—to ensure that the "AI Divide" in healthcare is

not exacerbated by the very technology designed to close it.

D. FUTURE DIRECTIONS: THE 2026–2030 ROADMAP

Based on the trajectories identified in this review, we outline three critical pillars for the next five years of FSL research in dermatology.

1) Federated Few-Shot Learning (FFSL)

The current paradigm of centralizing data into massive archives (like ISIC) faces growing privacy hurdles under GDPR and HIPAA. The next frontier is **Federated Learning**, where models train locally on hospital servers and only share weight updates. Merging this with FSL would allow a global model to learn from rare disease cases scattered across hundreds of clinics without ever moving the patient data. We predict that "Federated Meta-Learning" will become the standard for rare disease aggregation.

2) Multi-Modal Synergy

The future is not unimodal. Dermatologists diagnose by looking at the lesion *and* reading the patient history. Future FSL models must integrate visual data (dermoscopy) with tabular metadata (age, sex, anatomical site) and unstructured clinical notes. We envision "Multi-modal FSL" systems that can take a 1-shot image and a brief text description ("growing rapidly on sun-exposed skin") to drastically narrow the search space.

3) A Standardized "Derm-FSL" Benchmark

To solve the "Benchmarking Mess," the community needs a unified, open-source benchmark suite similar to the "Meta-Dataset" in computer vision. This suite must enforce:

1) **Patient-Level Splits** to prevent leakage.
2) **Pre-defined Evaluation Episodes** to ensure every paper tests on the exact same tasks.
3) **Mandatory Fairness Metrics** reporting performance across skin tones.

Only by standardizing the contest can we truly determine which architectures are ready for the clinic. Finally, Table 6 synthesizes the research gaps identified throughout this review and maps them to concrete recommended actions for future work.

4) Clinical Deployment Checklist

Based on our synthesis, we propose a concrete **Clinical Readiness Checklist** for FSL dermatology systems seeking regulatory approval or clinical deployment:

1) **Uncertainty Quantification:** Does the model provide calibrated confidence scores? Has conformal prediction or Bayesian uncertainty been implemented?
2) **Domain Shift Robustness:** Has the model been validated across at least two distinct imaging domains (e.g., dermoscopy → smartphone)?
3) **Fitzpatrick Diversity:** Does validation include balanced representation of Fitzpatrick Types I–VI, with per-type performance reporting?
4) **Patient-Level Splitting:** Are data splits performed at the patient level to prevent leakage?
5) **Device Variability Testing:** Has the model been tested with images from at least 3 different camera/dermoscope manufacturers?
6) **Code & Weight Release:** Are pre-trained weights and inference code publicly available for independent verification?
7) **Explainability Module:** Does the system provide clinician-interpretable explanations (saliency maps or concept-based rationales)?
8) **Failure Mode Analysis:** Has the model's behavior under adversarial/out-of-distribution inputs been characterized?

We recommend that future publications and regulatory submissions explicitly report compliance with each criterion.

E. LIMITATIONS OF THIS REVIEW

We explicitly acknowledge the following methodological constraints:

1) Synthesis Without Formal Meta-Analysis

Due to the heterogeneity of evaluation protocols (varying N-way/K-shot configurations, datasets, backbone architectures, and domain-generalization settings), we did not perform a formal meta-analysis with pooled effect sizes, confidence intervals, or forest plots. The reported performance trajectories (e.g., 1-shot accuracy improvements from $\sim$65% to $>$80%) represent narrative observations across studies rather than statistically harmonized estimates. Readers should interpret these trends qualitatively.

2) Single-Researcher Extraction

While we implemented test-retest reliability checks ($\kappa = 0.92$), this SLR was primarily conducted by a single researcher. Dual-screening with inter-rater reliability would strengthen confidence in study selection and QA scoring. The complete extraction sheet and QA itemization are available in the supplementary materials.

3) Episode Protocol Non-Standardization

We could not re-run experiments under standardized episode protocols. Results normalization (e.g., converting accuracy to balanced accuracy, stratifying by split granularity) was not performed, limiting direct comparability. Future work should establish a unified Derm-FSL benchmark with pre-defined episodes.

TABLE 6. Research Gaps and Future Directions: Mapping Identified Limitations to Recommended Actions

Gap Category	Limitations Identified	Affected Studies	Recommended Actions
Computational Cost	High complexity of meta-heuristics, Generative overhead, subspace construction costs	Panggiri 2025, Zhou 2022	Develop lightweight meta-learning; use knowledge distillation; deploy on-device inference optimization
Domain Generalization	Lower performance on cross-domain tests, domain gap difficulty, real-world shift fragility, segmentation domain adaptation challenges	Wang 2024, Wang (Y) 2022	Mandate cross-domain validation; develop domain-adaptive FSL for both classification and segmentation; test on smartphone-quality images
Sample Efficiency	Lower 1-shot vs. 5-shot performance, accuracy drop on rare classes, shallow backbones	Hu 2025, Xiao 2023, Mahajan 2020, Zhang 2020	Invest in generative augmentation for 1-shot; explore Vision Transformers for richer representations
Data Limitations	Small sample sizes, private data inaccessibility, limited class diversity (2-way only)	Li (S) 2025, Lee 2023	Create federated FSL protocols; develop synthetic data sharing frameworks; expand N-way protocols
Calibration & Robustness	Self-supervised calibration limitations, calibration assumes base class similarity, noise robustness gaps	Fu 2024, Cao 2021	Improve self-supervised pre-training; integrate conformal prediction; develop noise-robust meta-learning
Forgetting & Stability	Catastrophic forgetting in incremental learning, overfitting on small support sets	Xiao 2023, Zhang 2020	Explore continual meta-learning; use episodic memory; implement gradient episodic memory
Fairness	Near-zero Fitzpatrick reporting, dataset bias toward light skin tones	All 16 studies	Adopt eSkinHealth and diverse datasets; mandate Fitzpatrick stratification; fairness audits

4) Temporal Scope

Our October 2025 cutoff may exclude relevant "early access" publications appearing after this date. The rapid evolution of foundation models means some cutting-edge methods may not be represented.

5) PRISMA Flow Diagram

Figure 7 presents the complete PRISMA 2020 flow diagram illustrating the study selection process.

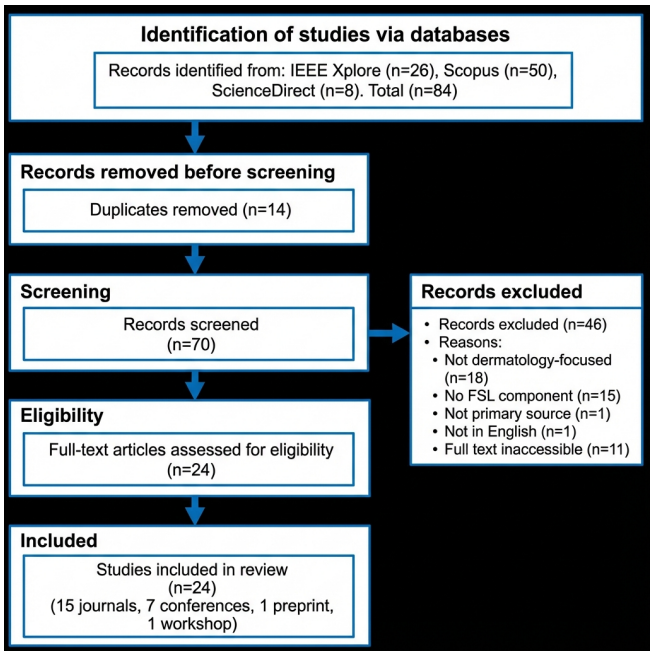

FIGURE 7. PRISMA 2020 flow diagram illustrating the systematic study selection process from identification through final inclusion.

VI. CONCLUSION

This Systematic Literature Review has synthesized the rapid evolution of Few-Shot Learning in dermatology from 2020 to 2025. Our analysis of the 16 primary studies leads to a dual verdict: the field has achieved **Methodological Maturity**, yet it remains **Clinically Unvalidated**.

Methodologically, we have witnessed a clear trajectory from simple Metric Learning (e.g., vanilla Prototypical Networks) to sophisticated, hybrid architectures. The "Third Wave" of Generative Foundation Models (2024–2025) has successfully pushed 5-shot accuracy on rare diseases above the critical 90% threshold. The integration of Vision Transformers and Hyperbolic Embeddings demonstrates that the community has successfully adapted the "Learning to Learn" paradigm to the nuances of dermatological texture and hierarchy.

Clinically, however, significant barriers prevent immediate deployment. The reliance on intra-domain validation (training and testing on high-quality PH2/ISIC images) masks the true fragility of these models when faced with real-world smartphone photos ("Web Atlas" quality). The lack of cross-domain benchmarking, the persisting "Reproducibility Crisis" (with only 6.25% code availability), and the complete absence of skin tone diversity reporting constitute a "Reality Gap" that must be bridged.

We conclude that FSL is the definitive solution to the

"Long-Tail" problem in dermatology, but its promise will only be realized if future research pivots from chasing incremental accuracy gains to ensuring robustness, fairness, and transparency.

DECLARATION OF GENERATIVE AI

During the preparation of this work, the author utilized Large Language Models (LLMs) to assist in checking LaTeX syntax, generate figure and summarizing bibliometric trends. The final content was reviewed and verified by the human author, who takes full responsibility for the study's scientific integrity.

.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their constructive feedback.

REFERENCES

[1] S. Li et al., "Dynamic Subcluster-Aware Network for Few-Shot Skin Disease Classification," IEEE Trans. Neural Netw. Learn. Syst., 2025.
[2] M. Noman et al., "FEGGNN: Feature-Enhanced Gated Graph Neural Network for Few-Shot Classification," Comput. Biol. Med., 2025.
[3] Y. Hu et al., "Hyperbolic Geometry-Driven Robustness Enhancement for Rare Skin Disease Diagnosis," IEEE J. Biomed. Health Inform., 2025.
[4] M. Özdemir et al., "Meta-Transfer Derm-Diagnosis: Exploring Few-Shot Learning and Transfer Learning for Skin Disease Classification in Long-Tail Distribution," IEEE J. Biomed. Health Inform., 2025.
[5] J. Panggiri et al., "Optimized Few-Shot Learning with Adversarial Feature Hallucination Networks for Multiclass Skin Disease Prediction Modeling," in Proc. ICoICT, 2025.
[6] W. Fu et al., "Boosting few-shot rare skin disease classification via self-supervision and distribution calibration," Biomed. Eng. Lett., 2024.
[7] T. Chen, Q. Liu, and J. Yang, "Few-Shot Classification with Multiscale Feature Fusion for Clinical Skin Disease Diagnosis," Clin. Cosmet. Investig. Dermatol., vol. 17, pp. 1101–1118, 2024.
[8] W. Wang et al., "Medical Tumor Image Classification Based on Few-Shot Learning," IEEE/ACM Trans. Comput. Biol. Bioinform., 2024.
[9] J. Xiao et al., "FS3DCIoT: A Few-Shot Incremental Learning Network for Dermatology," IEEE Trans. Consum. Electron., 2023.
[10] K. Lee et al., "Multi-Task and Few-Shot Learning-Based Fully Automatic Deep Learning Platform for Mobile Diagnosis of Skin Diseases," IEEE J. Biomed. Health Inform., vol. 27, no. 5, pp. 2408–2419, 2023.
[11] Y. Wang et al., "Cross-Domain Few-Shot Learning for Rare-Disease Skin Lesion Segmentation," in Proc. ICASSP, 2022.
[12] Y. Zhou et al., "Few-shot Learning Framework Based on Adaptive Subspace for Skin Disease Classification," in Proc. IEEE BIBM, 2022.
[13] Y. Cao et al., "An auxiliary tool for preliminary tests of skin cancer: A self-modifying meta-learning method," in Proc. ICBASE, 2021.
[14] W. Zhu et al., "Temperature network for few-shot learning with distribution-aware large-margin metric," Pattern Recognit., vol. 112, p. 107700, 2021. [Online]. Available: https://github.com/zwwews/TemperatureNetwork.git
[15] K. Mahajan et al., "Meta-DermDiagnosis: Few-Shot Skin Disease Identification using Meta-Learning," in CVPR Workshops, 2020.
[16] J. Zhang et al., "ST-MetaDiagnosis: Meta learning with Spatial Transform for Skin Disease," in Proc. IEEE BIBM, 2020.

DEDY VAN HAUTEN received the bachelor's degree from the Faculty of Informatics, Surya University Bogor, and the master's degree from the Faculty of Computer Science, University of Indonesia. His research interests include machine learning, deep learning, and AI in health.

MUHAMMAD HANNAN HUNAFA received the bachelor's degree from the Faculty of Informatics, Telkom University Purwokerto, and the master's degree from the Faculty of Computer Science, University of Indonesia. His research interests include machine learning, deep learning, and computer vision.

WISNU JATMIKO (Senior Member, IEEE) is one of the academic staff at the Faculty of Computer Science, head of Intelligent Robotics and System (IRoS) Laboratory, and Head of Artificial Intelligence Cluster Research at University of Indonesia. He obtained his Bachelor of Engineering degree and Magister of Computer Science degree from University of Indonesia in 1997 and 2000, respectively. In 2007, he received his Dr. Eng. degree from Micro-Nano System Engineering, Nagoya University, Japan; and starting from September 2017, became a Professor at the Faculty of Computer Science, University of Indonesia. He has served as Chair of The Institute of Electrical and Electronics Engineers (IEEE) Indonesia Section for the 2019 and 2020 periods. His current research interests include autonomous robots, biomedical, and computer vision.