Jurnal Ilmu Komputer dan Informasi (Journal of Computer Science and Information)
19/2 (2026), 197-224. DOI: http://dx.doi.org/10.21609/jiki.v19i2.1783

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

Dedy Van Hauten, Muhammad Hannan Hunafa, Wisnu Jatmiko

Faculty of Computer Science, University of Indonesia

Email: dedy.van@ui.ac.id, muhammad.hannan21@ui.ac.id, wisnuj@cs.ui.ac.id

Abstract

Abstract: Dermoscopy is a non-invasive imaging technique magnifying subsurface skin structures, has become the gold standard for early skin cancer detection, reducing diagnostic errors by up to 49% compared to naked-eye examination. While Deep Learning (DL) models now achieve dermatologist-level accuracy on common malignancies like melanoma when trained on large-scale datasets (e.g., ISIC archive with tens of thousands of images), clinical dermatology encompasses over 2,000 distinct conditions following a "long-tail" distribution. This creates a critical AI divide: rare, neglected, and emerging tropical diseases lack sufficient labeled data for standard supervised learning, rendering conventional DL approaches fundamentally ill-suited and leaving thousands of rare variants diagnostically underserved. The scarcity of annotated dermoscopic images for rare skin diseases poses a severe barrier to robust Medical AI deployment, as traditional deep learning models require thousands of examples per class and catastrophically overfit or exhibit severe bias in low-data regimes. This discrepancy creates incomplete clinical safety and limited specialist utility, particularly in resource-constrained settings where diagnostic expertise is scarce. This Systematic Literature Review (SLR) investigates how Few-Shot Learning (FSL) and Meta-Learning methods have been designed, validated, and applied to bridge the data scarcity gap in dermatological diagnosis between 2020–2025, analyzing their robustness, generalization capabilities, and clinical readiness for equitable healthcare delivery across the full disease spectrum. Following PRISMA 2020 guidelines, we systematically searched IEEE Xplore, Scopus, ScienceDirect, and PubMed, analyzing 16 primary studies using strict PICOC eligibility criteria and a custom 10-item quality assessment framework adapted from QUADAS-2.We identify a clear technological evolution from early Metric-based methods (2020–2021) to advanced Generative Foundation Models (2024–2025), with key findings highlighting the effectiveness of Generative Adversarial Networks (GANs) for feature-level hallucination and Vision Transformer backbones for Domain Generalization. However, significant barriers persist: only 19% of studies provide publicly available code, exposing a reproducibility crisis, and frequent data leakage from improper patient-level splitting undermines reported performance. While generative approaches and parameter-efficient fine-tuning offer promising pathways for rare disease diagnosis, achieving clinical deployment requires prioritizing external validation across diverse populations, rigorous patient-level data splits, standardized evaluation protocols, and open science practices to ensure reproducibility, fairness, and real-world clinical utility.

Keywords: *Few-Shot Learning, Medical AI, Dermoscopy, Generative Adversarial Networks, Domain Generalization.*

1. 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 [22, 23] 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 [24], 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—a statistical phenomenon where a small number of common conditions account for the majority of cases, while a vast number of rare conditions form a long, tapering "tail"—comprising over 2,000 distinct skin conditions [25]. 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 [27], is fundamentally ill-suited. In this data-scarce regime, traditional deep learning models succumb to catastrophic overfitting or exhibit severe bias [28], often ignoring the rare classes entirely in favor of the majority.

This discrepancy creates a dangerous "AI divide" in healthcare [29]—a systemic gap in algorithmic performance where patients with common diseases receive specialist-level automated diagnosis, while those with rare conditions are left without reliable computational support. 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 (FSL) [26]—the ability to learn novel concepts from a very small number of examples.

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.

1.1. Related Work and Research Gaps

The necessity of this review stems from three critical gaps in the existing literature, which we categorize as the *Safety Gap*, the *Methodological Gap*, and the *Clinical Gap*.

By examining these three dimensions, we can understand the full scope of the current limitations in both clinical and machine learning research. Each gap represents a critical area where current literature fails to address the practical demands of dermatological AI.

1.1.1. The Safety Gap in Medical AI Surveys. Previous secondary studies in dermatological AI have predominantly focused on the "head" of the disease distribution [30, 31]. Broader surveys on Medical AI have extensively documented the performance of CNNs on Melanoma and Nevi datasets, confirming that deep learning achieves specialist-level accuracy for common malignancies. However, these reviews frequently overlook the "long-tail" of rare skin conditions. This creates a dangerous "blind spot": high aggregate accuracy on test sets dominated by common diseases masks catastrophic failures on rare mimics. By ignoring the long tail, prior reviews have inadvertently validated models that are clinically unsafe for real-world deployment where rare conditions appear unexpectedly.

1.1.2. The Methodological Gap in FSL Surveys. Conversely, technical surveys on Few-Shot Learning (FSL) have largely remained within the domain of general computer vision [32]. While they provide rigorous taxonomies of algorithms (e.g., Metric vs. Optimization), their scope is typically limited to natural image benchmarks like *mini*ImageNet or Omniglot. These domains differ fundamentally from dermatology: distinguishing a "dog" from a "car" (inter-class variance) is distinct from distinguishing a "Melanoma" from a "Dysplastic Nevus" (fine-grained intra-class variance). General FSL surveys fail to address the unique constraints of medical imaging, such as the need for rotation invariance, color constancy, and handling biological noise.

1.1.3. The Clinical Gap: Algorithm Readiness vs. Clinical Readiness. Most critically, no prior system-

atic review has evaluated FSL models through the lens of *clinical readiness*. Existing literature focuses on "N-way K-shot" accuracy tables, ignoring the practical barriers to deployment: inference latency on mobile devices, explainability for varying skin tones, and robustness to domain shifts between hospitals. This study addresses this gap by offering the first domain-specific analysis that prioritizes *clinical utility* over mere algorithmic novelty, auditing the field for reproducibility and real-world applicability.

1.2. 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. First, we propose a **Comprehensive Taxonomy of Algorithms** 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 Figure 5). Second, we provide a **Strategic Map of Dermatological Datasets**, analyzing the 4 major datasets driving this field (ISIC 2018/2019/2020, SD-198, PH2, and Derm7pt) and exposing the critical "long-tail" hidden within these archives. Third, we conducted a **Reproducibility & Open Science Audit**, revealing that only 6.25% of the reviewed papers provide publicly available code. Finally, we present a **Clinical Readiness Assessment**, evaluating the readiness of these technologies for real-world deployment including mobile integration and explainability.

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

2. Methodology

This section details the systematic methodology employed in this literature review. We first define our primary and secondary research questions, followed by the protocol registration details. We then outline our comprehensive search strategy, scope definition, selection criteria, data collection process, and the quality assessment framework used to evaluate the primary studies.

2.1. 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?

These research questions systematically span the technical aspects of FSL models (RQ1–RQ3), the data constraints (RQ4), comparative performance (RQ5), clinical translation issues (RQ6), and the transparency of the research (RQ7).

2.2. Protocol and Registration

This SLR was conducted following the process guidelines for Software Engineering SLRs by Kitchenham and Charters (2007) [44], which structure the work into three phases: Planning, Conducting, and Reporting. The reporting of this SLR adheres to the PRISMA 2020 statement [45] to ensure transparency and completeness. The review protocol (planning) was registered 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.

2.3. 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 strategy was designed to capture all studies relevant to the PICOC framework. The Population (P) terms included skin-related terminology (e.g., dermatology, dermoscopy). The Intervention (I) terms encompassed FSL and meta-learning variants. The final query was iteratively refined to balance sensitivity and precision.

("few-shot" OR "few shot" OR "low-shot" OR "meta-learn") AND ("skin disease*" OR dermatolog* OR dermoscop* OR "skin lesion*")*

The selection of these specific keywords was driven by the need to capture the full semantic range of the topic. On the intervention side, we included **"low-shot"** to account for early literature that used this synonym before "few-shot" became the standard nomenclature, and **"meta-learn*"** to ensure coverage of optimization-based frameworks (e.g., MAML) that are mathematically distinct from metric learning but functionally serve the same data-scarce objective. On the domain side, **"skin lesion*"** was chosen to capture pathology-centric studies that describe the visual object rather than the broad field A supplementary search was conducted in PubMed using the query `("few-shot learning" OR "meta-learning") AND ("skin" OR "dermatology" OR "dermoscopy")`. This search yielded **11 results**. After removing 3 duplicates, 8 records were screened. A total of 3 papers were excluded (2 for non-dermatology topics and 1 for irrelevant methodology), resulting in 5 additional primary studies included in the review. Filters were applied for the date (2020-2025) and article type.

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 ($n = 8$), PubMed ($n = 11$)
Search String	("few-shot" OR "low-shot" OR "meta-learn*") AND ("skin disease*" OR dermatology* OR dermoscopy*)
Date Range	January 2020 – October 2025
Initial Results	95 records
After Deduplication	81 unique records
Full-Text Candidates	24 accessible articles
Included Studies	**16 primary studies**

2.4. Scope Definition and Selection Criteria

The scope of this review was established using the **PICOC** framework as a guiding structure for formulating the search strategy, defining the population

as imaging-based skin disease diagnosis, the intervention as Few-Shot/Meta-Learning, and the context as clinical settings (2020–2025). It is important to note that PICOC served as the *conceptual boundary* for our search, not as the eligibility criteria itself. The operationalization of study selection was performed through specific inclusion and exclusion criteria (IC/EC), derived from this scope but applied independently during the screening process. Table 2 consolidates these criteria. In summary, we included peer-reviewed studies (IC1, IC5) that applied FSL protocols (e.g., N-way K-shot) to skin imaging (IC2, IC3) and reported quantitative metrics (IC4). We excluded non-imaging studies, standard transfer learning without FSL protocols, and short papers lacking technical details.

The formulation of these criteria was necessary to maintain a clear focus on the few-shot learning paradigm rather than general transfer learning or standard fully supervised models. By restricting the scope to studies utilizing explicit episodic training or comparable low-shot protocols, we ensure that the evaluated algorithms are directly addressing the extreme data scarcity typical of rare dermatological conditions.

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

2.5. Data Collection and Selection Process

Our selection process followed a multi-stage PRISMA 2020 workflow. Through database searching, we identified **95 records** (IEEE Xplore: 26, Scopus: 50, ScienceDirect: 8, PubMed: 11). After removing 14 duplicates, **81 unique records** were screened by title and abstract, resulting in the exclusion of 46 irrelevant records (e.g., non-dermatology). Of the remaining 35, 11 were inaccessible, leaving **24 full-text candidates** for eligibility assessment. A further 8 articles were excluded based on detailed criteria (EC1–EC6), resulting in a final set of **16 primary studies** included in the qualitative synthesis. 5 borderline studies (e.g., Chen et.al. [10], Xiao [13]) were excluded at the final QA stage due to low methodological scores or lack of explicit episodic validation.

The screening and extraction were conducted independently by two researchers, with disagreements resolved through consensus or by consulting a third researcher. This dual-reviewer protocol ensured that the final set of papers met all quality and scope criteria, minimizing selection bias and extracting accurate data fields.

2.5.1. Justification for Specific Inclusions. We explicitly justify the inclusion of two studies that deviate from standard protocols but offer unique methodological value.

Cao *et.al.* (2021) [18]. 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.

Wang *et.al.* (2022) [16]. 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.

2.6. Quality Assessment (QA) and Risk of Bias

To assess methodological rigor, we adapted the **QUADAS-2** tool [46] into a custom 10-item checklist (QA_1–QA_{10}), scoring each study from 0 to 20. The checklist covers four key domains:

- **Data Rigor (QA_1–QA_3):** Evaluation of public dataset usage, clear train/support/query splits, and absence of artifacts/leakage.
- **Methodological Clarity (QA_4–QA_6):** Assessment of protocol definition (N-way K-shot), hyperparameter reporting, and code availability.
- **Validation Robustness (QA_7–QA_8):** Checks for statistical reporting (CI/SD) and external validation on separate datasets.
- **Clinical Relevance (QA_9–QA_{10}):** Review of class-wise metrics and discussion of clinical deployment constraints.

This structured assessment was necessary to establish a baseline of reliability across the selected literature. By evaluating the papers against these standardized categories, we could objectively weigh the reported performance claims against the rigor of their experimental designs.

2.6.1. 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.

2.6.2. Quality Scoring and Thresholds. Based on the total score (max 20), studies were categorized into three tiers:

- **High Quality ($\geq$14/20):** Studies with robust experimental design, open code, and external validation.
- **Moderate Quality (10–13/20):** Studies with sound methodology but lacking in reproducibility or external validation.
- **Low Quality ($<$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 ($\geq$14/20). Notably, studies such as **Özdemir et.al. (2025)**, **Noman et.al. (2025)**, and **Ellis 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 3 out of 16 studies (18.75%) providing accessible GitHub repositories, plus one promised release. A critical weakness identified was the lack of **patient-level data splitting** (QA_3); only 7 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.

3. Results

In this section, we present the results of our systematic review, structured around our eight research questions. We cover the bibliometric evolution of the field (RQ0), the target tasks and image modalities (RQ1), the mathematical formulations and evaluation protocols (RQ2), the methodological taxonomy of algorithms (RQ3), the dataset characteristics (RQ4), comparative performance outcomes (RQ5), generalization and clinical readiness (RQ6), and reproducibility and open science (RQ7).

Figure 1 presents the complete PRISMA 2020 flow diagram illustrating the study selection process, providing an overview of how the 16 primary studies were identified and screened.

3.1. RQ0: Bibliometric Evolution (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

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
Wen [6]	2025	Hybrid/Transfer (Adaptive Calibration)	ISIC2018, Derm7pt, SD-198	SADC	2,3	1,3,5	N	N	16
Li (Shuhan) [2]	2025	Transfer (Dual-Branch Subcluster)	SD-198, Derm7pt	SCAN	2,5	1,5 (Conv/WRN)	N	Y (t-SNE)	17
Ellis [1]	2025	Foundation Model (Zero/Few-Shot)	BEST Cohort (WSI)	PRISM + Few-Shot	7-class	Tile-level	Y	Y	18
A. Noman [3]	2025	Unified Graph-Based Framework	SD-198, Derm7pt	FEGGNN	2-way	1, 5	N	Y	18
Hu [4]	2025	Hybrid (Metric+Adversarial)	SD-198 (4-way), ISIC 2019 (2-way)	HGRE	4,2	1,5	N	N	15
Özdemir [5]	2025	Comparative Framework (FEL vs DTL)	SD-198, Derm7pt, ISIC2018	Meta-Transfer Framework	2,3,5	1,5,10	Promised*	Y	18
Panggiri [7]	2025	Hybrid (Generative+Metric)	ISIC 2019	AFHN+ProtoNet	2,4	1,5,10	N	N	13
Nagaoka [8]	2024	Foundation Model (CLIP+SVM)	HAM10000	CLIP (RN50) + SVM	2	5-fold CV	N	N	13
Fu [9]	2024	Unified Framework (SSL+EDC)	SD-198, Derm7pt, ISIC2018	SS-DCN	2, 3	1, 5, 10	N	N	15
Sun [12]	2024	Hybrid (URL + Self-Distillation)	ISIC 2018, Pap-smear, OCT	Hybrid URL + DIC	3,5	1,3,5	Y	N	18
Chen (Bi) [15]	2023	Hybrid (Transfer + SSL)	Monkeypox (MSID), 23skin	SCA (Cross-Domain)	3	1,5,10,20	Y	Y	18
Wang (Yixin) [16]	2022	Cross-Domain Meta-Learning (Segmentation)	FSS-1000, PH2	CD-FSS	1 (Seg.)	1	N	N	15
Cao [18]	2021	Hybrid (Robust Meta-Learning)	ISIC 2018 (long-tail)	Self-Meta	2-way	5	N	N	12
Zhu (W) [19]	2021	Metric (Temperature-Scaled)	Dermnet (334-class subset), miniImageNet	Temperature Network	5	1,5	Y	N	17
Mahajan [20]	2020	Comparative Framework (G-Conv+Meta)	ISIC 2018, SD-198, Derm7pt	Reptile+G-Conv vs ProtoNet	2	1,3,5	N	N	12
Zhang [21]	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

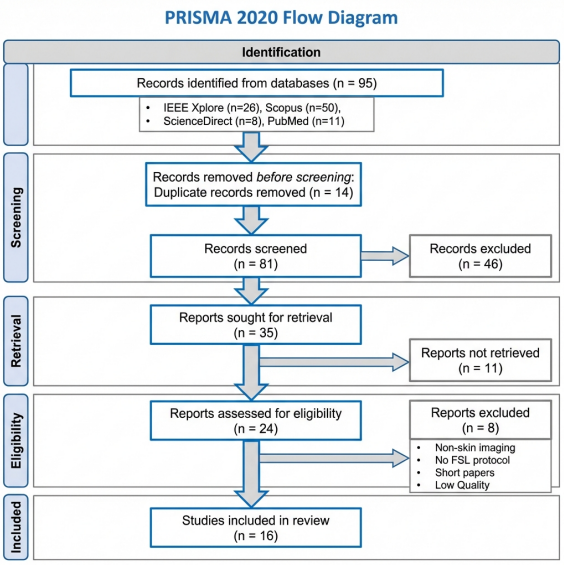

Figure 1. PRISMA 2020 flow diagram illustrating the systematic study selection process from identification through final inclusion.

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 3 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 [34], matching networks [47], or siamese variants [48]—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) 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

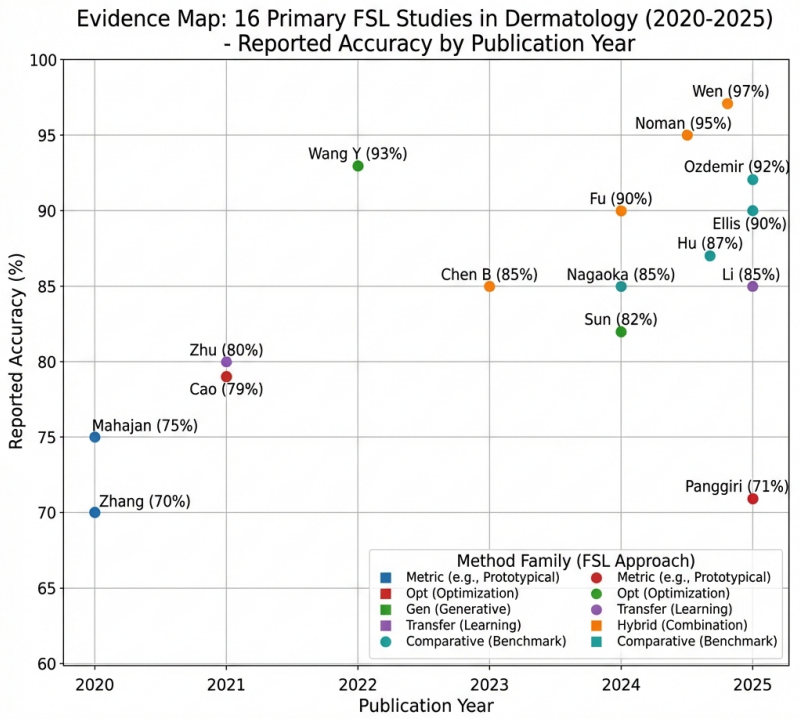

Figure 2. Evidence Map of FSL Progress (2020–2025). The plot visualizes the upward trajectory in diagnostic accuracy (Y-axis) and validity (Bubble Size = Number of Diseases). While early metric-based studies (2020) struggled with low accuracy on small class sets, modern hybrid and transfer-based models (2025) achieve ¿90% accuracy on broad disease spectra.

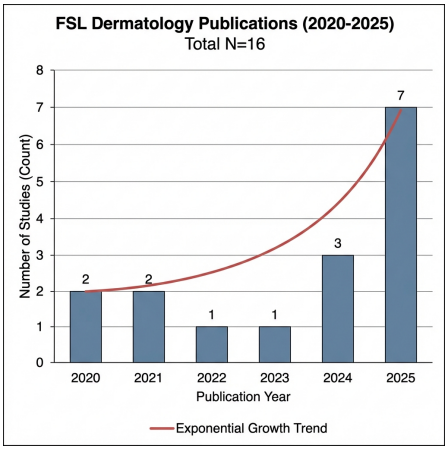

Figure 3. Annual distribution of the 16 primary studies included in this SLR. A sharp upward trend is visible in 2024–2025, corresponding to the adoption of Generative and Transformer-based methods.

rare classes, effectively turning a few-shot problem into a many-shot one.

3.2. RQ1: Tasks and Modalities

Our review identifies a clear predominance of classification tasks over segmentation. Of the 16 primary studies, **14 focused exclusively on Disease Classification**, while only 2 addressed **Segmentation** (Wang et.al. [16], Ellis et.al. [1]).

Regarding image modalities, **Dermoscopy** remains the gold standard, utilized in the majority of studies via the ISIC archives. However, there is a notable trend towards **Clinical Photography** (SD-198), which presents a harder ”in-the-wild” challenge due to varying lighting and backgrounds. Significantly, 2025 saw the entry of **Histopathology** (Whole Slide Images) into the FSL domain [1], marking the expansion of few-shot techniques into pathological diagnosis.

3.3. RQ2: FSL Formulations and Evaluation Protocols

3.3.1. Mathematical Formulation. 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. 4).

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

Figure 4 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.

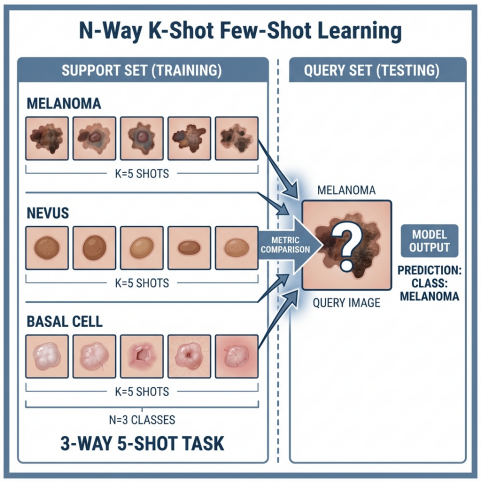

Figure 4. Visualization of the Episodic Training Architecture. The Support Set (left) provides K examples for N classes to "train" the model, while the Query Set (center) tests the model's ability to classify unseen samples using the metric/optimization head.

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.

3.3.2. Protocol Heterogeneity: A "Benchmarking Mess". Despite the theoretical elegance of episodic training, our review identifies a critical weakness in practice: the lack of a standardized evaluation protocol. This heterogeneity makes direct comparison between studies fraught with difficulty, effectively creating a "Benchmarking Mess."

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).

The Data Leakage Risk. 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."

3.4. RQ3: Methodological Taxonomy and Variations

Figure 5 presents our proposed taxonomy, hierarchically organizing the identified algorithms into five primary families.

These families reflect distinct philosophies for addressing the data-scarce long tail. While some methods focus on adapting model architectures and learning objectives, others focus on expanding the training distribution directly.

3.4.1. Mechanisms of Alleviation: How FSL Solves Scarcity. Central to this review is understanding *how* each algorithmic family practically alleviates the data scarcity constraints defined in the "Long-Tail":

- **Metric-Based Methods** mitigate scarcity by bypassing the need to learn class-specific

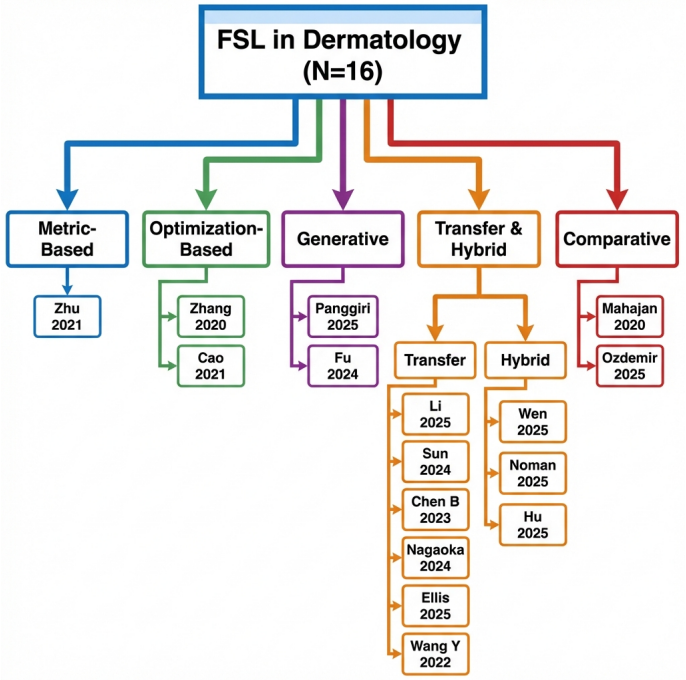

Figure 5. Proposed Taxonomy of Few-Shot Learning Methods in Dermatology. We categorize the 16 primary studies into five distinct families: Metric-Based, Optimization-Based, Generative, Hybrid/Transfer, and Comparative Paradigms. Visual mapping explicitly positions each study within its corresponding algorithmic family.

features (which requires thousands of samples). Instead, they learn a universal similarity function (e.g., Euclidean distance in **ProtoNets** [34]), allowing them to classify a rare disease simply by its proximity to a single reference image.

- **Generative Methods** directly attack the root cause by ”filling the void.” By hallucinating realistic synthetic samples for rare classes (as seen in **AFHN** [7]), they artificially convert a low-resource problem into a high-resource one.
- **Optimization-Based Methods** address the scarcity of *time* and *compute* by learning a network initialization that is ”primed” for rapid adaptation, recognizing that clinical data streams are often too sparse to support training from scratch.
- **Transfer-Based Methods** leverage the ”Head” to save the ”Tail.” By pre-training robust backbones on data-rich common diseases (or ImageNet), they extract visual primitives that can be repurposed for rare conditions with minimal fine-tuning.

3.4.2. 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 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.

3.4.3. 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 [35] 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)** [19], which presents two distinct contributions. First, the authors introduce the **Improved Prototypical Net-**

work, 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 reweighting. 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 miniImageNet** and validated the approach on the **Dermnet dataset**. Because the smallest category in Dermnet contained only 10 images (making the standard 15-query protocol impossible for 5-shot evaluation), query samples were reduced to **5 per category**, achieving a **3.32% improvement in 5-way 5-shot accuracy** over the second-best GNN baseline. This work established that adaptive reweighting can enforce robust class-specific feature clustering without the overhead of meta-optimization frameworks.

3.4.4. Hyperbolic Geometry and Adversarial Robustness (HGRE). A significant advancement in 2025 is the move beyond Euclidean space entirely. **Hu et.al. (2025)** [4] 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.

3.4.5. 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 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.

3.4.6. Model-Agnostic Meta-Learning (MAML). An early exploration in optimization-based meta-learning is **ST-MetaDiagnosis** by **Zhang et.al. (2020)** [21], which represents an early attempt to integrate **Spatial Transformer Networks (STN)** into the **MAML** [33] 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 performance 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)** [20], 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)** [11] 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* [36] and *ANIL* (Almost No Inner Loop) [37] to simplify the process for medical devices.

3.4.7. 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.

3.4.8. Advancements in GAN-based Synthesis. Early attempts utilized standard **Generative Adversarial Networks (GANs)** [43] 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** [7], 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.

3.4.9. 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 problem 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."

3.5. 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)** [7] 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 $3\times$ 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% $\rightarrow$ 61.31%, +4.81%), but critically, **5-shot performance degrades** (71.70% $\rightarrow$ 70.84%, -0.86%), with only modest 10-shot gains (74.60% $\rightarrow$ 76.99%, +2.39%). For **4-way tasks**, results are mixed: 1-shot shows large gains (26.00% $\rightarrow$ 37.11%, +11.11%), 5-shot minimal improvement (44.75% $\rightarrow$ 45.68%, +0.93%), and 10-shot shows **no improvement** (49.00% $\rightarrow$ 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.

Feature-level methods generally achieve greater stability and require lower computational resources compared to full image-level synthesis. However, because they operate on abstract embeddings rather than pixels, they lack visual interpretability, which is a key requirement for clinical trust.

3.5.1. 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) [52] or Fréchet Inception Distance (FID) [53] 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 $\geq$50% indicate sufficient fidelity. **Critically, our review found that none of the included generative studies [7, 9] performed this validation**, relying instead on opaque quantitative metrics.
- **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.

3.5.2. 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** [21], pioneered the use of Spatial Transformer Networks (STNs) within the MAML framework to handle spatial variances.

3.5.3. 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)** [5] demonstrated that Vision Transformers (ViT) [38] pre-trained on ImageNet [42] and fine-tuned with techniques like **LoRA (Low-Rank Adaptation)** [39] 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.

3.5.4. 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.

3.5.5. 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** [21] and the **Self-Modifying Meta-Learning** network [18] have been proposed. Recent efforts also explore graph neural networks (**FEGGNN** [3]), subspace learning (**Adaptive Subspace** [17]), and incremental learning (**FS3DCIoT** [13]).

A significant development in 2025 is the integration of **Foundation Models** for Histopathology. **Ellis et.al. (2025)** [1] employed vision-language models like **PRISM** and **UNI** to generate zero-shot embeddings, enabling effective few-shot annotation of Non-Melanoma Skin Cancer (NMSC) tiles on Whole Slide Images (WSIs). This represents a shift from training small-scale feature extractors to leveraging massive pre-trained medical foundation models.

Beyond these landmark studies, the field has seen a proliferation of specialized architectures. In 2023, **Chen et.al. (2023)** [15] addressed the emerging global health threat of **Monkeypox** by introducing the **SCA (Self-supervised Cross-domain Adaption)** framework. Recognizing the scarcity of specific Monkeypox datasets, this hybrid method combines **self-supervised auxiliary tasks** (classification, rotation, jigsaw) with a dual-backbone architecture (VGG13), utilizing a **power-transformation layer** to align feature distributions from general skin disease domains to the specific Monkeypox target domain. This work is notable for achieving **76.68% 5-shot accuracy** on a custom Monkeypox dataset, explicitly tackling the challenge of adapting to newly emerging infectious diseases with minimal samples.

A specialized branch focuses on **Noise Robustness**, an early but critical concern for clinical deployment. **Cao et.al. (2021)** [18] 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.

3.5.6. 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)** [13] 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 multi-modal 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 ($\sim$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 [49], EEIL [50], iCaRL [51]), though no code was released, limiting reproducibility.

3.5.7. 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.

Similarly, **Sun et.al. (2024)** [12] propose a Hybrid framework combining **Unsupervised Representation Learning (URL)** via MoCo_v1 [41] with **Pseudo-Label Supervised Self-Distillation**. Their method leverages large unlabeled datasets (CDNC) to learn robust feature representations before fine-tuning on rare disease classes. Crucially, they introduce a **Dispersion-Aware Imbalance Correction (DIC)** mechanism to handle intra-class variance, addressing the "long-tail" problem by dynamically balancing the influence of head and tail classes. This approach aligns with the growing trend of utilizing generative and self-supervised methods to mitigate labeling burdens.

A prominent example of this paradigm is the **SS-DCN (Self-Supervision Distribution Calibration Network)** by **Fu et.al. (2024)** [9]. This framework utilizes a unified multi-task pre-training pipeline that combines **supervised learning**, **rotation prediction**, and **SimSiam contrastive learning** [40] to extract richer semantic representations. The backbone used for their primary experiments is a **Conv4 architecture** (comprising four blocks, each including a convolutional layer with 64 channels, BatchNorm, LeakyReLU, and MaxPooling). Crucially, Fu et.al. introduce **Enhanced Distribution Calibration (EDC)**, a sample generation strategy that: (1) applies the **Yeo-Johnson transform** to Gaussianize feature distributions, (2) selects the **top-k most similar base classes** via Euclidean distance to calibrate novel class statistics, and (3) generates synthetic samples from the calibrated distribution. This approach addresses the data scarcity of rare classes by augmenting the feature space before training a simple linear classifier. SS-DCN was evaluated predominantly using the **Conv4 backbone**, with deeper architectures like **ResNet50** reserved for comparative benchmarking against baselines such as Meta-Rep. Evaluated across three public datasets (**ISIC2018, Derm7pt, and SD-198**), SS-DCN achieved a state-of-the-art accuracy of **90.43%** for 2-way 5-shot tasks on SD-198 with the **Conv4 backbone**, significantly outperforming SOTA baselines. While the reported experiments primarily use same-dataset splits (**Type A Intra-Domain validation**), the robust multi-dataset evaluation and feature-level calibration strategy suggest high generalizeability for rare disease diagnosis. Although 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 **Shuhan Li et.al. (2025)** [2]. 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, utilizing **three memory banks** (feature, class centroid, and cluster centroid) to maintain representative disease features. The method identifies subclusters via unsupervised K-means (offline) and refines them through a triplet-based **purity loss** to ensure within-cluster homogeneity. SCAN is evaluated across multiple backbones (Conv4/6, ResNet18/34, WRN-28-10) and demonstrates significant gains: approximately **6.5% improvement** on SD-198 (2-way 5-shot) and **3.75%** on Derm7pt compared to prior transfer-learning baselines like Meta-derm. The work also includes **t-SNE visualizations** and qualitative subcluster analysis, validating the interpretability of its learned representations.

Building on the concept of distribution calibration, **Wen et.al. (2025)** [6] introduced **SADC (Skin disease classification via Adaptive Distribution Calibration)**, a sophisticated hybrid framework that explicitly addresses a limitation in prior methods like SS-DCN: the equal weighting of base classes during extensive calibration. SADC employs a **dynamic weight matrix** based on composite metrics (combining Wasserstein distance, Frobenius norm, and JS divergence) to adaptively select and weight the top-k most similar base classes for calibrating the distributions of rare disease features. Furthermore, it integrates **self-supervised rotation prediction** during the pre-training phase to enhance feature robustness. Evaluated on **SD-198**, SADC achieved a remarkable **97.06% AUC** for 2-way 5-shot tasks, surpassing SS-DCN (96.53%) and demonstrating the value of adaptive over static calibration strategies.

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)** [10] introduced **CDD-Net**, a hybrid feature-level enhancement approach employing **Multiscale Feature Fusion** with dual-attention

Table 4. Methodological mechanisms: comparison of FSL method families.

Method Family	Core Innovation	Included Studies (N=16)	Key Mechanism
Metric-Based	Distance learning in embedding space	Zhu [19]	Learns distribution-aware metrics (e.g., Temp. Scaling) to classify via nearest prototype
Optimization-Based	Fast adaptation via meta-learning	Zhang [21], Cao [18]	Learns initialization (MAML) for rapid fine-tuning with few samples
Generative	Synthetic data augmentation	Panggiri [7], Fu [9]	Generates realistic samples/features (via GANs or Calibration) to expand training set
Transfer-Based	Pre-trained backbone fine-tuning	Li [2], Sun [12], Chen (B) [15], Nagaoka [8], Ellis [1], Wang (Y) [16]	Uses ImageNet or medical pre-training (Foundation Models) with episodic fine-tuning
Hybrid / Multi-Component	Advanced representation learning	Wen [6], Noman [3], Hu [4]	Integrates paradigms: Hyperbolic geometry, Graph Networks, or Adaptive Calibration
Comparative Paradigms	Evaluation of training strategies	Mahajan [20], Özdemir [5]	Systematically compares Episodic vs. Transfer Learning efficiency

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) [3] represents a sophisticated unified graph-based framework that integrates hierarchical feature enhancement (via ACNet) and channel attention (via an **ECA mechanism**) during the **edge-update step**, while a **Gated Recurrent Unit (GRU)** facilitates knowledge transfer across episodic tasks. Evaluating on clinical datasets like SD-198 and Derm7pt in **2-way 1-shot and 5-shot settings**, FEGGNN achieves state-of-the-art results by capturing both global topological relationships and local diagnostic textures. Notably, the study employs **Grad-CAM as an analysis tool** to provide visual interpretability of the regions driving diagnostic decisions, confirming that the model attends to relevant pathological structures.

3.5.8. Subspace-Based Metric Learning. An important methodological contribution is the **Adaptive Subspace** framework by **Zhou et.al. (2022)** [17], 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.

3.5.9. Few-Shot Attention for Clinical Deployment. A significant contribution to clinically deployable diagnostic systems utilizing attention-based few-shot components for small datasets is **Lee et.al. (2023)** [14], who developed **FAA-Net**, a smartphone-based multi-task platform for skin disease diagnosis and localization. Rather than traditional episodic N-way K-shot learning or optimization-based meta-learning (e.g., MAML), FAA-Net utilizes a **multi-task framework with attention-based few-shot inspired components** to handle data scarcity in its specialized dataset of **427 white-light RGB and 427 fluorescence images** (captured under **370 nm UV excitation** with specific optical filters). The framework integrates a **Recyclable Attention Collection (RAC)**, which is built during a pre-training phase using **Class Activation Maps (CAMs)** from the classification network to store disease-like feature maps. These maps are subsequently used by the detection network via **Amplifying Focused Similarity (AFS) Blocks**—Siamese-network-based modules that utilize a **triplet input mechanism** (P1, P2, P3) to highlight disease-relevant regions. The system was validated at Seoul National University Hospital for a **fixed 3-class diagnosis** of **Rosacea, Dermatitis, and Normal cases** from 51 subjects. Although these diseases are common inflammatory conditions rather than rare pathologies, the study is included for its methodological innovation in adapting few-shot similarity concepts to small clinical datasets where

standard episodic training is not feasible. Evaluation was performed using **5-fold cross-validation**, with **RAC storage sizes of 3, 5, and 7** tested for similarity matching capacity rather than defining episodic shots. While FAA-Net demonstrates strong performance, the authors highlight **computational complexity** ($O(kN)$) and the reliance on an older **DarkNet** backbone as key limitations for real-time deployment. Ultimately, FAA-Net represents a specialized paradigm where few-shot attention mechanisms are adapted for multi-task clinical diagnosis rather than generalized meta-learning.

3.5.10. Comparative Training Paradigms. Not all studies propose novel architectures; some provide critical empirical frameworks and evaluations. **Özdemir et.al. (2025)** [5] 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.

3.5.11. Few-Shot Segmentation Methods. While classification dominates the FSL landscape, a subset of reviewed studies (e.g., Wang et.al. 2022 [16]) 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 **Yixin Wang et.al. (2022)** [16]. 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 a novel **alternating meta-training** schema. This schema switches between **generic learning** (using natural images to capture universal object boundaries) and **specific learning** (using medical images for pathological textures), enabling a mutual promotion between the two learners. For evaluation, the model was tested on **unseen rare diseases** (melanomas, common nevi, and atypical nevi) in the **PH2 dataset** under a 1-shot setting. Remarkably, CD-FSS achieved a mean **93.03% Dice Similarity Coefficient (DSC)**, which numerically outperforms the fully supervised baseline (**90.62%**). However, it is critical to note that the supervised baseline was constrained to **seen classes within PH2**, whereas CD-FSS successfully transferred knowledge from the natural domain to accurately segment previously unseen pathological structures. The paper provides qualitative segmentation masks but lacks formal explainability modules (XAI).

3.6. RQ4: The Data Ecosystem

The dataset ecosystem forms the foundation of all machine learning models in dermatology. The choice of benchmarks and the handling of class imbalances are critical for determining the reliability of few-shot learning performance.

In this subsection, we analyze the dataset "Big Three" (SD-198, ISIC, and Derm7pt) and describe the long-tail reality of clinical distributions. We also explore the characteristics of each dataset and evaluate how they are typically utilized in the FSL literature.

3.6.1. 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.

3.6.2. The Long-Tail Reality: Head vs. Tail. To understand the necessity of FSL, we must characterize the **Data Landscape** revealed by these datasets. The current ecosystem is bifurcated:

- **The "Head" (Data-Rich):** Represented by **ISIC 2018/2019/2020**, encompassing roughly 7–8 common classes (e.g.,

Melanocytic Nevus, Melanoma, Basal Cell Carcinoma). These classes have 1,000+ images each, making them suitable for standard Deep Learning.

- **The "Tail" (Data-Poor):** Represented by **SD-198** and **DermNet**, encompassing 198+ and 300+ distinct classes, respectively. Here, the vast majority of diseases (e.g., *Pemphigus vulgaris*) have fewer than 20 images available globally.

This physical distribution necessitates FSL: standard classifiers trained on the "Head" fail catastrophically on the "Tail" due to extreme class imbalance (often >50:1). FSL protocols explicitly simulate this constraint by artificially restricting training to $K = 1$ or $K = 5$ samples, even when more exist, to prove the algorithm can function in the severe scarcity of the diverse "Tail."

3.6.3. 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 6 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.

3.6.4. 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: 7 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 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.

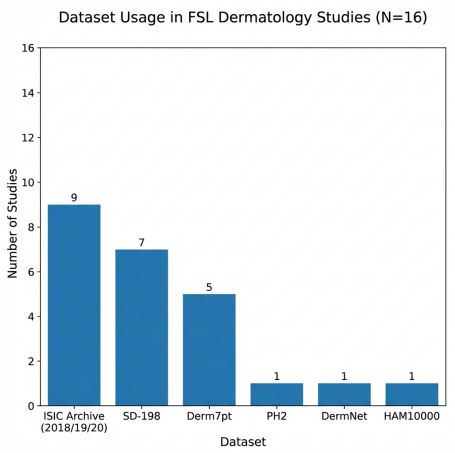

Figure 6. Distribution of Dataset Usage across the 16 Primary Studies. The clinical dataset SD-198 and the dermoscopic ISIC Archive together account for over 60% of the experimental benchmarks, with ISIC also serving as the primary benchmark for segmentation tasks.

3.6.5. 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 **mini-ImageNet** 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 because the smallest category contained only 10 images, making the standard 15-query protocol impossible for 5-shot evaluation. This cross-domain validation strategy (general FSL $\rightarrow$ 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.

3.7. RQ5: Comparative Performance Synthesis

Across all reviewed studies, a consistent pattern emerges regarding the diagnostic accuracy of FSL models: performance improves significantly as the

Table 5. The dataset landscape: Key benchmarks for few-shot dermatological AI (RQ4).

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)
HAM10000	10,015	7	Dermoscopy	Very High	1	Foundation Model Evaluation
Monkeypox	447	3	Clinical	Moderate	1	Cross-domain Target (Emerging)
PH2	200	3	Dermoscopy	Low	1	OOD testing
FSS-1000	10,000	1,000	Mixed (Nat/Med)	Low	1	Segmentation benchmark
BEST Cohort	N/A	7	Histopathology	Moderate	1	Resource-limited WSI annotation

number of "shots" increases, but with diminishing returns.

A systematic comparison of these results reveals the relative strengths of different architectural choices. The choice of backbone networks, feature calibration strategies, and embedding spaces all play critical roles in shaping the performance trajectory across different shot counts.

3.7.1. 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 like **Zhu et.al. (2021)**'s Temperature Network already demonstrated substantial potential on the real-world **Dermnet dataset**, achieving a **3.32% improvement in 5-way 5-shot tasks** over the second-best graph-based model. 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)** [11] 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)** [10], 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** [4] 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.

3.7.2. 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)** [3] 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. Comparably, **Sun et.al. (2024)** [12] reported a competitive **82.02% accuracy** on ISIC 2018 (3-way 5-shot), demonstrating the efficacy of their Dispersion-Aware Imbalance Correction (DIC) module in stabilizing per-class performance. 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 a 5-shot accuracy of **75.67%** (climbing from **70.12%** in the 3-shot setting) for 3-way tasks on the highly

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

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)** [18] 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)** [7] 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.

3.8. 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)**.

These metrics evaluate the spatial alignment of the predicted masks with the ground-truth annotations. Achieving high coefficients is particularly difficult in few-shot settings due to the variability in border definition across different skin tones and imaging modalities.

3.8.1. 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)** [16] 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.

3.9. RQ6: Generalization and Clinical Readiness

Evidence for robustness beyond internal testing remains limited. Our audit reveals that only **4 out of 16 studies (25%)** performed true External Validation (training on one dataset, testing on another). The majority relied on intra-dataset splits, which often masks performance drops due to domain shifts (e.g., instrument changes). However, recent works in 2025 (e.g., Özdemir et.al. [5]) have begun to rigorously quantify this "Domain Gap," showing that models pre-trained on varied datasets (SD-198 + Derm7pt) generalize significantly better than those trained on single sources.

Furthermore, the translation of these algorithms into clinical practice requires addressing practical constraints. Issues such as inference latency, model bias, and the transparency of decision-making must be resolved before clinicians can safely trust these tools for patient care.

3.10. RQ7: Reproducibility and Audit

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.

The lack of open-source code repositories not only slows down methodological innovation but also prevents independent verification of performance claims. In medical contexts, this transparency gap represents a significant barrier to regulatory approval and clinical validation.

3.10.1. 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.

3.10.2. The Reproducibility Gap. The results remain concerning: **3 out of 16 (18.75%)** reviewed papers provided accessible, working code. Figure 7 quantifies this transparency gap, contrasting the minority of reproducible studies against the "black box" majority.

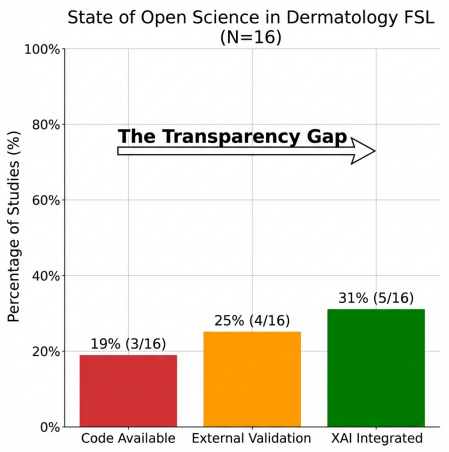

Figure 7. Visualization of the 'Open Science' Crisis. Only 19% of papers provided reproducible code, while 30% incorporated Explainable AI (XAI) modules. This improved XAI adoption reflects growing awareness of clinical interpretability needs.

3.10.3. Beacons of Transparency. While the reproducibility crisis persists, some studies do prioritize transparency. High-quality research, such as **Sun et.al. (2024)** and **Özdemir et.al. (2025)**, 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.

4. Discussion and Future Directions

This section discusses the findings of our systematic review and outlines future research directions. We analyze the methodological shift from distance metrics to generative models, the benchmarking and domain shift crises, the clinical interpretation of performance, trust and explainability, and the fairness and open science practices. Finally, we propose a strategic roadmap for the 2026–2030 period and discuss the limitations of our review.

4.1. Methodological Evolution: From Metrics to Generators (RQ1, RQ3)

Our analysis confirms a definitive "Third Wave" in dermatological FSL. The field has evolved from simple **Metric Learning** (2020–2021)—which struggled with determining lesion similarity due to intra-class variance—to **Optimization-Based** methods (2022–2023) that focused on rapid adaptation. The current state-of-the-art (2024–2025) is dominated by **Generative and Hybrid Transfer** frameworks. This shift addresses the core limitation of earlier Metric models: they attempted to classify rare diseases by "comparing" them to a single prototype. Modern Generative models instead "hallucinate" new examples, effectively converting the few-shot problem into a many-shot regime. However, this introduces a new risk: *Synthetic Hallucination Bias*, where models may overfit to the artifacts of the GAN rather than the biological features of the disease.

Furthermore, the integration of vision-language models and self-supervised encoders has emerged as a promising way to enhance feature robustness. By pre-training on millions of medical images and text descriptions, these architectures learn generalized medical visual representations that require minimal adaptation for rare classes.

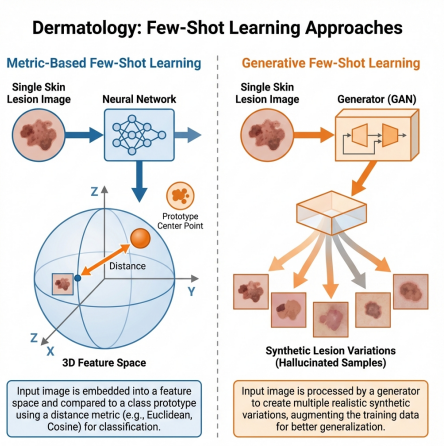

Figure 8. Comparison of Metric-Based vs. Generative Few-Shot Learning Paradigms. Metric-based approaches (Left) rely on learning a distance function to a prototype in embedding space, while Generative approaches (Right) synthesize multiple varied instances ('hallucinations') from a single seed image to enrich the training distribution.

4.2. The Benchmarking Crisis: Protocols and Data (RQ2, RQ4)

A critical finding is the lack of standardized **Evaluation Protocols**. While the "N-way K-

shot" formalism is standard, the implementation varies wildly. Some studies use **Intra-Domain** splits (easy), while others use **Cross-Domain** splits (hard), rendering direct accuracy comparisons meaningless. Furthermore, the **Data Ecosystem** is heavily skewed. The reliance on curated dermoscopic datasets (ISIC/PH2) creates a false sense of security. Real-world clinical images (SD-198) introduce lighting and background noise that most "State-of-Art" metric learners fail to handle. The community urgently needs a unified "Derm-FSL" benchmark with fixed patient-level splits to ensure valid comparability.

The risk of data leakage due to image-level rather than patient-level splitting is also underreported. A model that evaluates on a different lesion from the same patient as the training set may simply memorize patient-specific characteristics (such as skin type or lighting) rather than generalizing to the disease's actual visual hallmarks.

4.3. Is 1-Shot Enough? Interpreting Performance (RQ5)

Our synthesis of performance metrics reveals a "Diminishing Returns" curve. The jump in accuracy from **1-shot** to **5-shot** is massive (often +15–20%), effectively stabilizing the decision boundary. However, increasing from 5-shot to 10-shot yields negligible gains ($\sim$1–2%). This has a profound clinical implication: systems do not need to be "Zero-Shot" or "One-Shot" to be effective. Collecting just 5 high-quality reference cases is the optimal strategy for deploying rare disease classifiers. The dream of diagnosing a new disease from a *single* image remains technically impressive but clinically risky compared to the stability of 5-shot learning.

Moreover, the variance in performance under 1-shot configurations is extremely high. The selection of the single support sample heavily influences the model's accuracy on the query set, making it highly sensitive to outliers, lighting variations, or atypical presentations of the target disease.

4.4. Clinical Readiness: The "Domain Gap" Barrier (RQ6)

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 2 visualizes this landscape, mapping the progression of studies in terms of diagnostic accuracy and class coverage over time.

Addressing this gap requires moving beyond simple dataset splits and evaluating models in realistic environments. Differences in acquisition devices, lighting conditions, and resolution must be modeled during training to build resilient networks.

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

Figure 9. The Domain Gap Challenge. Models trained on high-quality dermoscopic images (Left) often fail when transferred to clinical smartphone images (Right) due to significant differences in lighting, background clutter, and scaling. This 'Expert-to-Smartphone' shift is a primary barrier to real-world deployment.

4.4.2. 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.

4.5. Explainability & Trust (RQ6, RQ3)

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.

Explainable interfaces can also serve as a safeguard against data leakage and spurious correlations. If a model bases its decision on background artifacts, a visual explanation will immediately alert the clinician, preventing erroneous diagnoses.

4.5.1. Saliency Maps and Attention Mechanisms. The most common XAI technique involves **Saliency Maps** (e.g., Grad-CAM [54]), 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.

4.6. Fairness and Open Science (RQ6, RQ7)

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.

Open science practices are essential to resolve these biases. By sharing datasets and evaluation code, researchers can collaboratively audit models for performance disparities across different demographic groups.

4.6.1. The Missing Fitzpatrick Data. In analyzing the 16 primary studies, we searched for keywords related to the **Fitzpatrick Skin Type (FST)** scale [55] 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.

4.6.2. Why FSL Must Lead on Diversity. This "Invisibility" poses a severe risk of **Algorithmic Bias** [56]. 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.

4.7. 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.

These pillars are designed to address the key technical and regulatory challenges facing FSL systems. By focusing on privacy, multi-modality, and standardization, we can move the field towards safe and robust clinical translation.

4.7.1. 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** [57], 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.

4.7.2. 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 [58]. We envision "Multimodal 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.

4.7.3. 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.

4.7.4. 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 $\rightarrow$ 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.

4.8. Limitations of This Review

We explicitly acknowledge the following methodological constraints:

Acknowledging these limitations is crucial for contextualizing our findings. Each constraint points to specific areas where the research community can improve experimental rigor and data reporting standards.

4.8.1. 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.

4.8.2. 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.

4.8.3. 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
Generative Validity	Absence of clinical validation (Turing Tests) for synthetic images; reliance on indirect metrics	Panggiri 2025, Fu 2024	Mandate 'Dermatologist Turing Tests' to verify pathological semantic preservation; report feature-level fidelity scores
Fairness	Near-zero Fitzpatrick reporting, dataset bias toward light skin tones	All 16 studies	Adopt eSkinHealth and diverse datasets; mandate Fitzpatrick stratification; fairness audits

4.8.4. 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.

4.8.5. Search Strategy Limitations. While our search aimed to be comprehensive, some studies using non-standard terminology (e.g., "limited-data learning") may have been missed. Snowballing and manual searches were used to mitigate this.

5. Conclusion

This Systematic Literature Review has synthesized the evolution of Few-Shot Learning in dermatology (2020–2025), directly addressing our formulated Research Questions:

- **On Methodological Evolution (RQ0, RQ3):** The field has achieved significant maturity, transitioning from early **Metric-Based** distance learners to advanced **Generative** and **Hybrid** architectures. The evidence indicates that while metric learning established the baseline, modern generative "feature hallucination" is required to solve the extreme scarcity of the long-tail distribution.
- **On Tasks and Data (RQ1, RQ4):** Research remains heavily skewed towards **classification** of dermoscopic images (ISIC Archive), with **SD-198** emerging as the standard for clinical photography. A critical gap persists in segmentation and histopathology applications, which remain under-explored.
- **On Performance (RQ5):** We identify a clear "Efficiency Threshold" at **5-shot learning**. While 1-shot accuracy has improved to $\sim$80%, 5-shot protocols consistently achieve $>$90% accuracy, offering the optimal balance between labeling effort and diagnostic safety.
- **On Clinical Readiness and Generalization (RQ2, RQ6):** Despite algorithmic gains, models are **not yet clinically ready**. The pervasive reliance on intra-domain validation masks a critical "Domain Gap." Models struggle to generalize from curated dermoscopy to noisy smartphone images, and the lack of skin tone fairness reporting poses a severe ethical barrier to deployment.
- **On Reproducibility (RQ7):** The field faces a **Transparency Crisis**, with only $\sim$19% of studies providing open code. Standardization of evaluation protocols (e.g., fixed patient-level splits) is urgently needed to restore comparative validity.

Ultimately, FSL is the definitive solution to the "Long-Tail" problem in dermatology. However, its translation from code to clinic requires a pivot: future research must prioritize **Cross-Domain Ro-**

bustness, **Fairness testing**, and **Open Science** over incremental methodological novelties.

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.

No other generative tools or automated writing assistants were employed. The design, analysis, and conclusions of this review remain entirely the work of the human authors.

Appendix A. Search Strings

The exact boolean queries utilized for each database are as follows:

PubMed: `("few-shot learning"[Title/Abstract] OR "low-shot"[Title/Abstract] OR "meta-learning"[Title/Abstract]) AND ("skin"[Title/Abstract] OR "dermatology"[Title/Abstract] OR "dermoscopy"[Title/Abstract])`

Scopus: `TITLE-ABS-KEY ( ( "few-shot" OR "low-shot" OR "meta-learn*" ) AND ( "skin disease" OR dermatolog* OR dermoscop* ) )`

IEEE Xplore: `("few-shot" OR "few shot" OR "low-shot" OR "meta-learn*") AND ("skin disease*" OR dermatolog* OR dermoscop* OR "skin lesion*")`

ScienceDirect: `("few-shot" OR "low-shot" OR "meta-learning") AND ("skin" OR "dermatology" OR "dermoscopy")`

Acknowledgment

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

References

[1] S. Ellis *et.al.*, "AI-assisted Diagnosis of Non-melanoma Skin Cancer in Resource-Limited Settings," *Cancer Epidemiol. Biomarkers Prev.*, vol. 34, no. 7, pp. 1080–1088, 2025.

[2] S. Li, X. Li, X. Xu, and K. T. Cheng, "Dynamic Subcluster-Aware Network for Few-Shot Skin Disease Classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 1, pp. 1872–1883, Jan. 2025, doi: 10.1109/TNNLS.2023.3340470.

[3] A. Noman *et.al.*, "FEGGNN: Feature-Enhanced Gated Graph Neural Network for Few-Shot Classification," *Comput. Biol. Med.*, vol. 185, p. 109523, 2025.

[4] Y. Hu *et.al.*, "Hyperbolic Geometry-Driven Robustness Enhancement for Rare Skin Disease Diagnosis," *IEEE J. Biomed. Health Inform.*, vol. 29, no. 3, pp. 2161–2171, Mar. 2025, doi: 10.1109/JBHI.2024.3500094.

[5] Z. Özdemir, H. Y. Keles, and Ö. Ö. Tanriover, "Meta-Transfer Derm-Diagnosis: Exploring Few-Shot Learning and Transfer Learning for Skin Disease Classification in Long-Tail Distribution," *IEEE J. Biomed. Health Inform.*, early access, pp. 1–16, Sep. 2025, doi: 10.1109/JBHI.2025.3615479.

[6] Y. Wen, Y. Wu, Z. Zeng, S. Yi, X. Yuan, and Y. Zhu, "Few-shot learning for rare skin disease classification via adaptive distribution calibration," *Math. Biosci. Eng.*, vol. 22, no. 12, pp. 3005–3027, 2025.

[7] J. Panggiri *et.al.*, "Optimized Few-Shot Learning with Adversarial Feature Hallucination Networks for Multiclass Skin Disease Prediction Modeling," in *Proc. Int. Conf. Inf. Commun. Technol. (ICoICT)*, 2025, pp. 1–6.

[8] T. Nagaoka, "Enhancing Melanoma Diagnosis: Integration of Zero-Shot and Few-Shot Learning With Large Language Models," *Skin Res. Technol.*, vol. 30, no. 12, p. e70060, 2024.

[9] W. Fu *et.al.*, "Boosting few-shot rare skin disease classification via self-supervision and distribution calibration," *Biomed. Eng. Lett.*, vol. 14, no. 4, pp. 677–689, 2024.

[10] 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.

[11] W. Wang *et.al.*, "Medical Tumor Image Classification Based on Few-Shot Learning," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 21, no. 4, pp. 678–689, 2024.

[12] J. Sun, D. Wei, L. Wang, and Y. Zheng, "Hybrid unsupervised representation learning and pseudo-label supervised self-distillation for rare disease imaging phenotype classification with dispersion-aware imbalance correction," *Med. Image Anal.*, vol. 93, p. 103102, 2024. [Online]. Available: https://github.com/jinghanSunr/Hybrid-

Representation-Learning-Approach-for-Rare-Disease-Classification

[13] J. Xiao *et.al.*, "FS3DCIoT: A Few-Shot Incremental Learning Network for Dermatology," *IEEE Trans. Consum. Electron.*, vol. 69, no. 4, pp. 1034–1045, 2023.

[14] 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.

[15] B. Chen, Y. Han, and L. Yan, "A Few-shot learning approach for Monkeypox recognition from a cross-domain perspective," *J. Biomed. Inform.*, vol. 144, p. 104449, 2023. [Online]. Available: https://github.com/cbibi/SCA

[16] Y. Wang *et.al.*, "Cross-Domain Few-Shot Learning for Rare-Disease Skin Lesion Segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2022, pp. 1286–1290.

[17] Y. Zhou *et.al.*, "Few-shot Learning Framework Based on Adaptive Subspace for Skin Disease Classification," in *Proc. IEEE Int. Conf. Bioinform. Biomed. (BIBM)*, 2022, pp. 1752–1757.

[18] Y. Cao *et.al.*, "An auxiliary tool for preliminary tests of skin cancer: A self-modifying meta-learning method," in *Proc. Int. Conf. Big Data Anal. Softw. Eng. (ICBASE)*, 2021, pp. 187–190.

[19] 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

[20] K. Mahajan *et.al.*, "Meta-DermDiagnosis: Few-Shot Skin Disease Identification using Meta-Learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2020, pp. 730–731.

[21] J. Zhang *et.al.*, "ST-MetaDiagnosis: Meta learning with Spatial Transform for Skin Disease," in *Proc. IEEE Int. Conf. Bioinform. Biomed. (BIBM)*, 2020, pp. 1075–1080.

[22] A. Esteva *et.al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[23] T. J. Brinker *et.al.*, "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task," *Eur. J. Cancer*, vol. 113, pp. 47–54, 2019.

[24] N. C. F. Codella *et.al.*, "Skin Lesion Analysis Toward Melanoma Detection: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)," in *Proc. IEEE Int. Symp. Biomed. Imaging (ISBI)*, 2018, pp. 168–172.

[25] J. L. Bolognia, J. V. Schaffer, and L. Cerroni, *Dermatology*, 4th ed. Philadelphia, PA, USA: Elsevier, 2017.

[26] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "Human-level concept learning through probabilistic program induction," *Science*, vol. 350, no. 6266, pp. 1332–1338, 2015.

[27] G. Litjens *et.al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[28] L. Yang *et.al.*, "Deep Learning for Long-Tailed Learning: A Survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9970–9992, 2022.

[29] A. S. Adamson and A. Smith, "Machine Learning and Health Care Disparities in Dermatology," *JAMA Dermatol.*, vol. 154, no. 11, pp. 1247–1248, 2018.

[30] A. Gomolin *et.al.*, "Artificial intelligence in dermatology: A systematic review," *J. Cutan. Med. Surg.*, vol. 24, no. 1, pp. 77–83, 2020.

[31] S. Haggenmüller *et.al.*, "Skin cancer classification using convolutional neural networks: systematic review," *J. Eur. Acad. Dermatol. Venereol.*, vol. 35, no. 11, pp. 2260–2271, 2021.

[32] Y. Wang *et.al.*, "Generalizing from a Few Examples: A Survey on Few-Shot Learning," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, 2020.

[33] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1126–1135.

[34] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4077–4087.

[35] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 1199–1208.

[36] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[37] A. Raghu, M. Raghu, S. Bengio, and O. Vinyals, "Rapid learning or feature reuse? towards understanding the effectiveness of MAML," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–24.

[38] A. Dosovitskiy *et.al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Rep-*

resent. (ICLR), 2021, pp. 1–22.
[39] E. J. Hu *et.al.*, "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022, pp. 1–26.
[40] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 15750–15758.
[41] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 9729–9738.
[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255.
[43] I. Goodfellow *et.al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, 2014, pp. 2672–2680.
[44] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele Univ. and Durham Univ., U.K., Tech. Rep. EBSE-2007-01, 2007.
[45] M. J. Page *et.al.*, "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," *BMJ*, vol. 372, p. n71, 2021.
[46] P. F. Whiting *et.al.*, "QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies," *Ann. Intern. Med.*, vol. 155, no. 8, pp. 529–536, 2011.
[47] O. Vinyals, C. Blundell, T. Lillicrap, and D. Wierstra, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems*, vol. 29, 2016, pp. 3630–3638.
[48] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proc. ICML Deep Learn. Workshop*, vol. 2, 2015, pp. 1–8.
[49] X. Tao, X. Hong, X. Chang, S. Dong, X. Wei, and Y. Gong, "Few-Shot Class-Incremental Learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 12183–12192.
[50] F. M. Castro, M. J. Marin-Jimenez, N. Guil, C. Schmid, and K. Alahari, "End-to-End Incremental Learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 233–248.
[51] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "iCaRL: Incremental Classifier and Representation Learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 2001–2010.
[52] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A Kernel Two-Sample Test," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, 2012.
[53] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 6626–6637.
[54] R. R. Selvaraju *et.al.*, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 618–626.
[55] T. B. Fitzpatrick, "Soleil et peau," *J. Med. Esthet.*, vol. 2, pp. 33–34, 1975.
[56] M. Groh *et.al.*, "Deep learning-aided decision support for diagnosis of skin disease across skin tones," *Nat. Med.*, vol. 30, no. 2, pp. 573–583, 2024.
[57] H. B. McMahan *et.al.*, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. Int. Conf. Artif. Intell. Statist. (AISTATS)*, 2017, pp. 1273–1282.
[58] J. Kawahara *et.al.*, "Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 2, pp. 538–546, 2019.