

Article

A Clinician-in-the-Loop Framework for Validating and Selecting Synthetic Paediatric Dermatology Images

Ali Tariq Nagi ^{1,*}, Chiara Bellatreccia ¹, Andrea Borghesi ^{1,*} , Arianna Dondi ² , Luca Pierantoni ² ,
Daniele Zama ² , Iria Neri ², Marcello Lanari ^{1,2} and Roberta Calegari ^{1,*} 

¹ Department of Computer Science and Engineering, University of Bologna, 40126 Bologna, Italy; chiara.bellatreccia2@unibo.it (C.B.); m.lanari@ausl.imola.bo.it (M.L.)

² IRCCS Azienda Ospedaliero-Universitaria di Bologna, 40138 Bologna, Italy; arianna.dondi2@unibo.it (A.D.); luca.pierantoni@aosp.bo.it (L.P.); daniele.zama@aosp.bo.it (D.Z.); iria.neri@unibo.it (I.N.)

* Correspondence: alitariq.nagi@studio.unibo.it (A.T.N.); andrea.borghesi3@unibo.it (A.B.); roberta.calegari@unibo.it (R.C.)

Abstract

Synthetic data are increasingly proposed as a strategy for addressing data scarcity and representation imbalance in medical AI, particularly for paediatric populations and darker skin tones. However, visually plausible synthetic images may still contain clinically implausible features or fairness-relevant inconsistencies that are not adequately captured by automatic image-quality metrics. In this study, we present and empirically evaluate a clinician-guided framework for validating and selecting synthetic paediatric dermatology images. The framework combines a clinician-facing evaluation platform with structured assessments of visual realism, mask quality, diagnostic plausibility, confidence, and skin-tone relevance. Four clinicians with complementary expertise in paediatrics and dermatology completed 282 assessments of 93 real and synthetic images. Synthetic images were often rated as visually realistic but showed lower inter-rater agreement and weaker mask-quality assessments than real images. Clinician realism and confidence ratings were then used to divide 30 synthetic images into 18 approved and 12 non-approved images. To assess downstream utility, we compared a real-only ResNet50 classifier with classifiers augmented using all synthetic images, clinician-approved synthetic images, or non-approved synthetic images. Across three patient-level experimental splits, the clinician-approved condition achieved the strongest overall classification performance and the largest gains for the under-represented Dark-Skin subgroup. Because the Dark-Skin subgroup contained only seven patients and the synthetic subsets differed in size and disease composition, these fairness results should be interpreted as exploratory. The present study therefore provides evidence for clinician-guided validation and data curation rather than for a completed iterative generator-retraining process. Future work will evaluate whether clinician feedback can also support repeated generative-model refinement in larger, multi-centre datasets.



Academic Editor: Roser Morante

Received: 16 June 2026

Revised: 27 July 2026

Accepted: 30 July 2026

Published: 1 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

Keywords: synthetic data; human-in-the-loop AI; pediatric dermatology; fairness-aware machine learning; medical image generation; clinical validation; healthcare AI; skin tone bias; generative AI; medical image classification

1. Introduction

Over the past decade, the use of AI models in medical imaging for computer vision tasks, including disease classification, has increased substantially [1]. AI systems have shown high potential to enhance disease diagnosis across multiple imaging modalities [2–5].

Dermatology is one of the domains where AI could offer substantial benefits [6–9]; however, training effective models requires large and diverse datasets. Acquiring such datasets is challenging due to the limited availability of images for rare diseases and the under-representation of specific populations, for example, along dimensions such as ethnicity and skin tone, as well as issues related to image quality, privacy, and ethics [1,10]. A proposed solution to data scarcity in the medical domain is the generation of synthetic imaging data using generative AI [1,11,12]. Recent work has explored the use of generative models to produce synthetic MRI, retinal, and other medical images [1,13–15]. Such synthetic images can be used for data augmentation in downstream classification tasks. However, performance remains limited by the difficulty of faithfully capturing both local and global clinical and biological features, their co-appearance patterns, and by the lack of robust qualitative and quantitative metrics to audit the clinical plausibility of synthetic images [1]. In addition, the use of synthetic medical imaging data in healthcare applications is hindered by concerns about bias amplification introduced by synthetic samples and by the absence of reliable methods to assess whether generated images preserve clinically meaningful patterns and temporal correlations [1,10]. Prior research has mainly relied on automatic measures such as FID, precision, and recall [16], which do not encode clinical knowledge and therefore fail to diagnose clinical implausibility or detect unrealistic, non-existing correlations between image features [1]. These metrics have also been shown to be misaligned with human perception in text-to-image generation [16]. Together, these limitations point to the need for human-in-the-loop evaluation methods that integrate clinician expertise into the synthetic image generation pipeline, both to assess clinical realism and to guide the refinement of generative models [1,17].

This paper proposes and empirically evaluates a clinician-guided framework for the validation and selection of synthetic paediatric dermatology images. The framework defines fairness-relevant subgroups and clinical plausibility criteria, provides a clinician-facing platform for structured evaluation of real and synthetic images across multiple clinically relevant dimensions (including visual realism, diagnostic plausibility, mask quality, and skin-tone representation), and specifies procedures for structured data collection, inter-rater agreement analysis, and transparent reporting. Within this framework, clinician feedback is treated as an integral component of synthetic-data curation rather than as a one-off validation step. The dataset and its preprocessing pipeline have been previously described in [18], where clinical images collected in a hospital setting are refined through patch extraction and masking procedures to isolate disease-relevant regions. We operationalise the framework in a pilot study involving paediatric dermatologists, analysing rating distributions, inter-rater agreement, confidence scores, and qualitative feedback to identify domain-specific failure modes and fairness-relevant ambiguities that are not captured by automated metrics. To assess the practical utility of clinician-guided validation, we evaluate whether clinician-approved synthetic images improve downstream dermatology classification and subgroup fairness compared with both real-only training and unfiltered synthetic augmentation. Although the broader framework also supports iterative refinement of the generative model through clinician feedback, the present study is limited to the empirical evaluation of the clinician-validation and synthetic-data selection stages, without performing generator retraining or multi-cycle optimisation.

Before proceeding, we briefly outline the structure of the paper. Section 2 reviews the background and related work, highlighting the research gap addressed by this study. Section 3 presents the proposed methodological framework. Section 4 reports the results of the clinician evaluation and downstream experiments. Section 5 discusses the findings, outlines the study limitations, and identifies directions for future work. Finally, Section 5 concludes the paper.

Research Questions

This study is guided by the following research questions (RQs):

- **RQ1:** To what extent do paediatric dermatologists perceive synthetic paediatric dermatology images as visually realistic, and how do these assessments compare with those of real images?
- **RQ2:** What patterns of clinician disagreement, uncertainty, and mask-quality deficiencies emerge from the structured evaluation of real and synthetic images?
- **RQ3:** Does training with clinician-approved synthetic images improve downstream dermatology classification performance compared with real-only training and training with unfiltered synthetic images?
- **RQ4:** Does the use of clinician-approved synthetic images reduce performance disparities between Fair-Skin and Dark-Skin subgroups in downstream dermatology classification?

For RQ4, performance-based fairness is assessed using subgroup gaps in accuracy, balanced accuracy, and macro F1, whereas complementary parity-based fairness is evaluated using disparate impact (DI), equalised odds ratio (EOR), and predictive rate ratio (PRR). Together, these research questions define the scope of the empirical evaluation, which focuses on clinician-guided validation, synthetic-image selection, downstream classification performance, and subgroup fairness.

Synthetic data generation has emerged as a promising strategy for addressing data scarcity, class imbalance, and privacy constraints in medical imaging. A recent review by Koetzier et al. [1] highlights the growing use of generative models across multiple imaging modalities and reports that synthetic images can improve downstream diagnostic performance when used for data augmentation. At the same time, the review emphasises that widely adopted automatic quality measures, such as FID, precision, and recall, primarily assess visual fidelity and diversity rather than clinical realism. Consequently, synthetic images that appear visually convincing may still contain anatomically implausible structures or spurious correlations capable of misleading both clinicians and learning algorithms. The authors therefore advocate the development of rigorous human-in-the-loop evaluation protocols that integrate clinical expertise throughout the synthetic-data lifecycle. Recent advances have increasingly incorporated clinician feedback into the optimisation of generative medical-imaging models. Wang et al. [19] introduced MINIM, a multimodal medical image-text foundation model that combines clinician assessment with reinforcement learning from human feedback (RLHF). Expert ratings are used to train a reward model that iteratively improves image quality, resulting in progressive gains in both perceived realism and downstream diagnostic performance across multiple imaging modalities. Although this work demonstrates the effectiveness of clinician-guided optimisation, the evaluation primarily focuses on image quality and image-report consistency, without analysing inter-rater disagreement, clinician confidence, fairness, or task-specific clinical attributes. Fairness has also become an important objective in synthetic medical-image generation. Ktena et al. [20] showed that conditional diffusion models can improve robustness and reduce subgroup performance disparities under distribution shift by conditioning image generation on diagnostic and demographic attributes. Similarly, Luo et al. [21] proposed FairDiffusion, a fairness-aware latent diffusion framework that explicitly optimises demographic equity during image generation through group-specific Bayesian perturbation strategies [21]. Both studies demonstrate that fairness can be addressed either through subgroup-aware data augmentation or directly within the generative process itself. However, their evaluations rely predominantly on quantitative performance and fairness metrics, without incorporating structured multi-clinician assessment, analysis of inter-rater agreement and confidence, or clinician-guided selection of individual synthetic images. Taken together, these studies

demonstrate significant progress toward clinically informed and fairness-aware synthetic medical-image generation. Nevertheless, current approaches primarily focus on generator optimisation, demographic conditioning, or downstream predictive performance. Comparatively less attention has been devoted to structured clinician-centred validation workflows that explicitly characterise agreement, uncertainty, and clinical plausibility before synthetic images are incorporated into downstream learning pipelines.

Closer to dermatology, Luschi et al. [10] proposed a framework for evaluating GAN-generated dermatoscopic images using a combination of quantitative image-quality metrics (FID, KID, precision, and recall) and expert assessment by dermatologists. Their results showed that highly realistic synthetic images are often difficult for specialists to distinguish from real ones, while inter-rater agreement remains relatively low. Although this study demonstrates the importance of complementing automated metrics with expert evaluation, it focuses primarily on image realism and origin discrimination, without systematically assessing clinical plausibility, lesion delineation, skin-tone representation, or downstream fairness. Within paediatric dermatology, Borghesi et al. [18] demonstrated that diffusion-based models can successfully generate realistic clinical skin images even from limited hospital datasets by combining lesion-centred patch extraction, segmentation masks, and disease-specific preprocessing. Their work established the technical feasibility of synthetic image generation under severe data scarcity but primarily addressed generative performance, leaving clinician-centred validation and fairness-aware evaluation largely unexplored. More recently, Wang et al. [22] introduced MAGIC, a clinician-guided framework for medically accurate skin-disease image generation. Rather than relying exclusively on automatic metrics, MAGIC incorporates dermatologist knowledge through structured disease-specific checklists describing lesion morphology, colour, texture, and anatomical location. These expert-derived criteria are subsequently transformed into optimisation signals for reward-based fine-tuning and direct preference optimisation, resulting in improved medical consistency and downstream classification performance. However, image-level evaluation during optimisation is largely AI-mediated, whereas dermatologist involvement is mainly limited to checklist design and final validation. The study does not explicitly analyse inter-clinician agreement, clinician confidence, paediatric populations, or fairness across skin-tone subgroups. A complementary expert-feedback paradigm was proposed by Sun et al. [17] for synthetic haematopathology image generation. Their pathologist-in-the-loop framework uses expert feedback to train a reward model that guides diffusion-model optimisation toward clinically plausible images. The resulting models improve both image quality and downstream predictive performance, demonstrating the value of integrating clinical expertise directly into the optimisation process. Nevertheless, clinician feedback is primarily binary, while inter-rater disagreement, confidence, segmentation quality, and subgroup fairness are not explicitly investigated. Collectively, these studies demonstrate that clinician expertise can substantially improve the quality and clinical usefulness of synthetic medical images, either through expert validation or feedback-guided optimisation. However, existing dermatology-oriented approaches largely emphasise generator refinement or medical realism, whereas comparatively less attention has been devoted to structured, multidimensional clinician evaluation capable of quantifying agreement, confidence, clinical plausibility, and fairness before synthetic images are incorporated into downstream learning pipelines.

Beyond the clinical domain, a growing body of work has addressed the methodological foundations of human evaluation for generative AI systems. Otani et al. [16] argue that human evaluation of text-to-image models should follow transparent and reproducible protocols, emphasising explicit reporting of experimental settings, annotator characteristics, sample selection, and inter-rater agreement. Their framework highlights

that automated image-quality metrics alone are insufficient to assess perceptual quality and advocates multidimensional human evaluation supported by rigorous experimental design. Although developed for general text-to-image generation, these recommendations are directly relevant to safety-critical medical applications, where clinical expertise represents an essential source of validation. Complementing this perspective, Schuff et al. [23] provide practical guidelines for designing and analysing human evaluation studies, covering research-question formulation, study design, questionnaire construction, statistical analysis, and reporting practices. While their framework targets NLP and human–computer interaction, its emphasis on reliable study design, appropriate statistical analysis, and transparent reporting provides methodological principles that readily transfer to clinician-based evaluation of medical images. Neither framework, however, addresses the specific challenges of expert-only clinical studies, where evaluations involve small numbers of highly specialised clinicians assessing multidimensional constructs such as diagnostic plausibility, lesion delineation, and skin-tone representation. These methodological developments also reflect a broader shift towards data governance and accountable AI. Recent work increasingly argues that synthetic medical data should be evaluated not only in terms of visual fidelity or downstream predictive performance but also through transparent documentation, structured expert involvement, and explicit quality-control procedures throughout the data lifecycle [1,16,17,23]. Within this perspective, clinician participation becomes a core component of synthetic-data curation rather than a post hoc validation step, supporting more transparent, accountable, and fairness-aware development of medical AI systems.

Taken together, recent advances have substantially improved synthetic medical-image generation by introducing clinician-guided optimisation, fairness-aware generative models, dermatology-specific image synthesis, and rigorous methodologies for human evaluation. However, these contributions have largely addressed individual aspects of the problem in isolation. Existing studies typically focus on generator optimisation, downstream predictive performance, or expert validation but rarely integrate these components into a unified clinician-centred evaluation framework. To the best of our knowledge, no previous work combines, within a paediatric dermatology setting, repeated blinded evaluation by multiple clinicians; multidimensional assessment of clinical realism, diagnostic plausibility, lesion delineation, and skin-tone representation; explicit analysis of inter-rater agreement and clinician confidence; clinician-guided selection of synthetic images; and downstream evaluation of classification performance and subgroup fairness. These limitations are particularly relevant in paediatric dermatology, where limited data availability, under-representation of darker skin tones, and the need for specialist expertise make both synthetic-data validation and fairness assessment especially challenging. This work addresses these gaps by proposing and empirically evaluating a clinician-in-the-loop framework that integrates structured clinician evaluation with synthetic-image curation and downstream fairness analysis. Rather than focusing on iterative optimisation of the generative model, the proposed framework investigates how clinician-guided validation and image selection can improve the quality and practical utility of synthetic data for fairness-sensitive paediatric dermatology applications.

2. Materials and Methods

2.1. Methodological Framework

In this section, we present our methodological framework for clinician-in-the-loop evaluation of synthetic paediatric dermatology images and describe how it is instantiated in a dedicated web platform. Our aim is not to introduce yet another software stack but to make explicit the design principles, evaluation tasks, and data flows that operationalise a

co-designed, human-in-the-loop methodology for synthetic data in a safety-critical clinical domain. Methodologically, our work is grounded in two strands of literature. First, we build on proposals for verifiable and reproducible human evaluation of generative models, which highlight the limitations of small-scale, poorly documented user studies and call for explicit protocols, reliability reporting, and transparent sampling strategies for text-to-image systems [16]. Second, we draw from best practices in human evaluation and user studies in NLP [23], which emphasise clear research questions, careful design of rating scales and experimental conditions, and appropriate statistical analysis and reporting. We adapt and extend these principles to the specific constraints of clinician-based evaluation in paediatric dermatology and to fairness-relevant aspects such as Fitzpatrick skin-type representation. We also connect to recent work on clinician validation and feedback-guided optimisation of synthetic medical images across both general medical-imaging and domain-specific settings [10,12,17]. While these studies demonstrate the value of expert-informed evaluation and generator refinement, their assessment protocols commonly rely on broad, binary, or AI-mediated notions of clinical plausibility. Concretely, our framework extends the methodological proposals by Otani et al. [16] and Schuff et al. [23] in three key ways. First, it specialises their largely domain-agnostic guidelines to an expert-only, safety-critical clinical setting with paediatric dermatologists. Second, it decomposes human evaluation into task-specific modules that jointly cover clinical plausibility, diagnosis, and fairness-relevant dimensions such as skin-tone representation. Third, it specifies how human feedback can be linked to iterative refinement decisions in the synthetic data pipeline, rather than treating expert judgements as a one-off validation step. In the present study, we empirically evaluate the clinician-facing validation component of this broader framework, while downstream iterative refinement remains specified at the methodological level and is deferred to future empirical work.

The dataset used in this study originates from clinical images collected at Sant'Orsola Hospital and has been previously described in [18]. As detailed in prior work, the raw dataset consists of hospital-acquired images that undergo multiple stages of preprocessing, including manual cleaning, patch extraction, and mask generation. In particular, variable-sized images are transformed into fixed-size patches focusing on disease-relevant skin regions, while segmentation masks are generated to identify pathological areas. This preprocessing pipeline is essential for enabling both synthetic image generation and clinician-based evaluation. While this preprocessing pipeline enables effective model training and evaluation, it may also introduce acquisition-related artefacts, selection biases, and structural correlations tied to imaging conditions or patch extraction strategies. These factors are particularly relevant in fairness-sensitive settings and are therefore explicitly considered in our analysis and discussion (Section 4).

2.2. Synthetic Image Generation Pipeline

The synthetic images evaluated in this study were generated using the DreamBooth–Stable Diffusion pipeline previously described in [24]. As the present work focuses on clinician-guided validation and downstream use of synthetic images rather than on generative-model development, only the main steps of the generation pipeline are summarised here. The source dataset comprised approximately 8000 non-dermoscopic clinical images from 273 paediatric patients collected at IRCCS Azienda Ospedaliero–Universitaria di Bologna. The dataset covered nine skin-disease categories and was acquired under routine clinical conditions using consumer-grade cameras, resulting in substantial variability in image size, illumination, focus, acquisition angle, anatomical coverage, and lesion visibility [24]. Prior to model training, images were processed using a lesion-centred patch-extraction pipeline. A sliding-window procedure generated fixed-size 256×256 -pixel

patches, which were filtered according to lesion coverage and visual quality using the corresponding segmentation masks [18]. Redundant overlapping patches were removed through non-maximum suppression, producing standardised crops centred on clinically relevant lesion regions. Skin tone was estimated using a modified individual typology angle (ITA) procedure. Lesion pixels were excluded from the computation using the segmentation masks, and the resulting ITA values were clustered with a six-component Gaussian mixture model, yielding the categories dark, brown, tan, intermediate, light, and very light. These labels were used to identify under-represented skin-tone groups and construct the generative training subsets. Synthetic generation focused on maculopapular exanthema, viral exanthema, and scabies, as these were the only disease categories represented in both the dark and brown skin-tone groups. For each disease–skin-tone combination, a dedicated mini-dataset was manually assembled, resulting in six training subsets containing between 14 and 29 representative image crops after removal of redundant samples from the same patient. Each mini dataset was used to fine-tune a pretrained stable diffusion model using DreamBooth, which associates a small image collection with a unique textual identifier while preserving the generative capabilities of the base diffusion model. Fine-tuning employed two instance prompts (“<unique_ID>” and “<unique_ID> human skin”), while a grid search explored learning rates (5×10^{-7} to 1×10^{-5}) and training steps (1000–5000, depending on dataset size). A batch size of one was adopted throughout training. For each disease and skin-tone subset, candidate models were selected through qualitative inspection of the generated images, considering output diversity, disease-specific realism, preservation of skin colour and texture, and fidelity to the corresponding real images. The selected models were subsequently used to generate disease- and skin-tone-specific synthetic image pools by varying both the hyperparameter configuration and the random generation seed to maximise output diversity. The resulting synthetic image pools formed the basis of the clinician evaluation reported in this work. Thirty synthetic images were included in the blinded clinician study. Importantly, clinician feedback was not used to retrain or optimise the DreamBooth models; instead, it was used exclusively to evaluate clinical plausibility, analyse agreement and confidence, and partition the synthetic images into clinician-approved and non-approved subsets for the downstream classification experiments. The main components and configuration of the synthetic image generation pipeline are summarised in Table 1.

2.3. Datasets and Experimental Splits

The complete dataset comprised approximately 8000 non-dermoscopic clinical images from 273 paediatric patients spanning nine skin-disease categories. Different subsets of this collection were used for generative-model fine-tuning, clinician evaluation, and downstream classification. Specifically, we distinguish four datasets: (i) the complete clinical image collection; (ii) the disease- and skin-tone-specific subsets used for DreamBooth fine-tuning; (iii) the real and synthetic images included in the clinician evaluation; and (iv) the patient-level training, validation, and test sets used for downstream classification. The complete image collection included iatrogenic drug-induced exanthema, maculopapular exanthema, morbilliform exanthema, polymorphous exanthema, viral exanthema, urticaria, pediculosis, scabies, and chickenpox. Images were acquired under routine clinical conditions using consumer-grade cameras, resulting in substantial variability in illumination, focus, image size, acquisition angle, anatomical coverage, and lesion visibility. Generative-model training was restricted to maculopapular exanthema, scabies, and viral exanthema, as these were the only diseases represented in both the dark and brown skin-tone groups. Separate mini datasets were created for each disease–skin-tone combination, yielding six training subsets containing between 14 and 29 representative image crops. The same three

disease categories were subsequently used in the downstream classification experiments to ensure consistency between synthetic-image generation, clinician evaluation, and model training. The clinician evaluation comprised 93 unique images, including 63 real and 30 synthetic samples, and generated a total of 282 independent image-level assessments. The real-image subset contained seven images from each of the nine disease categories, whereas the synthetic subset contained ten images from each of the three target diseases. Of the synthetic images, 19 represented the dark skin-tone group and 11 the brown skin-tone group. Four clinicians with complementary expertise in paediatrics and dermatology participated in the evaluation. Each image was independently assessed by between two and four clinicians: 13 images received two assessments, 64 received three assessments, and 16 received four assessments, resulting in the 282 completed evaluations analysed in this study. The composition of the clinician-evaluation dataset is summarised in Table 2.

Table 1. Summary of the synthetic image generation pipeline.

Component	Configuration
Source data	Approximately 8000 images from 273 paediatric patients.
Image type	Non-dermoscopic, consumer-grade clinical photographs.
Preprocessing	Sliding-window lesion-centred patch extraction.
Patch size	256 × 256 pixels.
Patch selection	Disease coverage, contrast filtering, and non-maximum suppression.
Skin-tone estimation	Modified individual typology angle (ITA) computation excluding lesion pixels.
Skin-tone grouping	Six-component Gaussian mixture model.
Target diseases	Maculopapular exanthema, viral exanthema, and scabies.
Target skin-tone groups	Dark and brown.
Generative training sets	Six mini datasets containing 14–29 crops each.
Generative model	Stable diffusion fine-tuned using DreamBooth.
Base generation resolution	512 × 512 pixels.
Learning rates tested	5×10^{-7} , 2×10^{-6} , 5×10^{-6} , and 1×10^{-5} .
Training steps tested	1000–5000, depending on mini dataset size.
Batch size	1.
Instance prompts	<unique_ID> and <unique_ID> human skin.
Model selection	Qualitative assessment of diversity, disease-specific accuracy, skin texture, colour, and fidelity.
Diversity strategy	Use of multiple selected models and varied random seeds.
Images evaluated by clinicians	30.
Use of clinician feedback	Post-generation validation and image selection only.

Table 2. Composition of the clinician-evaluation dataset. Rating records refer to completed individual clinician assessments.

Disease	Real Images	Synthetic Images	Total Images	Ratings
Maculopapular exanthema	7	10	17	50
Scabies	7	10	17	54
Viral exanthema	7	10	17	50
Chickenpox	7	0	7	21
Iatrogenic drug-induced exanthema	7	0	7	20
Morbilloform exanthema	7	0	7	24
Pediculosis	7	0	7	21
Polymorphous exanthema	7	0	7	21
Urticaria	7	0	7	21
Total	63	30	93	282

The synthetic evaluation set was balanced across disease categories, with ten images per disease, but was not balanced across intended skin-tone categories. In particular, all ten evaluated scabies images were generated for the dark skin-tone category. The disease- and skin-tone-specific composition should therefore be considered when interpreting

comparisons across conditions. The disease and skin-tone-specific composition of the synthetic evaluation set is shown in Table 3.

Table 3. Disease and intended skin-tone distribution of the synthetic images included in the clinician evaluation.

Disease	Dark	Brown	Total
Maculopapular exanthema	6	4	10
Scabies	10	0	10
Viral exanthema	3	7	10
Total	19	11	30

2.4. Conceptual Clinician-in-the-Loop Framework

Figure 1 provides a high-level view of our conceptual framework. The pipeline can be read as a cycle that starts from jointly defining clinical and fairness objectives and ends with informed changes to the synthetic data pipeline. The full loop in Figure 1 represents the broader conceptual framework; the empirical evaluation reported in this paper focuses on its validation-stage component rather than on a completed multi-iteration refinement cycle.

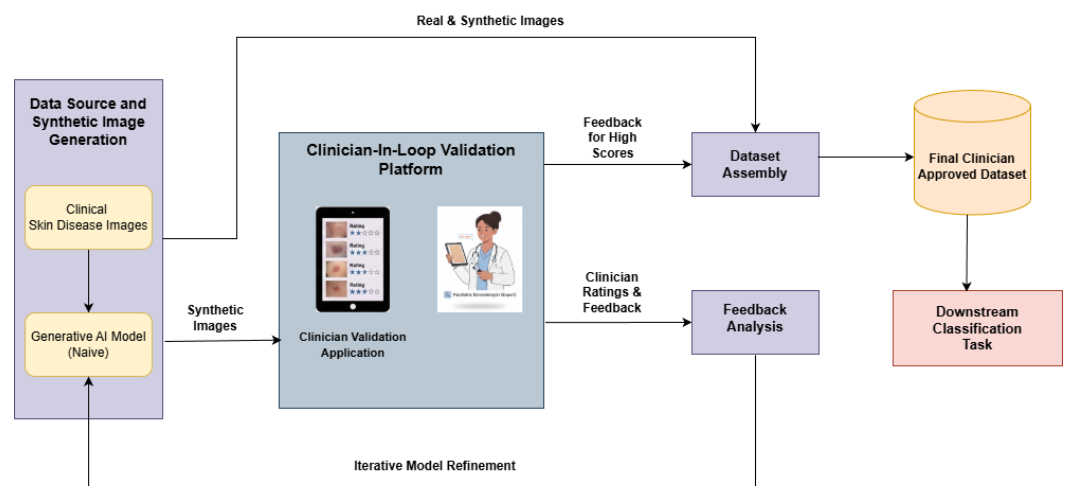


Figure 1. Conceptual clinician-in-the-loop framework for synthetic paediatric dermatology. This study evaluates the clinician-validation, feedback-aggregation, synthetic-image selection, dataset assembly, and downstream classification stages. The iterative model-refinement loop represents the broader framework and is left for future work.

We begin with a scoping phase in which paediatric dermatologists and researchers identify relevant diseases, patient subgroups, and fairness-relevant dimensions such as Fitzpatrick skin type. In this phase, we also co-define what counts as a *clinically plausible* and *ethically acceptable* synthetic image in this domain, clarifying which types of artefacts, anatomical inconsistencies, or combinations of attributes should be considered unacceptable. On this basis, a generative model produces candidate synthetic images conditioned on disease and fairness-relevant attributes, with a particular focus on under-represented subgroups. In parallel, real images and associated masks are pre-processed into crops following a patch extraction and mask generation pipeline that expose the lesion and surrounding skin as detailed in the study [18].

The implemented DreamBooth–Stable Diffusion generation procedure is described in Section 2.2. The clinician-in-the-loop framework itself is model-agnostic, so that its validation, feedback, and data-curation components can be applied to other generative architectures.

Clinicians then evaluate real and synthetic images through a set of structured tasks implemented in our web platform. Each task elicits multi-dimensional ratings (for example, realism, crop quality, Fitzpatrick plausibility), categorical labels (such as diagnosis or error tags), and optional free-text comments. Images are sampled at random and can be presented to multiple clinicians, allowing us to estimate inter-rater agreement and to analyse disagreement patterns, in line with recent methodological recommendations [16,23]. The collected feedback is aggregated and analysed to derive an error taxonomy that captures typical failure modes of synthetic images (for example, unrealistic skin tones for specific Fitzpatrick types, anatomically inconsistent lesions, misleading crop boundaries). We relate plausibility scores and error categories to fairness-relevant subgroups, identifying where synthetic images systematically underperform or misrepresent certain populations. Finally, the aggregated signals are fed back into the synthetic data pipeline: low-plausibility or ethically problematic images can be filtered out, conditioning strategies can be adjusted to better target underserved subgroups, and pre-processing steps on real images (such as cropping or masking) can be revised. In this way, human evaluation is not a one-off ‘sanity check’ but an integral part of an iterative refinement loop.

2.5. Operationalisation in the Clinician Evaluation Platform

The conceptual framework described above is operationalised through a web-based clinician evaluation platform. The platform is designed as a methodological instrument rather than as a technology showcase: its main purpose is to make it feasible, systematic, and reproducible to run human evaluation studies with paediatric dermatologists under realistic time and workload constraints. The platform embodies three design principles that directly reflect and extend methodological recommendations from the literature. First, we explicitly document who the evaluators are. Before participating, clinicians complete a short registration form capturing their paediatric dermatology expertise (specialisation, years of experience, affiliated hospital). This enables us to analyse how judgements vary with seniority and practice context and aligns with calls to report annotator characteristics in human evaluation [16,23]. Second, we decompose evaluation into a small set of well-defined tasks, each with task-specific questions and scales, rather than relying on a single global realism score. This responds to the observation that different notions of quality and plausibility are often conflated in prior work on synthetic medical images [10,17]. Third, we combine quantitative and qualitative feedback: for each task, clinicians provide ratings on Likert-style scales as well as optional free-text comments and error tags. This combination follows best practices for user studies [23] and is essential to uncover domain-specific failure modes that automatic metrics cannot detect.

Prior to the evaluation, clinicians were informed that they were participating in a research study focused on the assessment of paediatric dermatology images for methodological and quality evaluation purposes, including both real and synthetic samples. However, the provenance of individual images (i.e., whether real or synthetic) was not disclosed at the time of rating. Images were presented in randomised order via a web-based platform, and all evaluations were conducted independently to minimise potential sources of bias, including expectation and confirmation effects. Task instructions were deliberately framed around clinically relevant assessment criteria—such as realism, lesion delineation, diagnostic plausibility, and Fitzpatrick skin tone consistency—rather than on identifying whether an image was synthetic. This design choice was intended to ensure that clinician feedback reflected perceived clinical plausibility and utility, rather than an explicit attempt to detect synthetic artefacts. While this approach mitigates bias associated with awareness of image origin, residual bias cannot be entirely excluded, and this limitation should be considered when interpreting the results. Figure 2 shows three representative screens

from the platform, corresponding to crop and mask evaluation, synthetic batch evaluation with Fitzpatrick assessment, and diagnosis on real images. All screens follow a consistent pattern: a small number of focused questions, clearly labelled scales, and optional comment fields, so as to minimise cognitive load while preserving expressiveness.



Figure 2. Illustrative screenshots of the clinician evaluation platform for three of the tasks described in Section 2.6.

2.6. Evaluation Tasks and Sampling Strategy

Concretely, the platform organises the clinician's work into five evaluation tasks, which correspond to the five modules introduced in the system design. Each task targets a distinct aspect of clinical plausibility, disease recognisability, or fairness and produces structured signals that feed into aggregation and refinement. In what follows, we describe the tasks from a methodological perspective, rather than in terms of low-level user-interface details.

In the first task, clinicians assess the plausibility of real crop–mask pairs extracted from hospital images. Each screen presents a cropped image together with its associated segmentation mask, highlighting the lesion area. Clinicians judge whether the image and mask are of sufficient quality for clinical use, rate visual realism and the correctness of lesion delineation, and indicate their confidence. Optional comments allow them to point out typical segmentation errors, such as under- or over-segmentation or missing lesions. These judgements provide reference distributions for real data and reveal systematic issues in pre-processing that might compromise subsequent synthetic generation. The second task focuses on fairness-related aspects and disease recognisability in synthetic images. Clinicians are shown batches of synthetic crops generated for a target disease class. For each batch, they select the disease they believe is depicted (or “none of the above”), assess the Fitzpatrick skin type they perceive, and rate their confidence in both diagnosis and skin-type assessment. They can also comment on unrealistic tones, implausible lesion appearances, or inconsistencies across images in the batch. By repeatedly sampling batches across diseases, we obtain a plausibility-aware view of synthetic data quality and can detect systematic mismatches between intended and perceived skin-tone distributions. The third task elicits diagnostic judgements on real images from the hospital dataset. Here, the goal is to characterise inter-clinician variability and diagnostic difficulty on real cases. Images are sampled at random in batches; for each batch of images, clinicians select the most likely diagnosis from a predefined list and rate their confidence, with the option to comment on atypical or ambiguous presentations. These data later serve as a baseline when interpreting judgments on synthetic images and support the construction of an error taxonomy grounded in real-world diagnostic practice. In the fourth task, clinicians rate the quality of image crops in isolation. Each screen displays a single crop and asks

whether the lesion and relevant context are adequately visible, whether the technical quality (focus, lighting, absence of artefacts) is sufficient, and how confident the clinician is in the rating. Repeated sampling ensures that the same crop can be evaluated by multiple clinicians, enabling us to estimate agreement and to identify patterns of low-quality imagery across diseases and patient subgroups. Finally, the fifth task supports Fitzpatrick skin-type annotation of uncropped, hospital-sourced images. Clinicians can filter images by validation status or patient identifier and assign a Fitzpatrick skin type (I–VI), together with a confidence score and optional comments on borderline cases. These annotations provide a clinically grounded estimate of the skin-tone distribution in the underlying dataset, which is crucial for designing fairness-aware generation strategies and for analysing subgroup-specific patterns in the other tasks.

Across all tasks, images are sampled uniformly at random from predefined pools (for example, real crops, synthetic images by disease, full hospital images). The same image can be presented to multiple clinicians at different times, and the platform logs the full evaluation history for each image, including time stamps and a pseudonymised clinician identifier. This sampling and logging strategy directly addresses methodological concerns raised in prior work [16,23]: it enables us to compute inter-rater agreement, to study disagreement rather than discarding it, and to design analyses that respect the within-subject and between-subject structure of the data. Across all tasks, image provenance was concealed during evaluation, and clinicians were explicitly instructed to focus on clinical plausibility and task-specific criteria, rather than attempting to determine whether an image was real or synthetic.

The evaluation involved four clinicians with complementary expertise in paediatrics and dermatology. Each image was independently assessed by between two and four clinicians, with images presented in random order and with provenance blinded at rating time. Altogether, 13 images received two assessments, 64 received three assessments, and 16 received four assessments, resulting in 282 completed image-level evaluations. The numbers of completed assessments were unequal across clinicians: two clinicians completed 93 assessments each, one completed 78 assessments, and one completed 18 assessments.

For each image, Likert-scale ratings [16,23,25,26] (e.g., realism, mask accuracy, diagnostic confidence) were aggregated using the median to obtain a single image-level score, reflecting the ordinal nature of the data. Binary judgements (real versus generated) were instead aggregated via majority voting. Inter-rater reliability was assessed using ordinal Krippendorff's α [16,27], computed separately for each evaluation dimension and further stratified by disease category and image type (real vs. synthetic). Disease-level summaries were subsequently derived from these aggregated image-level scores. Importantly, inter-clinician disagreement was not treated as noise or annotation error to be eliminated. Rather, it was explicitly retained and analysed as a meaningful signal of epistemic uncertainty and clinical ambiguity. This perspective is operationalised through the examination of agreement patterns stratified by clinician confidence, enabling a more nuanced interpretation of variability in expert judgement.

2.7. From Raw Feedback to Pipeline Refinement

The last component of our framework concerns how the collected feedback is transformed into actionable insights for refining the synthetic data pipeline. Rather than treating ratings as opaque scores, we explicitly define how they are aggregated, how errors are classified, and how these results influence subsequent iterations. For each image (or batch of images), ratings are aggregated into plausibility scores along multiple dimensions, combining central tendencies with measures of dispersion and inter-rater agreement. This makes it possible to distinguish, for example, images that are consistently judged as moder-

ately plausible from images that trigger strong disagreement among clinicians. Free-text comments and error tags are analysed thematically to build an error taxonomy that captures recurring issues such as unrealistic skin tones in specific Fitzpatrick categories, anatomically inconsistent lesion shapes, misleading background artefacts, or inadequate cropping. These quantitative and qualitative summaries are then examined through a fairness lens by relating them to fairness-relevant subgroups such as skin tone and disease class. This allows us to identify where synthetic images systematically underperform or misrepresent certain populations and to prioritise corrective actions accordingly. Finally, the results feed back into the synthetic data pipeline in the form of concrete decisions: low-plausibility or ethically problematic images can be excluded from downstream use; generation prompts or conditioning variables can be modified to better target underserved subgroups; and pre-processing steps on real images can be revised where crop or mask quality is systematically low. By making this flow from raw clinician feedback to pipeline refinement explicit, and by grounding it in methodological guidance from the human evaluation literature, our framework turns the clinician evaluation platform into a central instrument for the design and governance of synthetic paediatric dermatology data, rather than a one-off validation tool.

Although the framework is iterative, it is not intended to operate as an open-ended loop. Convergence criteria can be defined in terms of the stabilisation of plausibility distributions across successive iterations, the improvement or stabilisation of inter-rater agreement trends, and retention decisions informed by clinician confidence. In particular, low-plausibility ratings provided with high confidence constitute strong signals for exclusion, whereas low-confidence ratings are treated as uncertain cases and may be retained for further inspection. In practice, the refinement process would be considered to have converged once newly generated batches no longer yield materially improved plausibility profiles or reductions in subgroup-specific failure modes. The current implementation supports clinician-facing data collection, structured rating workflows, logging, aggregation, and subgroup-aware analysis. These components are empirically evaluated in the present study. In contrast, downstream regeneration mechanisms, reward-based retraining, and multi-cycle optimisation are specified and supported at the framework level but were not empirically executed. As such, the present work should be understood as validating the evaluation and decision-making layer of the pipeline, while deferring full closed-loop optimisation to future work.

2.8. Downstream Classification and Fairness Evaluation

Although the complete source collection contained nine disease categories, the downstream classification experiment was restricted to maculopapular exanthema, scabies, and viral exanthema. These were the three diseases for which disease- and skin-tone-specific synthetic images had been generated and evaluated by clinicians. The downstream experiment therefore represents a three-class subset of the complete hospital dataset.

To evaluate whether clinician-guided validation translates into downstream predictive benefits, we conducted a classification experiment using a ResNet50 model pretrained on ImageNet. The downstream experiment was designed to isolate the contribution of clinician validation during synthetic-data selection. Within each patient-level split, all four experimental conditions used the same real-image training, validation, and test partitions together with identical network architecture, optimisation procedure, and training protocol. The only difference between conditions was the subset of synthetic images included during training. This design allows the contribution of clinician-guided synthetic-image selection to be evaluated independently of other experimental factors.

The classifier was trained under four conditions:

- **C0 Real only:** training using only real images.
- **C1 Real + all synthetic:** training using real images plus all evaluated synthetic images.
- **C2 Real + approved synthetic:** training using real images plus clinician-approved synthetic images.
- **C3 Real + non-approved synthetic:** training using real images plus synthetic images that did not satisfy the clinician-approval criteria.

C1 evaluates the effect of augmenting the training set with all evaluated synthetic images irrespective of clinician assessment. C2 evaluates whether restricting augmentation to clinician-approved synthetic images improves downstream classification performance and subgroup fairness. C3 provides a complementary comparison by using only synthetic images that did not satisfy the clinician-approval criteria.

Clinician approval was determined using the structured evaluations collected through the clinician-evaluation platform. Each synthetic image was independently evaluated by multiple paediatric dermatologists, who rated image realism and their confidence in that assessment using 7-point Likert scales. For each image, realism and confidence ratings were aggregated across clinicians by computing the *mean* score for each dimension. A synthetic image was classified as *clinician-approved* only when both its mean realism score and mean clinician-confidence score were at least 5. This criterion ensured that retained images were not only judged to be visually realistic but were also supported by relatively confident clinician assessments. Images satisfying both criteria formed the C2 synthetic subset, whereas all remaining evaluated synthetic images formed the C3 subset.

Overall, 30 synthetic images were evaluated by clinicians, of which 18 satisfied the clinician-approval criteria and 12 formed the non-approved subset (Table 4).

Table 4. Disease distribution of the evaluated synthetic images and the clinician-approved and non-approved subsets used in the downstream classification experiment.

Disease	C1: All	C2: Approved	C3: Non-Approved	Approval Rate
Maculopapular exanthema	10	6	4	60%
Scabies	10	4	6	40%
Viral exanthema	10	8	2	80%
Total	30	18	12	60%

The clinician-approval rate differed across disease categories, ranging from 40% for scabies to 80% for viral exanthema. Consequently, the C2 and C3 subsets differed not only in clinician-assessed quality but also in disease composition. The downstream comparison should therefore be interpreted as reflecting the combined effects of clinician-guided selection and synthetic subset composition. The four downstream classification conditions are summarised in Table 5.

Within each experimental run, all four conditions used the same real-image training, validation, and test partitions. The only difference between conditions was the synthetic subset added to the training data: no synthetic images in C0, 30 in C1, 18 in C2, and 12 in C3. Each experimental condition was repeated under three independent random seeds (42, 123, and 999). Within a given seed, C0–C3 used identical real-image training, validation, and test partitions and differed only in the synthetic subset added to training. Across seeds, the Fair-Skin patient partition was re-sampled, whereas the predefined Dark-Skin allocation remained fixed at three training patients, one validation patient, and three test patients.

Table 5. Summary of the synthetic-data training conditions used in the downstream classification experiments.

Condition	Synthetic Images	Subset Size	Purpose
C0	None	0	Real-image-only baseline.
C1	All evaluated synthetic images	30	Baseline synthetic augmentation without clinician filtering.
C2	Clinician-approved synthetic images	18	Evaluate whether clinician-guided selection improves classification performance and subgroup fairness.
C3	Non-approved synthetic images	12	Evaluate the contribution of synthetic images that did not satisfy the clinician-approval criteria.

ResNet50 Implementation and Training Details

Real images were partitioned at the patient level to prevent data leakage between the training, validation, and test sets. For the Fair-Skin subgroup, patients were randomly assigned to training, validation, and test partitions using proportions of 70%, 10%, and 20%, respectively. A relatively small validation partition was chosen to maximise the number of training images while retaining an independent set for model selection. Owing to the limited number of Dark-Skin patients (7 patients), a predefined patient-level split assigned three patients to training, one to validation, and three to testing. This predefined allocation was adopted to preserve patient independence while ensuring that training, validation, and test partitions each contained representative Dark-Skin cases.

No patient appeared in more than one partition. Synthetic images were included only in the training set, whereas the validation and test sets contained exclusively real images. This design ensures that synthetic images influence model training only, while model selection and final performance evaluation are performed exclusively on real clinical images. The same validation and test partitions were used for all four experimental conditions. Apart from the synthetic-image subset included during training, all experimental settings, preprocessing, optimisation parameters, validation procedures, and evaluation protocols were identical across the four conditions. The patient and real-image counts for the three experimental runs are reported in Table 6. The Fair-Skin and Dark-Skin subgroup counts across the three experimental runs are reported in Table 7.

The number of image patches substantially exceeded the number of patients because multiple lesion-centred patches could be extracted from an individual clinical photograph and multiple photographs could be available for the same patient. Nevertheless, all images and derived patches associated with a patient were assigned exclusively to one partition. Across the three runs, the real-image training set contained 82,739–94,378 images from 86–87 patients, the validation set contained 14,561–15,323 images from 12–13 patients, and the test set contained 18,937–31,200 images from 27–28 patients.

Table 6. Patient and real-image counts in the downstream classification partitions for the three experimental runs. Splitting was performed at the patient level. Within each run, the same real-image partitions were used for conditions C0–C3.

Run	Training		Validation		Test	
	Patients	Images	Patients	Images	Patients	Images
Run 1	87	92,295	12	14,561	28	21,782
Run 2	87	82,739	13	14,699	27	31,200
Run 3	86	94,378	13	15,323	28	18,937

Table 7. Fair-Skin and Dark-Skin patient and image counts in the downstream classification partitions.

Run	Subgroup	Training		Validation		Test	
		Patients	Images	Patients	Images	Patients	Images
Run 1	Dark-Skin	3	2470	1	38	3	1415
	Fair-Skin	84	89,825	11	14,523	25	20,367
Run 2	Dark-Skin	3	2470	1	38	3	1415
	Fair-Skin	84	80,269	12	14,661	24	29,785
Run 3	Dark-Skin	3	2470	1	38	3	1415
	Fair-Skin	83	91,908	12	15,285	25	17,522

The Dark-Skin partition remained fixed across all runs and comprised 2470 training images from three patients, 38 validation images from one patient, and 1415 test images from three patients. The Fair-Skin partition was re-sampled across seeds. It comprised 80,269–91,908 training images from 83–84 patients, 14,523–15,285 validation images from 11–12 patients, and 17,522–29,785 test images from 24–25 patients.

Synthetic images were included only in the training partition and did not contribute to checkpoint selection or final testing, both of which used exclusively real clinical images.

All images were loaded as three-channel RGB images and resized to 224×224 pixels. During training, online data augmentation consisted of random horizontal flipping with probability 0.5, random rotation within $\pm 10^\circ$, and colour jitter with brightness, contrast, saturation, and hue parameters of 0.1, 0.1, 0.1, and 0.02, respectively. Images were then converted to tensors and normalised using the ImageNet mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225]. Validation and test images underwent only resizing, tensor conversion, and ImageNet normalisation. The classifier was based on the Torchvision implementation of ResNet50 and was initialised with ImageNet-1K pretrained weights. The original classification layer was replaced by a task-specific head comprising a fully connected layer with 512 hidden units, a ReLU activation, dropout with probability 0.2, and a final fully connected layer with three output units. All parameters of the pretrained backbone and the classification head were updated during training (i.e., full-network fine-tuning). Models were trained using multiclass cross-entropy loss and stochastic gradient descent with an initial learning rate of 1×10^{-3} , momentum of 0.9, and weight decay of 5×10^{-4} . No class weighting, weighted sampling, or oversampling procedure was applied. A cosine-annealing learning-rate schedule with $T_{\max} = 100$ was used. Each model was trained for a fixed 100 epochs with a batch size of 32. Training samples were loaded as shuffled mini batches at every epoch, whereas validation and test samples were evaluated in deterministic order. After each epoch, model performance was evaluated on the real-image validation set. The checkpoint achieving the highest validation macro F1 score across the three disease classes was retained and subsequently evaluated on the held-out real-image test set. The test set was not used for checkpoint selection or hyperparameter optimisation. The random seeds were applied to Python, NumPy, PyTorch, and all available CUDA random-number generators, controlling model initialisation, mini batch ordering, and stochastic data augmentation.

Experiments were conducted on one node of the Booster partition of the LEONARDO supercomputer (BullSequana XH2135, supplied by ATOS), hosted by CINECA at the Bologna Technopole, Bologna, Italy [28]. Each Booster node is equipped with four NVIDIA A100 GPUs with 64 GB of memory per GPU; two GPUs and 64 GB of system memory were allocated for the present experiments. The software environment comprised Python 3.11.6, PyTorch 2.7.1, Torchvision 0.22.1, CUDA 12.6, and cuDNN 9.5.1. Data processing and evaluation were performed using scikit-learn 1.7.2, NumPy 2.3.5, pandas 2.3.2, and Pillow 11.3.0. Multi-GPU training was performed using PyTorch DataParallel.

The classification task included three disease categories: maculopapular exanthema, scabies, and viral exanthema. We evaluated overall performance using accuracy, balanced accuracy, and macro F1. Balanced accuracy was computed as the average recall across disease classes. Balanced accuracy is particularly appropriate for imbalanced classification problems because it assigns equal importance to each class, rather than being dominated by the majority classes [29]. Macro F1 was computed as the unweighted mean of the disease-wise F1 scores. Macro F1 is widely used in imbalanced multiclass classification because it gives equal importance to each class regardless of class frequency [29]. Macro F1 complements balanced accuracy by incorporating both precision and recall, thereby penalising models that achieve high recall at the expense of excessive false positives. To assess subgroup fairness, all metrics were additionally computed separately for Fair-Skin and Dark-Skin test subsets.

$$\text{Balanced Accuracy} = \frac{1}{K} \sum_{i=1}^K \text{Recall}_i = \frac{1}{K} \sum_{i=1}^K \frac{TP_i}{TP_i + FN_i} \quad (1)$$

$$\text{Macro F1} = \frac{1}{K} \sum_{i=1}^K F1_i \quad (2)$$

where K denotes the number of disease classes [29].

We also computed subgroup performance gaps between Fair-Skin and Dark-Skin groups:

$$\text{Accuracy Gap} = |\text{Accuracy}_{\text{Fair}} - \text{Accuracy}_{\text{Dark}}|, \quad (3)$$

$$\text{Balanced Accuracy Gap} = |\text{BA}_{\text{Fair}} - \text{BA}_{\text{Dark}}|, \quad (4)$$

$$\text{Macro F1 Gap} = |\text{Macro F1}_{\text{Fair}} - \text{Macro F1}_{\text{Dark}}|. \quad (5)$$

Lower values for these gap metrics indicate smaller subgroup performance disparities. These metrics can be interpreted as performance-based group fairness measures because they quantify differences in predictive performance between demographic subgroups. Recent reviews of clinical AI fairness classify fairness metrics based on subgroup differences in predictive performance (e.g., accuracy, sensitivity, specificity, and related performance measures) as performance-based group fairness metrics [30]. In addition to subgroup performance-gap metrics, we report disparate impact (DI), equalised odds ratio (EOR), and predictive rate ratio (PRR), which are widely used fairness metrics in algorithmic fairness evaluation [31,32]. These metrics have also been employed in recent fairness analyses for AI-based skin disease diagnosis under data scarcity and representation imbalance settings [24]. DI evaluates parity in positive prediction rates between groups, EOR evaluates parity in true-positive and false-positive rates, and PRR evaluates parity in positive predictive value between groups.

3. Results

This section introduces the empirical evaluation of the first and core module of the proposed clinician-in-the-loop framework. This module represents the critical fairness-sensitive component of the pipeline, as it directly governs the validation of synthetic images generated to address representational gaps in the original dataset, with a specific focus on under-represented skin tones. It is important to note that the evaluation is performed on preprocessed image patches derived from the original dataset [18], which may influence both clinician perception and the statistical properties of the data. By subjecting these generated images to expert clinical evaluation, this module operationalises fairness enhancement not as an abstract design goal but as an empirically testable property grounded in clinician judgement. Accordingly, the results reported here should be interpreted as

an empirical assessment of the framework's validation gate rather than as evidence of completed end-to-end iterative optimisation of the generator. Here, fairness is conceptualised as subgroup-aware validation of synthetic images generated for under-represented skin-tone categories. This perspective is grounded in prior work emphasising fairness across multiple subgroups and the challenges posed by imbalanced subgroup data [33,34]. Accordingly, fairness is addressed at the level of data curation and validation, rather than being framed as evidence of downstream predictive parity across populations.

The evaluation reported here is based on the assessments of four clinicians with complementary expertise in paediatrics and dermatology with high levels of expertise whose clinical experience ranges from 11 to 39 years. The clinician pool is limited in absolute size but reflects the practical constraints inherent to specialist-only paediatric dermatology evaluation, where expert assessment requires highly specialised training, and annotation is constrained by factors such as cost and reliance on scarce domain experts [10,16]. Consistent with this context, prior human-in-the-loop studies in the medical domain have often relied on small cohorts of expert annotators (e.g., 1, 3, 5, or 7 specialists) for the evaluation of synthetic images, with no established consensus on an optimal annotator pool size [17,35–37]. Within such settings, the key methodological consideration is not achieving crowd-scale annotation but rather determining whether repeated expert-only assessments can yield reliable and informative governance signals. Accordingly, the study is designed around repeated ratings, image-level multi-rater evaluation, inter-rater agreement analysis, and confidence stratification, rather than large annotator counts typical of general-purpose human evaluation. This design prioritises depth and consistency of expert judgement over breadth of annotation. The clinicians' evaluations focus on synthetic images generated for three selected diseases, chosen based on both their clinical relevance and the presence of under-represented skin-tone categories in the source data. By concentrating on these conditions, the analysis explicitly examines whether the generative model can produce clinically plausible images in regions of the data space where real-world coverage is limited. This directly addresses a central fairness concern in paediatric dermatology datasets, namely, the under-representation of certain skin-tone subgroups.

The remaining modules of the framework operate downstream of this evaluation component and are designed to consume its outputs—such as aggregated ratings, agreement measures, and identified failure modes—without modifying the fairness-relevant decision mechanisms assessed here. As a result, their behaviour does not affect the validity of the fairness analysis presented in this section. We therefore provide a complete and self-contained evaluation of the system at its critical fairness decision point while deferring the empirical assessment of downstream modules, which primarily concern integration and deployment, to future work.

Across the three diseases for which synthetic images were available (maculopapular exanthema, scabies, and viral exanthema), clinicians showed a marked tendency to accept generated images as real (Table 8). The false real rate (FRR, i.e., the probability that a medician mistakes a generated image as real) ranged from 0.42 to 0.50, meaning that approximately half of the individual evaluations of synthetic images resulted in a “real” judgement. Although confidence intervals are relatively wide, they consistently exclude very low values, indicating that a small number of outliers does not drive this effect but reflects a systematic difficulty in distinguishing synthetic from real images. At the same time, inter-rater agreement on the binary real vs. generated decision was low, with Krippendorff's α between 0.18 and 0.31 across these diseases. This combination of high FRR and low agreement suggests that clinicians not only struggle to reliably detect synthetic images but also disagree substantially about which specific images appear real.

In other words, synthetic images frequently “pass” as real for at least some experts but not in a consistent way across the clinician pool.

Table 8. False real rate (FRR) on synthetic images and inter-rater agreement (Krippendorff’s α) by disease. Each disease category included ten unique synthetic images.

Disease	FRR [95% CI]	Krippendorff’s α
Maculopapular exanthema	0.48 [0.28, 0.68]	0.22
Scabies	0.50 [0.31, 0.69]	0.31
Viral exanthema	0.42 [0.23, 0.64]	0.18

Disease-specific differences in FRR were relatively modest, ranging from 0.42 for viral exanthema to 0.50 for scabies (Table 8). Although scabies exhibited the highest tendency to be perceived as real and viral exanthema the lowest, the corresponding confidence intervals overlapped substantially, indicating that these differences should be interpreted as descriptive rather than conclusive. Overall, the results suggest that the difficulty of distinguishing synthetic from real images is broadly consistent across the three evaluated diseases rather than being driven by a single disease category.

Realism ratings on the 7-point Likert scale provide a more fine-grained view of this phenomenon (Table 9). For the three diseases with synthetic images, generated samples received a high proportion of ratings in the upper realism band (5–7), ranging from 52.6% to 66.7%, which is broadly comparable to the distributions observed for real images of the same diseases. This confirms that a substantial fraction of synthetic images is perceived as visually realistic. However, the associated agreement patterns differ markedly: for real images, realism ratings exhibit moderate agreement (Krippendorff’s $\alpha \approx 0.39$ –0.46), whereas for generated images, agreement drops to low values (Krippendorff’s $\alpha \approx 0.08$ –0.21). Thus, realism is not a uniform property of the synthetic images: some are judged highly realistic by most clinicians, while others elicit divergent opinions. For diseases without synthetic images, real-image realism follows a consistent pattern, with the majority of ratings in the 5–7 band and moderate agreement (Krippendorff’s $\alpha \approx 0.40$ –0.47), providing a stable reference for human perception on this dataset.

Table 9. Distribution of realism ratings (7-point Likert scale) and inter-rater agreement per disease and image type. Percentages indicate the proportion of ratings falling into low (1–3), neutral (4), and high (5–7) realism bands. Krippendorff’s α is computed on the ordinal scale.

Disease	Type	1–3 (%)	4 (%)	5–7 (%)	Krippendorff’s α
Maculopapular exanthema	Real	21.4	18.6	60.0	0.41
	Generated	19.0	14.3	66.7	0.12
Scabies	Real	27.3	20.5	52.3	0.46
	Generated	18.2	22.7	59.1	0.21
Viral exanthema	Real	25.0	19.4	55.6	0.39
	Generated	21.1	26.3	52.6	0.08
Urticaria	Real	24.6	21.7	53.6	0.44
Pediculosis	Real	29.8	18.4	51.8	0.42
Chickenpox	Real	23.1	20.0	56.9	0.47
Morbilliform exanthema	Real	26.4	17.6	56.0	0.45
Iatrogenic drug-induced exanthema	Real	28.9	19.3	51.8	0.43
Polymorphous exanthema	Real	27.1	20.8	52.1	0.40

The real-image categories that did not have synthetic counterparts (urticaria, pediculosis, chickenpox, morbilliform exanthema, iatrogenic drug-induced exanthema, and polymorphous exanthema) served as reference conditions for clinician evaluation. Across these

diseases, realism ratings remained consistently concentrated in the upper rating band (5–7), with moderate inter-rater agreement (Krippendorff's $\alpha \approx 0.40$ – 0.47), indicating stable clinician perception across the underlying clinical dataset. No disease emerged as a clear outlier in terms of realism ratings or agreement. These reference evaluations therefore provide a stable baseline against which the greater variability observed for synthetic images can be interpreted.

A similar qualitative pattern emerges for the evaluation of segmentation masks (Table 10). For real images, most mask accuracy ratings fall in the 5–7 band (55–66% across diseases) and agreement is again moderate (Krippendorff's $\alpha \approx 0.47$ – 0.54), indicating that clinicians broadly concur on the correctness of lesion delineations. In contrast, the masks associated with generated images are rated more critically: the share of high ratings (5–7) drops to 36.9–47.6%, with a corresponding increase in low and neutral ratings, and inter-rater agreement decreases to Krippendorff's $\alpha \approx 0.14$ – 0.22 . These results indicate that even when the underlying synthetic image can appear realistic, the quality of the associated segmentation may be perceived as less reliable and more contentious. From a methodological perspective, this illustrates the value of jointly evaluating visual realism and lesion delineation, as relying on image-level impressions alone would miss a clinically relevant weakness of the synthetic data pipeline.

Table 10. Distribution of mask accuracy ratings (7-point Likert scale) and inter-rater agreement per disease and image type. Percentages indicate the proportion of ratings falling into low (1–3), neutral (4), and high (5–7) accuracy bands. Krippendorff's α is computed on the ordinal scale.

Disease	Type	1–3 (%)	4 (%)	5–7 (%)	Krippendorff's α
Maculopapular exanthema	Real	18.6	23.3	58.1	0.48
	Generated	28.6	23.8	47.6	0.19
Scabies	Real	16.0	20.5	63.5	0.52
	Generated	31.8	27.3	40.9	0.22
Viral exanthema	Real	19.4	22.2	58.3	0.46
	Generated	36.8	26.3	36.9	0.14
Urticaria	Real	17.9	21.5	60.6	0.50
Pediculosis	Real	20.2	19.3	60.5	0.49
Chickenpox	Real	15.4	18.5	66.1	0.54
Morbilliform exanthema	Real	18.8	20.0	61.2	0.51
Iatrogenic drug-induced exanthema	Real	21.7	22.9	55.4	0.47
Polymorphous exanthema	Real	19.6	21.7	58.7	0.48

A similar pattern was observed for mask-quality assessments. Real images consistently received predominantly high ratings across all disease categories, with moderate inter-rater agreement ($\alpha \approx 0.47$ – 0.54). This stability suggests that the lower ratings and weaker agreement observed for synthetic images are unlikely to reflect disease-specific characteristics of the evaluation dataset alone but instead are associated with properties of the generated images and their corresponding masks.

Finally, we examined how inter-rater agreement varies with clinicians' self-reported confidence (Table 11). Agreement increases monotonically with confidence, from $\alpha = 0.17$ in the low-confidence band (1–3) to $\alpha = 0.44$ in the high-confidence band (5–7). This pattern suggests that disagreement is concentrated precisely where clinicians feel uncertain, rather than reflecting arbitrary or noisy use of the scales. High-confidence evaluations reach agreement levels comparable to those observed for real-image realism and mask accuracy, supporting the internal coherence of the clinician-in-the-loop evaluation. At the same time, the presence of many low- and neutral-confidence ratings, particularly for synthetic images

and masks, highlights where further refinement of the generative model and data curation is most urgently needed.

Table 11. Global inter-rater agreement (Krippendorff's α) stratified by self-reported confidence level. Confidence ratings are grouped into low (1–3), neutral (4), and high (5–7) bands. Agreement is computed across all diseases, image types, and evaluation tasks.

Confidence Band	Confidence Scale	Krippendorff's α
Low confidence	1–3	0.17
Neutral confidence	4	0.26
High confidence	5–7	0.44

Downstream Classification Results

Because the dataset is substantially imbalanced across disease categories, balanced accuracy and macro F1 provide a more informative assessment than standard accuracy alone [29]. Balanced accuracy evaluates average disease-wise recall, whereas macro F1 evaluates average disease-wise precision and recall, ensuring that all disease categories contribute equally to model evaluation regardless of class prevalence [29]. The overall test performance across the three experimental runs is reported in Table 12. Skin-tone-specific classification performance across the four training conditions is reported in Table 13, while the corresponding subgroup performance gaps and parity-based fairness metrics are summarised in Table 14.

Table 12. Overall test performance across three seeds. Values are mean $\pm$ standard deviation. Best values are shown in bold.

Condition	Accuracy	Balanced Accuracy	Macro F1
C0 Real only	0.613 $\pm$ 0.037	0.470 $\pm$ 0.019	0.462 $\pm$ 0.029
C1 Real + all synthetic	0.616 $\pm$ 0.043	0.472 $\pm$ 0.014	0.466 $\pm$ 0.028
C2 Real + approved synthetic	0.630 $\pm$ 0.035	0.485 $\pm$ 0.038	0.479 $\pm$ 0.050
C3 Real + non-approved synthetic	0.608 $\pm$ 0.047	0.475 $\pm$ 0.032	0.467 $\pm$ 0.046

Table 13. Skin-tone subgroup performance across three seeds. Values are mean $\pm$ standard deviation. Bold values indicate the best performance within each skin-tone subgroup for each metric.

Group	Condition	Accuracy	Balanced Accuracy	Macro F1
Fair Skin	C0 Real only	0.638 $\pm$ 0.051	0.482 $\pm$ 0.016	0.467 $\pm$ 0.035
Fair Skin	C1 Real + all synthetic	0.639 $\pm$ 0.055	0.471 $\pm$ 0.019	0.461 $\pm$ 0.043
Fair Skin	C2 Real + approved synthetic	0.650 $\pm$ 0.051	0.484 $\pm$ 0.043	0.476 $\pm$ 0.061
Fair Skin	C3 Real + non-approved synthetic	0.625 $\pm$ 0.056	0.479 $\pm$ 0.027	0.464 $\pm$ 0.053
Dark Skin	C0 Real only	0.237 $\pm$ 0.100	0.188 $\pm$ 0.058	0.229 $\pm$ 0.100
Dark Skin	C1 Real + all synthetic	0.287 $\pm$ 0.092	0.320 $\pm$ 0.034	0.286 $\pm$ 0.097
Dark Skin	C2 Real + approved synthetic	0.351 $\pm$ 0.140	0.348 $\pm$ 0.069	0.313 $\pm$ 0.128
Dark Skin	C3 Real + non-approved synthetic	0.353 $\pm$ 0.083	0.258 $\pm$ 0.043	0.297 $\pm$ 0.085

Table 14. Fairness metrics and subgroup performance gaps. Values are mean $\pm$ standard deviation. For gap metrics, lower is better. For DI and PRR, values closer to 1 are preferred. For EOR, higher/closer to 1 is preferred. Bold values indicate the best result for each metric according to these criteria.

Condition	Accuracy Gap	BA Gap	Macro F1 Gap	DI	EOR	PRR
C0 Real only	0.402 $\pm$ 0.151	0.294 $\pm$ 0.056	0.239 $\pm$ 0.110	1.027 $\pm$ 0.242	0.240 $\pm$ 0.070	4.709 $\pm$ 4.002
C1 Real + all synthetic	0.352 $\pm$ 0.142	0.151 $\pm$ 0.021	0.175 $\pm$ 0.126	1.249 $\pm$ 0.408	0.458 $\pm$ 0.101	13.181 $\pm$ 17.187
C2 Real + approved synthetic	0.299 $\pm$ 0.191	0.136 $\pm$ 0.061	0.163 $\pm$ 0.154	1.142 $\pm$ 0.353	0.483 $\pm$ 0.072	10.709 $\pm$ 14.504
C3 Real + non-approved synthetic	0.272 $\pm$ 0.110	0.221 $\pm$ 0.023	0.167 $\pm$ 0.098	1.117 $\pm$ 0.347	0.292 $\pm$ 0.053	10.412 $\pm$ 14.154

The subgroup gap metrics (accuracy gap, balanced accuracy gap, and macro F1 gap) correspond to performance-based group fairness measures. In contrast, DI, EOR, and PRR correspond to parity-based fairness metrics that evaluate equality of prediction behaviour across demographic groups [30–32]. Together, these complementary metrics provide insight into both predictive-performance disparities and prediction-parity disparities between Fair-Skin and Dark-Skin populations.

Disease-wise results show that the clinician-approved synthetic condition (C2) achieved the highest mean F1 score for all three disease categories (Table 15), suggesting that clinician-guided image selection consistently benefits downstream classification rather than improving only a single disease class. Across the three experimental runs, C2 also achieved the strongest overall classification performance, improving accuracy, balanced accuracy, and macro F1 compared with the real-only baseline (Table 12). The largest gains were observed for the under-represented Dark-Skin subgroup, where balanced accuracy and macro F1 increased substantially relative to C0. Although these improvements should be interpreted cautiously given the small subgroup size, they suggest that clinician-approved synthetic images provide disease-relevant information for the population most affected by the original representation imbalance. The fairness analysis showed a consistent reduction in subgroup performance disparities under C2. Compared with the real-only baseline, the balanced accuracy gap decreased by approximately 54% and the macro F1 gap by approximately 32%, while C2 achieved the highest equalised odds ratio among the evaluated conditions. These results indicate that clinician-guided synthetic-image selection is associated with both improved predictive performance and reduced subgroup performance disparities.

Table 15. Disease-wise overall F1 scores across three seeds. Values are mean $\pm$ standard deviation. Bold values indicate the highest mean F1 score for each disease.

Disease	C0 Real Only	C1 All Synthetic	C2 Approved Synthetic	C3 Non-Approved Synthetic
Maculopapular exanthema	0.132 $\pm$ 0.075	0.141 $\pm$ 0.094	0.149 $\pm$ 0.113	0.142 $\pm$ 0.105
Scabies	0.575 $\pm$ 0.131	0.563 $\pm$ 0.112	0.585 $\pm$ 0.140	0.581 $\pm$ 0.150
Viral exanthema	0.679 $\pm$ 0.070	0.693 $\pm$ 0.060	0.703 $\pm$ 0.026	0.677 $\pm$ 0.031

The parity-based fairness metrics require a more nuanced interpretation. Although disparate impact (DI) was closest to one under the real-only condition, this did not coincide with the strongest subgroup performance, highlighting that prediction-rate parity alone is insufficient to characterise fairness in clinical classification. Similarly, predictive rate ratio (PRR) exhibited high variability because precision estimates became unstable for disease–subgroup combinations with very small denominators, particularly for maculopapular exanthema. PRR is therefore reported for completeness but is not used to support the principal fairness conclusions.

4. Discussion

The findings should be interpreted within the scope of the present study, which evaluates clinician-guided validation and synthetic-image selection rather than iterative optimisation of the underlying generative model. Clinician feedback was used to assess the clinical plausibility of generated images and guide their selection for downstream classification but not to retrain the DreamBooth–Stable Diffusion model. Accordingly, the results provide evidence for clinician-guided validation as an effective data-curation mechanism rather than for a fully closed-loop optimisation process. Despite this scope, the study demonstrates that structured clinician evaluation provides information that cannot be captured by automated image-quality metrics alone. Although many synthetic images

were perceived as visually realistic, lower inter-rater agreement and variations in clinician confidence revealed clinically meaningful ambiguities that would otherwise remain hidden. These findings indicate that multidimensional expert assessment provides complementary signals for identifying synthetic images suitable for downstream use and for exposing failure modes requiring further investigation. More broadly, the proposed framework positions clinician evaluation as an integral component of synthetic-data governance rather than a post hoc quality-control step. Multidimensional ratings, inter-rater agreement, and subgroup-aware analyses support transparent retention and exclusion decisions, aligning with recent recommendations for human oversight, accountability, and trustworthy validation of synthetic medical data in safety-critical healthcare applications [1,38,39].

4.1. Comparison with Related Studies

Our findings reinforce previous evidence that visual realism alone is insufficient to establish the clinical reliability of synthetic dermatology images. Luschi et al. [10] similarly reported that highly realistic synthetic dermoscopic images are often difficult for dermatologists to distinguish from real ones despite limited inter-rater agreement. Consistent with these observations, our synthetic paediatric dermatology images frequently received high realism ratings while simultaneously exhibiting lower agreement than real images. Extending previous work, we show that evaluating additional dimensions—including mask quality, diagnostic plausibility, clinician confidence, and skin-tone representation—reveals clinically relevant ambiguities that are not captured by realism alone. Our results also support the value of multidimensional clinician evaluation over feedback strategies based primarily on global image quality or predefined clinical attributes. Previous approaches such as MINIM [19], MAGIC [22], and the framework of Sun et al. [17] demonstrate that expert feedback can successfully guide generative-model optimisation. However, these studies rely predominantly on image-quality assessment, binary expert feedback, or AI-mediated evaluation. In contrast, our framework combines repeated blinded evaluation by multiple paediatric dermatologists with explicit analysis of inter-rater agreement and clinician confidence. The observed discrepancy between high realism scores and weaker mask-quality assessments further suggests that visually convincing images may still fail to satisfy clinically relevant quality criteria. The downstream classification results are also consistent with recent studies reporting that synthetic data can improve predictive performance and subgroup fairness [20,21]. Unlike previous approaches, however, fairness in our study is not pursued through demographic conditioning or direct optimisation of the generative model. Instead, clinician feedback is used as a structured data-curation mechanism to identify synthetic images that are most appropriate for downstream learning. The improved balanced accuracy and macro F1 observed for the clinician-approved subset, particularly within the under-represented Dark-Skin subgroup, suggest that clinician-guided image selection can complement existing fairness-aware generative approaches. The comparison between the four training conditions further highlights the importance of clinician-guided data curation. Augmenting the training set with all available synthetic images produced only modest improvements, whereas restricting augmentation to clinician-approved images yielded the strongest overall performance. Although the approved and non-approved subsets differed in both size and composition, these findings indicate that the clinical quality of synthetic images is at least as important as their quantity. Consequently, clinician validation should be viewed not only as an evaluation procedure but also as a practical mechanism for improving the downstream utility of synthetic medical data. Overall, the proposed framework complements existing clinician-guided and fairness-aware generative approaches. Rather than focusing on iterative optimisation of the generator, it demonstrates that structured clinician evaluation, uncertainty-aware interpretation, and expert-guided synthetic-image

selection constitute effective data-curation strategies that improve the practical utility of synthetic images for fairness-sensitive paediatric dermatology applications.

4.2. Implications of the Downstream Fairness Evaluation

The downstream classification experiments demonstrate that clinician-guided synthetic-image selection can improve both predictive performance and subgroup fairness. Whereas the clinician evaluation identified differences in realism, mask quality, confidence, and agreement, the downstream analysis shows that these signals also have practical value for model development. In particular, restricting augmentation to clinician-approved synthetic images consistently outperformed both real-only training and indiscriminate synthetic augmentation, indicating that expert-guided data curation can improve the utility of synthetic datasets. The largest gains were observed for the under-represented Dark-Skin subgroup, which represents the primary fairness motivation of this study. Clinician-approved synthetic images substantially improved balanced accuracy and macro F1 for this subgroup while simultaneously reducing subgroup performance gaps. These findings are consistent with recent work emphasising performance-based fairness in clinical AI, where equitable predictive performance across demographic groups is often more clinically meaningful than prediction parity alone [24,30]. The comparison between the four training conditions further highlights the importance of synthetic-image quality. Including all evaluated synthetic images produced only modest improvements, whereas restricting augmentation to clinician-approved images yielded the strongest overall performance and the greatest reduction in subgroup disparities. Although the approved and non-approved subsets differed in both size and composition, these results suggest that careful clinician-guided filtering is more beneficial than indiscriminate synthetic-data augmentation. The complementary fairness metrics also require careful interpretation. While the clinician-approved condition achieved the strongest equalised odds ratio, the real-only baseline produced a disparate impact value closest to one despite substantially poorer performance on the Dark-Skin subgroup. This illustrates a well-recognised limitation of parity-based fairness metrics: similar prediction rates do not necessarily imply comparable diagnostic performance or clinical utility. Accordingly, the conclusions of this study are based primarily on subgroup performance and performance-gap metrics, whereas parity-based measures provide complementary information. Predictive rate ratio (PRR) is reported for completeness but is not used to support the main conclusions because its dependence on precision makes it unstable when subgroup precision values approach zero. Overall, these findings indicate that clinician validation can serve not only as a quality-assurance mechanism but also as an effective fairness-aware data-curation strategy. Although the present study does not implement iterative retraining of the generative model, it demonstrates that clinician feedback can be successfully translated into downstream improvements by guiding the selection of synthetic images used for model training. This validation-to-classification workflow provides a practical foundation for future clinician-in-the-loop pipelines in which expert feedback may also support iterative refinement of the generative model.

4.3. Study Limitations

Several limitations should be considered when interpreting the results of this study. First, the empirical evaluation was conducted using data collected at a single institution following a specific acquisition and preprocessing pipeline. Although patient-level splitting prevented data leakage, institution-specific imaging conditions, patch extraction, and mask-generation procedures may have introduced biases that were learned by both the generative and classification models. Consequently, the findings may not generalise to images acquired under different clinical protocols, imaging conditions, or patient popula-

tions. Validation on independent multi-centre paediatric dermatology datasets is therefore required. Second, the clinician evaluation involved four experts and a relatively small set of synthetic images (30 images from three disease categories). Although repeated independent assessments enabled the analysis of inter-rater agreement and clinician confidence, the limited numbers of clinicians, diseases, and evaluated images reduce statistical power and the generalisability of the observed rating patterns. Moreover, all clinicians were affiliated with the same hospital system, which may limit the representativeness of the evaluation. Future studies should involve larger and geographically diverse expert panels together with broader disease coverage. Third, the fairness analysis was constrained by the limited representation of darker skin tones in the source dataset. The Dark-Skin subgroup comprised only seven patients, making subgroup performance estimates sensitive to individual cases. Furthermore, grouping Fitzpatrick skin types into Fair-Skin and Dark-Skin categories simplifies a clinically continuous and partially subjective characteristic. Consequently, the reported fairness improvements should be interpreted as exploratory and validated using larger cohorts spanning all Fitzpatrick skin types. Fourth, the comparison among the downstream training conditions does not fully isolate the effect of clinician-guided image selection from differences in synthetic subset size and disease composition. Although all experimental settings were identical across conditions, the superior performance of the clinician-approved subset may reflect the combined influence of image quality, subset composition, and dataset size. In addition, the approval criterion considered only realism and clinician confidence, without explicitly incorporating diagnostic correctness, mask quality, or skin-tone consistency. Future work should evaluate alternative selection criteria using balanced synthetic subsets. Fifth, the downstream evaluation was performed using a single classification architecture (ResNet50), one patient-level splitting strategy, and three random seeds. Although identical training and evaluation protocols were adopted across all experimental conditions, the observed improvements should be confirmed using additional architectures, repeated patient-level resampling, larger test sets, and formal statistical significance analyses. Finally, the present study evaluates the clinician-validation and synthetic-image selection stages of the proposed framework rather than a fully iterative generation–evaluation–refinement process. Clinician feedback was used to curate synthetic images for downstream classification but was not employed to optimise the generative model through reward learning or preference optimisation. Likewise, the study prioritised clinician-based validation over representation-level analyses such as embedding-space alignment or domain-discrimination experiments. Accordingly, the results should be interpreted as evidence supporting clinician-guided validation and synthetic-data curation, rather than demonstrating convergence of an iterative clinician-in-the-loop optimisation framework or complete statistical alignment between synthetic and real image distributions.

4.4. Future Directions

The proposed framework should first be validated on larger and more diverse paediatric dermatology datasets collected across multiple institutions, imaging devices, and clinical protocols. Broader clinical validation involving independent expert panels and a wider range of diseases, anatomical locations, age groups, and Fitzpatrick skin types will be essential to assess the generalisability of the observed clinician-rating patterns and subgroup-fairness improvements. Such studies will also enable more robust analyses of intersectional effects involving skin tone, disease, age, sex, and acquisition conditions. A second direction concerns the evolution of the current validation-to-classification workflow into a fully iterative clinician-in-the-loop optimisation framework. In this setting, clinician assessments of realism, diagnostic plausibility, mask quality, skin-tone consistency, confi-

dence, and agreement could be translated into reward or preference signals for iterative refinement of the generative model. Future studies should also investigate principled convergence criteria based on the stabilisation of clinician ratings, inter-rater agreement, subgroup performance, and diminishing performance gains across successive refinement cycles [40–42]. Finally, future work should combine clinician-centred evaluation with complementary technical validation and investigate the deployment of clinician-guided synthetic-data curation in routine clinical workflows. Representation-level analyses, such as embedding-space alignment and domain-discrimination models, could provide additional evidence regarding the similarity between real and synthetic data while remaining complementary to expert clinical assessment. At the same time, practical aspects including image provenance, documentation of clinician decisions, management of expert disagreement, and post-deployment monitoring of subgroup performance should be incorporated to support transparent, accountable, and reusable clinician-in-the-loop pipelines for safety-critical medical AI.

5. Conclusions

This study presented and empirically evaluated a clinician-guided framework for the validation and selection of synthetic paediatric dermatology images. The proposed methodology integrates structured clinician assessment with synthetic-data curation to evaluate visual realism, diagnostic plausibility, mask quality, clinician confidence, and skin-tone representation. Our results show that visual realism alone is insufficient to establish the clinical suitability of synthetic images. Although many generated images were perceived as realistic, multidimensional clinician evaluation revealed lower inter-rater agreement and weaker mask-quality assessments than those observed for real images. Moreover, restricting data augmentation to clinician-approved synthetic images improved downstream classification performance and reduced subgroup performance disparities, particularly for the under-represented Dark-Skin population. Overall, these findings demonstrate that structured clinician evaluation provides an effective mechanism for fairness-aware synthetic-data curation, improving the downstream utility of synthetic images without requiring direct optimisation of the generative model. Future work will validate the framework on larger multi-centre datasets and investigate iterative clinician-in-the-loop pipelines in which expert feedback directly supports generative-model refinement.

Author Contributions: Conceptualization, A.T.N., C.B., A.B., A.D., L.P., D.Z., I.N., M.L. and R.C.; methodology, A.T.N., C.B., A.B. and R.C.; software, A.T.N.; validation, A.T.N., C.B., A.B., A.D., L.P., D.Z., I.N., M.L. and R.C.; formal analysis, A.T.N., C.B., A.B., A.D., L.P., D.Z., I.N., M.L. and R.C.; investigation, A.T.N.; resources, C.B., A.B., A.D., L.P., D.Z., I.N., M.L. and R.C.; data curation, A.T.N. and C.B.; writing—original draft preparation, A.T.N.; writing—review and editing, C.B., A.B., A.D., L.P., D.Z., I.N., M.L. and R.C.; visualization, A.T.N.; supervision, A.B., M.L. and R.C.; project administration, A.B., M.L. and R.C. All authors have read and agreed to the published version of the manuscript. Except for Ali Tariq Nagi, who performed all experiments and led the experimental work, all remaining authors contributed equally to the conception, interpretation of the results, critical revision, and final approval of the manuscript.

Funding: This research received no external funding

Institutional Review Board Statement: The study protocol was reviewed and approved by the Comitato Etico Area Vasta Emilia Centro (AVEC) (protocol number 341/2023/Oss/AOUBo, date of approval 26 July 2026), which granted a consent waiver.

Informed Consent Statement: All images were anonymized, avoiding identifiable features such as faces unless clinically necessary. Informed consent, including a specific photographic release form, was obtained from parents or legal guardians before image acquisition.

Data Availability Statement: The experiments use a real-world, non-public dermatology image dataset collected at a public Italian hospital as part of routine clinical care. Due to patient privacy and institutional data governance constraints, the dataset cannot be publicly released; all images were anonymized prior to analysis.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this paper:

AI	Artificial Intelligence
GAN	Generative Adversarial Network
FRR	False Real Rate
NLP	Natural Language Processing
TP	True Positive
FP	False Positive
FN	False Negative
BA	Balanced Accuracy
DI	Disparate Impact Ratio
EOR	Equalised Odds Ratio
PRR	Predictive Rate Ratio
F1	F1 Score
CI	Confidence Interval

References

1. Koetzier, L.R.; Wu, J.; Mastrodicasa, D.; Lutz, A.; Chung, M.; Koszek, W.A.; Pratap, J.; Chaudhari, A.S.; Rajpurkar, P.; Lungren, M.P.; et al. Generating synthetic data for medical imaging. *Radiology* **2024**, *312*, e232471.
2. Cartocci, A.; Luschi, A.; Tognetti, L.; Cinotti, E.; Farnetani, F.; Lallas, A.; Paoli, J.; Longo, C.; Moscarella, E.; Todorovic, D.; et al. Comparative analysis of ai models for atypical pigmented facial lesion diagnosis. *Bioengineering* **2024**, *11*, 1036.
3. Li, Z.; Koban, K.C.; Schenck, T.L.; Giunta, R.E.; Li, Q.; Sun, Y. Artificial intelligence in dermatology image analysis: Current developments and future trends. *J. Clin. Med.* **2022**, *11*, 6826.
4. Nagi, A.T.; Awan, M.J.; Mohammed, M.A.; Mahmoud, A.; Majumdar, A.; Thinnukool, O. Performance analysis for COVID-19 diagnosis using custom and state-of-the-art deep learning models. *Appl. Sci.* **2022**, *12*, 6364.
5. Tognetti, L.; Miracapillo, C.; Leonardelli, S.; Luschi, A.; Iadanza, E.; Cevenini, G.; Rubegni, P.; Cartocci, A. Deep learning techniques for the dermoscopic differential diagnosis of benign/malignant melanocytic skin lesions: From the past to the present. *Bioengineering* **2024**, *11*, 758.
6. Furriel, B.C.; Oliveira, B.D.; Prôa, R.; Paiva, J.Q.; Loureiro, R.M.; Calixto, W.P.; Reis, M.R.; Giavina-Bianchi, M. Artificial intelligence for skin cancer detection and classification for clinical environment: A systematic review. *Front. Med.* **2024**, *10*, 1305954.
7. Martínez-Vargas, E.; Mora-Jiménez, J.; Arguedas-Chacón, S.; Hernández-López, J.; Zavaleta-Monestel, E. The Emerging Role of Artificial Intelligence in Dermatology: A Systematic Review of Its Clinical Applications. *Dermato* **2025**, *5*, 9.
8. Omiye, J.A.; Gui, H.; Daneshjou, R.; Cai, Z.R.; Muralidharan, V. Principles, applications, and future of artificial intelligence in dermatology. *Front. Med.* **2023**, *10*, 1278232.
9. Zbrzezny, A.M.; Krzywicki, T. Artificial intelligence in dermatology: A review of methods, clinical applications, and perspectives. *Appl. Sci.* **2025**, *15*, 7856.
10. Luschi, A.; Tognetti, L.; Cartocci, A.; Cevenini, G.; Rubegni, P.; Iadanza, E. Advancing synthetic data for dermatology: GAN comparison with multi-metric and expert validation approach. *Health Technol.* **2025**, *15*, 553–562.
11. Jain, S.; Seth, G.; Paruthi, A.; Soni, U.; Kumar, G. Synthetic data augmentation for surface defect detection and classification using deep learning. *J. Intell. Manuf.* **2022**, *33*, 1007–1020.
12. Wang, X.; Xie, L.; Dong, C.; Shan, Y. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021*; IEEE: Piscataway, NJ, USA, 2021; pp. 1905–1914.
13. Costa, P.; Galdran, A.; Meyer, M.I.; Niemeijer, M.; Abràmoff, M.; Mendonça, A.M.; Campilho, A. End-to-end adversarial retinal image synthesis. *IEEE Trans. Med Imaging* **2017**, *37*, 781–791.
14. Quiros, A.C.; Murray-Smith, R.; Yuan, K. PathologyGAN: Learning deep representations of cancer tissue. *arXiv* **2019**, arXiv:1907.02644.

15. Shin, H.C.; Tenenholtz, N.A.; Rogers, J.K.; Schwarz, C.G.; Senjem, M.L.; Gunter, J.L.; Andriole, K.P.; Michalski, M. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *Proceedings of the International Workshop on Simulation and Synthesis in Medical Imaging*; Springer: Cham, Switzerland, 2018; pp. 1–11.
16. Otani, M.; Togashi, R.; Sawai, Y.; Ishigami, R.; Nakashima, Y.; Rahtu, E.; Heikkilä, J.; Satoh, S. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023*; IEEE: Piscataway, NJ, USA, 2023; pp. 14277–14286.
17. Sun, S.; Goldgof, G.; Butte, A.; Alaa, A.M. Aligning synthetic medical images with clinical knowledge using human feedback. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 13408–13428.
18. Borghesi, A.; Calejari, R. Generation of clinical skin images with pathology with scarce data. In *AI for Health Equity and Fairness: Leveraging AI to Address Social Determinants of Health*; Springer: Cham, Switzerland, 2024; pp. 47–64.
19. Wang, J.; Wang, K.; Yu, Y.; Lu, Y.; Xiao, W.; Sun, Z.; Liu, F.; Zou, Z.; Gao, Y.; Yang, L.; et al. Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nat. Med.* **2025**, *31*, 609–617.
20. Ktena, I.; Wiles, O.; Albuquerque, I.; Rebuffi, S.A.; Tanno, R.; Roy, A.G.; Azizi, S.; Belgrave, D.; Kohli, P.; Cemgil, T.; et al. Generative models improve fairness of medical classifiers under distribution shifts. *Nat. Med.* **2024**, *30*, 1166–1173.
21. Luo, Y.; Khan, M.O.; Wen, C.; Afzal, M.M.; Wuermeling, T.F.; Shi, M.; Tian, Y.; Fang, Y.; Wang, M. FairDiffusion: Enhancing equity in latent diffusion models via fair Bayesian perturbation. *Sci. Adv.* **2025**, *11*, eads4593.
22. Wang, J.; Zhang, Y.; Ding, Z.; Hamm, J. Doctor approved: Generating medically accurate skin disease images through AI-expert feedback. *Adv. Neural Inf. Process. Syst.* **2026**, *38*, 19971–20003.
23. Schuff, H.; Vanderlyn, L.; Adel, H.; Vu, N.T. How to do human evaluation: A brief introduction to user studies in NLP. *Nat. Lang. Eng.* **2023**, *29*, 1199–1222.
24. Bellatreccia, C.; Zama, D.; Dondi, A.; Pierantoni, L.; Laura, A.; Neri, I.; Lanari, M.; Borghesi, A.; Calejari, R.; Addressing Bias and Data Scarcity in AI-Based Skin Disease Diagnosis with Non-Dermoscopic Images. *CEUR Workshop Proc.* **2025**, 3961, 1–14.
25. Joshi, A.; Kale, S.; Chandel, S.; Pal, D.K. Likert scale: Explored and explained. *Br. J. Appl. Sci. Technol.* **2015**, *7*, 396–403.
26. Bertram, D. *Likert Scales: CPSC 681 Topic Report*; University of Calgary: Calgary, AB, Canada, 2006; pp. 1–11. Available online: <http://poincare.matf.bg.ac.rs/~kristina/topic-dane-likert.pdf> (accessed on 29 July 2026).
27. McHugh, M.L. Interrater reliability: The kappa statistic. *Biochem. Medica* **2012**, *22*, 276–282.
28. CINECA. LEONARDO: The New Pre-Exascale Tier-0 EuroHPC Supercomputer. Available online: <https://www.hpc.cineca.it/systems/hardware/leonardo/> (accessed on 29 July 2026).
29. Gösgens, M.; Zhiyanov, A.; Tikhonov, A.; Prokhorenkova, L. Good classification measures and how to find them. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 17136–17147.
30. Liu, M.; Ning, Y.; Teixayavong, S.; Liu, X.; Mertens, M.; Shang, Y.; Li, X.; Miao, D.; Liao, J.; Xu, J.; et al. A scoping review and evidence gap analysis of clinical AI fairness. *npj Digit. Med.* **2025**, *8*, 360.
31. Feldman, M.; Friedler, S.A.; Moeller, J.; Scheidegger, C.; Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, NY, USA, 2015; pp. 259–268.
32. Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; Wallach, H. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, PMLR 80, Stockholm Sweden, 10–15 July 2018; pp. 60–69.
33. Shui, C.; Xu, G.; Chen, Q.; Li, J.; Ling, C.X.; Arbel, T.; Wang, B.; Gagné, C. On learning fairness and accuracy on multiple subgroups. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 34121–34135.
34. Tarzanagh, D.A.; Hou, B.; Tong, B.; Long, Q.; Shen, L. Fairness-aware class imbalanced learning on multiple subgroups. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, PMLR 216, Pittsburgh, PA, USA, 31 July–4 August 2023; pp. 2123–2133.
35. Schuit, G.; Parra, D.; Besa, C. Perceptual Evaluation of GANs and Diffusion Models for Generating X-Rays. In *Proceedings of the International Workshop on Human-AI Collaboration*; Springer: Cham, Switzerland, 2025; pp. 93–101.
36. Myong, Y.; Yoon, D.; Kim, B.S.; Kim, Y.G.; Sim, Y.; Lee, S.; Yoon, J.; Cho, M.; Kim, S. Evaluating diagnostic content of AI-generated chest radiography: A multi-center visual Turing test. *PLoS ONE* **2023**, *18*, e0279349.
37. Graf, R.; Platzek, P.S.; Riedel, E.O.; Kim, S.H.; Lenhart, N.; Ramschütz, C.; Paprottka, K.J.; Kertels, O.R.; Möller, H.K.; Atad, M.; et al. Generating synthetic high-resolution spinal STIR and T1w images from T2w FSE and low-resolution axial Dixon. *Eur. Radiol.* **2025**, *35*, 1761–1771.
38. Giuffrè, M.; Shung, D.L. Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy. *npj Digit. Med.* **2023**, *6*, 186.
39. Gonzales, A.; Guruswamy, G.; Smith, S.R. Synthetic data in health care: A narrative review. *PLoS Digit. Health* **2023**, *2*, e0000082.
40. Christiano, P.F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; Amodei, D. Deep reinforcement learning from human preferences. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4302–4310.

41. Settles, B. *Active Learning Literature Survey*; Computer Sciences Technical Report 1648; University of Wisconsin-Madison: Madison, WI, USA, 2009.
42. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B Methodol.* **1977**, *39*, 1–22.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.