IMPERIAL

IMPERIAL COLLEGE LONDON
FACULTY OF MEDICINE
DEPARTMENT OF METABOLISM, DIGESTION AND REPRODUCTION
DIVISION OF SYSTEMS MEDICINE

Artificial Intelligence Tools for Clinical Decision Support in Dermatology

Abdurrahim Yilmaz

Supervisors:

Dr. Burak Temelkuran

Dr. Joram M. Posma

Prof. Zoltan Takats

August 2026

A thesis submitted in fulfillment of the requirements for the degree of *Philosophiae Doctor*

Dedication

To my family, mentors, and everyone who has supported me on this journey.

Declaration of Originality

I, Abdurrahim Yilmaz, declare that the work presented in this thesis is my own. Where the work of others has been used, it has been clearly acknowledged and appropriately referenced. A complete list of references is provided in the bibliography.

Copyright

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Under this license, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This permission is granted on the condition that you credit the author and do not use the work, or any derivative work, for a commercial purpose.

When reusing or sharing this work, ensure you make the license terms clear to others by naming the license and linking to the license text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this license or permitted under UK Copyright Law.

Declarations

The work presented here is based on research carried out by the author, with the following support:

- Section 3.2: Professor Gülsüm Gençoğlan and Professor Şirin Pekcan Yaşar provided the dataset and contributed to data annotation.
- Section 3.3: Data collection was carried out in collaboration with Furkan Yüceyalçın.
- Section 4.3: AI training and data annotation were carried out in collaboration with Rana Gürsoy, Barış Kızılyaprak, Professor Gülsüm Gençoğlan, and Professor Ayşe Esra Koku Aksu. Professor Gülsüm Gençoğlan and Professor Ayşe Esra Koku Aksu also contributed clinical data collected from their clinics.
- Section 4.4: Professor Ayşe Esra Koku Aksu and Dr. Duygu Yamen annotated the evaluation data.
- Section 4.5: AI training and data collection were carried out in collaboration with Furkan Yüceyalçın and Dr. Rahmetullah Varol.
- Section 5.2: Data annotation was carried out by Professor Gülsüm Gençoğlan, Dr. Ozan Erdem, Dr. Ece Gökyayla, Dr. Ayda Acar, Dr. Burç Buğra Dağtaş, and Dr. Dilara İlhan Erdil.
- Section 5.3: AI model testing was carried out in collaboration with Furkan Yüceyalçın.

In summary, my main contributions to this work include the conception and design of the overall thesis framework, development of the research questions, coordination of the studies, dataset structuring, synthetic data generation strategy, AI model development and evaluation, benchmarking design, analysis and interpretation of the results, and writing of the thesis.

I hereby certify that all material in this thesis that is not my own work has been properly acknowledged.

Abdurrahim Yilmaz

The formatting of this thesis was adapted from the thesis template by R. Robinson.

Acknowledgments

I am deeply grateful to my supervisors, Dr. Burak Temelkuran, Dr. Joram M. Posma, and Professor Zoltan Takats, for their guidance, expertise, and encouragement throughout this doctoral work. Their support shaped not only the scientific direction of this thesis but also my growth as an independent researcher.

I would also like to express my sincere gratitude to Professor Gülsüm Gençoğlan, who has been a continuous supervisor and mentor in dermatology over the past eight years. Her clinical insight, support, and encouragement have been invaluable throughout my journey, extending far beyond research, including her guidance during my own experience with atopic dermatitis.

I am grateful to all my collaborators, clinical partners, annotators, clinicians, patients, and research teams whose contributions made this work possible. I would like to sincerely thank the clinicians who contributed to this work with their time, expertise, and clinical insight. I am especially grateful to Dr. Ozan Erdem, Professor Ayşe Esra Koku Aksu, and Professor Mehmet Salih Gürel for their valuable support in clinical data collection, annotation, interpretation, and guidance throughout the related studies. Their contributions were essential in grounding this research in real dermatologic practice.

I would also like to thank Dr. Yasin Çotur, Ömer Kürşad Katırcı, Dr. Ayhan Aktaş, Ufuk Görkem Kırabalı, Dr. Rahmetullah Varol, Dr. Zeynep Karavelioğlu Varol, and Professor Hüseyin Üvet for their support and contributions during this journey.

This work was supported by the Sir Michael Uren Medicine Scholarship through the President's PhD Scholarship at Imperial College London (2023–2026). I am sincerely grateful for this support, which made this research possible and provided an exceptional environment in which to undertake my doctoral studies.

I owe special thanks to my friends Amar Rai, Jasmin Werner, and Dr. Jinshi Zhao for their companionship, encouragement, and support throughout this process. I am especially thankful to my brother, Dr. Ali Anıl Demirçalı, for his constant belief in me, patience, and support. My deepest thanks go to my father Recep, my mother Şükran, my sisters Reyhan and Tuba, my supervisors, and my friends, whose presence made this journey possible and meaningful. Finally, I thank my wife, Peyruze, for her endless love, care, and support.

Abstract

Dermatology is inherently multimodal, requiring the integration of clinical photographs, dermoscopy, microscopy, textual patient information, and longitudinal follow-up. Artificial intelligence (AI) has considerable potential to support clinical decision-making and enhance workflow efficiency. Recent advances in language models further extend these capabilities by enabling the joint interpretation of textual and visual clinical information and supporting more context-aware reasoning. However, the development and deployment of clinically useful dermatology AI remain constrained by fragmented datasets, scarce image–text supervision, privacy concerns, and the high computational demands of state-of-the-art models.

This thesis develops a data-centric and resource-efficient AI framework for dermatology, spanning dataset construction, model development, benchmarking, and privacy-preserving deployment. First, it establishes structured data foundations through DERM12345, a multisource dermoscopic dataset comprising 12,345 images organized into 40 clinically meaningful subclasses, and DermaSynth, a synthetic multimodal dermatology dataset generated from open-access datasets, PubMed-derived resources, and instruction-following strategies to provide scalable image–text supervision for dermatology vision-language models (VLMs). Together, these resources support fine-grained skin lesion analysis and multimodal learning. Second, it investigates a range of AI methods for dermatological applications, beginning with convolutional neural network (CNN) baselines on DERM12345 and progressing to artifact-aware object detection for fungal hyphae localization in potassium hydroxide (KOH) microscopy, local language models for longitudinal pemphigus follow-up, and DermatoLlama, a dermatology-specific vision-language model designed for report generation and multimodal reasoning under resource constraints. Third, it introduces evaluation infrastructure, including AnnotatorMed, a locally deployable hierarchical annotation tool; DermaBench, a clinician-annotated visual question-answering (VQA) benchmark integrating clinical and dermoscopic images; and hierarchical foundation-model benchmarking on DERM12345.

Together, these contributions define an end-to-end pathway for explainable, multimodal, resource-efficient, and privacy-preserving dermatology AI. The central argument of this thesis is that clinically credible dermatology AI must be not only accurate but also interpretable, evaluable, data-efficient, and compatible with real-world healthcare constraints.

Contents

Declarations 1

Acknowledgments 2

Abstract 3

Contents 4

List of Abbreviations 11

List of Figures 14

List of Tables 20

1 Introduction 23
- 1.1 Motivation and Research Objectives . . . 24
- 1.2 Key Contributions . . . 26
- 1.3 Publications . . . 27
- 1.4 Overview . . . 29

2 Background 32
- 2.1 Overview . . . 33
- 2.2 Dermatology as a Clinical and Multimodal Data Domain . . . 34
  - 2.2.1 Dermatologic Disease and Clinical Burden . . . 34
  - 2.2.2 The Clinical Dermatology Workflow . . . 35
  - 2.2.3 Dermatologic Morphology . . . 35
  - 2.2.4 Clinical Photography . . . 36
  - 2.2.5 Dermoscopy . . . 36
  - 2.2.6 Microscopy . . . 37
  - 2.2.7 Histopathology . . . 37
  - 2.2.8 Textual and Longitudinal Clinical Data . . . 38
- 2.3 Fundamentals of Artificial Intelligence . . . 38
  - 2.3.1 Artificial Intelligence, Machine Learning, and Deep Learning . . . 38
  - 2.3.2 Transfer Learning . . . 40
  - 2.3.3 Foundation Models . . . 40
  - 2.3.4 Model Parameters and Hyperparameters . . . 40

2.3.5 Data Augmentation . . . 41
2.3.6 Cross-Validation . . . 41
2.3.7 Computer Vision . . . 41
2.3.7.1 Convolutional Neural Networks . . . 42
2.3.7.2 Vision Transformers . . . 42
2.3.7.3 Image Classification . . . 43
2.3.7.4 Object Detection . . . 43
2.3.7.5 Image Segmentation . . . 43
2.3.7.6 Image Generation . . . 44
2.3.8 Natural Language Processing and Language Models . . . 44
2.3.8.1 The Transformer Architecture . . . 45
2.3.8.2 Language Model Pretraining . . . 45
2.3.8.3 Large and Small Language Models . . . 46
2.3.8.4 Retrieval-Augmented Generation . . . 46
2.3.8.5 Local Deployment and Privacy . . . 47
2.4 AI in Dermatology . . . 47
2.4.1 Evolution of Artificial Intelligence in Dermatology . . . 47
2.4.2 Data Foundations for Dermatology Artificial Intelligence . . . 50
2.4.2.1 Major Dermoscopic Datasets . . . 51
2.4.2.2 Clinical Image Datasets . . . 52
2.4.2.3 Multimodal Dermatology Data . . . 52
2.4.2.4 Image–Text Pairs and Visual Question Answering Data . . . 52
2.4.2.5 Synthetic Data . . . 53
2.4.2.6 Train, Validation, and Test Sets . . . 53
2.4.3 Explainability and Interpretability in Dermatology AI . . . 53
2.5 Multimodal Language Models and Vision-Language Models . . . 55
2.5.1 Parameter-Efficient Fine-Tuning . . . 57
2.6 Model Evaluation and Benchmarking . . . 57
2.6.1 Confusion Matrix . . . 58
2.6.2 Main Metrics . . . 58
2.6.3 Object Detection Metrics . . . 59
2.6.4 Segmentation Metrics . . . 60
2.6.5 Language Generation Metrics . . . 60
2.6.6 Visual Question Answering Metrics . . . 60
2.6.7 Evaluation Beyond Accuracy . . . 60
2.7 Data Annotation for Medical and Multimodal Artificial Intelligence . . . 62
2.7.1 Image-Level Diagnostic Annotation . . . 63
2.7.2 Metadata and Structured Clinical Annotation . . . 63
2.7.3 Hierarchical Annotation . . . 64

2.7.4 Object Detection Annotation . . . 65
2.7.5 Segmentation Annotation . . . 65
2.7.6 Visual Question Answering Annotation . . . 66
2.7.7 Text and Longitudinal Annotation . . . 67
2.7.8 Multimodal Annotation . . . 67
2.7.9 Synthetic and Model-Assisted Annotation . . . 68
2.7.10 Annotation Formats and Data Schemas . . . 68
2.7.11 Annotation Quality Control . . . 70
2.7.12 Privacy, Provenance, and Auditability . . . 70
2.8 Challenges for Clinical Deployment . . . 71
2.8.1 Final Notes on Deployment Challenges for Multimodal Dermatology AI . . . 73
2.9 Synthesis of the Literature and Research Gaps . . . 75
2.10 Chapter Summary . . . 76

3 Data Foundations for Dermatology AI 78
3.1 Overview . . . 79
3.1.1 Dermoscopic Image Datasets . . . 80
3.1.2 Multimodal Dermatology Datasets . . . 81
3.2 DERM12345: Dermoscopic Image Dataset . . . 84
3.2.1 Methods . . . 86
3.2.1.1 Data collection from digital dermoscopy devices . . . 86
3.2.1.2 Data collection from mobile phones and hand-held dermatoscopes . . . 86
3.2.1.3 Data selection and quality control . . . 88
3.2.1.4 Taxonomy Tree . . . 89
3.2.2 Data Records . . . 91
3.2.2.1 Dataset metadata . . . 91
3.2.3 Technical Validation . . . 93
3.2.4 Usage Notes . . . 93
3.2.5 Code availability . . . 98
3.3 DermaSynth: Synthetic Image–Text Supervision for Dermatology . . . 99
3.3.1 Background . . . 99
3.3.2 Methods . . . 100
3.3.2.1 Data Sources and Preprocessing: Open-Access PubMed Articles . . . 100
3.3.2.2 Data Collection and Preprocessing: Open-Access Datasets for Dermatology . . . 101
3.3.2.3 Synthetic Data Generation . . . 101
3.3.3 Data Records . . . 104
3.3.4 Technical Validation . . . 105
3.3.5 Usage Notes . . . 105

3.3.6 Data availability . . . 107
3.3.7 Code availability . . . 108

4 AI in Dermatology 109
4.1 Introduction . . . 110
4.2 Baseline CNN Results on DERM12345 . . . 112
4.3 Artifact-Aware Object Detection for KOH Microscopy . . . 114
4.3.1 Introduction . . . 114
4.3.2 Materials and Methods . . . 116
4.3.2.1 Clinical Sample Collection and Microscopic Imaging . . . 116
4.3.2.2 Detection Architecture Overview and Experimental Design . . . 118
4.3.2.3 Artificial Intelligence Model and Training Procedure . . . 119
4.3.2.4 Baseline Model Evaluation . . . 120
4.3.3 Results . . . 121
4.3.4 Discussion . . . 122
4.4 Privacy-Preserving Local Language Models for Pemphigus Follow-Up . . . 127
4.4.1 Introduction . . . 127
4.4.2 Methods . . . 128
4.4.2.1 Study Design and Dataset . . . 128
4.4.2.2 Feature Schema and Ground-Truth Annotation . . . 129
4.4.2.3 Question Design and Inference Procedure . . . 131
4.4.2.4 Model Deployment . . . 132
4.4.2.5 Evaluation Metrics . . . 132
4.4.3 Results . . . 133
4.4.4 Discussion . . . 135
4.5 DermatoLlama: Resource-Efficient Dermatology Vision-Language Modeling . . . 139
4.5.1 Introduction . . . 139
4.5.2 Methods . . . 142
4.5.2.1 Data . . . 142
4.5.2.2 Evaluation Data . . . 142
4.5.2.3 Model Training and Evaluation . . . 143
4.5.2.4 License Issues . . . 144
4.5.3 Results . . . 145
4.5.3.1 SCALEMED Framework Implementation . . . 145
4.5.3.2 Model Performance on Benchmark Evaluations . . . 145
4.5.3.3 Resources Released . . . 148
4.5.4 Discussion . . . 150
4.5.5 Data availability . . . 156
4.5.6 Code availability . . . 156

5 Benchmarks and Evaluations in Dermatology **157**
5.1 AnnotatorMed: Local Annotation Infrastructure for Benchmark Construction . . . 158
5.1.1 Overview . . . 158
5.1.2 AnnotatorMed . . . 158
5.2 DermaBench: Clinician-Annotated Dermatology Visual Question Answering . . . 161
5.2.1 Background and Summary . . . 161
5.2.2 Methods . . . 163
5.2.2.1 Dataset Composition . . . 163
5.2.2.2 Annotation Framework: AnnotatorMed . . . 167
5.2.2.3 Consensus and Quality Control . . . 167
5.2.2.4 Annotation Efficiency and UX Performance . . . 167
5.2.3 Data Records . . . 168
5.2.4 Technical Validation . . . 169
5.2.4.1 Validation of Source Datasets (DDI and DERM12345) . . . 170
5.2.4.2 Validation of Annotation Quality . . . 171
5.2.5 Usage Notes . . . 171
5.2.6 Data availability . . . 174
5.3 Hierarchical Benchmarking of Dermatology Foundation Models . . . 175
5.3.1 Introduction . . . 175
5.3.2 Methods . . . 176
5.3.2.1 Study Design and Dataset . . . 176
5.3.2.2 Taxonomy Mapping . . . 178
5.3.2.3 Foundation Models and Embedding Extraction . . . 179
5.3.2.4 Adapter Training and Optimization . . . 181
5.3.2.5 Hyperparameter Optimization . . . 182
5.3.2.6 Hierarchical Performance Evaluation . . . 182
5.3.3 Results . . . 183
5.3.3.1 Hierarchical Performance Benchmark . . . 183
5.3.3.2 Qualitative Analysis of Embedding Spaces . . . 186
5.3.3.3 Fine-Grained Error Analysis . . . 186
5.3.4 Discussion . . . 189
5.3.5 Data availability . . . 191

6 Conclusion **192**
6.1 Conclusions and Limitations . . . 193
6.2 Future Work . . . 196
6.3 Final Remarks . . . 198

Bibliography **199**

Appendices **218**

A DermaSynth Appendix **220**
A.1 Prompts . . . 220
A.1.1 Question Generation Prompt . . . 220
A.1.2 API Prompt Instruction . . . 221
A.2 Example Questions . . . 222
A.3 Data Generation Prompts . . . 223
A.3.1 Image-Grounded Prompt Template for Open-Access Datasets . . . 223
A.3.2 Prompt Template for Training QA Generation (BIOMEDICA) . . . 224
A.3.3 Prompt Template for Instruction-Following QA Synthesis . . . 226
A.3.4 Prompt Template for Knowledge-Base QA Synthesis . . . 228
A.3.5 Prompt Template for Structured Case Synthesis (Anamnesis) . . . 230
A.4 Wikipedia API Scraping . . . 232

B Follow-up Study of Pemphigus Disease **233**
B.1 Main Prompt . . . 233
B.2 Clinical Features . . . 233

C DermatoLlama **238**
C.1 PubMed Scraping Tool . . . 238
C.2 Training Parameters of DermatoLlama for LoRA . . . 239
C.3 Prompt Template for Clinical Image MCQA Generation (BIOMEDICA) . . . 241
C.4 BIOMEDICA Dataset Filtering . . . 243
C.5 DermatoLlama Example Reports . . . 244

D Benchmarking of Foundation Models in Dermatology **249**
D.1 Comprehensive Performance Tables . . . 249
D.1.1 40-Subclass-Trained Hierarchical Experiment . . . 249
D.1.2 Four-Superclass-Trained Model Evaluation . . . 260
D.2 Extended Visualizations . . . 261
D.2.1 Embedding Space Analysis (t-SNE) . . . 261
D.2.2 Fine-Grained Confusion Matrices . . . 264
D.2.3 Hierarchical ROC Curves . . . 267

E Supplementary Resources **268**
E.1 Supplementary Movie . . . 268
E.1.1 Supplementary Movie 1: DermatoLlama Demonstration . . . 268
E.2 Supplementary Data . . . 268
E.2.1 Supplementary Data 1: Seed Tasks for Instruction-Following Question Answering Synthesis . . . 268

E.2.2 Supplementary Data 2: Seed Tasks for Knowledge-Based Question Answering Synthesis . . . 268
E.2.3 Supplementary Data 3: Seed Tasks for Structured Case Synthesis (Anamnesis) . 269
E.2.4 Supplementary Data 4: Disease List for Structured Case Synthesis . . . 269
E.2.5 Supplementary Data 5: Anamnesis Question Template . . . 269
E.2.6 Supplementary Data 6: DERM12345 Dataset Annotation Question List for DermaBench . . . 269
E.2.7 Supplementary Data 7: DDI Dataset Annotation Guidelines for DermaBench . . . 269
E.2.8 Supplementary Data 8: DDI Dataset Annotation Question List for DermaBench . . 269

List of Abbreviations

AI Artificial Intelligence

AP Average Precision

API Application Programming Interface

AUC Area Under the Curve

BLEU Bilingual Evaluation Understudy

CBC Complete Blood Count

CLI Command-Line Interface

CNN Convolutional Neural Network

COCO Common Objects in Context

CSV Comma-Separated Values

DDI Diverse Dermatology Images

DETR Detection Transformer

DSLR Digital Single-Lens Reflex

EHR Electronic Health Record

FN False Negative

FP False Positive

GPU Graphics Processing Unit

ICC Intraclass Correlation Coefficient

IF Instruction-Following

IoU Intersection over Union

INH Isoniazid

IRB Institutional Review Board

ISIC International Skin Imaging Collaboration

IVIG Intravenous Immunoglobulin

JPEG Joint Photographic Experts Group

JSON JavaScript Object Notation

JSONL JSON Lines

KNN K-Nearest Neighbors

KOH Potassium Hydroxide

LLM Large Language Model

LoRA Low-Rank Adaptation

LR Logistic Regression

LFT Liver Function Test

mAP Mean Average Precision

MCQA Multiple-Choice Question Answering

MLP Multilayer Perceptron

MMF Mycophenolate Mofetil

MTX Methotrexate

NLP Natural Language Processing

OA Open Access

PCA Principal Component Analysis

PHI Protected Health Information

PM PubMed

PMC PubMed Central

PR-AUC Area Under the Precision–Recall Curve

QLoRA Quantized Low-Rank Adaptation

RAG Retrieval-Augmented Generation

RAM Random-Access Memory

RF Random Forest

RLHF Reinforcement Learning from Human Feedback

ROC Receiver Operating Characteristic

ROUGE Recall-Oriented Understudy for Gisting Evaluation

RT-DETR Real-Time Detection Transformer

SCALEMED Scalable Clinical Assistants and LEarning for MEDicine

SD Standard Deviation

SLM Small Language Model

SOTA State of the Art

SVM Support Vector Machine

TF–IDF Term Frequency–Inverse Document Frequency

t-SNE t-Distributed Stochastic Neighbor Embedding

TN True Negative

TP True Positive

UI User Interface

UX User Experience

VLM Vision-Language Model

VQA Visual Question Answering

VRAM Video Random-Access Memory

XGBoost Extreme Gradient Boosting

List of Figures

1.1 Overview of the thesis structure and main contributions across three connected layers: data foundations, AI model development, and benchmarking. Chapter 3 introduces DERM12345 and DermaSynth as structured dermoscopic and multimodal datasets. Chapter 4 presents representative AI approaches ranging from image classification and object detection to local language models and DermatoLlama. Chapter 5 focuses on benchmarking and evaluation through DermaBench and hierarchical foundation model testing. Together, these components define a pathway from structured dermatology data to resource-efficient AI models and clinically meaningful evaluation. 30

3.1 An overview of this study. (A) Skin lesion images are collected using a digital dermatoscope, a DSLR camera, or a mobile phone with a dermatoscope attachment. (B) The collected images undergo an initial screening based on inclusion and exclusion criteria, followed by expert annotation (histopathology, follow-up, or consensus) and a final quality assurance step. (C) Schematic representation of the skin lesion dataset. 87

3.2 A representative image from each subclass. 90

3.3 An overview of the taxonomy tree. The first level includes melanocytic and nonmelanocytic with benign, indeterminate, and malignant groups. The second level comprises banal, dysplastic, and lentigo for melanocytic benign; melanoma for melanocytic malignant; keratinocytic, fibro-histiocytic, and vascular for nonmelanocytic benign; keratinocytic for nonmelanocytic-indeterminate; and keratinocytic, fibro-histiocytic, and vascular for nonmelanocytic malignant. The third level represents the subclasses associated with their respective main and superclasses. Our taxonomy tree consists of 5 superclasses, 15 main classes, and 40 subclasses. 92

3.4 Overview of the synthetic data creation process for DermaSynth. A state-of-the-art large language model (Gemini 2.0 Flash) was used to generate synthetic and clinically relevant image–text pairs. 100

3.5 A figure from the DERM12345 dataset with a field-specific question, a dataset-specific question, and the corresponding Gemini 2.0 Flash answers. 104

3.6 The 20 most common root verb–noun pairs in the question set. 106

3.7 Input images, a simple prompt, and responses generated by the original Llama model and DermatoLlama. 107

4.1 Schematic overview of the study workflow: clinical sample collection, KOH preparation, microscopic imaging, expert annotation, and artificial intelligence modeling steps. 117

4.2 (a, b) Representative KOH microscopy images illustrating the multiclass annotation strategy. Fungal hyphae are marked with green bounding boxes and confounding artifacts are marked with purple boxes. . . . 118
4.3 Overview of the fungal detection pipeline. KOH microscopy images are preprocessed and passed through the RT-DETR architecture, which combines a CNN backbone, hybrid encoder, attention-based query selection, and a transformer decoder to generate bounding box predictions for fungal elements and confounding artifacts. . . . 120
4.4 Qualitative examples of model predictions in KOH microscopy images. (a–c) Ground truth annotations showing fungal elements (green bounding boxes) and artifacts (purple bounding boxes). (d–f) Model predictions with confidence scores, including examples of correct discrimination between fungal structures (blue bounding boxes) and artifact mimics (yellow bounding boxes). (g–i) Overlay of ground truth and model predictions illustrating correct detections and representative misclassifications. . . . 125
4.5 Overview of the study workflow: longitudinal pemphigus follow-up notes are aggregated into patient-level records and processed locally by a privacy-preserving SLM to retrieve structured clinical features and generate longitudinal summaries, which are then evaluated based on expert annotations. . . . 129
4.6 Qualitative comparison between clinician-written ground-truth reports and AI-generated reports: Each example consists of two parts, with the clinician-written ground-truth report shown above and the AI-generated response shown below. The figure contains text-only excerpts annotated for qualitative analysis. Correct or clinically useful information is highlighted in green, explanatory content that does not align with the instruction is highlighted in yellow, and clear errors are highlighted in red. (1) Example of a generally good AI report: although disease activity is not detected correctly, the explanation is helpful, treatments and comorbidities are well captured, and the warning statements are useful for clinician review. (2) Example of a poor AI report: despite clear explanations and reasonable coverage of treatments and comorbidities, disease activity is not detected correctly, limiting overall report quality. Both AI-generated reports are longer than the ground-truth reports, as the AI model provides additional explanatory commentary. Clinicians prepared the ground-truth reports for automated metric calculation (BLEU and ROUGE), with primary comparisons conducted based on the original reports. This figure is included solely to support qualitative visualization and interpretation of model behavior. . . . 136
4.7 Overview of the SCALEMED framework. a, Data collection pipeline of DermaSynth, comprising 1.2 million data samples derived from PubMed, open-access datasets, and instruction-following methods. b, DermaSynth dataset size and data sources. c, DermatoLlama was fine-tuned from Llama-3.2-11B-Vision using the LoRA optimization method. d, DermatoLlama achieves accuracy comparable to GPT-4o and Gemini 2.0 Flash and surpasses all evaluated models in text generation across all evaluation sets. e, AnnotatorMed is a free and open-source tool for clinical data annotation. . . . 140

4.8 Overview of synthetic data generation workflow. a, 46,057 open-access dermatology images with metadata were used to generate metadata-informed, image-focused QA pairs. b, 41,829 open-access PubMed images with captions and inline mentions, comprising 37,565 images used to generate descriptive QA pairs for training and 4,264 images used to generate MCQAs for evaluation. c, 1,015,241 synthetic instruction-following samples were generated based on self-instruct, knowledge base, and anamnesis to cover clinical reasoning. . . . 146
4.9 Comparison of results. (a, b) Eval1: classification and text-generation performance of DermatoLlama; (c, d) Eval2: classification and text-generation performance of DermatoLlama; (e, f) Eval3: classification and text-generation performance of DermatoLlama; (g, h) Eval2: scatter plots of accuracy and ROUGE-L versus parameter count illustrate the relationship between model size and performance, highlighting DermatoLlama's competitive performance despite its relatively small parameter count. . . . 149
4.10 Example reporting results for dermoscopic cases: ajb (acral junctional banal nevus), acd (acral compound dysplastic nevus), mel (melanoma), sk (seborrheic keratosis), and scc (squamous cell carcinoma). For each case, the ground-truth report is shown on the left, the DermatoLlama output in the center, and the base Llama output on the right. DermatoLlama's descriptions align more closely with the reference reports than those generated by the base Llama model. . . . 155

5.1 AnnotatorMed interface displaying hierarchical annotation questions on the right side, guiding systematic categorization of a dermatologic skin lesion image shown on the left. An optional text field in the bottom-right corner allows clinicians to add notes. . . . 159
5.2 Dynamic hierarchical annotation interface in AnnotatorMed. Selecting "No" hides irrelevant follow-up questions, whereas selecting "Yes" reveals subsequent subquestions, streamlining the clinician workflow and reducing annotation time. . . . 159
5.3 AnnotatorMed can be adapted to different evaluation needs, such as ranking model outputs or comparing alternative reports. This flexibility makes it suitable for clinician feedback collection, model output review, and reinforcement learning from human feedback (RLHF)-style workflows. . . . 160
5.4 Dermoscopic images from DERM12345 and clinical images from DDI were annotated using expert-developed hierarchical question sets. Annotations were performed in AnnotatorMed, resulting in a clinician-annotated dermatology VQA benchmark comprising 856 images and 36,071 annotations. . . . 164

5.5 Experimental pipeline for benchmarking foundation models on the DERM12345 dataset [115]. High-resolution skin lesion images are preprocessed and passed through 13 evaluated model variants representing 10 foundation model families, used as frozen feature extractors and grouped into three domains: General Computer Vision (green), General Medical (purple), and Dermatology-Specific (yellow). Numbers in parentheses indicate the corresponding embedding dimensions. The extracted embeddings are used to train lightweight downstream classifiers, including K-Nearest Neighbors (KNN), Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), Multilayer Perceptron (MLP), and Extreme Gradient Boosting (XGBoost). Performance is evaluated hierarchically by mapping the predicted subclass labels from 40 subclasses to coarser diagnostic levels (main classes, superclasses, and binary malignancy), enabling assessment of model performance across the full clinical taxonomy. 177

5.6 Hierarchical performance benchmark across four diagnostic label sets for fine-grained analysis: superclass (2), superclass (4), main class (15), and subclass (40). Bar charts show the weighted F1-scores achieved by the six adapter classifiers (KNN, LR, SVM, RF, MLP, XGBoost) for each foundation model. Foundation models are grouped by domain using background colors: general computer vision models are shown in green, general medical models in purple, and dermatology-specific models in yellow. MedImageInsight achieves the highest performance across the hierarchy, while MedSigLIP remains among the strongest general medical models, particularly at broader hierarchy levels. 184

5.7 Projections are colored by the 15 main classes of the DERM12345 taxonomy. The panels display projections for MedSigLIP, DINOv3, MONET, MedImageInsight, Derm Foundation, and PanDerm (Large). Each point represents a skin lesion. The projections illustrate the two-dimensional t-SNE organization of the embeddings, showing the separation of distinct peripheral classes (e.g., Vascular) and the dense clustering of the melanocytic region. . . 187

5.8 Multi-level confusion matrix analysis for MedImageInsight (MLP Adapter). Left column: Fine-grained confusion matrices at the subclass (top-left) and main class (bottom-left) levels, revealing substantial semantic overlap between visually similar lesion categories. In the subclass confusion matrix, the row and column indices (0–39) correspond to the subclass indices defined in Table 5.5, which provides the mapping from each subclass index to its subclass, main class, and superclass. The main class matrix illustrates substantial overlap between related melanocytic classes, with frequent confusion between dysplastic and banal nevi. Right column: Aggregated confusion matrices at progressively coarser levels (superclass (4), malignancy, and superclass (2)), demonstrating increasingly distinct block-diagonal patterns and improved discrimination of broad diagnostic categories. This progression highlights the degradation in performance from coarse-grained to fine-grained classification within the hierarchical taxonomy. The color intensity in each confusion matrix cell represents the normalized proportion of samples for a given true class that are predicted as the corresponding class. Darker blue indicates a higher proportion of samples, while lighter blue indicates a lower proportion. Cells along the main diagonal represent correct classifications, whereas darker cells outside the diagonal indicate more frequent misclassification between particular classes. The color bar shows the corresponding normalized values, ranging from 0 to 1. 188

C.1 Example of DermatoLlama-generated report using the test split from the DERM12345 dataset. The input image (left) is accompanied by the ground truth diagnosis, the clinical question, and the corresponding report generated by DermatoLlama (right). 244

C.2 Example of DermatoLlama-generated report using the test split from the DERM12345 dataset. The input image (left) is accompanied by the ground truth diagnosis, the clinical question, and the corresponding report generated by DermatoLlama (right). 245

C.3 Example of DermatoLlama-generated report using the test split from the DERM12345 dataset. The input image (left) is accompanied by the ground truth diagnosis, the clinical question, and the corresponding report generated by DermatoLlama (right). 246

C.4 Example of DermatoLlama-generated report using the test split from the DERM12345 dataset. The input image (left) is accompanied by the ground truth diagnosis, the clinical question, and the corresponding report generated by DermatoLlama (right). 247

C.5 Example of DermatoLlama-generated report using the test split from the DERM12345 dataset. The input image (left) is accompanied by the ground truth diagnosis, the clinical question, and the corresponding report generated by DermatoLlama (right). 248

D.1 t-SNE projections of embeddings for general vision models (DINOv2 (Base), DINOv2 (Giant), DINOv3, CLIP (Base), CLIP (Large), ResNet-50). Points are colored by main class. 261

D.2 t-SNE projections for general medical models (BiomedCLIP, MedImageInsight, MedSigLIP). 262

D.3 t-SNE projections for dermatology-specific models (MONET, PanDerm (Base), PanDerm (Large), Derm Foundation). 263

D.4 Confusion matrices for general vision models. Normalized confusion matrices evaluated on the main class (15) task. The panels show results for (a) DINOv2 (Giant), (b) DINOv3, (c) CLIP (Large), and (d) ResNet-50. 264
D.5 Confusion matrices for general medical models. Normalized confusion matrices evaluated on the main class (15) task. The panels show results for (a) BiomedCLIP, (b) MedSigLIP, and (c) MedImageInsight. 265
D.6 Confusion matrices for dermatology-specific models. Normalized confusion matrices evaluated on the main class (15) task. The panels show results for (a) MONET, (b) PanDerm (Base), (c) PanDerm (Large), and (d) Derm Foundation. 266
D.7 One-vs-rest ROC curves at the main class (15) hierarchy level for the 40-subclass-trained experiment. The four panels show one-vs-rest ROC curves for (a) MedSigLIP, (b) Derm Foundation, (c) MONET, and (d) MedImageInsight. 267

List of Tables

1.1 Overview of the thesis structure, associated studies, and the main problems addressed in each chapter. . . . 31

2.1 Major dermatology data modalities and their corresponding clinical and artificial intelligence applications. . . . 38
2.2 Representative evaluation metrics for dermatology artificial intelligence tasks. . . . 62

3.1 Specifications of the largest open-access dermoscopic skin lesion datasets in the literature. 84
3.2 The DERM12345 dataset contains 12,345 dermoscopic images categorized into 5 superclasses, 15 main classes, and 40 subclasses. . . . 85
3.3 Characteristics of the DermaSynth dataset. . . . 103

4.1 Descriptive statistics of the microscopy dataset and annotation categories used for model training and evaluation. . . . 117
4.2 Object-level detection and image-level classification performance. All metrics are reported as mean ± standard deviation (SD) across five fold-level models evaluated on the independent held-out test set (254 images: 89 positive and 165 negative). A case was classified as positive when at least one fungal element was detected with a confidence score exceeding 0.25. . . . 122
4.3 List of fields and their data types used for feature extraction. . . . 129
4.4 Summary of study and deployment statistics. . . . 132
4.5 Summary of key performance results (mean ± SD across 10 repeated inference runs). . 133
4.6 Summary of report characteristics, expert evaluation of AI-generated reports, preference, and reliability analyses. . . . 134
4.7 Classification and text-generation performance of DermatoLlama with different variants and SOTA models, GPT-4o and Gemini 2.0 Flash. OA: Open-access images with QA pairs, PM: Open-access PubMed images with QA pairs, IF: Text-based instruction-following samples. The highest metric across all models per evaluation set is given in **bold**, and the second-highest value is underlined. The 10k, 50k, 100k, and 200k designations are nominal, rounded configuration labels rather than exact sample counts. . . . 151

5.1 Comparison of DermaBench with representative medical VQA corpora. Various[1]: clinical, dermoscopic, and pathology images. . . . 164
5.2 Hierarchy of the clinical image-analysis workflow used for the DDI component of DermaBench. 165
5.3 Hierarchy of the image-analysis workflow used for the DERM12345 dermoscopic component of DermaBench. . . . 166

5.4 Mapping between SkinCon concepts and their corresponding DermaBench question categories. . . . 171
5.5 DERM12345 Hierarchical Taxonomy Map . . . 178
5.6 Foundation Model Specifications and Weight Sources . . . 180
5.7 Hyperparameter Search Space . . . 182
5.8 Fine-grained Foundation Model Leaderboard for the 40-subclass-trained hierarchical experiment. Weighted F1-score (%) achieved by the best-performing adapter, selected independently for each foundation model and evaluated hierarchy level. All adapters were trained to predict the 40 fine-grained subclasses, and their predictions were evaluated hierarchically at progressively coarser levels: malignancy (2) (benign vs. malignant); superclass (2) (melanocytic vs. nonmelanocytic); superclass (4) (melanocytic benign, melanocytic malignant, nonmelanocytic benign, and nonmelanocytic malignant); main class (15) (15 main diagnostic classes); and subclass (40) (40 fine-grained lesion subtypes). Models are sorted by their performance at the finest level of granularity, subclass (40). **Bold** values indicate the best performance, and *italic* values indicate the second-best performance. Malig., malignancy; SC, superclass; MC, main class; Sub., subclass. . . . 185
5.9 Coarse-level Foundation Model Leaderboard for the superclass (4)-trained hierarchical experiment. Weighted F1-score (%) achieved by the best-performing adapter, selected independently for each foundation model and evaluated hierarchy level. All adapters were trained to predict the four superclass labels, and their predictions were evaluated hierarchically at progressively coarser levels: malignancy (2) (binary benign vs. malignant); superclass (2) (melanocytic vs. nonmelanocytic); and superclass (4) (melanocytic benign, melanocytic malignant, nonmelanocytic benign, and nonmelanocytic malignant). Models are sorted by their performance on the superclass (4) task. **Bold** values indicate the best performance, and *italic* values indicate the second-best performance. . . . 186

D.1 Hierarchical evaluation of the 40-subclass-trained model: weighted F1-score for malignancy (2) classification (%). . . . 250
D.2 Hierarchical evaluation of the 40-subclass-trained model: weighted F1-score for superclass (2) classification (%). . . . 251
D.3 Hierarchical evaluation of the 40-subclass-trained model: weighted F1-score for superclass (4) classification (%). . . . 252
D.4 Hierarchical evaluation of the 40-subclass-trained model: weighted F1-score for main class (15) classification (%). . . . 253
D.5 Hierarchical evaluation of the 40-subclass-trained model: weighted F1-score for subclass (40) classification (%). . . . 254
D.6 Hierarchical evaluation of the 40-subclass-trained model: balanced accuracy for malignancy (2) classification (%). . . . 255

D.7 Hierarchical evaluation of the 40-subclass-trained model: balanced accuracy for superclass (2) classification (%). 256
D.8 Hierarchical evaluation of the 40-subclass-trained model: balanced accuracy for superclass (4) classification (%). 257
D.9 Hierarchical evaluation of the 40-subclass-trained model: balanced accuracy for main class (15) classification (%). 258
D.10 Hierarchical evaluation of the 40-subclass-trained model: balanced accuracy for subclass (40) classification (%). 259
D.11 Four-superclass-trained model evaluation: weighted F1-score on the held-out test set (%). 260
D.12 Four-superclass-trained model evaluation: balanced accuracy on the held-out test set (%). 260

Chapter 1

Introduction

Chapter 1 introduces the thesis by outlining its background, motivation, aims, and original contributions. It provides an overview of the clinical and technical context of the work and explains how the studies presented in the following chapters address key challenges in dermatology artificial intelligence (AI).

1.1 Motivation and Research Objectives

Dermatologic diseases represent one of the most common causes of non-fatal disease burden worldwide, affecting hundreds of millions of individuals across all age groups and healthcare systems [1, 2, 3]. Skin diseases range from inflammatory and infectious conditions to autoimmune disorders and skin cancers, creating a substantial clinical, psychological, and economic burden globally [4, 5, 6, 7, 8, 9, 10]. Early and accurate diagnosis is particularly important because delayed recognition of malignant, infectious, or chronic dermatologic diseases may lead to disease progression, unnecessary biopsies, avoidable morbidity, prolonged ineffective treatment, or increased healthcare costs [11, 12]. These challenges, together with the highly visual and multimodal nature of dermatology, have made the field one of the major targets for artificial intelligence-based clinical decision support.

Dermatology is one of the most visually driven specialties in medicine, relying heavily on pattern recognition across clinical photographs, dermoscopy, microscopy, histopathology, and longitudinal patient follow-up. This visual and multimodal nature makes dermatology particularly well suited for advances in computer vision, natural language processing (NLP), and multimodal learning. In recent years, deep learning systems have demonstrated strong performance in tasks such as skin lesion classification, melanoma detection, and medical image interpretation [13, 14]. However, despite rapid progress and impressive benchmark accuracies, several critical limitations continue to hinder the translation of dermatology AI into real-world clinical workflows.

One of the most important challenges is the lack of sufficiently structured and clinically representative datasets. Many existing dermatology datasets are limited to broad classification tasks, often focusing on binary melanoma detection or a small number of diagnostic classes. Such datasets are insufficient for specialist-level dermatology AI because real clinical reasoning is hierarchical and fine-grained. Dermatologists distinguish between numerous visually similar entities, including melanoma subtypes, dysplastic nevi, banal nevi, keratinocytic lesions, inflammatory diseases, vascular lesions, infections, and rare malignancies. Existing datasets also frequently lack multimodal supervision, such as image–text pairs, morphology annotations, and longitudinal clinical context, restricting the development of advanced vision-language systems.

Another major limitation is the predominance of black-box AI systems. Classical convolutional neural networks (CNNs) often provide only classification probabilities without meaningful clinical explanation. While such systems may achieve high benchmark accuracy, they may fail to provide interpretable evidence or clinically useful reasoning. Existing explainability approaches, such as saliency maps and Grad-CAM visualizations, often only highlight image regions contributing to a prediction through heatmaps, without explaining the underlying dermatologic reasoning. In real clinical practice, dermatologists reason through interpretable morphological findings such as irregular pigment networks, a blue-whitish veil, regression structures, follicular pigmentation, vascular patterns, or scaling. Therefore, clinically useful dermatology AI systems should ideally move beyond “where the model looked” toward “why the lesion may represent

melanoma or another disease entity." For example, rather than only producing a heatmap over part of a lesion, a more clinically meaningful system would explain that "the lesion is suspicious for melanoma because of the asymmetric blue-whitish veil observed in the upper-left region together with irregular pigmentation and regression structures."

The emergence of large language models (LLMs) and multimodal vision-language models (VLMs) has introduced new possibilities for dermatology AI, including report generation, visual question answering (VQA), multimodal reasoning, and longitudinal clinical summarization [15]. However, these systems introduce additional challenges related to computational cost, privacy, data governance, hallucination, and real-world deployability. Most state-of-the-art multimodal models rely on cloud-based inference and massive computational infrastructure, which may be impractical or unsuitable in healthcare environments where patient confidentiality and local deployment are essential. Furthermore, dermatology lacks large-scale, high-quality image–text datasets comparable to those available in radiology or general computer vision, limiting the development of specialized dermatology vision-language models.

Benchmarking and evaluation also remain major unresolved problems in dermatology AI. Existing benchmarks often reduce diagnosis to flat classification tasks or binary screening, which fail to capture the hierarchical and multimodal nature of dermatologic reasoning. Moreover, many studies evaluate models only through retrospective accuracy metrics, without considering explainability, fairness, report quality, longitudinal reasoning, deployment constraints, or human-AI interaction. Accordingly, there is a growing need for clinically grounded benchmark datasets and evaluation frameworks that assess dermatology AI systems beyond conventional classification performance.

This thesis is motivated by these interconnected challenges. Rather than focusing on a single dermatologic disease or isolated AI task, this work approaches dermatology AI as a broader methodological and infrastructural problem. The thesis investigates how structured datasets, synthetic multimodal supervision, explainable and resource-efficient models, privacy-preserving deployment strategies, and clinically meaningful benchmarks can collectively support the development of next-generation dermatology AI systems. Through representative methodological case studies spanning dermoscopic lesion analysis, fungal microscopy, longitudinal clinical follow-up, multimodal reasoning, and hierarchical benchmarking, this thesis aims to contribute toward more explainable, deployable, and clinically useful dermatology AI.

The primary objective of this thesis is to investigate the development of clinically grounded, resource-efficient, explainable, and privacy-preserving artificial intelligence systems for dermatology by addressing major methodological challenges across datasets, model development, multimodal learning, and benchmarking. To achieve this goal, the thesis first aims to develop structured dermatology datasets that better reflect the hierarchical nature of dermatologic diagnosis, including fine-grained dermoscopic subclassification and multimodal image–text supervision. A second objective is to explore scalable synthetic data generation strategies for dermatology vision-language models through the integration of open-access datasets, PubMed Central (PMC) resources, instruction-following methods, and synthetic visual question-answering (VQA) pipelines. The thesis further aims to investigate multiple representative AI paradigms in

dermatology, including CNNs for dermoscopic classification, object detection-based systems for fungal microscopy, privacy-preserving local language models for longitudinal dermatologic follow-up, compact multimodal language models for focused dermatology tasks, and large-scale dermatology-specific multimodal language models optimized for local deployment and computational efficiency. Another central objective is to evaluate these systems using clinically meaningful and hierarchical benchmarking strategies rather than relying solely on conventional classification metrics. This includes the development of clinician-annotated visual question-answering benchmarks and hierarchical foundation-model evaluation. As part of this benchmarking layer, this thesis also introduces AnnotatorMed, a locally deployable, free, commercially usable data annotation tool designed to support hierarchical medical data annotation and scalable clinician-centered dataset construction. Overall, the work presented in this thesis represents a number of key steps toward the wider adoption of explainable, resource-efficient, privacy-preserving, and multimodal artificial intelligence systems in dermatology.

1.2 Key Contributions

The key contributions of this thesis are:

1. DERM12345 Dataset: This thesis presents DERM12345, a large dermoscopic skin lesion dataset with 12,345 images and 40 subclasses. This contribution is mainly explained in Chapter 3, presented with CNN results in Chapter 4, and used for benchmarking in Chapter 5.

2. DermaSynth Dataset: This thesis introduces DermaSynth, a synthetic dermatology image–text dataset generated from open-access dermatology datasets, PubMed Central resources, and instruction-following methods. This contribution is mainly explained in Chapter 3 and used for model development in Chapter 4.

3. CNN Baseline Models: This thesis evaluates conventional CNN-based models on DERM12345, showing the limitations of black-box image classification for fine-grained dermatology AI. This contribution is explained in Chapter 4.

4. Object Detection Model: This thesis develops an object detection-based AI system for fungal hyphae localization in potassium hydroxide (KOH) microscopy, providing localized visual evidence rather than only image-level classification. This contribution is explained in Chapter 4.

5. Local Language Model for Follow-up: This thesis evaluates a privacy-preserving local language model for longitudinal pemphigus follow-up, showing how clinical records can be summarized without cloud-based inference. This contribution is explained in Chapter 4.

6. Dermatology Vision-Language Model: This thesis presents DermatoLlama, a dermatology-specific multimodal language model trained using DermaSynth, with emphasis on resource efficiency, privacy-preserving deployment, and report generation. This contribution is explained in Chapter 4 and evaluated in Chapter 5.

7. AnnotatorMed: This thesis presents AnnotatorMed, a locally deployable, free, commercially usable annotation tool with hierarchical data annotation features. This contribution is explained in Chapter 5 as part of the benchmarking and evaluation layer.

8. DermaBench: This thesis introduces DermaBench, a clinician-annotated benchmark for dermatology visual question answering and reasoning. This contribution is explained in Chapter 5.

9. Foundation Model Benchmarking on the DERM12345 Dataset: This thesis uses DERM12345 to evaluate foundation models from different domains (general, medical, and dermatology-specific) across multiple levels of dermatologic diagnostic granularity, highlighting the importance of hierarchical benchmarking. This contribution is presented in Chapter 5.

10. SCALEMED: Overall, this thesis provides a connected framework for developing dermatology AI systems based on multimodal language models that are more explainable, resource-efficient, privacy-preserving, and clinically meaningful. This broader synthesis is discussed across Chapters 3–5.

1.3 Publications

The work presented in this thesis has been accepted for presentation at peer-reviewed conferences or for publication in peer-reviewed journals, or is currently under submission or review. The publications directly related to this thesis are listed below, followed by additional publications that are less closely related but were produced during the same research period.

Main publications:

- **A. Yilmaz**, S. P. Yasar, G. Gencoglan, B. Temelkuran. "DERM12345: A Large, Multisource Dermatoscopic Skin Lesion Dataset with 40 Subclasses," Scientific Data, https://doi.org/10.1038/s41597-024-04104-3
- **A. Yilmaz**, F. Yuceyalcin, R. Varol, E. Gokyayla, O. Erdem, D. Choi, A. A. Demircali, G. Gencoglan, J. M. Posma, B. Temelkuran. "DermaSynth: Rich Synthetic Image-Text Pairs Using Open Access Dermatology Datasets," arXiv, https://doi.org/10.48550/arXiv.2502.00196, under review at Scientific Data

- R. Gursoy*, **A. Yilmaz***, B. Kizilyaprak, E. Caglar, B. Temelkuran, H. Uvet, A. E. K. Aksu, G. Gencoglan. "Artifact-Aware Fungal Detection in Dermatophytosis: A Transformer-Based Approach for KOH Microscopy," Bioengineering, https://doi.org/10.3390/bioengineering13050591, *** Equal first authorship**
- **A. Yilmaz**, A. E. K. Aksu, D. Yamen, V. A. Erdemir, M. S. Gurel, G. Gencoglan, J. M. Posma, B. Temelkuran. "Privacy-Preserving Local Language Models for Longitudinal Data Retrieval in Chronic Dermatologic Disease: Implementation in Pemphigus Patients," arXiv, https://doi.org/10.48550/arXiv.2605.25020, under review at Scientific Reports
- S. Zhang, **A. Yilmaz**, G. Gencoglan, B. Temelkuran. "Fine-Tuning a Small Vision Language Model Using Synthetic Data for Explaining Bacterial Skin Disease Images," Diagnostics, https://doi.org/10.3390/diagnostics16040603, **Corresponding Author**
- **A. Yilmaz**, F. Yuceyalcin, R. Varol, E. Gokyayla, O. Erdem, D. Choi, A. A. Demircali, G. Gencoglan, J. M. Posma, B. Temelkuran. "Resource-efficient medical vision language model for dermatology via a synthetic data generation framework," medRxiv, https://doi.org/10.1101/2025.05.17.25327785, under review at IEEE Transactions on Medical Imaging
- **A. Yilmaz**, O. Erdem, E. Gokyayla, A. Acar, B. B. Dagtas, D. I. Erdil, G. Gencoglan, B. Temelkuran. "DermaBench: A Clinician-Annotated Benchmark Dataset for Dermatology Visual Question Answering and Reasoning," arXiv, https://doi.org/10.48550/arXiv.2601.14084, under review at Scientific Data
- F. Yuceyalcin*, **A. Yilmaz***, B. Temelkuran. "A Hierarchical Benchmark of Foundation Models for Dermatology," arXiv, https://doi.org/10.48550/arXiv.2601.12382, under review at Scientific Reports, *** Equal first authorship**
- **A. Yilmaz**, S. A. Aydin, D. Temur, F. Yuceyalcin, B. Deniz, R. Varol, O. Gokoz, G. Gencoglan, H. Uvet, G. Elcin. "An ensemble deep learning approach to detect tumors on Mohs micrographic surgery slides," arXiv, https://doi.org/10.48550/arXiv.2504.05219, under review at Signal, Image and Video Processing
- **A. Yilmaz**, F. Yuceyalcin, R. Varol, E. Gokyayla, O. Erdem, D. Choi, A. A. Demircali, G. Gencoglan, J. M. Posma, B. Temelkuran. "Dermatology-Specific Multimodal Large Language Model," 27th Prof. Dr. A. Lutfu Tat Symposium, Ankara, Türkiye, 2025, p. 6
- A. S. Sahin, **A. Yilmaz**, B. B. Dagtas, O. Erdem, M. S. Gurel. "Diagnostic Performance of Artificial Intelligence-Assisted Trichoscopy in the Differentiation of Cicatricial and Non-Cicatricial Alopecia," 14th Ord. Prof. Dr. Hulusi Behçet Dermatology Spring Symposium, Mugla, Türkiye, 2026, **3rd Best Oral Presentation Award**

Other publications:

- Center for AI Safety, Scale AI, **HLE Contributors Consortium**. "A benchmark of expert-level academic questions to assess AI capabilities." Nature, https://doi.org/10.1038/s41586-025-09962-4
- O. Erdem, **A. Yilmaz**, A. S. Sahin, B. B. Dagtas, E. Gokyayla, M. A. Kayiran, V. A. Erdemir, M. S. Gurel. "Performance of Vision-Language Models Compared with 252 Medical Students on Text-only and Image-based Dermatology Examinations," under review at Scientific Reports, **Corresponding Author**
- O. Erdem, V. A. Erdemir, B. B. Dagtas, A. E. K. Aksu, S. S. Ertekin, **A. Yilmaz**, E. Gokyayla, G. Gencoglan, F. Goktay, M. S. Gurel. "Cross-Polarized Macro Photography versus Polarized Dermoscopy in Basal Cell Carcinoma: A Blinded Paired-Image Comparison," Clinical, Cosmetic and Investigational Dermatology, https://doi.org/10.2147/CCID.S605974
- **A. Yilmaz**, K. Gem, M. Kalebasi, R. Varol, Z. O. Gencoglan, Y. Samoylenko, H. K. Tosyali, G. Okcu, H. Uvet. "An automated hip fracture detection, classification system on pelvic radiographs and comparison with 35 clinicians," Scientific Reports, https://doi.org/10.1038/s41598-025-98852-w

1.4 Overview

An overview of the thesis structure, including the main chapters, associated studies, and the key problems addressed in each chapter, is presented in Figure 1.1 and Table 1.1. The structure of this thesis is organized as follows:

Chapter 1 – Introduction: This chapter introduces the motivation, aims, research objectives, and key contributions of the thesis. It explains why dermatology AI needs more structured datasets, more explainable models, resource-efficient multimodal systems, privacy-preserving deployment strategies, and clinically meaningful benchmarks.

Chapter 2 – Background: This chapter reviews the development of AI in dermatology, from early image processing and traditional machine learning to CNNs, object detection, foundation models, language models, and multimodal vision-language models. It provides the literature background for the methodological contributions presented in the following chapters.

Chapter 3 – Data Foundations for Dermatology AI: This chapter focuses on dataset development. It presents DERM12345 as a large hierarchical dermoscopic dataset and DermaSynth as a synthetic multimodal dermatology image–text dataset. The chapter explains how these datasets address the lack of fine-grained labels and scalable multimodal supervision in dermatology AI.

Chapter 4 – AI in Dermatology: This chapter focuses on model development. It starts with CNN-based baseline models on DERM12345 and then presents representative methodological case studies, including fungal hyphae object detection, local language models for pemphigus follow-up, and DermatoLlama for dermatology-wide vision-language reasoning.

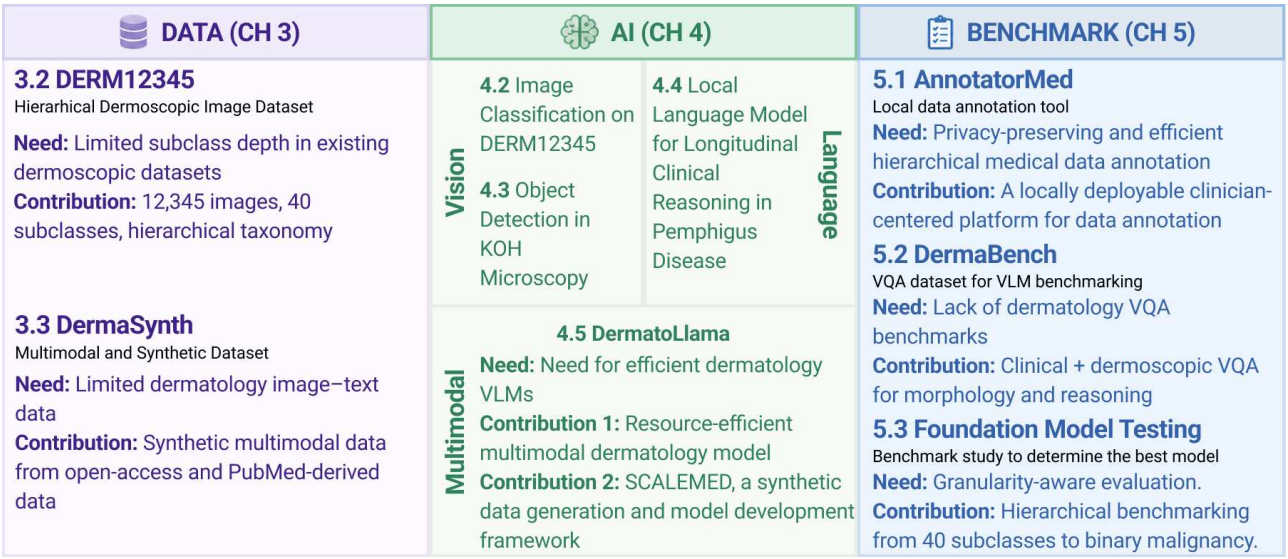


Figure 1.1: Overview of the thesis structure and main contributions across three connected layers: data foundations, AI model development, and benchmarking. Chapter 3 introduces DERM12345 and DermaSynth as structured dermoscopic and multimodal datasets. Chapter 4 presents representative AI approaches ranging from image classification and object detection to local language models and DermatoLlama. Chapter 5 focuses on benchmarking and evaluation through DermaBench and hierarchical foundation model testing. Together, these components define a pathway from structured dermatology data to resource-efficient AI models and clinically meaningful evaluation.

Chapter 5 – Benchmarks and Evaluations in Dermatology: This chapter focuses on evaluation and benchmarking. It explains the use of DERM12345 as a hierarchical benchmark, presents foundation-model testing on DERM12345, introduces DermaBench for dermatology VQA evaluation, and describes AnnotatorMed as a locally deployable hierarchical annotation tool supporting clinician-centered benchmark construction.

Chapter 6 – Conclusion: This chapter summarizes the main findings, limitations, and implications of this thesis. It discusses how the work contributes toward more explainable, resource-efficient, privacy-preserving, and multimodal AI systems in dermatology, and outlines future directions for real-world deployment, agentic systems, and advanced dermatology AI evaluation.

Table 1.1: Overview of the thesis structure, associated studies, and the main problems addressed in each chapter.

Chapter	**Main Focus**	**Related Studies / Short Names**	**What It Addresses**
Chapter 1	Introduction	–	Introduces the motivation, aims, thesis structure, and overall contribution.
Chapter 2	Background	–	Reviews the evolution of AI in dermatology from image processing and CNNs to language models and vision-language models.
Chapter 3	Data Foundations	DERM12345, DermaSynth	Addresses the lack of hierarchical dermoscopic datasets and scalable multimodal image–text data in dermatology AI.
Chapter 4	AI Models	CNN Baseline Results on DERM12345, Fungal Detection in KOH Microscopy, Local LLMs for Pemphigus, DermatoLlama/SCALEMED	Addresses black-box classification, localization, privacy-preserving longitudinal reasoning, and resource-efficient dermatology VLMs.
Chapter 5	Benchmarks and Evaluation	AnnotatorMed, DermaBench, Foundation Model Benchmarking	Addresses scalable annotation infrastructure, hierarchical evaluation, foundation-model benchmarking, and clinician-annotated VQA evaluation.
Chapter 6	Conclusion	–	Summarizes the thesis findings, limitations, and future directions for explainable, resource-efficient, privacy-preserving dermatology AI.

Chapter 2

Background

This chapter introduces the technical and clinical context for the remainder of the thesis. It situates the work at the intersection of dermatology, multimodal machine learning, and privacy-preserving clinical AI deployment.

2.1 Overview

This chapter provides the clinical and technical background required to understand the datasets, artificial intelligence models, and benchmarking methods presented in the subsequent chapters. As this thesis is translational and intended to be accessible to both clinicians and artificial intelligence researchers, the chapter introduces fundamental concepts from dermatology, medical imaging, machine learning, computer vision, natural language processing, and multimodal modeling. It also reviews the principal opportunities and limitations associated with the use of artificial intelligence in dermatologic practice.

Dermatology is particularly suitable for artificial intelligence because visual information plays a central role in diagnosis. Dermatologists interpret clinical photographs, dermoscopic images, microscopy preparations, and histopathological slides together with patient history, symptoms, anatomical location, lesion evolution, laboratory findings, and treatment response. The relevance of each information source depends on the clinical task. Dermoscopic images are particularly important for pigmented lesion and skin cancer assessment, while clinical photographs and textual information are essential for the analysis of common inflammatory, infectious, autoimmune, and chronic dermatologic diseases. Microscopy may be required for infections, whereas histopathology remains important for confirming many malignant, inflammatory, and diagnostically uncertain conditions.

The history of dermatology artificial intelligence reflects the wider development of medical artificial intelligence. Early approaches relied on image preprocessing, lesion segmentation, and manually designed features. These were followed by traditional machine-learning classifiers, such as support vector machines and random forests. Convolutional neural networks later enabled models to learn visual features directly from images and became the dominant method for skin lesion classification. More recently, vision transformers, self-supervised foundation models, language models, and vision-language models have expanded the field from image classification toward visual question answering, report generation, longitudinal record summarization, and multimodal clinical reasoning [16].

Despite this progress, several important challenges remain. Dermatology datasets are frequently small, imbalanced, retrospectively collected, and restricted to a limited number of diagnostic classes. Many models produce black-box predictions without clinically meaningful explanations. Large language and vision-language models may require substantial computational resources and cloud-based infrastructure, raising concerns regarding privacy, cost, accessibility, and governance. Furthermore, conventional benchmarks often focus on flat classification accuracy and may not adequately assess diagnostic hierarchy, morphology, reasoning, uncertainty, fairness, or deployment feasibility.

This chapter is organized around these challenges. Section 2.2 introduces dermatology as a clinical, visual, multimodal, and longitudinal domain, while Section 2.3 presents the fundamental concepts of artificial intelligence and machine learning. Section 2.4 reviews the historical evolution of AI in dermatology. Section 2.5 introduces language models and multimodal vision-language models. Section 2.6 describes model evaluation and benchmarking. Section 2.7 discusses data annotation for medical and multimodal

artificial intelligence, while Section 2.8 reviews the major challenges associated with clinical deployment. The chapter concludes by identifying the research gaps addressed in Chapters 3, 4, and 5.

2.2 Dermatology as a Clinical and Multimodal Data Domain

2.2.1 Dermatologic Disease and Clinical Burden

Dermatologic diseases are among the most common health conditions worldwide and represent a major source of non-fatal disease burden across all age groups [1, 2, 3]. They include inflammatory diseases, infections, autoimmune disorders, pigmentary conditions, hair and nail diseases, benign tumors, premalignant lesions, and skin cancers. Some dermatologic diseases are self-limiting, while others require lifelong follow-up or urgent intervention. Although many skin conditions are not immediately life-threatening, they can produce pain, itching, sleep disturbance, disfigurement, psychological distress, social stigma, loss of productivity, and considerable healthcare expenditure [11, 12].

Common inflammatory diseases include acne, atopic dermatitis, psoriasis, rosacea, contact dermatitis, and urticaria. Infectious dermatologic diseases include bacterial, fungal, viral, and parasitic conditions. Autoimmune and chronic diseases include pemphigus, pemphigoid, cutaneous lupus, vitiligo, and scleroderma. Skin cancers include melanoma, basal cell carcinoma, and squamous cell carcinoma. The clinical significance of each disease differs, but accurate diagnosis is important in all cases because delayed recognition may result in disease progression, inappropriate treatment, avoidable biopsies, increased morbidity, or unnecessary healthcare costs [11, 12].

Skin cancer represents a particularly important area of dermatology artificial intelligence. Melanoma is less common than keratinocyte cancers but has a greater potential for metastasis and mortality. Early detection is therefore essential [17]. Basal cell carcinoma and squamous cell carcinoma are more common and may cause substantial local destruction, particularly when located on the face or other functionally important anatomical sites [18, 19]. Artificial intelligence systems have consequently been developed for melanoma screening, benign–malignant classification, multiclass skin lesion diagnosis, biopsy prioritization, and clinical decision support.

However, dermatology extends considerably beyond skin cancer. Common inflammatory and infectious diseases are frequently assessed from clinical photographs, lesion morphology, symptoms, distribution, patient history, and treatment response. Chronic diseases require longitudinal assessment across repeated visits. Therefore, an artificial intelligence framework focused only on dermoscopic skin lesion classification cannot address the full spectrum of dermatologic practice.

2.2.2 The Clinical Dermatology Workflow

Dermatologic assessment usually begins with the patient history, or anamnesis. The clinician asks about the onset, duration, progression, symptoms, previous treatment, medication exposure, systemic disease, family history, occupational exposure, travel, and possible triggers. The relevance of these questions depends on the presenting condition. A changing pigmented lesion may require information about growth, bleeding, and melanoma risk factors, whereas an inflammatory eruption may require details about itching, medication exposure, infection, or contact allergens [20].

The visual examination then considers lesion morphology and distribution. Dermatologists assess whether lesions are solitary or multiple, localized or generalized, symmetric or asymmetric, flexural or extensor, dermatomal, acral, photodistributed, or associated with mucosal involvement. The examination may also include palpation, assessment of lymph nodes, hair and nail inspection, and evaluation of systemic signs.

Additional diagnostic procedures may be used when visual assessment is insufficient. Dermoscopy allows examination of subsurface structures. Skin scraping and KOH microscopy may be used for suspected fungal infection. Bacterial, viral, or fungal cultures may be required in infectious diseases. Patch testing may be used for allergic contact dermatitis. Histopathological examination following biopsy is often required for malignancy or diagnostically uncertain inflammatory conditions.

Longitudinal follow-up is also central to dermatology. A patient with multiple melanocytic nevi may undergo serial digital dermoscopy. A patient with psoriasis, pemphigus, or atopic dermatitis may accumulate years of treatment records, laboratory results, medication changes, adverse events, and disease activity assessments. The clinical decision at a particular visit may depend on events that occurred months or years earlier. This makes dermatology both a visual and longitudinal information-processing specialty.

2.2.3 Dermatologic Morphology

Dermatologic diagnosis relies on a standard descriptive vocabulary. Primary elementary lesions describe the initial morphological expression of disease. Common primary lesions include macules, patches, papules, plaques, nodules, tumors, vesicles, bullae, pustules, wheals, cysts, petechiae, purpura, and telangiectasia.

Secondary features arise through disease progression, trauma, scratching, infection, or healing. These include scale, crust, erosion, ulceration, fissuring, excoriation, lichenification, atrophy, necrosis, hyperkeratosis, and scarring. Surface characteristics may include smooth, verrucous, papillomatous, dome-shaped, umbilicated, exophytic, friable, shiny, greasy, or hyperkeratotic appearances.

Color also provides important diagnostic information. Lesions may be erythematous, violaceous, brown, black, blue, gray, yellow, white, or skin-colored. The interpretation of color depends on skin tone, lesion depth, vascularity, pigment distribution, and imaging conditions. A blue color may reflect deeper melanin, while white areas may represent fibrosis, depigmentation, scale, or regression. Morphology must therefore

be interpreted in context rather than as an isolated feature.

For artificial intelligence, this vocabulary provides a bridge between images and clinical language. A conventional classifier may predict a disease label, whereas a more advanced multimodal model may describe the lesion as an irregularly bordered pigmented plaque containing gray and blue structures. Such descriptions can provide clinically interpretable intermediate information and support more transparent evaluation.

2.2.4 Clinical Photography

Clinical photography captures the macroscopic appearance of the skin as seen during routine examination. It may show a single lesion, multiple lesions, a regional distribution, or a large body area. Clinical images are widely used in teledermatology, clinical documentation, education, treatment monitoring, and research.

Clinical photography provides information that may be absent from dermoscopy, including anatomical context, lesion distribution, surrounding skin, lesion multiplicity, and body-site relationships. It is particularly important for inflammatory and infectious diseases, in which diagnosis may depend on pattern and distribution rather than the appearance of one isolated lesion.

However, clinical images are difficult to standardize. They may vary in distance, angle, focus, illumination, color balance, camera type, background, compression, and framing. Clothing, jewelry, rulers, markings, hair, tattoos, and other artifacts may also be present. Skin tone may interact with lighting and camera processing, altering the apparent color and contrast of lesions. These factors create domain shifts that can affect model performance.

2.2.5 Dermoscopy

Dermoscopy is a non-invasive imaging technique that uses magnification and polarized or non-polarized illumination to reduce surface reflection and reveal subsurface skin structures [21]. It is widely used for the assessment of pigmented lesions and skin cancer. Dermoscopy can reveal pigment networks, dots, globules, streaks, regression structures, blue-whitish veil, vascular patterns, follicular structures, keratin, ulceration, and other features that may not be visible to the naked eye.

Dermoscopy improves diagnostic accuracy when used by trained clinicians, but interpretation depends on experience [22]. Several structured approaches have been proposed, including pattern analysis, the ABCD rule, the seven-point checklist, the Menzies method, and two-step pattern analysis [23, 24, 25, 26]. These approaches help clinicians distinguish melanocytic from nonmelanocytic lesions and assess whether a lesion is benign, suspicious, or malignant.

Dermoscopic images are often more standardized than clinical photographs because the lesion fills most of the image and acquisition conditions are controlled by the dermatoscope. This makes dermoscopy well suited to computer vision. Nevertheless, device type, magnification, polarization, image borders, air

bubbles, gel, hair, markers, and color calibration can still introduce variation.

For skin lesion analysis, the entire dermoscopic image usually corresponds to one lesion. Image-level classification is therefore a natural starting point. Object detection is not generally required to locate the lesion when it already occupies the central field. However, segmentation or region-level annotation may still be valuable for localizing structures such as blue-whitish veil, atypical vessels, regression, irregular network, or ulceration.

2.2.6 Microscopy

Microscopy is used in several areas of dermatology, including the diagnosis of fungal infection, parasitic disease, hair disorders, and dermatopathology. KOH microscopy is commonly used to detect fungal structures in skin, hair, or nail samples. KOH dissolves keratin and improves the visibility of fungal hyphae and spores.

Microscopic analysis differs from dermoscopic classification. A microscopy image may contain several small target structures distributed across a large field. Artifacts such as fibers, keratin edges, air bubbles, scratches, and debris may resemble the target. In this setting, object detection or segmentation may be more appropriate than image-level classification because the clinical task requires localization of fungal elements.

Microscopy also creates opportunities for automation. A manually operated microscope requires the observer to search multiple fields, which can be time-consuming and affected by fatigue. Automated slide scanning combined with artificial intelligence could allow large numbers of microscopic fields to be processed systematically. However, variation in sample preparation, microscope type, magnification, focus, and laboratory workflow must be considered.

2.2.7 Histopathology

Histopathology remains an important reference standard for many dermatologic diagnoses, particularly malignant and diagnostically uncertain lesions. Tissue is sampled through biopsy, processed, stained, and examined microscopically. Histopathology can reveal cellular architecture and tissue-level patterns that are not available from clinical or dermoscopic images.

Artificial intelligence has been applied to histopathological classification, segmentation, cell detection, and biomarker analysis. However, histopathology differs from non-invasive dermatologic imaging. It requires tissue removal and specialist interpretation, and whole-slide images are substantially larger than ordinary photographs. Although histopathology is not the main focus of this thesis, it is relevant because diagnostic confirmation, especially for malignant lesions, often depends on biopsy findings.

2.2.8 Textual and Longitudinal Clinical Data

Patient history and electronic health record (EHR) data provide information that cannot be obtained from images alone. Relevant information may include age, symptoms, duration, lesion evolution, prior diagnoses, treatments, medication doses, laboratory values, pathology results, comorbidities, adverse events, and current disease activity.

For chronic dermatologic diseases, this information may be distributed across dozens of clinical notes written by different clinicians. Records may contain duplicated text, inconsistent terminology, incomplete dates, copied medication lists, or contradictory statements. Reviewing these records manually can be burdensome and may result in clinically important details being missed.

Natural language processing and language models can assist with information extraction, summarization, and longitudinal analysis. However, clinical records are highly sensitive and may contain protected health information (PHI). Local or institutionally controlled model deployment is therefore particularly relevant.

Data modality	Typical clinical use	Information represented	Relevant AI tasks
Clinical photography	General dermatology, teledermatology, monitoring	Macroscopic morphology, anatomical context, lesion distribution and surrounding skin	Classification, segmentation, visual question answering, report generation
Dermoscopy	Pigmented lesions and skin cancer assessment	Subsurface pigmentary, vascular, follicular and keratin-related structures	Classification, segmentation, concept detection, hierarchical diagnosis
Microscopy	Fungal, parasitic and hair-related assessment	Small biological structures within complex fields	Object detection, segmentation, screening and quantification
Histopathology	Diagnostic confirmation and tumor assessment	Cellular and tissue architecture	Whole-slide classification, cell detection, segmentation and grading
Clinical text and EHR data	History, follow-up, treatment and safety monitoring	Symptoms, diagnoses, medication exposure, laboratory values and longitudinal course	Information extraction, summarization, question answering and decision support

Table 2.1: Major dermatology data modalities and their corresponding clinical and artificial intelligence applications.

2.3 Fundamentals of Artificial Intelligence

2.3.1 Artificial Intelligence, Machine Learning, and Deep Learning

Artificial intelligence is a broad term describing computational systems designed to perform tasks associated with human intelligence, including perception, reasoning, prediction, language understanding, and

decision support.

Machine learning is a subset of artificial intelligence in which a system learns patterns from data rather than following only manually programmed rules. A machine-learning model maps input variables to an output. In dermatology, the input may be an image or a clinical record, and the output may be a diagnosis, lesion location, summary, or risk score.

Deep learning is a subset of machine learning based on neural networks with multiple computational layers. Deep networks can learn complex representations directly from raw data. Convolutional neural networks are designed for images, while transformer architectures are widely used for text, images, and multimodal information.

Supervised Learning: In supervised learning, each training example has an associated target label. A dermoscopic image may be labeled as melanoma, nevus, or basal cell carcinoma. The model learns to minimize a loss function that measures the difference between its prediction and the target.

For multiclass classification, cross-entropy loss is commonly used:

$$\mathscr{L}_{\mathrm{CE}} = -\sum_{c=1}^{C} y_c \log(\hat{p}_c), \tag{2.1}$$

where C is the number of classes, y_c is the reference label, and $\hat{p}_c$ is the model-predicted probability for class c.

Supervised learning can achieve strong performance when large, accurately labeled datasets are available. Its main limitation is dependence on expert annotation. In medicine, labels may be expensive, uncertain, or influenced by clinical practice.

Unsupervised and Self-Supervised Learning: Unsupervised learning identifies structure in data without using diagnostic labels. Clustering methods group similar observations, while dimensionality-reduction methods produce lower-dimensional representations.

Self-supervised learning creates learning objectives from the data itself. In images, a model may learn by reconstructing masked regions or comparing different augmented views of the same image. In language, a model may predict missing or subsequent tokens.

Self-supervised learning is valuable in medicine because large quantities of unlabeled data may be available. A pretrained representation can later be adapted to a smaller labeled dataset. Models such as DINO, masked autoencoders, and contrastive image–text models illustrate this approach [27].

2.3.2 Transfer Learning

Transfer learning reuses knowledge learned from one dataset or task for another [28]. A CNN pretrained on ImageNet may be adapted to dermoscopic images. Early layers may already detect general visual features such as edges, textures, and shapes, reducing the quantity of dermatology-specific data required.

A model can be used as a frozen feature extractor or fine-tuned. When frozen, its weights remain unchanged and only a downstream classifier is trained. During fine-tuning, some or all pretrained weights are updated. Full fine-tuning can improve domain adaptation but requires more computation and may overfit small datasets.

2.3.3 Foundation Models

A foundation model is pretrained on large and diverse datasets and can be adapted to many downstream tasks [29, 30]. Vision foundation models learn general visual representations, medical foundation models learn from biomedical data, and dermatology-specific foundation models are pretrained on skin images.

Foundation models can reduce the need to train a model from the beginning. They may be evaluated through linear probing, lightweight adapters, few-shot learning, or fine-tuning. However, strong performance on one task does not guarantee performance on another. A model that captures broad semantic categories may perform well for malignancy screening but struggle with fine-grained subclass discrimination.

2.3.4 Model Parameters and Hyperparameters

Model parameters are learned during training. In a neural network, these include weights and biases. Hyperparameters are selected before or during experimentation and include learning rate, batch size, number of epochs, architecture, regularization strength, and augmentation strategy.

The learning rate determines the size of parameter updates. A learning rate that is too high may prevent convergence, while one that is too low may produce slow or suboptimal training. Batch size determines how many examples are processed before an update. An epoch corresponds to one complete pass through the training set.

Hyperparameter selection must be performed using validation data rather than the final test set. Repeated tuning on the test set can indirectly leak information and invalidate the final evaluation.

Overfitting, Underfitting, and Generalization: Overfitting occurs when a model learns the training data too closely and performs poorly on unseen cases. It may memorize noise, artifacts, or dataset-specific shortcuts. Underfitting occurs when a model is too simple or insufficiently trained to capture relevant patterns.

Strategies for reducing overfitting include data augmentation, regularization, dropout, early stopping,

transfer learning, cross-validation, and larger datasets. However, a model can still fail under domain shift even when internal test performance is strong.

Generalization refers to performance on data outside the training distribution. In dermatology, this includes new patients, devices, hospitals, skin tones, anatomical sites, and disease prevalence. External validation is therefore essential for clinically intended systems.

2.3.5 Data Augmentation

Data augmentation creates modified training examples while preserving the intended label. Common image augmentations include rotation, flipping, scaling, cropping, translation, brightness adjustment, and color transformation.

Augmentation should reflect clinically plausible variation. Excessive color manipulation may remove diagnostically relevant information. Aggressive transformation may distort thin fungal hyphae, lesion borders, or dermoscopic structures. Augmentation strategy must therefore be matched to the clinical task.

Text augmentation may involve paraphrasing, question reformulation, or synthetic instruction generation. These approaches can increase linguistic diversity but may also introduce factual drift.

2.3.6 Cross-Validation

Cross-validation estimates model stability by dividing the development dataset into multiple folds. In k-fold cross-validation, the data are divided into k subsets. The model is trained on $k-1$ folds and validated on the remaining fold. This process is repeated until each fold has served as validation data.

Cross-validation is useful for small datasets, but the split should remain patient- or lesion-aware. If related images are placed in different folds, leakage may occur. A fully independent test set should remain untouched until final evaluation.

2.3.7 Computer Vision

Traditional Image Processing: Early dermatology image-analysis systems relied on manually designed processing pipelines. These typically included image normalization, hair removal, lesion segmentation, feature extraction, and classification.

Hand-crafted features were often inspired by clinical criteria. Shape features measured asymmetry, area, compactness, and border irregularity. Color features measured color distributions and the number of distinct colors. Texture features included local binary patterns, wavelets, and gray-level co-occurrence matrices. Dermoscopic structures such as pigment network, dots, globules, streaks, and blue-white structures were sometimes detected through specialized algorithms.

These systems had the advantage of explicit feature definitions, but their performance depended strongly on preprocessing and segmentation quality. Hand-crafted features could not easily represent the full complexity of dermatologic morphology.

Traditional Machine Learning: Traditional machine-learning methods use predefined features as inputs. Common classifiers include logistic regression, support vector machines, decision trees, random forests, k-nearest neighbors, and shallow neural networks.

A support vector machine identifies a boundary that separates classes in feature space. Kernel functions allow nonlinear boundaries. Random forests combine multiple decision trees and can model complex interactions. Logistic regression provides a probabilistic linear classifier and remains useful for interpretable baselines.

Traditional machine-learning models can perform well with limited data, particularly when meaningful features are available. However, feature engineering requires domain expertise and may not generalize across datasets.

2.3.7.1 Convolutional Neural Networks

Convolutional neural networks learn image features through convolutional filters. A convolutional layer applies a set of learnable filters across an image, producing feature maps. Early layers typically detect edges and local textures, while deeper layers capture more complex visual patterns.

Pooling layers reduce spatial resolution and increase tolerance to small translations. Activation functions introduce nonlinearity. Fully connected or global pooling layers convert learned features into class predictions.

Architectures such as AlexNet, VGG, Inception, ResNet, DenseNet, Xception, and EfficientNet have been applied to medical images [31]. CNNs transformed dermatology image classification because they learn visual representations directly from images. However, they may also learn shortcuts such as ruler markings, image borders, acquisition artifacts, or correlations between devices and diagnoses. Their predictions are often difficult to interpret clinically.

2.3.7.2 Vision Transformers

Vision transformers apply transformer-based attention mechanisms to images [32]. An image is divided into patches, and each patch is represented as a token. Self-attention allows the model to learn relationships between distant regions.

Unlike a conventional convolution, which primarily processes local neighborhoods, self-attention can model global relationships. This may be useful for overall lesion asymmetry, spatial distribution, or relationships between multiple structures. However, transformers often require large datasets or pretraining.

Hybrid architectures combine convolutional feature extraction with transformer-based global reasoning.

These models are particularly relevant for high-resolution microscopy and medical images in which both local detail and global context are important.

2.3.7.3 Image Classification

Image classification assigns one or more labels to an entire image. In dermatology, common tasks include benign–malignant classification, melanoma detection, multiclass lesion diagnosis, inflammatory disease classification, and infection screening.

Binary classification simplifies the output to two categories. Multiclass classification assigns one class from several alternatives. Multilabel classification allows several simultaneous labels, which may be appropriate when multiple morphological features are present.

Classification is suitable when the target occupies most of the image, as in many dermoscopic datasets. However, it does not specify where the relevant feature is located. A high malignancy probability does not explain whether the model detected an irregular network, blue-whitish veil, or another feature.

2.3.7.4 Object Detection

Object detection predicts both the class and location of one or more objects, usually using bounding boxes. Two-stage detectors, such as Faster R-CNN, first propose candidate regions and then classify them. One-stage detectors, such as YOLO, predict boxes and classes directly. Transformer-based detectors, such as Detection Transformer (DETR) and Real-Time Detection Transformer (RT-DETR), use learnable object queries and attention mechanisms [33, 34].

Object detection is useful when several small targets may occur within one image. In dermatology, applications include detecting fungal hyphae, acne lesions, hair follicles, ulcers, parasites, or multiple lesions in a body image [35].

Bounding boxes provide more interpretable evidence than image-level classification because the user can inspect the predicted region. However, bounding boxes provide only approximate localization and may not capture the exact shape of an irregular structure.

2.3.7.5 Image Segmentation

Segmentation assigns a label to each image pixel. Binary segmentation may separate a lesion from background skin. Multiclass segmentation may identify several structures. Instance segmentation distinguishes individual objects.

Traditional segmentation methods include thresholding, edge detection, region growing, active contours, watershed algorithms, and graph-based methods. Deep-learning segmentation is commonly performed using architectures such as U-Net and its variants [36].

In dermoscopy, segmentation can delineate the overall lesion or specific concepts such as blue-whitish veil, regression, pigment network, vascular structures, or ulceration. Such region-level labels can support more

precise explainability than heatmaps or bounding boxes.

Segmentation requires detailed manual annotation and can be time-consuming. Inter-annotator variability may also be high when feature boundaries are uncertain. Consensus protocols and annotation guidelines are therefore important.

2.3.7.6 Image Generation

Generative models learn to create new images. Generative adversarial networks use a generator and discriminator trained in competition.

Synthetic dermatology images may support augmentation, rare-disease simulation, or privacy-preserving research. However, image realism does not guarantee clinical validity. A generated lesion may contain anatomically inconsistent or diagnostically misleading structures. Generated medical images must therefore be reviewed by domain experts and should not be used as evaluation ground truth without rigorous validation.

2.3.8 Natural Language Processing and Language Models

Natural Language Processing: Natural language processing is the computational analysis and generation of human language. Traditional clinical NLP methods included rule-based systems, dictionaries, regular expressions, bag-of-words representations, and statistical classifiers.

Clinical NLP tasks include named-entity recognition, diagnosis extraction, medication extraction, coding, document classification, question answering, and summarization. Medical language is challenging because it contains abbreviations, specialist terminology, negation, uncertainty, duplicated documentation, and institution-specific conventions.

Text Representation and Tokenization: Computers process text as numerical representations. Earlier methods represented documents using word counts or term frequency–inverse document frequency (TF–IDF) vectors. Neural methods learn dense vector representations known as embeddings.

Modern language models divide text into tokens. A token may represent a word, part of a word, number, or punctuation mark. Tokenization affects context length and computational requirements. Medical terms may be split into several subword tokens when they are uncommon in the general training corpus.

Recurrent Neural Networks:

Before transformers, recurrent neural networks and long short-term memory networks were widely used for sequential data [37, 38]. They process text token by token and maintain a hidden state representing previous information.

Recurrent models can capture sequence order but are difficult to parallelize and may struggle with long-range dependencies. Transformers largely replaced them in large-scale language modeling.

2.3.8.1 The Transformer Architecture

The transformer uses self-attention to model relationships between tokens [39]. Each token is projected into query, key, and value representations. Attention weights determine how much information one token receives from others.

Scaled dot-product attention is commonly expressed as:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2.2}$$

where Q, K, and V represent query, key, and value matrices, and d_k is the key dimension.

Transformers can process sequences in parallel and model long-range relationships. Encoder-only models are often used for text understanding, decoder-only models for generative language modeling, and encoder–decoder models for sequence-to-sequence tasks.

2.3.8.2 Language Model Pretraining

A language model estimates the probability of text sequences. Autoregressive models predict the next token based on preceding tokens:

$$P(x_1,\ldots,x_T) = \prod_{t=1}^{T} P(x_t \mid x_1,\ldots,x_{t-1}). \tag{2.3}$$

Masked language models instead predict hidden tokens from surrounding context. Large-scale pretraining allows models to learn grammar, factual associations, common reasoning patterns, and task-general representations.

Pretraining data may include web pages, books, scientific articles, code, and other text. General models may have limited exposure to specialist dermatology language or may contain outdated or conflicting medical information.

Instruction Tuning: A pretrained language model predicts text but may not reliably follow user requests. Instruction tuning adapts the model using prompt–response examples. These examples teach the model to answer questions, summarize text, classify information, or follow specified output formats [40].

Instruction data may be written by humans or generated synthetically by a larger teacher model. Domain-specific instruction tuning can adapt a general model to medical tasks. However, the quality of the resulting model depends on the diversity, accuracy, and structure of the instruction data.

2.3.8.3 Large and Small Language Models

Large language models contain many parameters and are trained on extensive corpora. Model size can improve capability, but it also increases memory, computation, latency, and deployment cost.

Small language models (SLMs) use fewer parameters and may be suitable for local deployment. Their performance may be improved through task-specific prompting, fine-tuning, retrieval, or structured output constraints. A smaller specialized model may be preferable to a large general model when privacy, cost, and predictable behavior are priorities.

The boundary between large and small language models is not fixed. In this thesis, the term small or resource-efficient model refers primarily to models that can be deployed on accessible local hardware rather than requiring large cloud infrastructure.

Long-Context Clinical Reasoning: Longitudinal patient records may exceed the context length of conventional language models. Long-context models can process larger token sequences, allowing multiple years of clinical notes to be analyzed together.

However, longer context does not guarantee accurate retrieval. Models may overlook information in the middle of long documents, confuse dates, combine separate events, or repeat incorrect information. Longitudinal evaluation should therefore test categorical, numerical, and date-based retrieval separately.

Structured prompts can improve reliability by defining the desired feature, accepted output format, and interpretation rules. Retrieval systems can also divide long records into relevant sections before generation.

2.3.8.4 Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) combines a language model with an external information source [41]. A user query is used to retrieve relevant documents or passages, which are then provided to the model as context.

In dermatology, retrieval could provide current guidelines, drug safety information, textbook descriptions, or patient-specific historical notes. Retrieval can reduce reliance on the model's internal memory and improve provenance. However, the final output remains dependent on retrieval quality and model interpretation.

Clinical Summarization: Clinical summarization condenses records into a shorter representation. A useful dermatology summary may include diagnosis, phenotype, prior treatments, response, adverse events, current medication, latest laboratory values, disease activity, and follow-up plan.

Summaries must balance completeness and conciseness. Overly short summaries may omit important information, while overly long summaries may fail to reduce workload. Evaluation should therefore consider factual accuracy, completeness, temporal consistency, clinical usefulness, and length.

Hallucination and Factual Errors: A hallucination is generated information that is unsupported by the input or reliable evidence. Medical hallucinations can include invented diagnoses, treatments, dates, laboratory

values, or causal relationships.

Language models may also produce extraction errors without fabricating entirely new information. They may choose an incorrect value from the record, confuse a suspected diagnosis with a confirmed diagnosis, or interpret absence of documentation as absence of disease.

Mitigation strategies include retrieval, structured prompting, constrained outputs, expert review, model specialization, self-verification, and refusal mechanisms. No method completely removes risk.

2.3.8.5 Local Deployment and Privacy

Cloud-based language models transmit data to externally hosted infrastructure. This may be incompatible with institutional policies or data-protection requirements. Local deployment keeps model inference within institutional hardware.

Local deployment can improve data control, reduce dependence on network access, and provide predictable operating costs. However, it does not automatically guarantee privacy or compliance. Access control, encryption, audit logs, secure storage, software updates, and governance remain necessary.

2.4 AI in Dermatology

2.4.1 Evolution of Artificial Intelligence in Dermatology

AI in dermatology has evolved from early quantitative prognostic models and computer-aided image analysis systems to contemporary deep learning, foundation models, generative AI, and multimodal vision-language systems. Although dermatology AI is frequently associated with melanoma and skin cancer diagnosis, its clinical scope is now considerably broader. AI has been explored in melanoma diagnosis and prognosis, nonmelanoma skin cancers, inflammatory and immune-mediated dermatoses, infectious diseases, hair and nail disorders, wounds and ulcers, dermatopathology, cosmetic dermatology, dermatologic surgery, teledermatology, education, and patient-facing digital tools. This expansion reflects both the visual nature of dermatology and the increasing availability of clinical images, dermoscopic images, microscopy data, histopathology slides, longitudinal records, and textual clinical information.

Historically, much of the early work in dermatology AI focused on melanoma, because melanoma is clinically significant, visually challenging, and suitable for image-based computational analysis. An early line of research used tumor-progression variables to estimate survival among patients with stage I melanoma [42]. Subsequent work moved toward computerized analysis of pigmented lesions using clinical and dermoscopic images. Early studies investigated color, texture, border, and shape features for the classification of melanocytic lesions [43]. By the late 1990s, epiluminescence microscopy-based (dermoscopy) systems combined computerized image analysis with artificial neural networks to classify pigmented skin lesions [44]. These studies established the feasibility of automated lesion analysis before the deep learning era, but

they were often limited by small datasets, handcrafted features, heterogeneous acquisition settings, and limited external validation.

The pre-deep-learning era was characterized by classical computer-aided diagnosis pipelines. A typical pre-deep-learning system first cleaned and delineated the lesion, converted the image into a selected set of quantitative descriptors, and then passed those descriptors to a statistical or machine-learning classifier. Features were commonly derived from asymmetry, border irregularity, color variation, diameter, texture, and dermoscopic structures. Classical classifiers such as logistic regression, discriminant analysis, support vector machines, decision trees, k-nearest neighbors, and early neural networks were then used to distinguish melanoma from benign melanocytic lesions or other pigmented lesions [45, 46, 47]. These studies also exposed weaknesses that remain familiar today, including nonrepresentative sampling, incomplete verification of reference diagnoses, overfitting, limited reproducibility testing, and insufficiently independent validation.

The introduction of deep learning, particularly convolutional neural networks, marked a major shift in dermatology AI. Instead of relying primarily on manually engineered features such as ABCD (asymmetry, border, color, diameter) features [48], CNNs enabled models to learn hierarchical visual representations directly from image data. This led to rapid progress in skin lesion classification, especially in melanoma detection and benign–malignant discrimination [49, 50]. Public datasets and benchmarking initiatives further accelerated this transition. The International Skin Imaging Collaboration (ISIC) challenge series, beginning in 2016, provided standardized tasks for lesion segmentation, dermoscopic feature detection, and disease classification, creating a common framework for comparing algorithms across dermoscopic datasets [51, 52, 53]. This benchmark-driven period helped transform dermatology AI from isolated single-center studies into a more reproducible and internationally comparable research field.

Despite this progress, skin cancer has continued to dominate the dermatology AI literature. This is partly because dermoscopic images are relatively standardized compared with many other dermatologic image types, and because melanoma diagnosis provides a clinically important binary or multiclass classification task. However, dermatology is not limited to tumor diagnosis. AI applications have expanded into inflammatory dermatoses, including psoriasis [54], eczema [55], atopic dermatitis [56], and acne [57]; infectious diseases [58], including fungal disease [59] and parasitic infestations [60]; and chronic disease monitoring, including ulcers and wounds [61]. These applications often require different forms of data and evaluation, such as severity scoring, longitudinal assessment, treatment response prediction, and localization of clinically relevant structures rather than simple image-level classification.

AI has also become increasingly relevant in dermatologic microscopy and advanced imaging. In addition to clinical photography and dermoscopy, AI has been applied to reflectance confocal microscopy [62], optical coherence tomography [63], high-frequency ultrasound [64], multispectral imaging [65], KOH microscopy [66], and dermatopathology [67]. These modalities extend dermatology AI beyond surface-level photographic assessment toward more localized, microscopic, cellular, and tissue-level analysis. Reflectance confocal microscopy is particularly relevant because it enables non-invasive, near-histological visualization of the

skin, while AI-based image analysis may reduce subjectivity and support the interpretation of complex image patterns. Similarly, AI applied to microscopic and histopathological images may support the detection of fungal elements, inflammatory patterns, tumor structures, and other diagnostically relevant findings.

Beyond skin cancer and imaging, AI has begun to enter subspecialized areas of dermatology. Within trichology, AI is being used for automated measurement and severity scoring, longitudinal monitoring, treatment-response prediction, molecular stratification, analysis of patient-reported data, and assistance with transplantation procedures [68, 69, 70]. In onychology, models have been studied for recognizing fungal infection and pigmented lesions, automating severity assessment, interpreting capillary patterns, and detecting nail changes that may indicate systemic illness [71]. Dermatologic surgery and cosmetic dermatology are also emerging areas [67, 72]. These areas remain less mature than skin cancer AI, but they illustrate the broadening clinical scope of the field.

More recently, generative AI has introduced a new direction in dermatology. Generative models can be used to create synthetic dermatology images [73], generate image–text pairs [74, 75], and support dermatology education [76]. In principle, such approaches may help address data scarcity, class imbalance, and the limited availability of annotated rare disease images. However, generative AI also introduces important risks. Synthetic images may contain unrealistic artifacts, reproduce biases from the training data, or generate clinically misleading representations. Therefore, synthetic dermatology data should be considered a form of training supervision or augmentation rather than clinician-verified ground truth. Its value depends on careful quality control, transparent provenance, and robust downstream evaluation.

In parallel, natural language processing and large language models have created new opportunities for dermatology AI. Earlier NLP applications focused on extracting information from clinical notes, coding diagnoses, analyzing patient-generated text, and supporting literature retrieval [77]. Contemporary LLMs have expanded these possibilities to patient education, clinical summarization, question answering, guideline interpretation, report generation, and communication support. Closed-source commercial LLMs have received particular attention because of their accessibility and strong general language capabilities [78]. However, their use in dermatology also raises concerns regarding hallucination, lack of clinical grounding, limited reproducibility, opaque model updates, privacy, and the inability to inspect training data. These limitations are especially important in medical contexts, where fluent language may be mistaken for reliable clinical reasoning.

The current direction of dermatology AI is increasingly multimodal. Vision-language models and foundation models aim to integrate images with clinical descriptions, dermoscopic findings, patient history, anatomical site, metadata, and textual reasoning. This better reflects real dermatologic practice, where diagnosis is rarely based on an image alone. Multimodal systems may support visual question answering, structured reporting, differential diagnosis generation, localization of relevant findings, and longitudinal follow-up. However, these models also require more rigorous evaluation than conventional classifiers, because their outputs may include free-text reasoning, explanations, recommendations, or image-grounded descriptions. Accuracy alone is therefore insufficient; evaluation must also consider calibration, clinical relevance,

interpretability, safety, bias, and generalizability.

Overall, the evolution of AI in dermatology can be understood as a progression from early prognostic models and handcrafted image analysis systems, to CNN-based classifiers, benchmark-driven deep learning, advanced dermatologic imaging, generative AI, language models, and multimodal models. This historical trajectory has shifted the central question of the field. Early studies primarily asked whether computers could classify skin lesions. Contemporary dermatology AI must ask whether models can generalize across populations, support clinically meaningful decisions, provide interpretable outputs, preserve privacy, operate under real-world resource constraints, and integrate safely into clinical workflows. This thesis is positioned within this later stage of dermatology AI, where progress depends not only on model performance, but also on structured data foundations, clinically meaningful benchmarking, explainability, and responsible deployment.

2.4.2 Data Foundations for Dermatology Artificial Intelligence

The Role of Data in Artificial Intelligence: Artificial intelligence models learn patterns from data. Consequently, the characteristics of a dataset determine which patterns a model can learn and how reliably the model can be evaluated. A model trained only on melanoma and benign nevus images cannot be expected to recognize the full range of dermatologic disease. Similarly, a model trained on a small number of broad labels cannot be expected to distinguish fine-grained subclasses.

Dataset design is therefore part of the scientific problem rather than a purely technical preparation step. Important considerations include the clinical task, inclusion and exclusion criteria, diagnostic confirmation, image quality, annotation protocol, demographic diversity, disease prevalence, data provenance, licensing, and train–test splitting.

Dataset quality is not determined by size alone. A smaller dataset with reliable labels, clinically meaningful metadata, and patient-level splits may provide a more trustworthy benchmark than a larger but weakly labeled dataset. Conversely, large-scale datasets can be valuable for pretraining and representation learning even when their labels are less detailed. The appropriate balance depends on the purpose of the dataset.

Diagnostic Labels and Ground Truth: Ground truth refers to the reference information against which a model is trained or evaluated. In dermatology, ground truth may be established through histopathology, expert consensus, longitudinal stability, laboratory confirmation, or a combination of these sources.

Histopathological confirmation provides strong evidence for many tumors but is not available or clinically justified for every benign lesion. Requiring biopsy confirmation for all benign lesions would create selection bias because lesions selected for biopsy are often atypical. Long-term follow-up and expert consensus may therefore provide appropriate confirmation for clearly benign conditions.

Different sources of ground truth carry different levels of certainty. A benchmark should document how each diagnosis was established. Where uncertainty remains, labels may include confidence levels or

indeterminate categories rather than forcing a single definitive diagnosis.

Hierarchical Diagnostic Structure: Dermatologic diagnosis is hierarchical. For example, a clinician may first decide whether a lesion is melanocytic or nonmelanocytic, then whether it appears benign or malignant, and finally identify a specific diagnosis or subtype. A flat label does not capture these relationships.

Hierarchical labels provide several benefits. First, they support training at different levels of granularity. Second, they enable clinically meaningful error analysis. Confusing two benign nevus subtypes is different from confusing melanoma with a benign lesion. Third, hierarchical evaluation can reveal whether a model understands broad diagnostic groups but struggles with fine-grained distinctions.

A hierarchical taxonomy can be represented as a tree in which fine-grained subclasses belong to parent main classes and superclasses. Model probabilities can then be aggregated upward through the hierarchy. This allows the same model to be evaluated for subclass diagnosis, broader lesion groups, and malignancy risk.

Skin Tone, Fairness, and Representation: Skin tone affects the visual presentation of many dermatologic conditions. Erythema, pigmentation, pallor, cyanosis, and inflammation may appear differently across skin tones. If darker skin tones are underrepresented in training data, model performance may be unequal across patient groups [79].

Fitzpatrick phototype is frequently used as a practical skin-type annotation, although it was originally designed to describe response to ultraviolet exposure rather than visual skin color.

Fairness cannot be established by dataset diversity alone. Models must be evaluated across relevant subgroups, and differences in disease prevalence, image quality, anatomical site, and acquisition conditions must be considered. Subgroup analysis should also account for small sample sizes and uncertainty in skin-tone annotations.

2.4.2.1 Major Dermoscopic Datasets

Public dermoscopic datasets have been central to the development of skin lesion artificial intelligence. PH^2 provided an early benchmark containing common nevi, atypical nevi, and melanoma [80]. Derm7pt combined clinical and dermoscopic images with seven-point checklist information [81]. HAM10000 provided more than ten thousand dermoscopic images across seven common diagnostic categories and became widely used for multiclass classification [82]. BCN20000 expanded large-scale multiclass dermoscopy and included additional lesion categories [83].

The International Skin Imaging Collaboration archive has supported multiple challenges involving lesion segmentation, feature detection, and classification [84]. These challenges accelerated methodological development and improved reproducibility by providing shared datasets and evaluation protocols.

However, many dermoscopic datasets contain a limited number of broad classes. They may also contain substantial class imbalance, repeated images, multiple images from the same patient, or heterogeneous

diagnostic confirmation. These limitations are particularly important when models are intended for fine-grained clinical decision support.

2.4.2.2 Clinical Image Datasets

Clinical image datasets support research in teledermatology, inflammatory disease, infectious disease, fairness, and general skin disease classification. PAD-UFES-20 combines clinical images with patient metadata [85]. Fitzpatrick17k includes a large collection of clinical images with skin type-related annotations [86]. SCIN contains patient-contributed clinical images and associated symptom or demographic information [87]. The Diverse Dermatology Images (DDI) dataset was designed to support evaluation across different skin tones and includes pathology-confirmed diagnostic information [79].

Clinical image datasets are valuable because they represent a broader range of dermatologic diseases than dermoscopic datasets. However, they are often more heterogeneous in image quality and may contain uncertain labels, duplicated online images, or incomplete metadata. Dataset provenance and annotation reliability must therefore be carefully documented.

2.4.2.3 Multimodal Dermatology Data

A multimodal dataset combines two or more information types, such as clinical images, dermoscopic images, text, histopathology, laboratory results, or metadata. Multimodal data can support models that more closely reflect clinical reasoning.

For example, a clinical photograph may show a scaly plaque, while the history indicates chronic itching and flexural distribution. Neither source alone may be sufficient for diagnosis. Similarly, a pigmented lesion may be better interpreted when dermoscopy is combined with age, anatomical site, lesion evolution, and previous imaging.

Large-scale multimodal data are difficult to collect. Images and text may be stored in different systems. Reports may not be linked consistently to images. Clinical records contain sensitive information. Expert annotation is expensive. These factors create a major bottleneck for multimodal dermatology artificial intelligence.

2.4.2.4 Image–Text Pairs and Visual Question Answering Data

Vision-language models require paired visual and textual information. This may include an image and caption, image and report, image and question–answer pair, or image and multi-turn conversation.

Image–text pairs allow a model to learn associations between visual structures and language. Visual question answering data can teach a model to answer targeted questions, such as whether a lesion is pigmented, which anatomical site is visible, or what diagnosis is most likely. Open-ended questions can support descriptive reasoning, while single-select and multi-select questions provide more structured evaluation.

Dermatology has fewer naturally paired image–report datasets than radiology. Clinical photographs and dermoscopic images are often stored with only a short diagnosis or sparse metadata. As a result, synthetic data generation has become a practical strategy for expanding multimodal supervision.

2.4.2.5 Synthetic Data

Synthetic data are artificially generated or transformed data used to supplement real-world observations. In multimodal learning, a teacher language or vision-language model may generate questions, answers, captions, summaries, or clinical scenarios from images and metadata.

Synthetic data can increase scale, diversify instruction formats, and reduce manual annotation burden. Metadata-grounded prompts can constrain generation to known information and reduce hallucination. Published figure captions and inline text can provide stronger grounding than label-only prompts.

However, synthetic data inherit the weaknesses of the source information and teacher model. Generated text may contain factual errors, unsupported assumptions, stylistic repetition, or simplified reasoning. Synthetic data may also overrepresent the linguistic preferences of the teacher model. Therefore, filtering, provenance tracking, deduplication, and expert validation remain important.

Synthetic images can also be generated using generative adversarial networks or diffusion models. They may help augment rare conditions or underrepresented skin tones, but visually realistic images are not necessarily clinically accurate. Synthetic image morphology must be reviewed carefully before such data are used for training or evaluation.

2.4.2.6 Train, Validation, and Test Sets

A machine-learning dataset is commonly divided into training, validation, and test subsets. The training set is used to optimize model parameters. The validation set is used for model selection and hyperparameter tuning. The test set is reserved for final evaluation.

The split must reflect the unit of independence. In dermatology, several images may originate from the same patient or lesion. If images from the same patient appear in both training and test sets, the model may recognize patient- or lesion-specific features, resulting in overly optimistic performance. Patient-level or lesion-level splitting is therefore preferred where identifiers are available.

External testing uses data from a different institution, population, device, or time period. It provides stronger evidence of generalizability than a random split from the same dataset. Prospective evaluation provides additional evidence by testing the model on cases acquired after deployment into a clinical workflow.

2.4.3 Explainability and Interpretability in Dermatology AI

Interpretability and Explainability: Interpretability generally refers to the extent to which a user can understand how a model operates or why it produces an output. Explainability refers to methods that provide

human-understandable information about a model prediction. The terms are often used interchangeably, although they describe related but distinct concepts [88].

Explainability is particularly important in medicine because model outputs may influence diagnosis, treatment, biopsy decisions, or referral. A clinician needs to understand whether a model relied on medically plausible evidence or an irrelevant artifact.

Black-Box Predictions: A conventional CNN may produce a probability distribution over disease classes. For example, it may assign a high probability to melanoma. This output is useful but incomplete. It does not indicate which structures were detected or why melanoma was favored over a benign nevus.

Black-box models may also exploit shortcuts. A model could associate a ruler, marker, image border, or acquisition device with malignancy if these artifacts are correlated with biopsy decisions in the training data. High test accuracy does not guarantee that the model uses clinically meaningful evidence.

Saliency Maps: Saliency methods estimate how strongly individual pixels influence a prediction. A basic saliency map may compute the gradient of the predicted score with respect to input pixels. Regions with large gradients are displayed as important.

Saliency maps are easy to generate but can be noisy and unstable. Similar maps may be produced for different classes, and small model changes may alter the visualization. A saliency map indicates sensitivity rather than causal clinical reasoning.

Grad-CAM: Gradient-weighted class activation mapping, or Grad-CAM, uses gradients flowing into a convolutional layer to produce a heatmap of regions associated with a prediction [89]. It is widely used in medical imaging because it can be applied to trained CNNs without changing the architecture.

Grad-CAM can indicate whether a model focused on the lesion rather than the background. However, its spatial resolution is limited by the selected feature map. More importantly, the heatmap does not explain the dermatologic concept represented in the highlighted region.

For example, a heatmap over the upper-left region of a lesion does not state whether the model identified blue-whitish veil, regression, irregular pigmentation, or an artifact. Clinically meaningful explainability should therefore move from “where the model looked” toward “what feature the model identified and why that feature supports the diagnosis.”

Attention Maps: Transformer models produce attention weights representing relationships between tokens or image patches. These weights are sometimes visualized as explanations. However, attention weights do not necessarily correspond to clinical importance or causal contribution. Attention visualizations should therefore be interpreted cautiously and supplemented with other methods.

Localization- and Segmentation-Based Explanations: Object detection and segmentation can provide more concrete explanations by identifying visible structures. A fungal detection model can place a bounding box around a suspected hypha. A dermoscopic model can segment a blue-whitish veil region or localize

atypical vessels.

Such explanations align more closely with clinical evidence. They allow the clinician to verify whether the identified structure is present and relevant. However, localization requires detailed annotation and does not by itself explain how multiple features were combined into a final diagnosis.

Concept-Based Explanations: Concept-based models predict human-understandable features before making a diagnosis. In dermoscopy, these concepts may include pigment network, dots, globules, streaks, regression, blue-whitish veil, ulceration, and vascular structures.

A concept bottleneck model may first predict these features and then use them to predict the disease. This approach provides intermediate outputs that can be audited. However, concept labels are expensive to collect, may be subjective, and may not capture all information used by expert clinicians.

Language-Based Explanations: Language models and vision-language models can generate descriptive explanations. Instead of outputting only "melanoma," a model could state that the lesion is suspicious because of asymmetry, irregular pigmentation, blue-whitish veil, and regression structures.

Language-based explanations can be intuitive and clinically useful, but they introduce an important risk: fluency can be mistaken for correctness. A model may generate a plausible explanation for an incorrect diagnosis or describe a feature that is not present. Explanations must therefore be grounded in image evidence and evaluated separately from final-answer accuracy.

Uncertainty and Calibration: A model should communicate uncertainty, particularly in ambiguous or low-quality cases. Confidence scores are useful only if they are calibrated. A calibrated model produces probabilities that correspond to observed frequencies. For example, among cases assigned approximately 80% probability, roughly 80% should be correct.

Neural networks may be overconfident, especially under domain shift. Calibration methods include temperature scaling, ensemble methods, Bayesian approximations, and conformal prediction [90, 91]. Uncertainty should complement, rather than replace, clinician oversight.

2.5 Multimodal Language Models and Vision-Language Models

Multimodal Learning: Multimodal learning combines information from different data types. A multimodal medical model may integrate images, text, laboratory values, audio, or structured metadata.

A vision-language model is a multimodal model that processes images and text. In this thesis, the term VLM is used for models that accept visual and textual input and generate or classify language-based outputs.

VLMs are relevant to dermatology because clinical reasoning often combines what is visible with what is known from the patient history. They can support image captioning, visual question answering, report

generation, differential diagnosis, patient-facing explanations, and multimodal decision support.

Contrastive Image–Text Models: Contrastive models learn aligned image and text embeddings. During training, matching image–text pairs are encouraged to have similar representations, while non-matching pairs are separated. CLIP is a widely used example [92].

Once trained, these models can perform zero-shot classification by comparing an image embedding with text prompts. They can also provide embeddings for retrieval or downstream classifiers.

Contrastive models are efficient and useful for representation learning, but they do not naturally generate long explanations. Their performance depends strongly on the quality and specificity of text prompts.

Generative Vision-Language Models: Generative VLMs usually contain a visual encoder, a projection or adapter module, and a language model. The visual encoder converts an image into features. The adapter maps these features into a representation that the language model can process. The language model then generates text conditioned on the visual and textual inputs.

Some models keep the visual encoder and language model mostly frozen while training only the connector. Others fine-tune several components. The architecture and training strategy affect computational requirements and domain adaptation.

Visual Question Answering: Visual question answering requires a model to answer a question about an image. Questions may be closed-ended, single-select, multi-select, or open-ended.

A simple dermatology VQA task might ask whether the image is clinical or dermoscopic. More detailed questions may ask about lesion distribution, color, primary morphology, dermoscopic structures, differential diagnosis, or image quality.

VQA can evaluate more than final diagnosis. It can test whether a model recognizes intermediate clinical features. However, benchmark questions must be clinically valid and unambiguous. Synthetic questions may be useful at scale but should be validated carefully.

Image Captioning and Report Generation: Image captioning produces a description of an image. Medical report generation aims to produce a structured and clinically meaningful interpretation.

A dermatology report may include image modality, lesion location, morphology, color, border, surface, dermoscopic structures, impression, differential diagnosis, and uncertainty. The required report style depends on the intended user. A specialist report differs from a patient-facing explanation.

Report generation can improve interpretability, but it also creates new evaluation challenges. A fluent report may contain incorrect details. Report correctness, completeness, conciseness, and clinical safety must therefore be evaluated independently.

Image–Text Training Data: VLM development requires image–text pairs. Sources include manually written reports, scientific figure captions, textbook descriptions, clinical metadata, question–answer annotations,

and synthetic instruction data.

The trust level of these sources varies. Expert-written annotations provide strong clinical grounding but are expensive. Published captions are often reliable but may focus on selected or unusual cases. Metadata-grounded synthetic data provide scale but may contain simplified descriptions. Fully synthetic clinical scenarios provide broad instruction coverage but carry greater hallucination risk.

A robust training corpus can combine sources while preserving provenance and quality indicators. Higher-trust data may be weighted differently or used for final alignment and evaluation.

2.5.1 Parameter-Efficient Fine-Tuning

Full fine-tuning updates all model parameters and can require substantial memory. Parameter-efficient fine-tuning approaches, such as Low-Rank Adaptation (LoRA), update only a small number of additional parameters.

LoRA represents weight updates using low-rank matrices [93]. LoRA reduces memory requirements and allows separate adapters to be trained for different tasks. Quantized LoRA further reduces memory by loading the base model at lower numerical precision [94].

Parameter-efficient fine-tuning is suitable for resource-constrained medical research. However, it may provide less complete domain adaptation than large-scale pretraining or full continued pretraining.

Multimodal Hallucination: A VLM may produce text that is inconsistent with the image. It may describe a feature that is absent, infer unsupported patient information, or generate a confident diagnosis from a low-quality image.

Multimodal hallucination can arise from language priors overriding visual evidence. A model may generate the most common description associated with a diagnosis rather than accurately analyzing the image.

Grounding methods include region-level supervision, object detection, segmentation, concept annotations, image-quality assessment, and evidence-linked reporting. Benchmarks should test whether generated explanations correspond to visible structures.

2.6 Model Evaluation and Benchmarking

Purpose of Evaluation: Model evaluation estimates how well an artificial intelligence system performs on unseen data. The evaluation protocol should match the intended clinical use. A screening model should be assessed for sensitivity and false-negative risk. A microscopy detector should be assessed for localization and case-level diagnosis. A language model should be assessed for factual extraction and summary quality. A VLM should be assessed for visual understanding, answer correctness, report quality, and hallucination.

No single metric captures all aspects of clinical utility. Multiple complementary metrics and qualitative

analyses are usually required.

2.6.1 Confusion Matrix

For binary classification, predictions can be divided into:

- true positives (TP): positive cases correctly identified;
- true negatives (TN): negative cases correctly identified;
- false positives (FP): negative cases incorrectly predicted as positive;
- false negatives (FN): positive cases incorrectly predicted as negative.

These values form the confusion matrix and are used to calculate classification metrics.

2.6.2 Main Metrics

Accuracy: Accuracy measures the proportion of correct predictions:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.4}$$

Accuracy is intuitive but can be misleading in imbalanced datasets. If malignant lesions represent a small proportion of the data, a model may achieve high accuracy by predicting most cases as benign.

Sensitivity and Specificity: Sensitivity, also called recall or true-positive rate, measures the proportion of positive cases correctly identified:

$$\text{Sensitivity} = \frac{TP}{TP + FN}. \tag{2.5}$$

Specificity measures the proportion of negative cases correctly identified:

$$\text{Specificity} = \frac{TN}{TN + FP}. \tag{2.6}$$

Sensitivity is particularly important when missed disease is dangerous, such as melanoma or fungal infection. Specificity is important when false positives may lead to unnecessary biopsy, referral, or treatment.

Precision and F1-score: Precision, also called positive predictive value, measures the proportion of positive predictions that are correct:

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{2.7}$$

The F1-score is the harmonic mean of precision and recall:

$$F_1 = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{2.8}$$

For multiclass tasks, metrics may be macro-averaged, micro-averaged, or weighted. Macro averaging treats each class equally, while weighted averaging accounts for class frequency.

Balanced Accuracy: Balanced accuracy averages sensitivity across classes. For binary classification:

$$\text{Balanced Accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}. \tag{2.9}$$

It is useful when class distributions are imbalanced.

Receiver Operating Characteristic Curve and Area Under the Curve: A receiver operating characteristic (ROC) curve plots sensitivity against false-positive rate across decision thresholds. The area under the ROC curve (AUC) measures ranking performance. A value of 0.5 indicates random ranking, while 1.0 indicates perfect separation.

Area under the precision–recall curve (PR-AUC) may be more informative when the positive class is rare. Neither ROC-AUC nor PR-AUC identifies the best clinical threshold by itself.

2.6.3 Object Detection Metrics

Object detection requires both correct classification and spatial overlap. Intersection over Union measures overlap between predicted and reference bounding boxes:

$$\text{IoU} = \frac{\left|B_{\text{pred}} \cap B_{\text{gt}}\right|}{\left|B_{\text{pred}} \cup B_{\text{gt}}\right|}. \tag{2.10}$$

A detection is considered correct when the predicted class matches the ground-truth class and the intersection over union (IoU) exceeds a predefined threshold. Average Precision (AP) summarizes the precision–recall curve for a given class. Mean Average Precision (mAP) is then calculated by averaging AP across classes. These metrics may be reported at a single IoU threshold, such as $\text{IoU} = 0.50$, or averaged across multiple thresholds, commonly from 0.50 to 0.95.

Object-level performance should be distinguished from image-level clinical performance. A detector may miss one of several objects while still correctly classifying the image as positive. Conversely, one false-

positive object may cause an incorrect image-level diagnosis if a one-object-positive rule is used.

2.6.4 Segmentation Metrics

The Dice similarity coefficient measures overlap between predicted and reference masks:

$$\mathrm{Dice} = \frac{2|M_{\mathrm{pred}} \cap M_{\mathrm{gt}}|}{|M_{\mathrm{pred}}| + |M_{\mathrm{gt}}|}. \tag{2.11}$$

Pixel-level IoU is also commonly used:

$$\mathrm{IoU}_{\mathrm{mask}} = \frac{|M_{\mathrm{pred}} \cap M_{\mathrm{gt}}|}{|M_{\mathrm{pred}} \cup M_{\mathrm{gt}}|}. \tag{2.12}$$

Boundary metrics may be important when exact lesion borders or thin structures are clinically relevant.

2.6.5 Language Generation Metrics

Bilingual Evaluation Understudy (BLEU) measures n-gram precision between generated and reference text. Recall-Oriented Understudy for Gisting Evaluation (ROUGE) measures overlap with emphasis on recall. ROUGE-L uses the longest common subsequence. BERTScore, a BERT-based evaluation metric, compares contextual token embeddings and can capture semantic similarity beyond exact word overlap [95].

These metrics are useful for reproducible comparison but have important limitations in medicine. A clinically correct response may use different wording from the reference report and therefore receive a low score, while a fluent but medically incorrect response may still achieve a moderate score if the wording is similar. Automated text-generation metrics should therefore be interpreted alongside expert clinical evaluation.

2.6.6 Visual Question Answering Metrics

Closed-ended VQA can be evaluated using accuracy or exact match. Open-ended VQA may require normalization, semantic matching, embedding similarity, or human assessment.

Evaluation should distinguish between final diagnosis, morphology recognition, localization, reasoning, and report generation. A model may answer the diagnosis correctly for the wrong reason. Intermediate questions help determine whether the model understands clinically relevant features.

2.6.7 Evaluation Beyond Accuracy

Expert Evaluation: Clinician evaluation is important for generated reports and explanations. Experts may rate accuracy, completeness, conciseness, usefulness, safety, and hallucination. Pairwise preference can

compare model-generated and clinician-generated outputs.

Expert evaluation also has limitations. Ratings may vary between reviewers, and blinding may be difficult if model style is recognizable. Consensus guidelines, multiple reviewers, repeated evaluation, and inter-rater reliability analysis can improve robustness.

Hierarchical Evaluation: Hierarchical evaluation measures performance at several diagnostic levels. Fine-grained predictions can be aggregated into parent categories. For example, melanoma subtypes can be aggregated into a melanoma class, and malignant subclasses into a malignancy category.

This reveals whether the model preserves broad clinical relationships when the exact subclass is incorrect. It also identifies a granularity gap: strong performance at broad screening levels may coexist with weak fine-grained discrimination.

Calibration and Selective Prediction: Calibration evaluates whether confidence corresponds to correctness. Common measures include expected calibration error and reliability diagrams.

Selective prediction allows a model to abstain when uncertainty is high. Performance can be analyzed as coverage decreases. In medicine, a safe model may refer uncertain cases rather than provide a forced answer.

Fairness Evaluation: Fairness evaluation compares performance across relevant groups, including skin tone, sex, age, anatomical site, device, and institution. Differences should be interpreted with confidence intervals and sample-size considerations.

Fairness includes more than equal accuracy. False negatives and false positives may have different consequences across groups. Image quality and disease prevalence may also act as confounders.

Robustness and Domain Shift: Robustness measures performance under changes in image quality, compression, illumination, device, artifacts, or clinical setting. Stress testing may include blurred images, color shifts, occlusion, or adversarial perturbation.

Domain shift occurs when deployment data differ from training data. External validation and temporal validation are important for identifying this problem. Performance monitoring is required after deployment because clinical populations and workflows change over time.

Benchmark Design: A benchmark is a standardized dataset and evaluation protocol used to compare models. A clinically meaningful benchmark should define the intended task, source population, annotation procedure, split strategy, metrics, and limitations.

A benchmark should avoid leakage, ambiguous questions, duplicate cases, and unstable reference answers. For VLM evaluation, it should contain questions that reflect real clinical reasoning rather than only superficial image attributes.

Benchmark performance should not be interpreted as proof of clinical readiness. A benchmark is an

approximation of capability under specified conditions.

Task	Common metrics	Main interpretation
Classification	Accuracy, sensitivity, specificity, precision, F1-score, ROC-AUC, PR-AUC	Correctness, screening performance, false-positive and false-negative trade-offs
Object detection	Precision, recall, F1-score, IoU, AP, mAP	Ability to detect and localize individual structures
Segmentation	Dice, pixel IoU, boundary distance	Pixel-level overlap and boundary accuracy
Text generation	BLEU, ROUGE, BERTScore, expert rating	Lexical overlap, semantic similarity, clinical quality and usefulness
Visual question answering	Exact match, accuracy, semantic matching, expert review	Visual understanding, morphology recognition and reasoning
Calibration	Expected calibration error, Brier score, reliability diagram	Agreement between predicted confidence and observed correctness
Fairness	Subgroup sensitivity, specificity, F1-score and calibration	Consistency of performance across patient and acquisition groups

Table 2.2: Representative evaluation metrics for dermatology artificial intelligence tasks.

2.7 Data Annotation for Medical and Multimodal Artificial Intelligence

Data annotation is the process of converting raw clinical data into structured, machine-readable information that can be used for model training, validation, and benchmarking. In medical artificial intelligence, annotation may involve assigning a diagnosis to an image, identifying an anatomical structure, drawing a bounding box, creating a segmentation mask, writing a clinical description, or linking multiple data modalities at the patient or lesion level.

The appropriate annotation strategy depends on the clinical question and intended model output. A skin lesion classifier may require one diagnostic label per image, whereas a fungal microscopy detector requires spatial annotation of individual hyphae. A vision-language model may require image–text pairs, question–answer annotations, or multi-turn conversations. A longitudinal language model may require patient-level features, dates, treatment histories, and summary reports.

Before annotation begins, the unit of annotation should be clearly defined. Depending on the study, the unit may be:

- an image;

- a lesion;
- a patient;
- a clinical visit;
- a microscopic object;
- an image region;
- a text document;
- a question–answer pair; or
- a linked multimodal case.

Failure to define the annotation unit can result in inconsistent labels and data leakage. For example, multiple images may represent one lesion, and several lesions may belong to one patient.

2.7.1 Image-Level Diagnostic Annotation

Image-level annotation assigns one or more labels to an entire image. In a single-label classification task, each image is assigned one diagnosis. In a multilabel task, an image may receive several simultaneous labels, such as morphology, anatomical site, color, and surface features.

Image-level diagnostic labels may be obtained through histopathology, laboratory confirmation, expert consensus, longitudinal follow-up, or clinical records. The method of diagnostic confirmation should be recorded because not all labels have the same level of certainty. A pathology-confirmed melanoma label, for example, differs in evidential strength from a benign nevus label based on clinical stability.

Image-level labels are straightforward to use for classification but provide limited information about why the diagnosis was assigned. They do not indicate which visual structures are present or where they are located.

2.7.2 Metadata and Structured Clinical Annotation

Metadata describe the context, origin, or clinical characteristics of a sample. Common dermatology metadata fields include:

- image identifier;
- patient and lesion identifiers;
- image modality;
- diagnosis and diagnostic hierarchy;

- benign, malignant, or indeterminate status;
- anatomical site;
- patient age and sex;
- skin type or skin tone;
- image-acquisition device;
- source institution;
- diagnostic-confirmation method;
- train, validation, or test split;
- data license and provenance; and
- annotation status or confidence.

Metadata annotation is valuable because it enables filtering, subgroup analysis, hierarchical modeling, fairness assessment, and reproducible dataset splitting. However, metadata fields should use clearly defined vocabularies and consistent missing-value conventions. For example, "unknown," "not assessed," and "not applicable" should not be treated as equivalent without explicit definition.

2.7.3 Hierarchical Annotation

Hierarchical annotation represents information at multiple related levels. In dermatology, a lesion may be annotated first as melanocytic or nonmelanocytic, then as benign or malignant, and finally as a specific subclass. Hierarchical annotation can also be used for clinical questions in which selecting one answer activates relevant follow-up questions.

A hierarchical annotation workflow may be represented as:

$$\text{Question} \rightarrow \text{Option} \rightarrow \text{Subquestion} \rightarrow \text{Suboption} \tag{2.13}$$

This structure reduces unnecessary annotation by showing only questions relevant to the preceding response. For example, detailed dermoscopic criteria for melanocytic lesions need not be displayed when a lesion has already been classified as clearly nonmelanocytic.

Hierarchical annotations preserve clinical relationships between labels and support evaluation at different levels of diagnostic granularity. They also require careful schema design because an error at an early branch may prevent relevant downstream questions from being displayed.

2.7.4 Object Detection Annotation

Object detection annotation identifies individual structures using bounding boxes. Each annotation typically contains:

- the object class;
- the coordinates of the bounding box;
- the associated image identifier; and
- optional confidence, difficulty, or annotator information.

A rectangular bounding box can be represented by its upper-left coordinate and its width and height:

$$B = (x_{\min}, y_{\min}, w, h), \tag{2.14}$$

or by two opposing corners:

$$B = (x_{\min}, y_{\min}, x_{\max}, y_{\max}). \tag{2.15}$$

Object detection annotation is suitable when the relevant target occupies only part of the image or when several targets may be present. In dermatology, examples include fungal hyphae, acne lesions, ulcers, parasites, vessels, or multiple lesions within a body image.

Clinically relevant mimics may also be annotated as separate classes. For example, fibers and keratin edges that resemble fungal hyphae can be labeled as artifacts rather than treated only as background. This allows the model to learn an explicit distinction between the target and common sources of false-positive prediction.

Bounding boxes are faster to create than segmentation masks but provide only approximate spatial information. Thin, branching, or irregular structures may not be represented precisely by rectangular boxes.

2.7.5 Segmentation Annotation

Segmentation annotation assigns labels at the pixel level. A segmentation mask identifies the exact region occupied by a lesion or structure. The main forms include:

- semantic segmentation, in which all pixels belonging to the same class share one label;
- instance segmentation, in which separate objects of the same class are distinguished;
- panoptic segmentation, which combines semantic and instance information; and

- concept segmentation, in which specific clinically meaningful structures are delineated.

In dermoscopy, segmentation may be used to delineate the overall lesion or individual features such as blue-whitish veil, regression structures, pigment network, ulceration, or atypical vascular regions. For example, a segmentation mask could indicate the exact blue-whitish veil region in the upper-left portion of a lesion, enabling the model's explanation to be linked to visible clinical evidence.

Segmentation masks may be created using freehand contours, polygons, brush-based tools, or interactive segmentation systems. They may be stored as raster masks, polygons, or run-length encoded representations.

Pixel-level annotation is more precise than bounding-box annotation but is considerably more time-consuming. Some dermatologic structures have uncertain or gradual boundaries, which may lead to disagreement between annotators. Annotation guidelines should define whether uncertain transition regions are included or excluded.

2.7.6 Visual Question Answering Annotation

Visual question answering annotation pairs an image with one or more questions and reference answers. VQA data can assess visual recognition, morphology, clinical reasoning, and language generation.

Question formats may include:

- single-select questions;
- multi-select questions;
- short open-ended questions;
- long-form descriptive questions;
- confidence questions; and
- multi-turn conversational questions.

A dermatology VQA item may ask about image modality, image quality, lesion multiplicity, anatomical site, morphology, color, surface features, dermoscopic structures, diagnosis, or confidence. Open-ended questions may request a structured clinical report or explanation.

The reference answer should be sufficiently specific to support reliable evaluation while allowing clinically acceptable variation. A narrow exact-match answer may penalize valid synonyms, whereas an excessively broad reference may make evaluation unreliable. Controlled vocabularies and clinician-written consensus guidelines can improve consistency.

VQA datasets should also avoid answer leakage. If a question directly contains the diagnosis or if file

names reveal labels, a model may answer without analyzing the image. Questions should be designed to test the intended visual or clinical capability.

2.7.7 Text and Longitudinal Annotation

Clinical text annotation may involve identifying diagnoses, medications, dates, laboratory values, symptoms, treatment responses, adverse events, and disease activity. An annotation can mark a text span directly or assign a structured value to the complete document.

Longitudinal annotation requires temporal linkage. The same treatment may be started, discontinued, restarted, or adjusted across several visits. An annotation schema should distinguish current status from historical exposure and should record the date or period associated with an event.

Text fields may be:

- categorical, such as whether rituximab was ever used;
- numerical, such as the current corticosteroid dose;
- date-based, such as the most recent flare date; or
- open-ended, such as a longitudinal clinical summary.

Clinical definitions should be documented explicitly. For example, a suspected phenotype may not satisfy the annotation criteria for a confirmed phenotype. Such distinctions are important when evaluating language models.

2.7.8 Multimodal Annotation

Multimodal annotation links information across images, text, metadata, laboratory results, pathology, or longitudinal records. The central challenge is ensuring that all linked components refer to the same patient, lesion, specimen, and time point.

A multimodal case may include:

- a clinical photograph;
- a dermoscopic image;
- a pathology result;
- patient history;
- anatomical site;
- symptoms;

- diagnosis;
- treatment information; and
- follow-up outcome.

Unique patient, lesion, image, visit, and specimen identifiers should be preserved in de-identified form. Temporal relationships should also be recorded. Without reliable linkage, an image may be paired with an unrelated report or outcome, creating a clinically incorrect training example.

Multimodal annotations may contain variable missingness because not every patient undergoes every investigation. Missing values should be represented explicitly rather than inferred as negative findings.

2.7.9 Synthetic and Model-Assisted Annotation

Large language or vision-language models may assist annotation by generating draft captions, questions, answers, labels, or reports. Model-assisted annotation can reduce workload, but generated content should not be assumed correct.

Synthetic annotations should preserve provenance, including:

- source dataset;
- source metadata;
- generation model and version;
- prompt or seed task;
- generation date;
- filtering procedure; and
- human-review status.

Model-generated annotations may be used as training supervision, but clinician-authored or independently validated data are generally more appropriate for high-trust evaluation benchmarks. Training and evaluation data should be separated carefully to prevent the same teacher-generated style or content from appearing in both.

2.7.10 Annotation Formats and Data Schemas

Annotations must be stored in formats compatible with model-development and evaluation pipelines. Flat metadata are commonly stored in comma-separated values (CSV) files. CSV is simple and widely supported but is less suitable for nested questions, multiple objects, or multi-turn conversations.

JavaScript Object Notation (JSON) provides a flexible representation for hierarchical and multimodal data. A simplified VQA-style record may take the following form:

```
{
  "image_id": "image_001",
  "patient_id": "patient_001",
  "modality": "dermoscopic",
  "question_id": "Q18",
  "question": "What colors are present?",
  "answer": ["brown", "blue", "white"],
  "question_type": "multi-select",
  "annotator_id": "annotator_01"
}
```

JSON Lines (JSONL) stores one JSON object per line and is convenient for large datasets and streaming pipelines.

Object detection and segmentation annotations may use Common Objects in Context (COCO)-style JSON format, which separates image records, category definitions, and spatial annotations. A detection annotation may contain a bounding box, while a segmentation annotation may contain a polygon or run-length encoded mask.

Conversational VLM data may use a message-based structure:

```
{
  "image": "image_001.jpg",
  "messages": [
    {
      "role": "user",
      "content": "Describe the dermoscopic findings."
    },
    {
      "role": "assistant",
      "content": "The lesion shows..."
    }
  ]
}
```

Regardless of format, a dataset should include a schema definition explaining field names, data types, allowed values, missing-value conventions, hierarchy, and relationships between records.

2.7.11 Annotation Quality Control

Medical annotation should use predefined instructions and clinically standardized terminology. Quality-control procedures may include:

- annotator training;
- pilot annotation rounds;
- consensus guidelines;
- independent duplicate annotation;
- adjudication of disagreements;
- automated validation of required fields;
- duplicate and inconsistency detection;
- final clinical review; and
- version-controlled correction logs.

Agreement may be assessed using Cohen's kappa for two categorical annotators, Fleiss' kappa for several annotators, intraclass correlation for numerical ratings, or Dice and IoU for spatial masks. Agreement metrics should be interpreted with consideration of class prevalence and task difficulty.

Disagreement is not always an annotation failure. It may reflect true diagnostic ambiguity or uncertain visual boundaries. In such cases, consensus labels, uncertainty scores, multiple reference answers, or explicit indeterminate categories may be more appropriate than forcing artificial agreement.

2.7.12 Privacy, Provenance, and Auditability

Annotation platforms used for medical data should protect sensitive information. Local or on-premises deployment can keep images and records within institutional infrastructure. Access should be limited according to role, and annotation activity should be logged where appropriate.

Every annotation should ideally retain provenance information, including who created it, when it was created, which source data were used, whether it was modified, and whether consensus review was performed. This supports auditability and allows errors to be corrected without losing the history of the dataset.

Annotations should also be versioned. Changes to diagnostic labels, taxonomies, question schemas, or segmentation masks can affect model performance and benchmark comparability. Dataset releases should therefore include a version identifier and a record of major changes.

Overall, annotation is not merely a preparatory step. It defines the clinical information available to the model, the outputs the model can learn to generate, and the capabilities that can be evaluated. Carefully designed annotation schemas are therefore fundamental to the reliability, interpretability, and translational value of medical artificial intelligence.

2.8 Challenges for Clinical Deployment

Privacy and Data Governance: Dermatology images and clinical records are sensitive medical data. Clinical photographs may contain faces, tattoos, jewelry, anatomical landmarks, or other identifying information. Text records may contain direct identifiers and detailed longitudinal history.

Data must be collected, stored, accessed, and processed according to institutional and legal requirements. De-identification reduces risk but may be imperfect. Governance should define who can access data, how it can be used, how long it is retained, and how model outputs are audited.

Local deployment can reduce external data transfer but does not replace governance. Secure authentication, encryption, access logs, backups, and incident response remain necessary.

Computational Cost and Resource Efficiency: Large models may require expensive GPUs, substantial memory, long inference times, and specialized engineering support. These requirements can limit adoption in hospitals, clinics, laboratories, and low-resource settings.

Resource-efficient methods include transfer learning, frozen feature extraction, LoRA, quantization, pruning, distillation, and smaller task-specific models. A smaller specialized model may provide an appropriate balance between capability and deployability.

Efficiency should be evaluated using memory use, inference speed, energy consumption, hardware requirements, and total cost. Model size alone does not determine efficiency; long outputs and repeated reasoning steps can also increase cost.

Local and On-Premises Deployment: On-premises deployment places the model and data within institutional infrastructure. This may support privacy, lower latency, offline operation, and local customization.

However, local deployment creates maintenance responsibilities. Institutions must manage hardware, updates, cybersecurity, model versions, and monitoring. A local model can also become outdated if clinical knowledge or guidelines change.

A hybrid deployment may use local models for sensitive data and external services for non-sensitive retrieval or general computation. Such architectures require clear separation of data flows.

Bias and Equity: Bias can arise from data collection, disease prevalence, annotation, model design, threshold selection, and deployment context. A model trained predominantly on one population may not generalize to others.

Dermatology is especially sensitive to visual representation across skin tones. Fair deployment requires diverse data, subgroup evaluation, error analysis, and monitoring. Models should not be assumed fair because their training dataset contains some diverse examples.

Human Factors and Clinical Workflow: Clinical usefulness depends on how a model is integrated into the clinical workflow. An accurate model may be ignored if it is slow, difficult to access, or produces excessively long outputs. A less accurate model may cause harm if users overtrust it.

Interface design should communicate prediction, confidence, evidence, and limitations. Outputs should be concise and adapted to the user. A dermatologist, primary care physician, trainee, and patient may require different information.

Automation bias occurs when users accept incorrect AI recommendations. Alert fatigue occurs when systems produce too many low-value warnings. Training and usability testing are therefore essential.

Regulatory Considerations: Artificial intelligence software may be regulated as a medical device when it is intended to diagnose, monitor, predict, or guide treatment. Regulatory requirements depend on intended use, risk, jurisdiction, and degree of autonomy.

Developers may need to document data provenance, model design, validation, cybersecurity, human oversight, update procedures, and post-market monitoring. Adaptive models create additional challenges because performance may change after deployment.

Research models should be clearly distinguished from clinically approved systems. Strong retrospective performance does not establish safety or effectiveness for patient care.

Monitoring and Model Drift: Model performance may change over time because of new devices, changes in disease prevalence, altered referral patterns, or updated clinical practice. This is known as model drift.

Deployment should include monitoring of data quality, output distribution, calibration, subgroup performance, failure cases, and user feedback. Model versions should be recorded so that decisions can be audited.

Retraining or recalibration may be necessary, but updated models should undergo validation before replacement.

Clinical Evidence and Prospective Evaluation: Retrospective evaluation is appropriate for model development but provides limited evidence of real-world benefit. Prospective studies test models on newly acquired cases. Silent deployment can evaluate performance without exposing outputs to clinicians. Interventional studies can then assess whether AI changes clinician behavior, diagnostic accuracy, workflow, or patient outcomes.

Relevant outcomes may include time saved, missed malignancies, unnecessary biopsies, treatment delays, documentation quality, user trust, and patient safety. Clinical utility should be demonstrated in the workflow

for which the model is intended.

2.8.1 Final Notes on Deployment Challenges for Multimodal Dermatology AI

Multimodal artificial intelligence systems aim to combine complementary sources of medical information, such as clinical images, dermoscopy, microscopy, histopathology, laboratory values, patient history, and longitudinal electronic health records. In principle, this more closely reflects clinical decision-making than models based on a single image or isolated text document. A dermatologist, for example, may interpret a lesion by combining its clinical appearance, dermoscopic structures, anatomical location, symptoms, evolution, previous treatments, and histopathological findings. However, the integration of these modalities introduces substantial technical, clinical, and organizational challenges that extend beyond those encountered in unimodal systems.

A major challenge is that medical modalities are rarely collected in a standardized or synchronized manner. Images, reports, laboratory results, and clinical notes may be stored in separate systems, recorded at different time points, and linked using inconsistent identifiers. A dermoscopic image may be acquired during one visit, biopsied at a later date, and associated with a pathology report stored in another database. Clinical notes may describe several lesions without clearly indicating which image or biopsy corresponds to each lesion. Consequently, constructing reliable multimodal datasets requires accurate patient-, visit-, and lesion-level linkage. Incorrect linkage can create misleading training examples in which an image is paired with unrelated text or an incorrect diagnostic outcome.

Missing modalities represent another important difficulty. In routine care, not every patient undergoes dermoscopy, biopsy, laboratory testing, or long-term follow-up. A multimodal model trained only on complete cases may therefore learn from a selected population of more complex or severe cases. Conversely, a model that requires every modality at inference may be unusable when one source is unavailable. Clinically deployable systems should therefore be designed to operate with variable combinations of inputs and should communicate how missing information affects confidence. Approaches such as modality masking, missing-data-aware training, and separate specialist modules may improve robustness, but they require careful validation.

The quality and meaning of each modality may also vary substantially. Clinical photographs differ in illumination, framing, distance, device type, and color balance. Dermoscopic images vary across acquisition systems, polarization modes, magnification levels, and interface media. Clinical text contains abbreviations, copied content, uncertainty, negation, and institution-specific terminology. Histopathological reports may contain detailed differential diagnoses or only short final conclusions. Combining these sources does not automatically eliminate their individual weaknesses; instead, errors or biases from one modality may propagate through the entire system.

Multimodal models must also learn when information sources agree, complement one another, or conflict. For example, an image may appear consistent with a benign lesion, while the clinical history describes

rapid change or bleeding. A pathology report may contradict an earlier clinical diagnosis. A model should not simply average these signals. It should recognize their different levels of reliability, timing, and clinical relevance. This requires temporal reasoning, source attribution, and uncertainty handling, which remain difficult for current multimodal systems.

Another challenge is visual and textual grounding. Vision-language models can generate fluent explanations, but the generated text may not correspond to structures that are actually visible in the image. A model may state that a lesion contains a blue-whitish veil, atypical vessels, or ulceration even when these features are absent. Clinically useful systems should link generated statements to identifiable visual evidence through bounding boxes, segmentation masks, region-level annotations, or concept predictions. For example, rather than providing only a melanoma probability or an attention heatmap, a grounded system should be able to identify and segment the region containing a blue-whitish veil and explain how this feature contributed to the diagnostic interpretation. Such grounding improves auditability but requires detailed and resource-intensive annotations.

The fusion strategy also affects model performance and interpretability. Early fusion combines modalities near the model input, while late fusion combines predictions or representations generated by separate modality-specific models. Intermediate fusion integrates information at multiple representation levels. Early fusion may capture complex interactions but may be sensitive to missing inputs and difficult to interpret. Late fusion is more modular and allows specialized models to be evaluated separately, but it may fail to capture subtle cross-modal relationships. No single strategy is optimal for every medical task, and the selected architecture should reflect the intended clinical workflow.

Computational requirements present a further barrier. Multimodal models may contain separate image encoders, language models, projection layers, retrieval systems, and specialist modules. Their training can require large image–text datasets and extensive GPU resources. Inference may also become slow when high-resolution images, long patient records, and multiple reasoning steps are processed together. These constraints are especially relevant in hospitals and research environments with limited infrastructure. Parameter-efficient fine-tuning, quantization, model compression, knowledge distillation, and modular specialist models can reduce resource requirements, but any reduction in model size must be assessed for its effect on clinical performance and safety.

Evaluation is also more difficult for multimodal models. Conventional accuracy metrics do not determine whether a model used the image, the text, or a spurious correlation. A system may answer correctly using language priors while ignoring the image. Benchmarks should also test performance when modalities are missing, contradictory, low quality, or temporally misaligned.

Human factors are equally important. Multimodal outputs can become excessively detailed and may increase rather than reduce cognitive workload. Clinicians generally require concise, structured, and actionable information rather than unrestricted model-generated narratives. Interfaces should distinguish information observed directly in the image from information retrieved from the record or inferred by the

model. They should also expose uncertainty, missing modalities, conflicting evidence, and the provenance of retrieved information.

Regulatory assessment of multimodal AI may be more complex than that of a single-purpose classifier because the system may perform several functions, use several data sources, and change behavior according to the available context. Intended use, target population, required inputs, output interpretation, human oversight, and model-update procedures must therefore be clearly defined. Modular systems may require both component-level validation and evaluation of the complete integrated workflow.

Overall, multimodal AI offers an opportunity to model medical reasoning more faithfully than isolated image or text systems. However, combining modalities does not automatically produce a safer or more clinically useful model. Successful deployment requires reliable data linkage, robustness to missing information, evidence-grounded outputs, resource-efficient architectures, privacy-preserving infrastructure, clinically appropriate evaluation, and integration into real workflows. These requirements are particularly relevant to dermatology, where visual findings, morphology, patient history, and longitudinal information must often be interpreted together.

2.9 Synthesis of the Literature and Research Gaps

The literature reviewed in this chapter demonstrates substantial progress in dermatology artificial intelligence, but also identifies several unresolved gaps.

The first gap concerns data structure. Existing dermoscopic datasets have enabled major advances in melanoma and skin lesion classification, but many contain a limited number of broad classes. They may not support fine-grained hierarchical evaluation or detailed failure analysis. Clinical image datasets provide broader disease coverage but often contain heterogeneous image quality, limited diagnostic verification, or incomplete metadata.

The second gap concerns multimodal supervision. Dermatologic diagnosis depends on images, morphology, history, symptoms, and longitudinal information, yet large-scale paired image–text datasets remain scarce. This limits the development of dermatology-specific vision-language models. Synthetic data provide a scalable approach, but require grounding, filtering, and validation.

The third gap concerns explainability. Conventional CNN classifiers usually provide labels and probabilities. Post-hoc methods such as Grad-CAM show which region influenced a prediction but do not identify the clinical concept present in that region. More useful systems should connect predictions to localized, segmented, or language-described evidence.

The fourth gap concerns longitudinal reasoning. Most dermatology AI studies focus on single images or isolated questions. Chronic disease management requires analysis of long-term records, treatment history, adverse events, and disease activity. Privacy-preserving local language models represent a promising approach, but require task-specific evaluation and governance.

The fifth gap concerns resource efficiency. Many advanced VLMs depend on large proprietary systems and cloud infrastructure. This limits reproducibility, accessibility, and use with sensitive data. Smaller models, local deployment, synthetic supervision, and parameter-efficient fine-tuning may improve practical accessibility.

The sixth gap concerns benchmarking. Standard classification benchmarks cannot fully evaluate hierarchical diagnosis, visual reasoning, report generation, uncertainty, or multimodal understanding. Clinically grounded benchmarks should include both structured and open-ended questions, expert annotations, diagnostic hierarchy, skin-tone diversity, and detailed error analysis.

These gaps provide the rationale for the subsequent contribution chapters. Chapter 3 develops data foundations for hierarchical dermoscopic analysis and multimodal dermatology modeling. Chapter 4 investigates representative artificial intelligence methods, progressing from CNN classification to object detection, local language models, and dermatology vision-language models. Chapter 5 presents hierarchical and multimodal evaluation approaches for dermatology artificial intelligence systems.

2.10 Chapter Summary

This chapter introduced the clinical and technical foundations of artificial intelligence in dermatology. Dermatology was described as a visual, multimodal, and longitudinal specialty that relies on clinical photography, dermoscopy, microscopy, histopathology, and textual patient records. The chapter reviewed the role of diagnostic morphology, patient history, imaging modality, and longitudinal context in clinical decision-making.

The chapter then introduced the data foundations of dermatology artificial intelligence. Public clinical and dermoscopic datasets have enabled major advances, but limitations remain in diagnostic granularity, class balance, demographic representation, multimodal supervision, and standardized evaluation. Hierarchical labels, patient-level splitting, diagnostic confirmation, image–text pairs, and synthetic data were identified as important components of modern dataset design.

Fundamental machine-learning concepts were explained, including supervised learning, self-supervised learning, transfer learning, foundation models, cross-validation, augmentation, overfitting, and generalization. Computer vision methods were reviewed from traditional image processing and machine learning to CNNs, vision transformers, object detection, and segmentation.

Explainability was discussed as a progression from black-box probabilities and saliency heatmaps toward localization, segmentation, concept-based prediction, and grounded language explanations. The distinction between showing where a model attended and explaining which clinical feature supports a diagnosis was emphasized.

Natural language processing, transformers, large and small language models, longitudinal clinical summarization, retrieval-augmented generation, hallucination, and local deployment were then introduced.

Vision-language models were described as systems capable of integrating image and text data for visual question answering, report generation, and multimodal reasoning.

Finally, the chapter reviewed evaluation metrics and clinical deployment challenges. Trustworthy dermatology artificial intelligence requires appropriate train–test separation, task-specific metrics, hierarchical benchmarking, expert evaluation, fairness analysis, uncertainty assessment, privacy protection, resource efficiency, workflow integration, and prospective clinical validation. These foundations support the datasets, models, and benchmarks developed in the following chapters.

Chapter 3

Data Foundations for Dermatology AI

This chapter begins with an overview of the datasets and data infrastructure developed for dermatology artificial intelligence (Section 3.1). First, DERM12345 is presented as a large, hierarchical dermoscopic skin lesion dataset designed to support fine-grained and hierarchical dermatology AI research (Section 3.2). Next, DermaSynth is introduced as a synthetic multimodal dermatology image–text dataset constructed using open-access dermatology datasets, PubMed Central resources, and instruction-following approaches to facilitate vision-language model development (Section 3.3).

Parts of this chapter are reprinted and adapted from the following publications:

- (Section 3.2) **A. Yilmaz**, S. P. Yasar, G. Gencoglan, B. Temelkuran. "DERM12345: A Large, Multisource Dermatoscopic Skin Lesion Dataset with 40 Subclasses," Scientific Data, https://doi.org/10.1038/s41597-024-04104-3
- (Section 3.3) **A. Yilmaz**, F. Yuceyalcin, R. Varol, E. Gokyayla, O. Erdem, D. Choi, A. A. Demircali, G. Gencoglan, J. M. Posma, B. Temelkuran. "DermaSynth: Rich Synthetic Image-Text Pairs Using Open Access Dermatology Datasets," arXiv, https://doi.org/10.48550/arXiv.2502.00196

3.1 Overview

Dermatology artificial intelligence relies heavily on the availability of high-quality datasets tailored to different clinical workflows. In dermatology, two major data paradigms are particularly important. First, dermoscopic imaging represents one of the most important modalities for skin cancer diagnosis and pigmented lesion analysis. Dermoscopy enables visualization of subsurface skin structures, significantly improving diagnostic accuracy compared to naked-eye examination. Consequently, many dermatology AI studies have focused on dermoscopic image analysis for melanoma detection and skin lesion classification. Nevertheless, most existing dermoscopic datasets contain only a limited number of diagnostic categories, restricting the development of fine-grained and hierarchical AI systems. To address this limitation, DERM12345 was developed as a large hierarchical dermoscopic skin lesion dataset containing 12,345 high-resolution dermoscopic images collected in Türkiye, at the crossroads of Europe and Asia, from a population with diverse skin types. The dataset introduces a hierarchical taxonomy with 5 superclasses, 15 main classes, and 40 subclasses, specifically designed to support advanced dermatologic AI research, reduce false predictions, and improve fine-grained failure analysis.

While dermoscopic datasets are essential for skin cancer AI, dermatology extends far beyond lesion classification. General dermatologic diseases often require multimodal reasoning involving clinical photographs, patient history, pathology reports, and longitudinal follow-up. This creates a major challenge for dermatology vision-language model development because large image–text pair datasets remain scarce. Unlike radiology or pathology, dermatologic images are frequently associated with limited textual information and lack formal accompanying reports. To address this limitation, DermaSynth was developed as a synthetic multimodal dermatology image–text dataset generated from open-access dermatology datasets, PubMed Central resources, and instruction-following approaches. Using clinically relevant prompting strategies, metadata-informed question construction, and Gemini 2.0 Flash, DermaSynth generated 204,809 synthetic image–text pairs from 83,622 dermatology images, including both dermoscopic and clinical images. For the open-access image pipeline, available metadata were used to inform question and prompt construction, while the answer-generation instruction explicitly directed the model to base its responses on visible image characteristics rather than the supplied metadata.

Together, DERM12345 and DermaSynth address two complementary but interconnected challenges in dermatology AI. DERM12345 focuses on hierarchical dermoscopic skin lesion analysis and skin cancer-related AI development, while DermaSynth addresses the broader need for scalable multimodal supervision in dermatology vision-language modeling. Collectively, the datasets and infrastructure presented in this chapter establish the data foundations required for the explainable, resource-efficient, and multimodal dermatology AI systems developed in Chapter 4 and the benchmarking strategies presented in Chapter 5.

3.1.1 Dermoscopic Image Datasets

Dermoscopy, also known as dermatoscopy or epiluminescence microscopy, is the inspection of the skin using magnifying lenses and polarized or non-polarized filtered illumination, allowing examination of lesions invisible to the naked eye. It is the most extensively used and standardized diagnostic procedure in clinics to diagnose skin lesions. Various types of dermatoscopes are actively used in routine clinical practice. Hand-held dermatoscopes, for example, have become widely available, affordable, and commonly used, particularly in resource-limited settings. However, dermatologists' diagnostic performance in dermoscopic skin lesion identification is closely linked to their training and prior clinical experience, highlighting inherent variability in human decision-making. The prevalence and distribution of malignant and benign lesions, which vary by population, geographic factors, and skin type, also affect dermatologists' performance. Addressing this gap, artificial intelligence models trained on specialized datasets can reduce the experience gap that results from the human factor during the clinical examination of these lesions [96].

One of the earliest studies addressed the classification of three classes: melanoma, nevus, and seborrheic keratosis [51, 52]. HAM10000, published in 2018, contains 10,015 dermoscopic images across seven common diagnostic categories [82]. BCN20000 later expanded the taxonomy to eight classes by including squamous cell carcinoma, a common lesion not represented in HAM10000 [83]. New datasets continue to be published in open-access repositories such as the International Skin Imaging Collaboration (ISIC) Archive [84]. Enlarging open-access datasets supports the development of more robust and intelligent AI algorithms.

Skin lesion datasets can be grouped into four categories: clinical image datasets, histopathology image datasets, dermoscopic image datasets, and multimodal datasets. Clinical image datasets such as the Interactive Atlas of Dermatoscopy and the Dermofit Image Library are available for a fee [97, 98]. The MED-NODE Dataset [99], the Asan and Hallym Dataset [100], the SD-198/SD-260 Datasets [101, 102, 103], PAD-UFES-20 [85], SCIN [87], Atlas Dermatologica [104], DermaAmin [105], and The SLICE-3D dataset with over 400,000 images (1,218 images with diagnoses) [106] are all open-access. Fitzpatrick 17k Dataset is a dataset sourced from Atlas Dermatologica and DermaAmin, with the addition of Fitzpatrick skin type information [86]. DDI dataset also presents skin images with their skin type [79]. The Cancer Genome Atlas includes a histopathology image dataset of skin lesions [107]. Clinical, dermoscopic, and histopathology images can all be found on DermNet NZ for free (high-resolution images for a fee) [108]. Study results indicate that both clinicians and AI systems achieve higher diagnostic accuracy using dermoscopic image datasets than clinical image datasets, highlighting the importance of dermoscopic imaging for skin lesion diagnosis [109]. Histopathology images require tissue excision from the patient. For this reason, histopathology image datasets are less practical for routine image acquisition than dermoscopic datasets.

The PH^2 dataset was published for benchmarking dermoscopic images and includes three lesion classes (common nevus, atypical nevus, and melanoma (mel)) [110]. The Derm7pt dataset, containing 2,000 dermoscopic and clinical images across five lesion classes (basal cell carcinoma (bcc), mel, miscellaneous,

nevus (nv), and seborrheic keratosis (sk)), was published along with the 7-point checklist pattern analysis and its automation by using artificial neural networks [81]. The ISIC Archive, one of the largest dermoscopic image repositories used in scientific research, contained 549,571 public images and 1,203,225 total images (accessed on 10 May 2026). This archive contains various datasets, including ISIC 2016, with 1,279 images and two lesion classes (benign and malignant) [51]; ISIC 2017, with 2,000 images and three lesion classes (mel, nv, and sk) [52]; ISIC 2018 [53], with 12,500 images and three tasks: lesion segmentation, attribute detection, and disease classification, with the classification task comprising 10,015 HAM10000 images across seven lesion classes (akiec, bcc, benign keratosis-like (bkl), dermatofibroma (df), mel, nv, and vascular (vasc)) [82]; BCN20000, with 18,946 images and eight lesion classes (actinic keratosis (ak), bcc, df, mel, nv, squamous cell carcinoma (scc), bkl, and vasc) [83]; the ISIC 2020 Patient-Centric dataset, with 44,108 images and two lesion classes (benign and mel) [111]; and the PROVe-AI dataset, with 603 images and 15 lesion classes [112]. Lastly, the dataset collected in Argentina with 10 subclasses (ak, bcc, df, mel, nv, scc, sk, solar lentigo (sl), lichenoid keratosis (lk), and vasc) was added to the ISIC Archive [113]. Between 2016 and 2025, numerous competitions were conducted for tasks such as lesion segmentation, feature extraction, and lesion classification, using these ISIC datasets.

3.1.2 Multimodal Dermatology Datasets

While dermoscopic image datasets have played a central role in the development of skin lesion and skin cancer AI, dermatology as a clinical discipline is not limited to isolated image interpretation. Many dermatologic diseases require the integration of multiple sources of information, including clinical photographs, dermoscopic images, anatomical site, patient history, symptoms, lesion evolution, histopathology, laboratory findings, treatment response, and longitudinal follow-up. Therefore, the next generation of dermatology AI systems requires datasets that go beyond image-label pairs and support multimodal learning.

Multimodal dermatology datasets can take several forms. Some datasets combine clinical and dermoscopic images from the same lesion, allowing models to learn complementary visual information from macroscopic and dermoscopic views [114]. Others include clinical/dermoscopic images and metadata, such as age, sex, anatomical location, skin type, symptoms, or diagnostic confirmation method [82, 115, 79]. More recent datasets have begun to include image–text pairs, captions, visual question-answering annotations, or patient-facing clinical descriptions, enabling the development and evaluation of vision-language models [74, 116, 75]. These datasets are particularly important because multimodal language models require not only visual inputs, but also clinically meaningful textual supervision.

The importance of multimodal data is especially clear for general dermatologic diseases. Unlike dermoscopic skin cancer analysis, where the image often focuses on a single lesion, common dermatologic diseases may present with multiple lesions, complex distribution patterns, secondary morphological changes, and variable clinical histories. For example, distinguishing inflammatory, infectious, autoimmune, and drug-related eruptions may require information about duration, itch, pain, medication exposure, systemic symptoms, anatomical distribution, and previous treatment response. Image-only models may

therefore be insufficient for these tasks, whereas multimodal datasets can better approximate real clinical reasoning.

However, large-scale multimodal dermatology datasets remain limited. Unlike radiology, where imaging studies are often accompanied by formal reports, dermatology images are frequently stored with only short diagnostic labels or sparse metadata. Clinical photographs may not be linked to detailed patient history, dermoscopic images may not include narrative descriptions, and open-access repositories rarely provide full longitudinal context. In addition, medical data sharing is constrained by privacy concerns, patient-identifiable information embedded in images, institutional governance requirements, and heterogeneous licensing conditions. These factors make it difficult to collect large, diverse, and clinically rich image–text datasets directly from routine practice.

Existing dermatology VQA and image–text datasets have begun to address this gap, but they remain limited in scale, modality coverage, annotation depth, or clinical grounding. Some resources focus primarily on clinical images, while others are derived from online forums, textbook materials, scientific figures, or metadata-enriched public datasets. Although these resources are valuable, they may not fully capture the diversity of dermatologic diseases, skin tones, image acquisition settings, or clinically relevant reasoning patterns. Furthermore, many available datasets contain limited structured information about morphology, diagnostic uncertainty, or follow-up context.

Synthetic data generation has emerged as a practical strategy to address these limitations [117]. By using large language models or vision-language models to generate captions, question–answer pairs, reports, and instruction-following examples from existing images and metadata, synthetic data can expand multimodal supervision without requiring every image to be manually annotated in detail. Metadata-grounded prompting can help reduce hallucination by anchoring generated text to known diagnostic labels, modality information, anatomical location, or other available fields. PubMed (PM)-derived images and captions can also provide a higher-trust source of medical image–text supervision because they are linked to published clinical or scientific context.

Nevertheless, synthetic multimodal datasets must be developed carefully. Generated text or images may contain factual errors, overly simplified descriptions, unsupported diagnostic reasoning, or stylistic bias inherited from the teacher model. Therefore, synthetic datasets are best understood as scalable training resources rather than definitive clinical ground truth. Their value depends on source data quality, prompt design, filtering, provenance tracking, and downstream validation using independent and clinician-annotated benchmarks.

These limitations motivated the development of DermaSynth. DermaSynth was designed to address the shortage of large-scale dermatology image–text pairs by generating synthetic multimodal supervision from open-access dermatology datasets, PubMed-derived resources, and instruction-following approaches. By transforming existing clinical and dermoscopic image repositories into image–text pairs and VQA-style data, DermaSynth provides a scalable foundation for training dermatology vision-language models and

supports the broader transition from image-only dermatology AI toward multimodal clinical reasoning.

3.2 DERM12345: Dermoscopic Image Dataset

Skin lesion datasets provide essential information for understanding various skin conditions and developing effective diagnostic tools. They aid the artificial intelligence-based early detection of skin cancer, facilitate treatment planning, and contribute to medical education and research. Published large datasets have only partially covered the subclassification of skin lesions. This limitation highlights the need for more expansive and varied datasets to reduce false predictions and support more detailed failure analysis in skin lesion classification. This study presents a diverse dataset comprising 12,345 dermoscopic images with 40 subclasses of skin lesions, collected in Türkiye, a country at the crossroads of Europe and Asia with a population representing diverse skin types. Each subgroup contains high-resolution images and expert annotations, providing a strong and reliable basis for future research. The detailed analysis of each subgroup provided in this study facilitates targeted research endeavors and enhances the depth of understanding regarding the skin lesions. This dataset distinguishes itself through a diverse structure with its 5 superclasses, 15 main classes, 40 subclasses and 12,345 high-resolution dermoscopic images. The largest open-access (OA) datasets that contain significant dermoscopic images are shown in Table 3.1.

Table 3.1: Specifications of the largest open-access dermoscopic skin lesion datasets in the literature.

Dataset	Study	Type	Dataset Size	Class Size
PH^2	[110]	Dermoscopic	200	3
Derm7pt	[81]	Dermoscopic + Clinic	2,000	5
ISIC 2016	[51, 118]	Dermoscopic	1,279	2
ISIC 2017	[52]	Dermoscopic	2,000	3
ISIC 2018—HAM10000	[53, 82]	Dermoscopic	10,015	7
ISIC 2019	[52, 53, 82, 83]	Dermoscopic	33,569	8
ISIC 2020	[111]	Dermoscopic	44,108	2
BCN20000	[83]	Dermoscopic	18,946	8
Argentina	[113]	Dermoscopic + Clinic	1,616	10
PROVe-AI	[112, 119]	Dermoscopic	603	15
DERM12345	This section	Dermoscopic	12,345	40

These large datasets are crucial for the development of AI-based models for skin lesion classification purposes. In addition, the introduction of subclass annotations in skin image datasets has high potential to enhance studies in the AI field focusing on skin disease, for creating more trustworthy, robust, and intelligent systems.

This study presents a taxonomic tree of skin lesion classes and a dataset comprising 40 skin lesion classes annotated with their corresponding subclasses. Our comprehensive dataset, which contains a total of 12,345 images, was collected from Türkiye, where various skin types are encountered [120, 121, 122]. This dataset is one of the largest collections to date in terms of the total number of images in multiclass datasets.

Table 3.2: The DERM12345 dataset contains 12,345 dermoscopic images categorized into 5 superclasses, 15 main classes, and 40 subclasses.

Superclass	Main class	Subclass (Labels)	# of Images
Melanocytic Benign (10,043 Images)	Banal Compound (1,300 Images)	Acral (acb)	37
		Congenital (ccb)	418
		Miescher (mcb)	123
		Compound Nevus (cb)	722
	Banal Dermal (723 Images)	Blue (bdb)	163
		Dermal Nevus (db)	560
	Banal Junctional (1,792 Images)	Acral (ajb)	426
		Congenital (cjb)	143
		Junctional Nevus (jb)	1,223
	Lentigo (91 Images)	Ink Spot Lentigo (isl)	6
		Lentigo Simplex (ls)	25
		Solar Lentigo (sl)	60
	Dysplastic Compound (539 Images)	Acral (acd)	9
		Congenital (ccd)	32
		Compound Nevus (cd)	498
	Dysplastic Junctional (5,582 Images)	Acral (ajd)	288
		Spitz-Reed (srjd)	19
		Junctional Nevus (jd)	5,275
	Dysplastic Recurrent (16 Images)	Recurrent (rd)	16
Melanocytic Malignant (400 Images)	Melanoma (400 Images)	Acral Lentiginous (alm)	60
		Acral Nodular (anm)	76
		Lentigo Maligna (lm)	86
		Lentigo Maligna Melanoma (lmm)	25
		Melanoma (mel)	153
Nonmelanocytic Benign (1,077 Images)	Keratinocytic (615 Images)	Seborrheic Keratosis (sk)	607
		Lichenoid Keratosis (lk)	8
	Fibro-histiocytic (180 Images)	Dermatofibroma (df)	180
	Vascular (282 Images)	Angiokeratoma (angk)	39
		Hemangioma (ha)	166
		Lymphangioma (la)	56
		Pyogenic Granuloma (pg)	7
		Spider Angioma (sa)	14
Nonmelanocytic Indeterminate (58 Images)	Keratinocytic (58 Images)	Actinic Keratosis (ak)	58
Nonmelanocytic Malignant (767 Images)	Keratinocytic (747 Images)	Basal Cell Carcinoma (bcc)	423
		Bowen's Disease (bd)	37
		Cutaneous Horn (ch)	12
		Mammary Paget Disease (mpd)	9
		Squamous Cell Carcinoma (scc)	266
	Fibro-histiocytic (4 Images)	Dermatofibrosarcoma Protuberans (dfsp)	4
	Vascular (16 Images)	Kaposi Sarcoma (ks)	16
		Total	12,345

3.2.1 Methods

The DERM12345 dataset contains 12,345 high-resolution dermoscopic images from 1,627 patients, collected from 2008 to 2021 in the Department of Dermatology and Venereology at Celal Bayar University (Manisa, Türkiye), Istinye University (Istanbul, Türkiye), and the University of Health Sciences Haydarpasa Numune Research and Training Hospital (Istanbul, Türkiye) using MoleMax 3, MoleMax HD (Derma Medical Systems, Vienna, Austria), FotoFinder® videodermatoscope (FotoFinder Systems, Bad Birnbach, Germany), and 3Gen DermLite DL4 hand-held dermatoscopes (DermLite LLC, California, United States of America) with a connection kit for the iPhone 5 and iPhone 7 (Apple Inc., California, United States of America). Ethical approval of images was based on ethics review board protocols 20.478.486/1023 (Manisa Celal Bayar University, 24/11/2021). The requirement for informed consent was waived because the study was retrospective and the dataset contained anonymized data. The final dataset includes 40 skin lesion classes. Figure 3.1 [123] shows how the dataset was constructed. The collection process included capturing a wide variety of skin lesions using digital dermoscopy devices, high-resolution digital single-lens reflex (DSLR) cameras, or mobile phones with dermatoscope attachments. After the data collection, each image was initially screened in terms of image quality, magnification, and resolution. Then, the images were reviewed and annotated, and quality assurance was performed by two expert dermatologists.

3.2.1.1 Data collection from digital dermoscopy devices

Digital dermoscopy devices are computer-connected systems equipped with dermatoscopes that utilize polarized light, enabling a magnification ranging from 10× to 140×. This feature facilitates the thorough examination of deeper skin structures with a high level of detail. These dermoscopy devices are linked to a computer system equipped with specialized software for the purpose of capturing and storing images. By using this software, dermatologists can longitudinally monitor patients and their skin lesions using detailed information such as lesion location and class. Each lesion is captured using a standardized technique and consistent lighting conditions to minimize any potential variability. Dermoscopy is performed in contact-polarized mode using an interface medium to reduce surface reflection from overlying scale, with ultrasound gel applied when necessary. The dermatologists ensure the most accurate possible alignment of the field of view, with the lesion at the center and a small margin of surrounding healthy skin included to provide contextual information. To generate our dataset, we exported the data using their software tools in suitable formats such as HTML files. The raw data were subsequently retrieved from the raw files as lesion images and their metadata. The cases were selected from this system by two expert dermatologists according to their consensus benign diagnosis, follow-up, and excised lesions with a histopathologic report.

3.2.1.2 Data collection from mobile phones and hand-held dermatoscopes

Mobile phone-integrated dermatoscopes are low-cost and accessible devices that can provide state-of-the-art image quality. In this dataset, Apple iPhone 5 and 7, equipped with a 3Gen DermLite DL4 hand-held dermatoscope kit, were used to collect dermoscopic images. The DermLite DL4 attachment with mobile

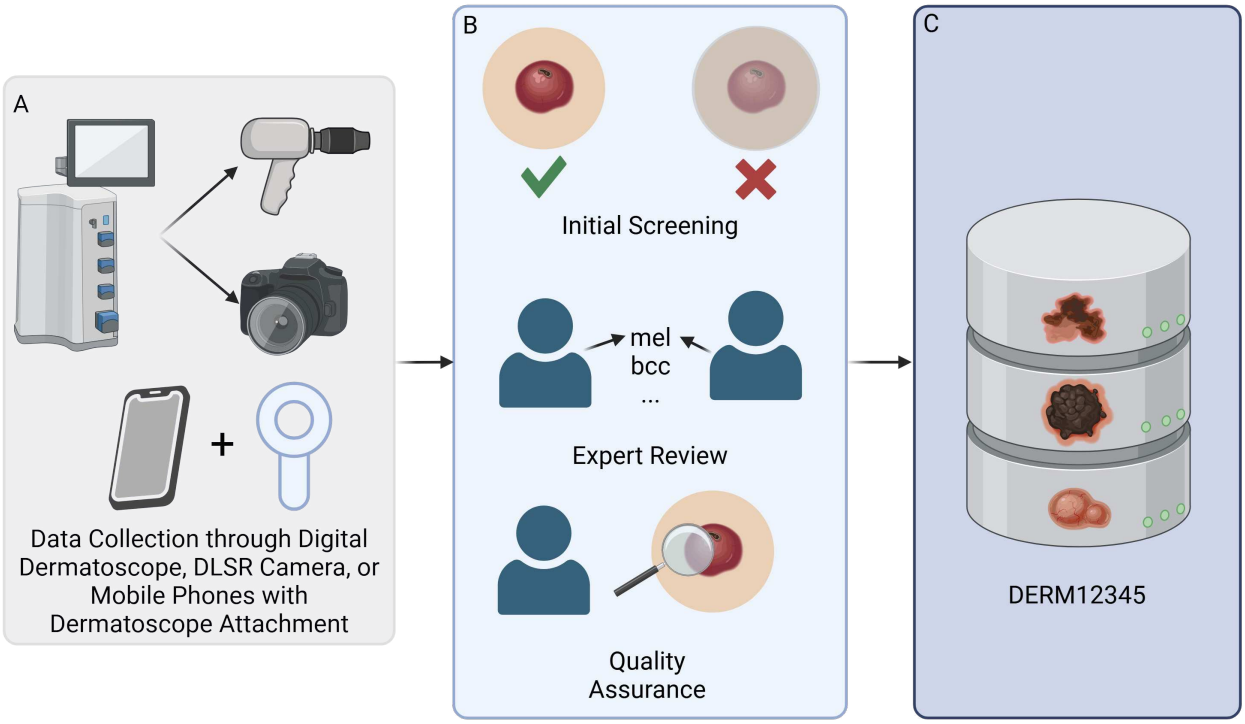


Figure 3.1: An overview of this study. (A) Skin lesion images are collected using a digital dermatoscope, a DSLR camera, or a mobile phone with a dermatoscope attachment. (B) The collected images undergo an initial screening based on inclusion and exclusion criteria, followed by expert annotation (histopathology, follow-up, or consensus) and a final quality assurance step. (C) Schematic representation of the skin lesion dataset.

devices enables the utilization of polarized light dermoscopy, which offers a maximum magnification of $10\times$. The use of high magnification is essential for thoroughly analyzing the intricate components of the skin. The mobile phone was utilized alongside dermatoscope attachments, facilitating the acquisition of high-resolution (such as at $3{,}840 \times 2{,}160$ pixels) skin lesion images directly onto the phone. The mobile images, along with their metadata, including lesion classes, were systematically documented and matched with their records on digital dermoscopy devices. A standardized image acquisition process was used to maintain consistent lighting and framing conditions for all images. The dermatologists conducted the imaging by positioning the lesions as centrally as possible in the field of view while also including a border of healthy skin to offer contextual information. When focusing or centering the mobile phone camera was challenging, the dermatologists took images at different angles. This usually caused the edges of the dermatoscopes to appear as black regions. Images with relatively large black regions were cropped to increase the standardization. Those with smaller black regions were left intact to avoid cropping out the parts of the image with lesion information.

3.2.1.3 Data selection and quality control

The raw dataset included not only dermoscopic images but also non-dermoscopic images and other images such as device test images. To accelerate the data selection procedure, the dermoscopic images were extracted by utilizing a script in Python (version: 3.11.5) that selected the images together with their metadata. The BeautifulSoup library (version: 4.12.2) was utilized for the automated extraction and categorization of images in HTML format. Data that could not be extracted or exported automatically in HTML format was hand-labeled and classified. The metadata was additionally stored manually in comma-separated values files. To remove duplicate images, their file sizes were compared by using a script coded in Python. These scripts found 280 duplicate images. Then, they were manually reviewed. Consequently, all dermoscopic images were retrieved and their corresponding information was stored in a CSV file. After automated extraction of the images, the careful selection of data was facilitated to maintain the integrity and usefulness of the dataset for both clinical reference and computational purposes. The criteria for selection were determined prior to the initiation of data collection and were consistently followed during the selection procedure.

Inclusion Criteria

- **Image Quality:** Images must be clear, in focus, and have sufficient quality. All images were reviewed manually to ensure that key diagnostic features of the skin lesions were discernible.
- **Diagnostic Confirmation:** Only images of lesions with a consensus clinical diagnosis by two expert dermatologists, longitudinal follow-up, or histopathological confirmation (where available) were included.

Exclusion Criteria

- **Poor Image Quality:** Images that were blurred, under- or overexposed, or had large artifacts. All

images were reviewed manually, and the images that could potentially interfere with diagnosis were excluded.

- **Incomplete Data:** Images lacking essential metadata such as the presumed diagnosis were not included in the dataset.
- **Ethical Concerns:** Images that contained identifiable patient information or did not have proper consent were excluded to uphold ethical standards.
- **Unsuitable Content:** Images containing out-of-scope skin diseases or artifacts were excluded, including but not limited to the following categories: nail lesions, mucosal lesions, cockade nevus, leukoplakia, sarcoidosis, psoriasis, verruca, cyst, keloid, sebaceous hyperplasia, wart, genital wart, lipoid proteinosis, subcorneal hemorrhage, pigmented purpuric dermatoses, mucosal melanosis, and some types of adenocarcinoma such as sebaceous carcinoma and eccrine gland lesions.

Selection Process The selection process involved multiple stages:

- **Initial Screening:** A preliminary review of images was conducted by trained engineers to remove images that obviously did not meet the inclusion and exclusion criteria. If more than one class of skin lesion was present in an image, the image was cropped.
- **Expert Review:** Two expert dermatologists then reviewed the remaining images to ensure that they met the clinical standards and diagnostic requirements.
- **Quality Assurance:** A final review was conducted by a panel comprising two dermatologists, Gülsüm Gençoğlan (GG) and Şirin Pekcan Yaşar (SPY), each with more than 20 years of experience in dermoscopy, to ensure that the images were of suitable quality for both clinical reference and computational tasks.

3.2.1.4 Taxonomy Tree

A three-level taxonomy tree was created to guide use of DERM12345, and the dataset was organized according to current dermatologic classification standards [124, 125, 126]. Benign melanocytic lesions with atypical features, diagnostic overlap with melanoma, or an established association with increased melanoma risk were classified into separate subclasses, including dysplastic nevi [127], congenital nevi [128], blue nevi [129], and recurrent nevi, which may mimic melanoma following incomplete removal. The skin lesions were classified according to specific anatomic sites, including acral (palms and soles), facial, and trunk/extremity locations. Combined lesions with multiple diagnoses were excluded if cropping into separate images was not possible. Skin lesions were initially grouped into two classes: melanocytic and nonmelanocytic. These two classes were further categorized into malignant, indeterminate, and benign groups, and five superclasses were created. These five superclasses were then classified into 15 main classes. Finally, the 15 main classes were divided into 40 subclasses. Benign melanocytic lesions were classified into banal with compound, dermal, and junctional main classes, lentigo with ink spot lentigo

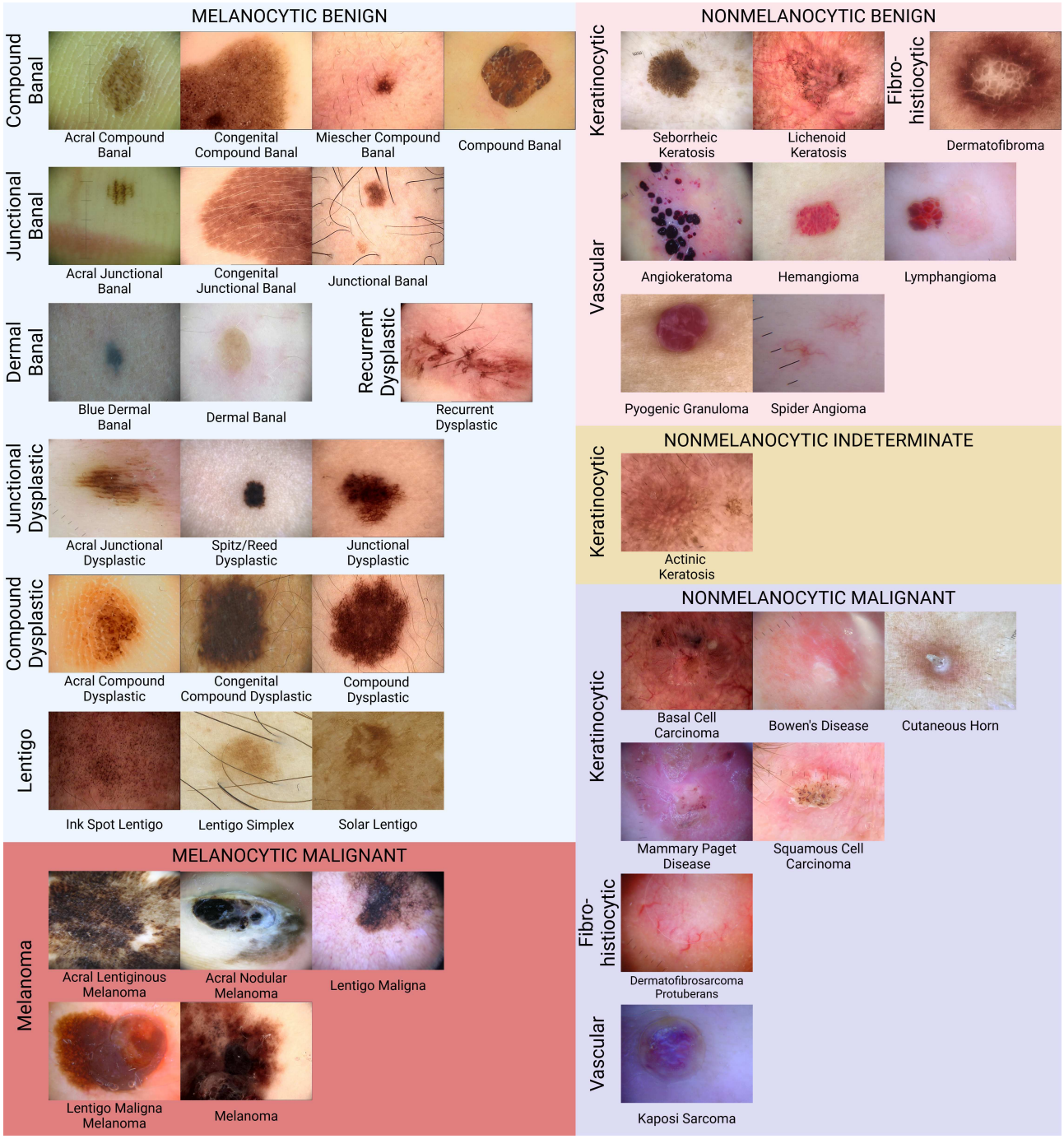


Figure 3.2: A representative image from each subclass.

(isl), lentigo simplex (ls), and solar lentigo (sl) subclasses, and dysplastic with compound, junctional, and recurrent (rd) main classes. Dysplastic dermal nevi were grouped with banal dermal nevus because of their similar patterns. Common nevus subclasses (acral, blue, Miescher, congenital, and Spitz/Reed nevus) under dermal, compound, and junctional groups were then classified into their subclasses if available. Malignant melanocytic lesions were classified into five subclasses: acral lentiginous melanoma (alm), acral nodular melanoma (anm), lentigo maligna (lm), lentigo maligna melanoma (lmm), and melanoma (mel). Benign nonmelanocytic lesions were classified into three main classes: keratinocytic lesions (subclasses: lichenoid keratosis (lk), seborrheic keratosis (sk)), fibro-histiocytic (subclasses: dermatofibroma (df)), and vascular (subclasses: angiokeratoma (angk), hemangioma (ha), lymphangioma (la), pyogenic granuloma (pg), and spider angioma (sa)). Malignant nonmelanocytic lesions were classified into three main classes: keratinocytic lesions (subclasses: basal cell carcinoma (bcc), Bowen's disease (bd), cutaneous horn (ch), mammary Paget disease (mpd), and squamous cell carcinoma (scc)), fibro-histiocytic (subclasses: dermatofibrosarcoma protuberans (dfsp)), and vascular (subclasses: Kaposi sarcoma (ks)). In addition, actinic keratosis (ak) was classified under nonmelanocytic-indeterminate subclasses. Figure 3.2 [130] shows a sample image of each subclass. The taxonomy tree and class sizes are shown in Figure 3.3 [131] and Table 3.2, respectively.

3.2.2 Data Records

The dataset is uploaded to the Harvard Dataverse [132] and is accessible at the following link `https://doi.org/10.7910/DVN/DAXZ7P`. It was released under a Creative Commons Attribution (CC BY) license. The images and metadata are accessible from this repository. This dataset will also be accessible through the ISIC Archive, which provides an advanced user interface and a command-line interface (CLI) for researchers. All images are in Joint Photographic Experts Group (JPEG) format, with their corresponding labels included in the metadata. The metadata file was generated as a CSV file for easy processing. The metadata contains file names and lesion classes with their detailed taxonomic identification. The dataset was split into training (9,860 images, 80%) and test (2,485 images, 20%) sets at the patient level to support standardized benchmarking. Class balancing was considered during the partitioning process to ensure that each set is representative of the overall dataset. For subclasses with low sample sizes, the images were manually labeled as training or test based on unique lesions to prevent data leakage (excluding dfsp and mpd, as there are few images of a single lesion in our dataset for these rare diseases). The subclass names of dysplastic and banal nevi are standardized as acronyms formed from the initials of their subclass and main-class names, such as (a)cral (j)unctional (b)anal nevus: ajb. Other subclass names are standardized as acronyms formed from their initial letters, such as (b)asal (c)ell (c)arcinoma: bcc.

3.2.2.1 Dataset metadata

The dataset metadata contains this information:

- **Unique Image ID:** Displays the unique image identification number of an image.

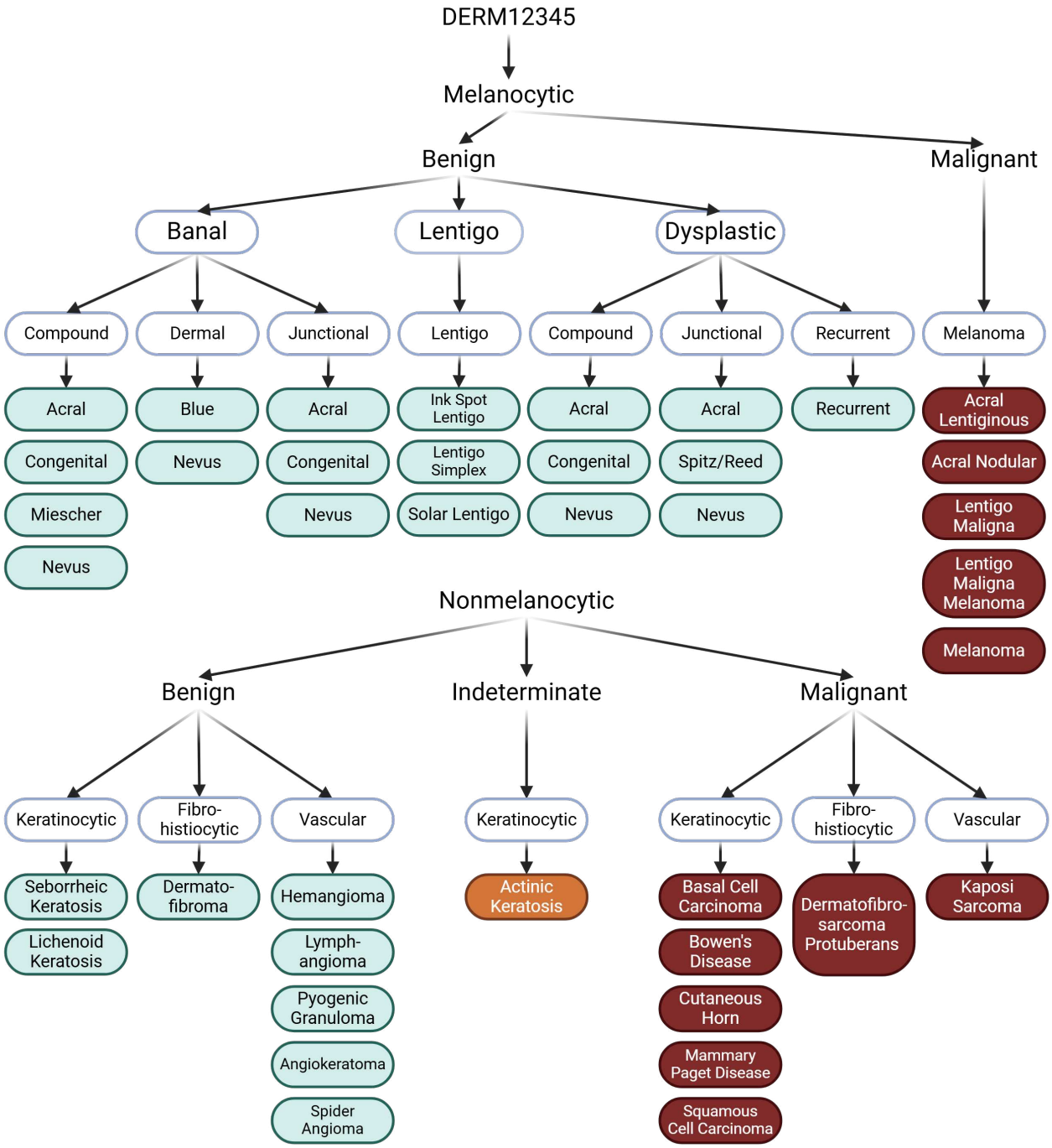


Figure 3.3: An overview of the taxonomy tree. The first level includes melanocytic and nonmelanocytic with benign, indeterminate, and malignant groups. The second level comprises banal, dysplastic, and lentigo for melanocytic benign; melanoma for melanocytic malignant; keratinocytic, fibro-histiocytic, and vascular for nonmelanocytic benign; keratinocytic for nonmelanocytic-indeterminate; and keratinocytic, fibro-histiocytic, and vascular for nonmelanocytic malignant. The third level represents the subclasses associated with their respective main and superclasses. Our taxonomy tree consists of 5 superclasses, 15 main classes, and 40 subclasses.

- **Patient ID:** Displays the unique patient identification number of the image.
- **Image Type:** Identifies the modality of the image.
- **Copyright License:** Specifies the copyright license for the image.
- **Split:** Indicates whether the image belongs to the training or test set.
- **Superclass:** Shows the first-level category of the image.
- **Malignancy:** Indicates the malignancy status of the condition shown in the image. Values are represented as benign, malignant, or indeterminate.
- **Main Class 1:** Displays the first component of the image's second-level category.
- **Main Class 2:** Displays the second component of the image's second-level category.
- **Subclass:** Shows the third-level category of the image.
- **Label:** Shows the label of the image.

3.2.3 Technical Validation

All images in this dataset were of patients presenting to the dermatology department with skin complaints. All lesions were classified and reviewed according to the rules of dermoscopic evaluation by two expert dermatologists in a consensus decision. All malignant skin lesions were biopsy-proven. Most images represented lesions followed longitudinally and were labeled using patient records. Nevi with over 2 years of digital dermoscopic follow-up were annotated with no change except nevoid involution. Dysplastic nevi (6,137 images) constituted a large part of the dataset. In routine clinical practice, dysplastic nevi undergo regular follow-up, whereas banal nevi (3,815 images) are generally not routinely monitored. Early forms of malignant tumors such as mpd and bd were included. Malignant forms of benign lesions with small sample sizes (such as dfsp, with four images, the counterpart of df) were also included in the dataset. The inclusion of subclasses with a small number of samples aims to achieve a more complete taxonomy tree. These samples can be used to develop new algorithms and can be combined with other datasets to achieve larger samples to identify these rare diseases. All benign images without histology or further follow-up were labeled by consensus between GG and SPY. GG (with a Ph.D. in Basic Oncology) and SPY are two expert dermatologists with more than 20 years of experience in dermatology.

3.2.4 Usage Notes

The dataset provided in this study represents the initial comprehensive collection of skin lesion data from Türkiye, serving as a distinctive and important asset for both the medical and machine learning fields. The launch of this dataset represents a notable progression, especially when considering the wide range of skin types it addresses and the comprehensive classification of skin lesions it contains. The dataset is valuable

due to the inclusion of various benign subclasses that closely resemble malignant tumors, distinguishing it from current collections and presenting a novel opportunity for the development of advanced diagnostic algorithms.

The absence of a rich subclassification may result in misclassification. For instance, congenital nevus (which is not categorized as a separate subclass in the published datasets) can often be mistaken for melanoma [128]. It is also important to identify a lesion as a congenital nevus, as congenital melanocytic nevi are associated with an increased risk of melanoma, particularly in larger and more extensive lesions [128]. We also included uncommon lesions that are difficult to collect or annotate such as Spitz/Reed nevus [133]. By using our detailed subclassification, this dataset enables researchers to develop and refine more reliable AI models capable of distinguishing benign from malignant ones, improving diagnostic accuracy and reducing the likelihood of misdiagnosis. The dataset includes 40 subclasses, allowing researchers to focus on more intelligent algorithms such as hierarchical learning [134].

The dataset exhibits potential to serve as a fundamental component in the advancement of effective algorithms for identifying malignant skin lesions. The ongoing obstacle in dermatologic diagnosis lies in the difficulty of differentiating between benign and malignant forms, particularly when they exhibit comparable visual characteristics. The incorporation of benign lesions that resemble malignant counterparts in this collection of data presents a potential for enhancing the precision of machine learning algorithms. Therefore, it is anticipated that this technology will facilitate progress in the early and accurate detection of diseases, potentially mitigating the occurrence of false-positive results and ultimately enhancing patient prognoses.

The inclusion of the dataset originating from Türkiye contributes to the expansion of the worldwide data repository, which has historically been underrepresented in this research domain. The geographical specificity of this information is of great use to academics who seek to create diagnostic models that are both resilient and efficient across various ethnicities and geographic areas. Additionally, it functions as a valuable resource for doctors who aim to comprehend the lesion diversity, which can be impacted by various factors such as regional environmental conditions and genetic predispositions.

Clinicians are encouraged to employ this dataset as a visual resource for the identification of skin lesions, while researchers are encouraged to utilize it for comparative analyses with datasets from other geographic locations. This dataset is of particular utility to machine learning practitioners for the purposes of training and verifying classification algorithms. The comprehensive depiction of complex instances within the dataset offers a rigorous platform for evaluating algorithms intended to distinguish between benign and malignant tumors.

In brief, this dataset serves to enhance the existing body of data pertaining to skin lesions, while also paving the way for the development of sophisticated diagnostic instruments. We anticipate that this tool will be a helpful resource in enhancing the precision and effectiveness of skin cancer detection and diagnosis among various demographic groups.

In this dataset, there are 40 subclasses. Brief descriptions of these subclasses are presented below for AI

researchers:

I. Melanocytic lesions: These lesions originate from melanocytes and are often pigmented in shades of brown, black, tan, or blue, reflecting varying concentrations and depths of melanin in the skin.

1. Benign: Non-cancerous melanocyte-origin moles, generally harmless.

   (a) Dysplastic Nevus: A dysplastic nevus, also known as an atypical nevus or Clark's nevus, is characterized by histologic features but may appear small and banal clinically. Dysplastic nevi demonstrate the following clinical features: usually greater than 5 mm, irregular borders, some with a pigmented and erythematous rim, and variegated pigmentation with a mixture of pink, light, and dark brown colors [135, 136]. This subclass contains compound nevus (cd) with acral (acd) and congenital (ccd) subclasses; junctional nevus (jd) with acral (ajd) and Spitz/Reed (srjd) subclasses; and recurrent nevus (rd).

   (b) Banal Nevus: A banal nevus, commonly known as a common mole, is a typical, non-cancerous skin mole that usually appears as a small, regular, round, or oval spot. It may be brown, tan, or pink, and may be either flat or slightly raised. This subclass contains compound nevus (cb) with acral (acb), congenital (ccb), and Miescher (mcb) subclasses; dermal nevus (db), including the blue (bdb) subclass; and junctional nevus (jb) with acral (ajb) and congenital (cjb) subclasses.

   (c) Lentigo: A pigmented, typically uniform lesion associated with increased epidermal melanin. This subclass includes ink spot lentigo (isl), lentigo simplex (ls), and solar lentigo (sl).

      i. Compound: Common mole, partially flat and partially elevated. Mix of dermal and epidermal melanocytes.

      ii. Dermal: Common mole. Raised bump, flesh-colored or brown mole. Located within the dermis. Generally benign. Unna nevus is included.

      iii. Junctional: Flat, typically dark or brown mole, located at the epidermis-dermis junction.

      iv. Recurrent nevus: Incompletely removed nevus. Mole regrowth at previously excised site. Typically benign but can be confused with melanoma.

         A. Acral nevus: Mole on palms or soles. Usually benign and characterized by its distinctive location on the body where the skin is thicker.

         B. Blue nevus: A deeply pigmented, blue-black mole originating from melanocytes in the deep dermis; usually benign.

         C. Congenital nevus: Mole present at birth that grows proportionally with the child. Sizes vary, potentially large. Increased melanoma risk later in life. Nevus spilus is included.

         D. Ink Spot Lentigo (isl): Small, dark brown to black spots, resembling ink spots.

E. Lentigo Simplex (ls): A small, flat, typically darker spot caused by a benign proliferation of melanocytes. It is also known as simple lentigo.

F. Miescher nevus: Reveals a pseudo network around hair follicles. Commonly benign, typically brown, and found on the face.

G. Solar Lentigo (sl): Larger and typically found on sun-exposed areas, solar lentigo may share clinical and histologic features with seborrheic keratosis. Although not primarily melanocytic, solar lentigo is influenced by melanocytic activity and is intentionally classified within the melanocytic-benign lentigo group.

H. Spitz/Reed nevus: A raised, pink, red, or brown mole that is often mistaken for melanoma. It is typically benign in children but may show a starburst pattern.

2. Malignant: Cancerous growths arising from melanocytes, with a risk of metastasis.

   (a) Melanoma: Aggressive skin cancer. Arises from melanocytes. High risk of metastasis.

      i. Acral nodular melanoma (anm): A nodular melanoma form on palms and soles. Amelanotic melanomas were also included.

      ii. Acral lentiginous melanoma (alm): A subtype of melanoma. Lentiginous form on palms and soles.

      iii. Lentigo maligna (lm): A larger lesion that is a precursor to lentigo maligna melanoma. A slow-growing skin cancer that appears as a flat, blotchy patch. It is an early form of melanoma, typically located on sun-exposed areas of the skin.

      iv. Lentigo maligna melanoma (lmm): A lesion with variable coloration, particularly blue or black. Blue or black areas may indicate that melanoma cells have reached deeper layers of the skin.

      v. Melanoma (mel): Melanoma can arise as a new dark spot on the skin or from an existing mole that changes in color, size, or feel. It can spread quickly to other parts of the body and requires early treatment.

II. Nonmelanocytic lesions: Lesions arising from keratinocytic, fibrohistiocytic, vascular, or other non-melanocytic tissues.

1. Benign: Non-cancerous growths from keratinocytes, fibroblasts, histiocytes, or vascular cells.

   (a) Keratinocytic:

      i. Seborrheic keratosis (sk): A waxy, wart-like, often brown or black growth. Commonly has a rough texture. Stucco keratosis, a variant located on legs, is included.

ii. Lichenoid keratosis (lk): An inflammatory, often regressing lesion that may arise in association with solar lentigo or seborrheic keratosis.

(b) Fibro-histiocytic:

i. Dermatofibroma (df): A firm, benign, typically brownish skin nodule that is common on the legs or arms.

(c) Vascular:

i. Angiokeratoma (angk): Small, dark red to purple, raised lesions with a rough surface. Composed of dilated blood vessels. Typically seen on the lower body or genitals.

ii. Hemangioma (ha): A bright red, raised birthmark composed of clustered blood vessels. It is common in infants and usually fades.

iii. Lymphangioma (la): Rare, soft, often translucent mass. Appears as a swelling on the skin or mucous membranes. Typically seen in infants.

iv. Pyogenic granuloma (pg): Small, reddish, raised lesion. Prone to bleeding and rapid growth. Often appears after skin injury.

v. Spider angioma (sa): Small, red, spiderweb-like lesion with a central red dot and radiating capillaries. Common on the face and upper body.

2. Indeterminate: Not exactly categorized as benign or malignant. This does not imply that these lesions are malignant.

(a) Keratinocytic:

i. Actinic keratosis (ak): Rough, scaly skin patch. Sun-induced, precancerous with a variable risk of progression to malignancy. Common on sun-exposed areas.

3. Malignant: Cancerous growths that can originate from keratinocytes, fibroblasts, histiocytes, and vascular cells.

(a) Keratinocytic:

i. Basal cell carcinoma (bcc): Appears as a shiny, pearly bump or a flat, scar-like lesion. Most common on sun-exposed areas. Slow growing and rarely metastasizes.

ii. Bowen's disease (bd): Early form of scc. Appears as a persistent, red, scaly patch. Can resemble eczema or psoriasis. Often occurs on sun-exposed skin.

iii. Cutaneous horn (ch): Hard, protruding growth resembling an animal horn. Composed of compacted keratin mostly located on the face, ears, and other sun-exposed areas. It

may be associated with benign, premalignant, and malignant lesions. Only lesions with a malignant base were included.

iv. Mammary Paget Disease (mpd): An intraepidermal adenocarcinoma associated with underlying breast carcinoma.

v. Squamous cell carcinoma (scc): A common malignancy that appears as a firm red nodule or a flat lesion with a scaly, crusted surface. It commonly develops in sun-exposed areas and may ulcerate or bleed.

(b) Fibro-histiocytic:

i. Dermatofibrosarcoma protuberans (dfsp): Rare, appears as a deep-seated, firm lump under the skin. Often starts as a small, painless nodule that gradually enlarges. Typically develops on the trunk or limbs.

(c) Vascular:

i. Kaposi sarcoma (ks): Presents as purple, red, or brown blotches or tumors on the skin.

3.2.5 Code availability

Code is available in the DERM12345 GitHub repository.

3.3 DermaSynth: Synthetic Image–Text Supervision for Dermatology

A major barrier to developing vision-language models (VLMs) in dermatology is the lack of large image–text pair datasets. We introduce DermaSynth, a dataset comprising approximately 1.2 million synthetic samples, including 204,809 image–text question–answer pairs derived from 83,622 images, as well as text-only and multi-turn samples. Using Gemini 2.0 Flash, we applied clinically relevant prompts and a self-instruct method to generate diverse and rich synthetic texts. Dataset metadata were used to inform the construction and selection of clinically relevant question prompts. During answer generation, metadata were included as contextual input, but the model was explicitly instructed not to reference or rely on them and to formulate its responses from visible image characteristics. The resulting dataset builds on open-access dermatologic image repositories (DERM12345, BCN20000, PAD-UFES-20, SCIN, and HIBA) distributed under permissive licenses. We anticipate that this dataset will support and accelerate AI research in dermatology. Data and code underlying this work are accessible in the DermaSynth GitHub repository.

3.3.1 Background

Instruction-following methods have emerged as a powerful technique for leveraging large language models as teacher models to generate synthetic data, which can then be used to optimize more specialized models. By prompting advanced models such as GPT-4 and GPT-4 vision, researchers can efficiently create high-quality instruction-following demonstrations that significantly boost model performance for both text [137] and vision cases [138, 139]. However, although several specialized models have begun to emerge for medical tasks [140, 141] and medical vision tasks [142, 143], VLMs in the medical domain remain relatively new and face several constraints, including privacy concerns and the limited availability of public datasets. Multimodal biomedical datasets suitable for instruction-following remain limited compared with the abundance of parallel image–text data available in the general domain [143]. A promising approach to mitigating these data limitations is self-instruct [144, 145], a method that enhances a pretrained language model's instruction-following capabilities by automatically generating new instructions from its own outputs.

Among medical fields that rely on image analysis, dermatology poses a unique data challenge. Unlike radiology or histopathology, where clinical reports or slide annotations often accompany images, dermatologic cases frequently lack formal textual documentation beyond brief physician annotations. This scarcity of paired image–text data complicates the training, fine-tuning, and evaluation of VLMs for dermatologic tasks, such as lesion classification or descriptive question answering. This gap hinders the broader research community, which is eager to explore how advanced language-and-vision models can transform clinical decision-making and improve patient outcomes.

To address these needs, we introduce DermaSynth, a collection of 1.2 million samples focusing on dermatology. Our dataset leverages instruction-following methods, combining state-of-the-art models with

expert-selected prompts to produce rich textual annotations for a variety of dermatology images. By bridging the gap between visual data and clinically relevant text, DermaSynth enables researchers to develop, test, and refine vision-based instruction-following models tailored to dermatology. This advancement enhances both model interpretability and diagnostic reliability in an underrepresented yet critical medical domain.

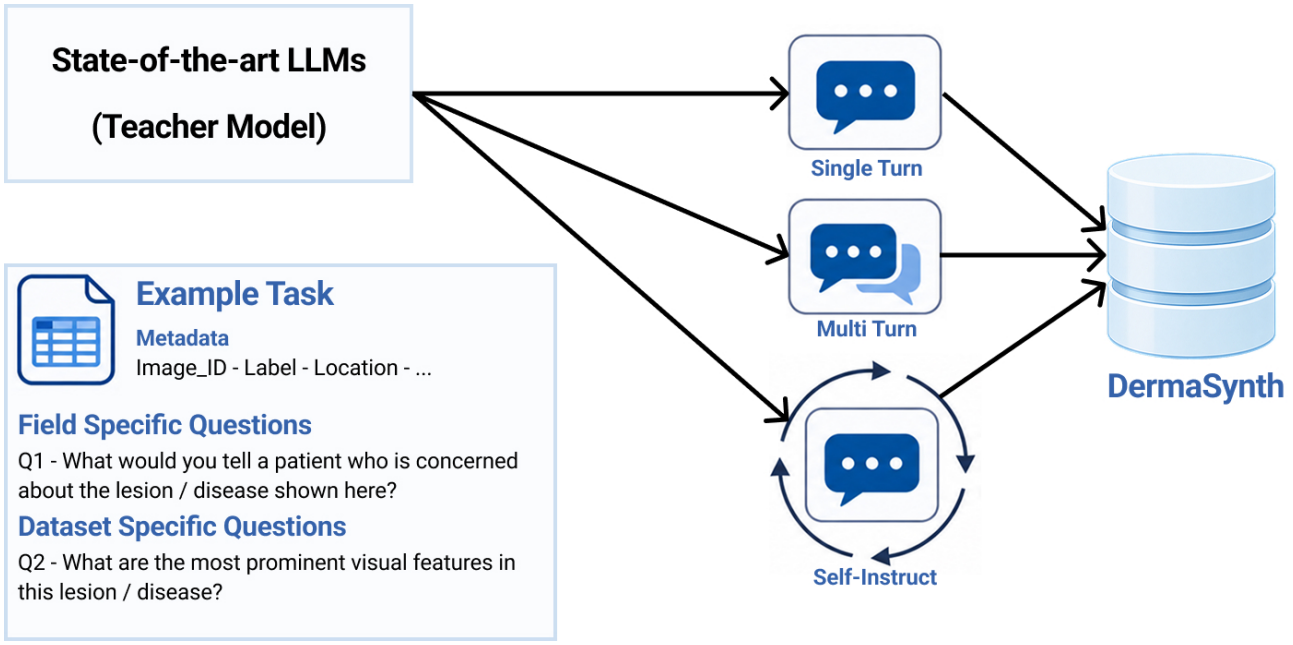


Figure 3.4: Overview of the synthetic data creation process for DermaSynth. A state-of-the-art large language model (Gemini 2.0 Flash) was used to generate synthetic and clinically relevant image–text pairs.

3.3.2 Methods

We curated DermaSynth [146] using multiple open-access dermatologic datasets and the BIOMEDICA dataset [147]. We used datasets with permissive CC BY 4.0 licenses, thereby excluding any images with restrictive terms. We generated general and dataset-specific questions using ChatGPT o1 pro, a state-of-the-art model (accessed on 20 December 2024). We then combined general prompts and metadata-informed question prompts to synthesize realistic question–answer pairs, refining outputs through a post-processing stage. The final dataset statistics, shown in Table 3.3, underscore the breadth and diversity of this resource, which is intended to facilitate a range of dermatology-focused AI applications.

3.3.2.1 Data Sources and Preprocessing: Open-Access PubMed Articles

We utilized the existing BIOMEDICA dataset [147], which contains image-caption pairs and metadata extracted from millions of open-access biomedical articles available on PubMed. We filtered this dataset using the keyword “Skin Lesion” to identify figures relevant to dermatology. We selected samples with permissive licenses allowing research and derivative works (e.g., CC0, CC BY). From this filtered subset, we extracted 56,258 dermatology-related images after excluding non-clinical images such as graphical figures.

Each image was paired with its associated textual context, primarily figure captions and relevant inline mentions identified within the BIOMEDICA metadata, forming the basis for generating image-grounded question–answer pairs. We also developed a PubMed Central scraping tool to collect up-to-date data for any field from PubMed Central (Appendix C.1).

3.3.2.2 Data Collection and Preprocessing: Open-Access Datasets for Dermatology

We identified open-access datasets relevant to our target domain, dermatology, prioritizing collections with clear licensing terms and comprehensive documentation. For dermatology, this primarily included image repositories covering various skin conditions and demographic groups to enhance model generalizability. When necessary, images were resized or converted to consistent resolutions, and mislabeled or corrupted files were removed. Pertinent metadata (such as lesion type, imaging modality, or diagnostic tags) was extracted and centralized for easy reference. This coordinated preprocessing produced a unified, high-quality dataset suitable for both human annotation and automated synthetic data generation within the Scalable Clinical Assistants and LEarning for MEDicine (SCALEMED) framework (Section 4.5).

In this study, we employed DERM12345 [132], BCN20000 [83], HIBA [113], PAD-UFES-20 [85], and SCIN [87] datasets, all of which have licenses that permit the intended use, including commercial use where applicable. These repositories collectively comprised 13,496 clinical images (30%) and 32,561 dermoscopic images (70%). Datasets or images with more restrictive licenses (e.g., CC BY-NC) were excluded to ensure that our final release aligns with open-access licenses. Some datasets were accessed via the isic-cli library, while others required downloads from institutional or project-specific websites.

3.3.2.3 Synthetic Data Generation

We employed multiple pipelines to generate a large and diverse synthetic dataset targeting various aspects of clinical and dermatologic knowledge and reasoning. The generation process primarily utilized Google's Gemini models (Gemini 2.0 Flash, from November 2024 to January 2025) via API calls. Gemini 2.0 Flash can generate high-quality synthetic data even from basic seed tasks.

VQA Generation from Image Datasets (Open Access & BIOMEDICA)

1. **Open-Access Dermatology Datasets:** For each image in open-access datasets (e.g., DERM12345), we first converted its structured metadata into concise natural language strings (e.g., "Metadata - Label: Melanoma, Age: 45, Sex: Female, Location: Forearm"). Predefined prompt templates included both "general" (e.g., "Describe the visual features in the image") and "dataset-specific" questions (e.g., "What is the diagnosis for this lesion located on the localization?") (Appendix A.3.1). For each image, one general and one dataset-specific question were randomly selected. These questions and metadata strings were sent to Gemini. The metadata were used to inform the construction, selection, and, where applicable, parameterization of question prompts. Although metadata were also supplied in a separate contextual block, the answer-generation instruction explicitly required Gemini not to reference or independently rely on that metadata context and to base its response on visible image

characteristics. We requested two synthetic QA pairs per image, formatted as JSON for easy parsing. The final dataset contained 92,114 synthetic QA pairs.

2. **BIOMEDICA Dataset:** For each of the 56,258 selected BIOMEDICA samples after the filtering step, we combined the image caption and associated contextual mentions into a single text block, which, together with the corresponding image, was provided to Gemini. Gemini rejected 14,429 samples due to its safety guardrails, leaving 41,829 samples. These remaining samples were then split into a training set of 37,565 samples (DermaSynth) and an evaluation set of 4,264 samples (Eval2 of the DermatoLlama study). We used two prompt styles: a training-style prompt requesting three descriptive question–answer pairs about visible characteristics based on the provided context, and an evaluation-style prompt, used to generate Eval2 in Section 4.5, requesting a multiple-choice question based on the image and context (Appendix A.3.2).

Instruction-Based Generation We applied self-instruction techniques to generate diverse text-based and conversational data, beginning with small sets of human-curated seed tasks. Seed tasks were initially generated using ChatGPT o1 pro and were subsequently reviewed and edited by expert clinicians to ensure relevance and accuracy, including varied perspectives such as "explain like I am 5." Our framework supports flexible integration of different seed tasks for diverse use cases, and we have designed it to be as comprehensive as possible.

1. Instruction-Following QA Synthesis: Starting with 190 seed tasks (Appendix E.2.1, 18 classification, 172 non-classification tasks, validated by three clinicians), we iteratively prompted Gemini (Appendix A.3.3). In each iteration, two human-generated and two synthetic samples served as context. Gemini generated new instructions, labeling them as classification or non-classification tasks. For classification tasks, Gemini created class labels and inputs; for non-classification, it generated input-output pairs. Outputs were parsed using regular expressions, and near-duplicates were filtered using semantic similarity. This process generated 50,027 instruction-following samples.

2. Knowledge-Base (KB) QA Synthesis: We developed a knowledge base covering approximately 2,700 skin-condition entries [148]. The knowledge base was sourced from Wikipedia's list of skin conditions via Wikipedia-API (Appendix A.4). We defined 40 seed-task templates (e.g., "List key triggers for <DISEASE_NAME>," "Describe the typical presentation of <DISEASE_NAME>") (Appendix E.2.2). For each disease, relevant templates and content sections were populated using disease information extracted from the KB and used to prompt Gemini. The supplied knowledge-base context served as the primary seed information; however, the generation prompt explicitly permitted Gemini to supplement incomplete context using its general dermatologic knowledge (Appendix A.3.4). Accordingly, these outputs should be interpreted as KB-seeded synthetic QA data rather than strictly source-grounded responses. This approach generated 72,400 KB-seeded synthetic QA pairs.

3. Structured Case Synthesis (Anamnesis): We used 137 clinician-validated seed dermatologic cases (Appendix E.2.3) based on the 137 most common dermatologic diseases (Appendix E.2.4). The cases

Data Source / Task	# Samples	Notes
Skin image–text pairs	112,695	37,565 skin images with three QA pairs
Open-Access Repositories	92,114	46,057 images with two QA pairs
DERM12345 (Train)	9,860	Dermoscopic
DERM12345 (Test)	2,485	Dermoscopic
BCN20000	18,946	Dermoscopic
PAD-UFES-20	2,292	Clinical
SCIN	10,858	Clinical
HIBA	1,616	346 clinical, 1,270 dermoscopic
Subtotal	46,057	13,496 clinical, 32,561 dermoscopic
Self-Instruct	50,027	190 seed tasks
Knowledge-based tasks	72,400	40 seed tasks, based on Wikipedia [148]
Multi-turn anamnesis	892,814	9,963 multi-turn cases and 137 seed tasks
Total	1,220,050	

Table 3.3: Characteristics of the DermaSynth dataset.

were designed using a standardized clinical questionnaire comprising 17 sections and 104 questions (Appendix E.2.5) to reflect common conditions and doctor–patient interactions. By providing Gemini with example cases, we requested the generation of new case studies adhering to a strict JSON schema (with sections such as base_information, questions_and_answers, and final_diagnosis) (Appendix A.3.5). Robust parsing handled variations in model outputs. This method generated 892,814 samples for 9,963 multi-turn case interactions.

Final Datasets Outputs from these pipelines were aggregated to form the DermaSynth datasets, comprising approximately 1.2 million synthetic samples. All experiments were conducted to remain within a modest budget. GPT-4o was evaluated through the OpenAI API, yielding an inference cost of $100. Synthetic data creation via the Gemini API involved roughly 400,000 batched prompt-completion cycles. At the current rate of $0.55 per million output tokens, total generation cost was about $200. Thus, the entire pipeline, including data generation, fine-tuning, and evaluation, was executed for well under $500, demonstrating that high-quality domain-specific models can be developed cost-effectively and made accessible to most academic and clinical research groups. Table 3.3 summarizes the sample counts and composition of each DermaSynth data source.

After generation, we post-processed the data to remove malformed, clinically implausible, or unsupported outputs. This was intended to improve consistency and clinical relevance. In addition, some images from the SCIN dataset [87] have additional symptom metadata (e.g., itching or scaling). To explore how symptom context affects model performance, we generated two variants of image–text pairs from these samples: one that explicitly includes symptom details and one that omits them. An example figure with questions and answers is shown in Figure 3.5.

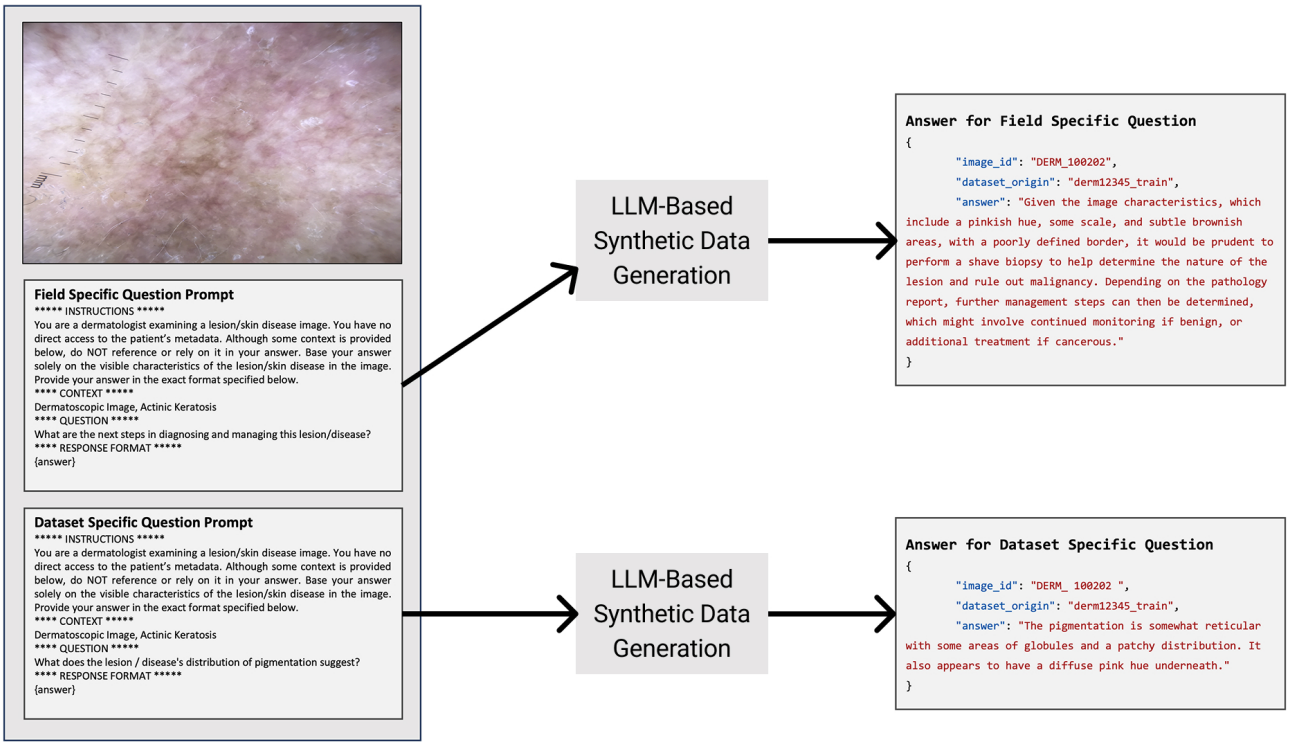


Figure 3.5: A figure from the DERM12345 dataset with a field-specific question, a dataset-specific question, and the corresponding Gemini 2.0 Flash answers.

3.3.3 Data Records

The dataset is uploaded to Hugging Face and is accessible at the following link https://huggingface.co/DermaVLM. It was released under a Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) license. All resources can also be accessed via https://github.com/abdurrahimyilmaz/DermaSynth and https://github.com/DermaVLM. The model is accessible at https://huggingface.co/abdurrahimyilmaz/DermatoLlama-0.1.

We release the following resources:

- **DermatoLlama 0.1:** A fine-tuned Llama-3.2-11B-Vision-Instruct model trained on 5,000 synthetic image–text pairs generated from images in the DERM12345 training split [115].
- **DermaSynth:** Approximately 1.2 million quality-checked synthetic samples, including 204,809 image–text pairs, released in Hugging Face Datasets and JSON formats.
- **Training Scripts:** Python utilities and a `Google Colab` notebook showcasing an end-to-end fine-tuning workflow with the `unsloth` library.

The dataset does not contain any images; they must be downloaded from the original sources.

3.3.4 Technical Validation

Question Prompt Generation. To ensure diverse, realistic prompts, we used a guiding instruction (Appendix A.1.1) that emphasized generating queries centered on the image rather than strictly referencing labels. This question generation prompt was used to generate the question set using the ChatGPT o1 pro model. The 20 most common root verb–noun pairs in the question set are shown in Figure 3.6, illustrating the diversity of the questions. Prompts spanned a wide range of question types, such as "What does this lesion look like?" and "Could this be *X*?" and were manually screened and selected for clarity and medical relevance. This manual selection of prompt examples provides a dermatologist's investigative approach, allowing the synthetic Q&A to cover nuanced scenarios without over-reliance on labels or assumptions.

Baseline Model For a proof-of-concept, we sampled 5,000 image–text pairs derived from the DERM12345 images [115]. We then fine-tuned a preliminary Llama-3.2-11B-Vision-Instruct model (FP16), DermatoLlama, using Python (v3.11.11) and the "unsloth" library (v2025.1.7) on a single NVIDIA A100 (40 GB) instance with 96 GB system RAM. This model is accessible from Hugging Face at `https://huggingface.co/abdurrahimyilmaz/DermatoLlama-0.1`. An example output of the original Llama model and the DermatoLlama model for a simple input is shown in Figure 3.7.

3.3.5 Usage Notes

Interest in integrating LLMs into clinical workflows is growing because these models can provide more interpretable, language-based outputs, particularly in dermatology [149, 150]. Here, models must not only achieve high success but also provide clinically interpretable, evidence-grounded outputs to build trust among clinicians and patients. Factors such as bias monitoring, fairness checks, and interpretable decision paths have become critical for ensuring that AI-driven systems comply with ethical and clinical standards. As researchers increasingly explore the use of VLMs in decision-support systems, placing emphasis on developing responsible AI practices will accelerate clinician engagement.

This study has several limitations. While synthetic data can help train and fine-tune LLMs, recent developments in dermatology may be absent from the teacher models' training data or fall beyond their knowledge cutoffs, resulting in incomplete or outdated synthetic content. Many state-of-the-art (SOTA) LLMs are not trained primarily on medical corpora and may therefore lack the latest developments in dermatology (e.g., Gemini 2.0 Flash had a knowledge cutoff of August 2024 as of February 2025). Models trained on such synthetic data may consequently struggle to capture and explain subtle morphological nuances or rare lesion types and cases. Expert review of synthetic content remains essential but is resource-intensive, underscoring the continued need for expert-driven curation and validation. Moreover, relying exclusively on synthetic datasets risks overlooking real-world variability in lesion presentation, lighting conditions, and patient demographics.

Beyond these limitations, several avenues can strengthen both the framework and its practical relevance in

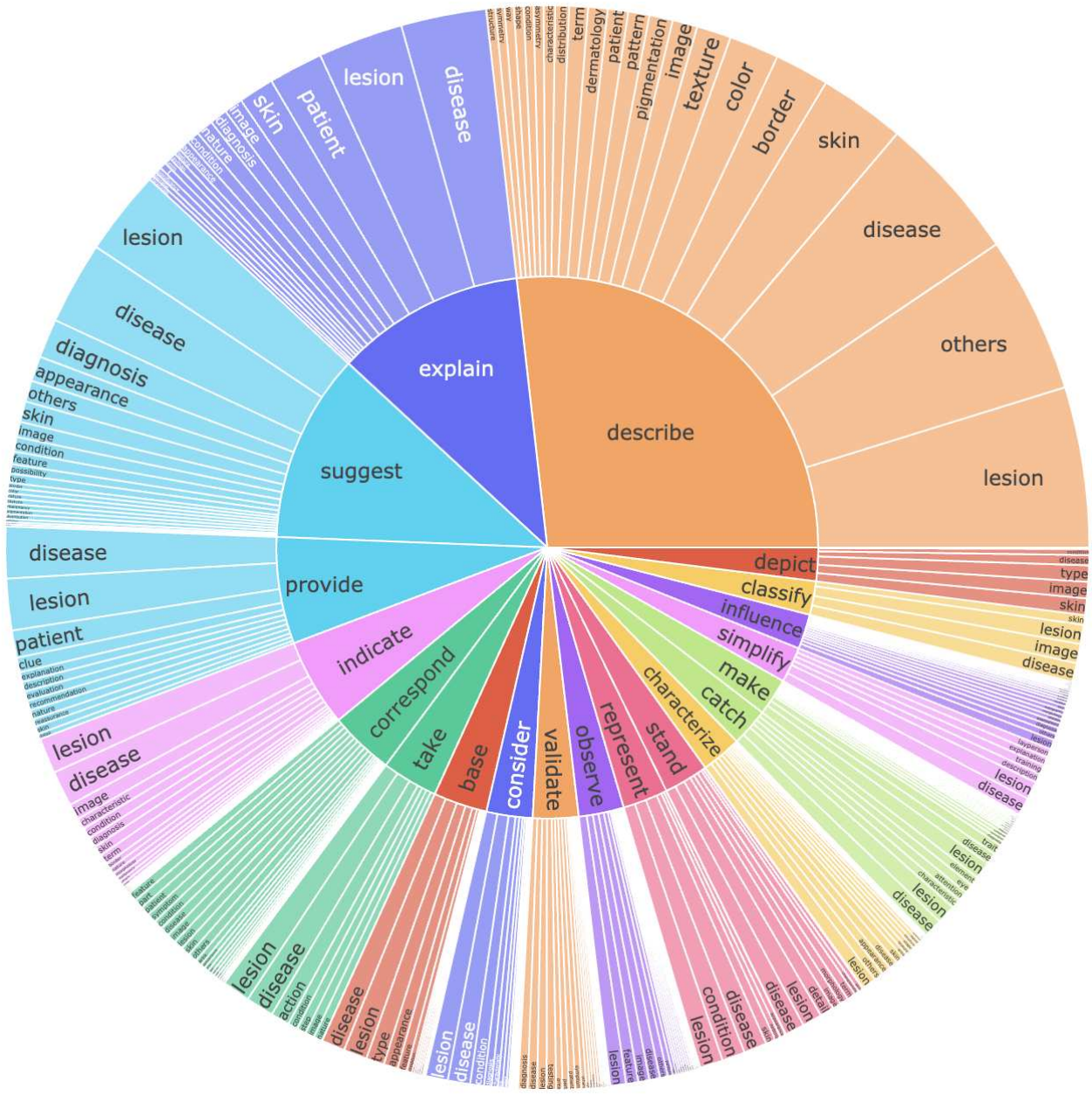


Figure 3.6: The 20 most common root verb–noun pairs in the question set.

the future. One direction is to design multi-agent systems, with each agent specializing in a subfield of dermatology, such as rare diseases or pigmentary disorders, to improve accuracy through collaborative validation. Another major consideration involves extended, multi-turn dialogs that integrate clinician feedback (e.g., via reinforcement learning from human feedback (RLHF) or retrieval-augmented generation (RAG)) to mimic diagnostic inquiries more accurately. In addition, training with larger and more diverse datasets can further refine performance on complex lesion types and underrepresented demographic groups, enabling models to adapt to the full spectrum of clinical scenarios.

Methods such as student–teacher model paradigms were employed in this work to facilitate more resource-

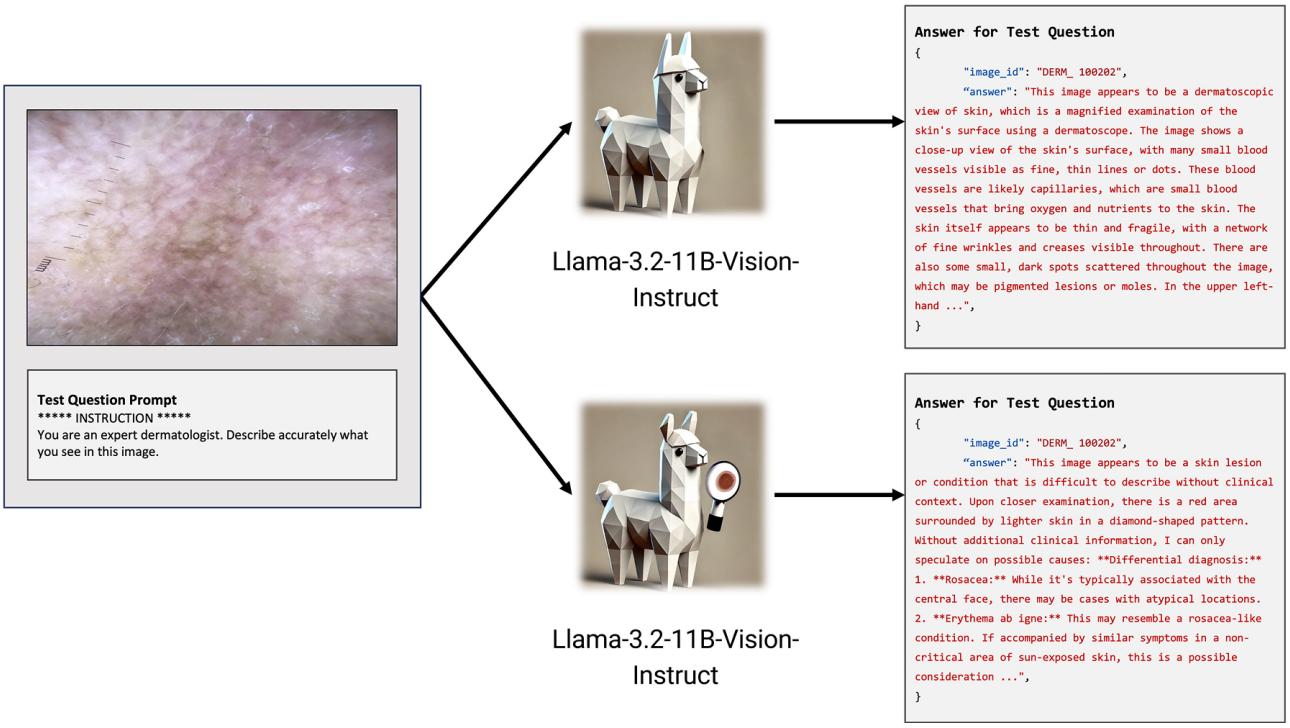


Figure 3.7: Input images, a simple prompt, and responses generated by the original Llama model and DermatoLlama.

efficient deployment. Broader quantitative and qualitative evaluations for these deployments are needed to determine how effectively these innovations align with evolving medical standards. Collaborations with dermatology experts will be critical: such partnerships not only offer in-depth insight into clinical outcomes, but also help to iteratively refine both synthetic datasets and the models tasked with interpreting them.

By transferring expertise from large instruction-tuned vision-language models to smaller, specialized networks, we released the DermaSynth dataset and the scalable DermatoLlama 0.1 model, both of which can be used with low-cost, open-source systems, enabling researchers to accelerate large language model research in dermatology. A further development of DermatoLlama 0.1, termed DermatoLlama 1.0 and trained at scale on the DermaSynth dataset, is described in Section 4.5.

License and Non-Commercial Use DermaSynth and DermatoLlama 0.1 are released strictly for academic and non-commercial use (CC BY-NC 4.0).

3.3.6 Data availability

The BIOMEDICA dataset is publicly available on BIOMEDICA (Hugging Face). The open-access dermatology datasets used in this study are available from their respective repositories: DERM12345 (Harvard Dataverse), BCN20000 (Figshare), SCIN (GitHub), HIBA (ISIC Archive), and PAD-UFES-20 (Mendeley Data).

3.3.7 Code availability

Code is available in the DermaSynth GitHub repository.

Chapter 4

AI in Dermatology

This chapter begins with an introduction to dermatology artificial intelligence and the scope of the presented studies (Section 4.1). It then presents baseline convolutional neural network results for DERM12345 across 40 skin lesion subclasses (Section 4.2). Next, an artifact-aware transformer-based object detection approach is developed for fungal detection in KOH microscopy images of dermatophytosis (Section 4.3). A privacy-preserving local language model framework is subsequently investigated for longitudinal data retrieval in patients with pemphigus (Section 4.4). Finally, DermatoLlama is presented as a resource-efficient dermatology vision-language model developed using a synthetic multimodal data generation framework (Section 4.5).

Parts of this chapter are reprinted and adapted from the following publications:

- (Section 4.2) **A. Yilmaz**, S. P. Yasar, G. Gencoglan, B. Temelkuran. "DERM12345: A Large, Multisource Dermatoscopic Skin Lesion Dataset with 40 Subclasses," Scientific Data, https://doi.org/10.1038/s41597-024-04104-3
- (Section 4.3) R. Gursoy*, **A. Yilmaz***, B. Kizilyaprak, E. Caglar, B. Temelkuran, H. Uvet, A. E. K. Aksu, G. Gencoglan. "Artifact-Aware Fungal Detection in Dermatophytosis: A Transformer-Based Approach for KOH Microscopy," Bioengineering, https://doi.org/10.3390/bioengineering13050591
- (Section 4.4) **A. Yilmaz**, A. E. K. Aksu, D. Yamen, V. A. Erdemir, M. S. Gurel, G. Gencoglan, J. M. Posma, B. Temelkuran. "Privacy-Preserving Local Language Models for Longitudinal Data Retrieval in Chronic Dermatologic Disease: Implementation in Pemphigus Patients," arXiv, https://doi.org/10.48550/arXiv.2605.25020
- (Section 4.5) **A. Yilmaz**, F. Yuceyalcin, R. Varol, E. Gokyayla, O. Erdem, D. Choi, A. A. Demircali, G. Gencoglan, J. M. Posma, B. Temelkuran. "Resource-efficient medical vision language model for dermatology via a synthetic data generation framework," medRxiv, https://doi.org/10.1101/2025.05.17.25327785

I supervised the study *Fine-Tuning a Small Vision Language Model Using Synthetic Data for Explaining Bacterial Skin Disease Images*, published in *Diagnostics* in 2026 [151]. The study was conducted by Shiwan Zhang and presents a compact vision-language model for dermatology that complements the larger DermatoLlama framework discussed in this thesis.

4.1 Introduction

The artificial intelligence studies presented in this chapter were designed to reflect a methodological progression from conventional image-based classification toward more explainable, multimodal, and clinically adaptable dermatology AI systems. Dermatology contains diverse diagnostic workflows, ranging from dermoscopic skin lesion assessment to microscopy-based infection detection and long-term follow-up of chronic inflammatory or autoimmune diseases. For this reason, a single AI method is unlikely to address the full spectrum of dermatologic practice. Instead, this chapter uses representative clinical use cases to examine different methodological directions in dermatology AI, including convolutional neural networks, object detection, text-based language models, and multimodal vision-language models.

The chapter begins with convolutional neural network-based classification using the DERM12345 dataset. DERM12345 was described in detail in the previous data-focused chapter (Section 3.2); here, it is used as a representative dermoscopic image dataset to evaluate classical deep learning approaches for skin lesion classification. These baseline CNN models provide an important starting point because they represent the dominant early paradigm in dermatology AI: image-level classification using black-box visual models. Although such models can learn diagnostically relevant visual patterns, their outputs are usually limited to class probabilities and do not provide detailed clinical explanations or reasoning.

The second methodological step moves from image-level classification to localization-based explainability. For this purpose, KOH microscopy for dermatophytosis was selected as a representative use case. Unlike dermoscopic skin lesion classification, where a single lesion image is typically classified as a whole, KOH microscopy requires the identification of small fungal hyphae within complex microscopic fields containing keratin debris, fibers, air bubbles, and other artifacts. Therefore, object detection is more appropriate than simple image-level classification for this task. By localizing fungal structures with bounding boxes and explicitly modeling artifacts as a separate class, this approach provides more interpretable visual evidence and better aligns with how clinicians examine microscopy slides.

In parallel with image-based and localization-based approaches, this chapter also explores the integration of language models into dermatology. Chronic dermatologic diseases often require longitudinal follow-up over many years, during which patients accumulate large volumes of clinical notes, treatment histories, laboratory results, disease activity descriptions, and adverse event documentation. Pemphigus was selected as a representative chronic dermatologic disease to evaluate whether an open-source, locally deployable text-based language model could retrieve structured information and generate clinically meaningful summaries from long-term follow-up records. This study extends the thesis beyond image analysis and introduces privacy-preserving language-based clinical reasoning as an important component of dermatology AI.

Finally, the chapter progresses toward multimodal language models, which represent a more advanced and flexible direction for dermatology AI. Dermatology is inherently multimodal: clinical reasoning often requires integrating visual findings with patient history, anatomical site, symptoms, clinical text, and longitudinal information. In this context, multimodal language models can move beyond classification by

generating descriptive reports, answering clinical questions, and providing multiple perspectives on the same image or case. This chapter therefore includes the development of dermatology-focused multimodal models, culminating in DermatoLlama, a dermatology-specific vision-language model trained using synthetic multimodal supervision. This final step reflects the central aim of the thesis: to move from black-box image classifiers toward more explainable, resource-efficient, privacy-preserving, and clinically useful dermatology AI systems.

4.2 Baseline CNN Results on DERM12345

The first methodological step in this chapter examines conventional convolutional neural network-based classification on dermoscopic images. This analysis establishes the baseline against which the later, more explainable and multimodal approaches in the chapter can be conceptually compared. CNN-based classifiers represent the classical deep learning paradigm in dermatology AI: an image is provided as input, and the model outputs a predicted diagnostic class or class probability. Although this approach has been widely used in skin lesion analysis, it provides limited insight into the visual evidence underlying the prediction and is therefore an appropriate starting point for the methodological progression explored in this chapter.

DERM12345, introduced in detail in Section 3.2, was used here as a representative fine-grained dermoscopic image dataset. In contrast to binary malignancy classification or small-category classification tasks, DERM12345 contains a hierarchical diagnostic structure with multiple levels of granularity. Evaluating CNNs on this dataset therefore provides not only a performance baseline, but also an initial assessment of the difficulty of conventional image-level classification when applied to a more detailed dermatologic taxonomy.

Three CNN architectures were trained and evaluated: ResNet-50 [152], Xception [153], and Inception-ResNet-v2 [154]. These architectures were selected as representative ImageNet-pretrained models commonly used in medical image classification. The purpose of this experiment was not to optimize state-of-the-art performance, but to establish a reproducible reference point for classical CNN-based dermoscopic classification using DERM12345.

All models were implemented using Python version 3.9.19 and TensorFlow version 2.10.1. Images were resized to either 224×224 pixels or 299×299 pixels, depending on the input requirements of the model architecture, and stored as NumPy arrays. To reduce overfitting and improve robustness, standard image augmentation strategies were applied, including rotation with a range of 45 degrees, zoom with a range of 0.2, width shift with a range of 0.2, and horizontal and vertical flipping. The models were initialized with ImageNet-pretrained weights [155], trained for 50 epochs with pretrained backbones frozen, then fine-tuned for 200 epochs using only the training partition; the held-out test partition was used only for final evaluation provided in the DERM12345 metadata.

The weighted accuracies on the test set were 0.50 for ResNet-50, 0.59 for Xception, and 0.58 for Inception-ResNet-v2. These results provide an important baseline for future work, but they also highlight the difficulty of fine-grained dermatologic classification. Unlike many existing dermoscopic datasets that focus on binary malignancy classification or a small number of diagnostic categories, DERM12345 contains 40 subclasses organized within a hierarchical taxonomy. Therefore, the relatively modest weighted accuracies should be interpreted in the context of a more challenging and clinically granular classification task.

These baseline results also illustrate one of the key limitations of conventional CNN-based dermatology

AI. Although CNNs can learn useful visual representations from dermoscopic images, their outputs are generally limited to predicted class labels or probabilities. They do not naturally explain which dermatologic structures support a prediction or why a lesion belongs to a particular diagnostic category. For this reason, the baseline CNN analysis serves as the starting point of this chapter: it establishes the classical black-box image classification paradigm before the thesis moves toward more explainable object detection, language-based longitudinal reasoning, and multimodal vision-language modeling approaches.

Together, these results establish DERM12345 as a challenging fine-grained dermoscopic classification dataset and provide a reproducible baseline for conventional black-box image classifiers. More advanced evaluation using modern foundation models on the same hierarchical dataset is presented in Section 5.3.

4.3 Artifact-Aware Object Detection for KOH Microscopy

Dermatophytosis is commonly assessed using KOH microscopy, yet accurate recognition of fungal hyphae is hindered by preparation-related artifacts, heterogeneous keratin clearance, and notable inter-observer variability. This study presents a transformer-based object detection framework using the RT-DETR architecture for precise, query-driven localization of fungal structures in high-resolution KOH images. A dataset of 2,540 routinely acquired microscopy images was manually annotated using a multiclass strategy that explicitly distinguishes fungal elements from confounding artifacts, enabling the model to actively suppress false detections arising from visually similar mimics. To assess architectural trade-offs, RT-DETR was benchmarked against two CNN-based detectors (YOLOv11 and Faster R-CNN) under identical training and inference conditions. Five-fold stratified cross-validation was performed, and each fold-level model was evaluated on the same independent held-out test set (254 images). Across the five evaluations, RT-DETR achieved a mean AP@0.50 of $89.73 \pm 1.48\%$, a mean recall of 0.831 ± 0.011, and a mean precision of 0.921 ± 0.014. At the image level, the model achieved a mean sensitivity of 0.989 ± 0.022 on the independent test set, with a mean of 0.2 ± 0.4 missed positive cases across the five evaluations. These results demonstrate the technical feasibility of a transformer-based artificial intelligence system as a decision-support aid for fungal region detection in KOH microscopy, pending prospective multicenter validation to establish clinical generalizability.

4.3.1 Introduction

Dermatophytosis is a common fungal infection of the nails and body. Early and accurate diagnosis is of critical clinical importance, as it directly influences treatment success [156]. Conventional diagnostic methods primarily include direct microscopic examination with KOH and fungal culture, both of which remain widely used in clinical practice. However, these approaches are limited by prolonged analysis times, relatively low sensitivity, and significant inter-observer variability, particularly in microscopic evaluation [157, 158].

The KOH preparation, one of the most frequently employed methods for microscopic confirmation, dissolves keratin and enhances the visibility of fungal hyphae and spores. However, direct KOH microscopy often presents diagnostic challenges due to residual keratin, air bubbles, debris, and artifactual structures that may resemble fungal elements, contributing to inter-observer variability and reduced sensitivity [159, 160]. Building a robust detection system requires addressing this inherent variability in samples prepared with KOH. In particular, artifacts that visually mimic fungal structures represent a major source of diagnostic ambiguity, not only for human observers but also for automated systems trained without explicit artifact-aware strategies. These artifacts also complicate automated analysis, particularly for object detection models that rely on accurate spatial localization [161, 162].

In recent years, AI and deep learning methodologies have increasingly been adopted to address diagnostic challenges in medical image analysis, particularly in settings where subtle, low-contrast structures must be identified within complex visual backgrounds [163, 164]. Through their ability to learn complex representations from large datasets, these models can detect subtle and low-contrast patterns that are not readily discernible under conventional visual inspection [165, 166]. Deep learning has achieved expert-level performance across multiple medical imaging domains, including radiology and digital pathology [167, 168]. In dermatology specifically, state-of-the-art models have demonstrated diagnostic accuracy comparable to, and in some cases exceeding, that of experienced dermatologists in the recognition of skin lesions [49, 169, 170].

To place fungal hyphae detection in the broader context of microscopic image analysis, recent deep learning applications in related domains are briefly reviewed below. Microscopic image analysis has likewise become an important focus within medical AI, where deep learning models have been successfully applied to detect small, morphologically variable, and low-contrast biological structures [171, 172]. Strong performance has also been demonstrated in blood-smear microscopy, including malaria parasite detection using convolutional neural network-based or hybrid architectures [173]. Convolutional and transformer-based models have further enabled reliable identification of parasitic forms and accurate detection of abnormal epithelial cells in cervical cytology [162, 174, 175]. In addition, hybrid deep feature fusion strategies, such as those implemented in the DeepCervix framework, have advanced automated cervical cell analysis [176]. Comparable progress in histopathology, including precise nuclei segmentation and classification, highlights the capacity of modern segmentation and localization frameworks to localize fine, irregular, and artifact-obscured targets within complex visual fields [161, 177]. These shared imaging characteristics (small object size, low contrast, complex textures, and the presence of distracting artifacts) closely mirror the challenges observed in KOH microscopy, making fungal hyphae a suitable and clinically relevant target for object detection approaches.

Deep learning-based studies on dermatophytosis have proliferated rapidly in recent years. Yilmaz et al. used CNN-based models on KOH microscopy images and achieved 96.2% accuracy in classifying infected versus healthy samples [66]. Jansen et al. employed a U-Net-based segmentation model on histologic nail sections, identifying fungal structures with a Dice score of 91% [178]. Decroos et al. developed a deep learning model for histopathologic sections and obtained 92.5% sensitivity, proving non-inferiority to conventional histopathologic diagnosis [179]. Koo et al. applied a YOLOv4-based object detection model to KOH microscopy images of superficial fungal infections, demonstrating that automated hyphae detection with bounding boxes is feasible in routine practice [180]. However, this study focused specifically on superficial mycoses, relied on an earlier CNN architecture, and did not address the substantial artifactual heterogeneity characteristic of dermatophytosis microscopy. Beyond single-architecture studies, recent research in biomedical object detection has emphasized the importance of comparative evaluation across different detection paradigms. Modern single-stage convolutional detectors such as YOLOv11 [181] and classical two-stage region-proposal networks such as Faster R-CNN [182] represent the principal CNN-based baselines against which transformer-based frameworks are benchmarked, yet their relative performance in

artifact-rich microscopy settings has not been systematically examined. Consequently, despite increasing interest in automated analysis, most available research still focuses on image-level classification or pixel-wise segmentation, and object-level detection approaches in routine KOH microscopy have not yet been systematically investigated. Crucially, although previous studies have addressed fungal detection using deep learning, relatively limited attention has been given to explicitly modeling artifactual structures as a distinct detection class in KOH microscopy datasets. Given the well-recognized diagnostic challenges posed by visually similar mimics, incorporating artifacts as a separate class may improve the robustness of automated detection systems in routine microscopy settings [183].

From a clinical perspective, spatial localization of fungal elements may offer additional information for estimating fungal burden and guiding treatment decisions, highlighting the insufficiency of image-level classification. While visually similar artifacts in KOH preparations frequently confound conventional local feature-based detectors, transformer-based architectures leverage global self-attention to effectively discriminate between biological structures and mimics, as evidenced by recent success in similar microscopic tasks [184, 175, 185].

In this study, we developed and evaluated an AI-based detection system for the automated and precise localization of fungal elements in routine KOH microscopy. All microscopic images were obtained using this routine preparation technique, thus capturing the heterogeneity, noise, and artifact patterns commonly encountered in real-world clinical practice, which are known to challenge the generalizability of AI-based systems [163, 165, 186]. Based on these considerations, we selected the RT-DETR (Real-Time Detection Transformer) model [33] as the object detection backbone. This study involved constructing an expert-verified and carefully curated dataset, where both fungal elements indicative of dermatophytosis and confounding artifactual structures were explicitly annotated as distinct classes. By training the RT-DETR model to simultaneously detect and distinguish these entities under standardized conditions, we assessed its performance using established detection metrics. Through this framework, we aimed to provide a fully automated and reproducible approach for the localization of fungal elements in routine KOH microscopy that remains robust against real-world artifactual interference.

4.3.2 Materials and Methods

The complete study pipeline from clinical specimen collection and KOH preparation, through high-resolution microscopic imaging and expert bounding-box annotation, to object detection model training, validation, and inference is schematically illustrated in Figure 4.1.

4.3.2.1 Clinical Sample Collection and Microscopic Imaging

The study involving human-derived specimens was approved by the Ethics Committee for Clinical Research of Istanbul Training and Research Hospital under protocol number 2011-KAEK-50, with approval number 20. Human scrapings were collected at Istanbul Research and Training Hospital under sterile conditions from patients with clinically and/or mycologically confirmed dermatophytosis. The specimens were incubated in

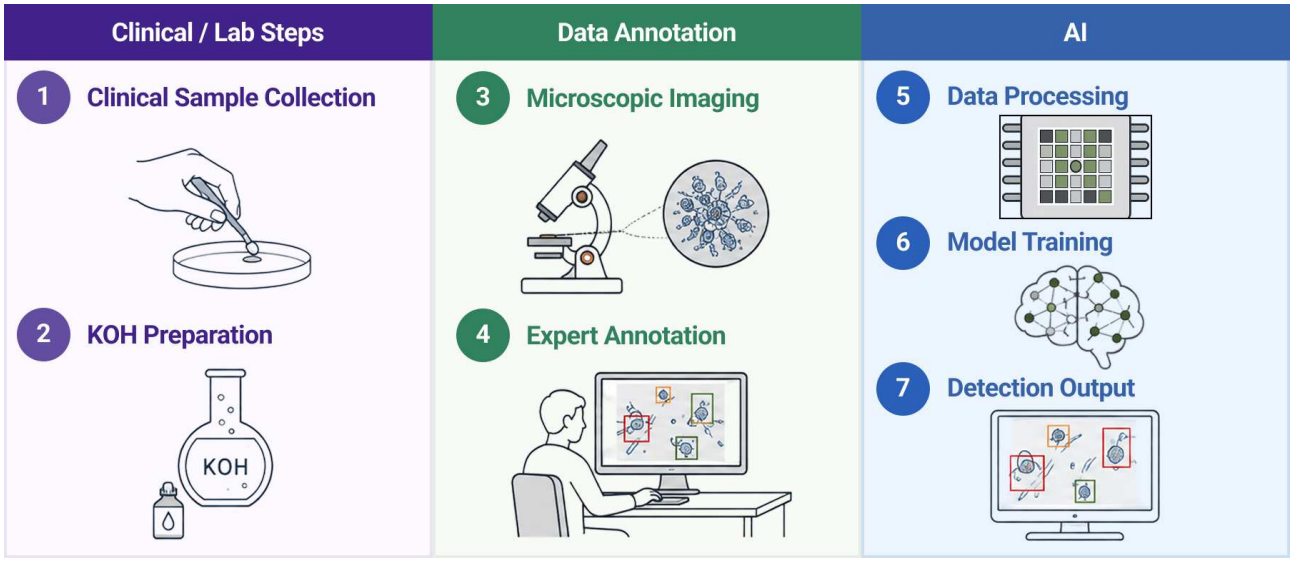


Figure 4.1: Schematic overview of the study workflow: clinical sample collection, KOH preparation, microscopic imaging, expert annotation, and artificial intelligence modeling steps.

Table 4.1: Descriptive statistics of the microscopy dataset and annotation categories used for model training and evaluation.

Category	Component	Count / Value
Images	Total Microscopic Frames	2,540
Resolution	Pixel Dimensions	$2{,}048 \times 2{,}048$
Annotations	Fungal Elements (Positive Class)	631
	Confounding Artifacts (Negative Class)	381

10–20% KOH solution to dissolve keratin, thereby enhancing the visibility of fungal hyphae and spores; nail specimens were incubated for 12 hours, whereas skin specimens were incubated for 1 hour. The cleared preparations were subsequently examined and imaged at high resolution using light microscopes.

A total of 2,540 microscopic images ($2{,}048 \times 2{,}048$ pixels) were acquired. These images were obtained from 100 clinical specimens collected from 60 patients diagnosed with dermatophytosis. Multiple unique, non-overlapping microscopic fields were captured from each specimen during routine microscopic examination, resulting in the total set of 2,540 high-resolution microscopy images used for model development and evaluation. All images were reviewed and subjected to an initial quality-control assessment by two clinicians (Gülsüm Gençoğlan (G.G.) and Barış Kızılyaprak (B.K.)), who subsequently performed the manual annotation. All annotations were reviewed by both clinicians, who were experienced in dermatologic microscopy. In cases of disagreement, the images were jointly re-evaluated and a consensus decision was reached. The descriptive statistics and class distribution of the resulting dataset are summarized in Table 4.1.

The annotation process adopted a multiclass strategy to address visual ambiguity in KOH microscopy. In addition to fungal elements, clinically relevant artifactual structures (such as keratin debris and fibers)

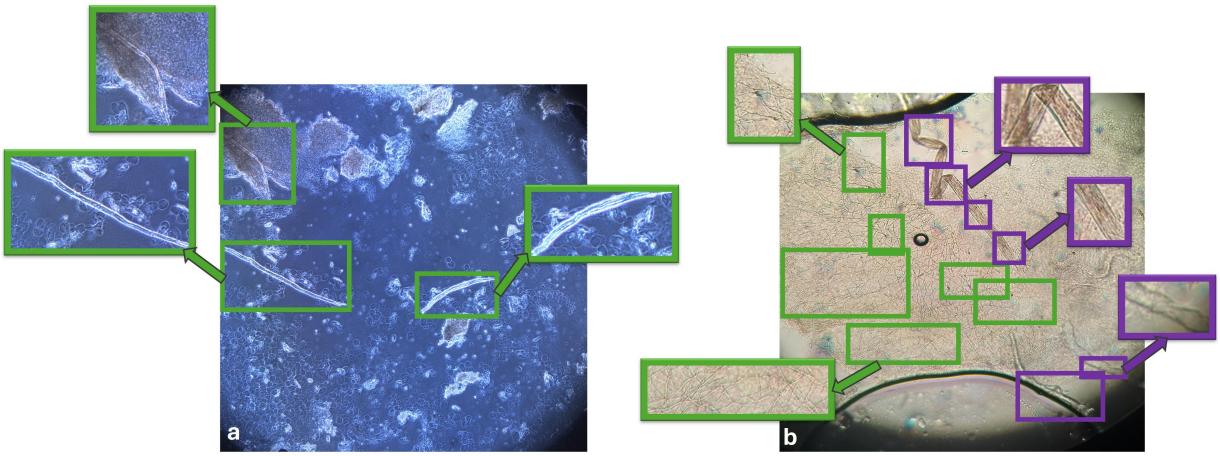


Figure 4.2: (a, b) Representative KOH microscopy images illustrating the multiclass annotation strategy. Fungal hyphae are marked with green bounding boxes and confounding artifacts are marked with purple boxes.

were annotated as a separate class to facilitate model discrimination between true fungal structures and visually similar mimics. By deliberately labeling these mimics rather than treating them as background, the strategy was designed to teach the model to actively discriminate between true fungal structures and the diagnostic challenges frequently encountered in routine KOH microscopy. A representative example of these multiclass expert annotations is shown in Figure 4.2.

4.3.2.2 Detection Architecture Overview and Experimental Design

Object detection in KOH microscopy presents a distinctive set of visual challenges: fungal hyphae are thin, low-contrast, and morphologically variable, while preparation-related artifacts, including keratin debris, synthetic fibers, refractive edges, and air bubbles, frequently resemble genuine fungal structures at the spatial scales relevant to bounding-box detection. To assess whether the design paradigm of the detection architecture meaningfully influences performance under these conditions, three distinct model families were evaluated: a transformer-based end-to-end detector (RT-DETR), a contemporary single-stage convolutional detector (YOLOv11), and a classical two-stage region-proposal network (Faster R-CNN).

RT-DETR [33] was selected as the primary model based on its hybrid CNN–Transformer encoder–decoder architecture, which integrates multi-scale convolutional feature extraction with global self-attention across the full image field. This design enables the model to contextualize each candidate detection within the surrounding tissue environment, a property hypothesized to be advantageous when discriminating structurally similar fungal and non-fungal structures. The attention-driven, query-based detection head performs direct bounding-box regression and classification without non-maximum suppression, removing a source of post-processing sensitivity that can affect recall in densely annotated microscopic fields. Training procedures, hyperparameter configuration, and augmentation strategies for RT-DETR are described in detail below.

YOLOv11 was included as a contemporary convolutional baseline, representing a modern single-stage

detection paradigm widely used in real-time computer vision applications [181]. Its architecture combines efficient convolutional feature extraction with an enhanced feature-pyramid neck, enabling rapid inference while maintaining competitive localization accuracy. Including YOLOv11 allows assessment of whether transformer-based global attention mechanisms provide measurable advantages over modern convolutional feature hierarchies in the context of artifact-rich KOH microscopy.

Faster R-CNN was included as a classical two-stage region-proposal detector that historically formed the foundation of many medical object detection systems [182]. Its region-proposal network first generates candidate object regions, which are subsequently refined and classified by a second-stage detection head. This architecture has been widely applied to tasks involving spatially sparse and morphologically heterogeneous objects. Evaluating Faster R-CNN therefore provides a reference baseline representing earlier convolutional detection frameworks.

4.3.2.3 Artificial Intelligence Model and Training Procedure

All three architectures (RT-DETR, YOLOv11, and Faster R-CNN) were trained on identical dataset partitions using the same two-class annotation scheme (fungal elements and confounding artifacts). The following describes the architectural configuration, training procedure, and inference settings specific to RT-DETR as the primary model.

Fungal element detection in KOH microscopy images was performed using the RT-DETR model, an end-to-end transformer-based object detector. The model employs an attention-driven query selection strategy within a hybrid CNN–Transformer encoder–decoder pipeline, enabling direct bounding box regression and classification. Combined with multi-scale feature fusion and efficient query refinement, the proposed architecture facilitates the detection of small, thin, and irregularly shaped fungal structures, such as hyphae and arthroconidia, while simultaneously distinguishing them from confounding artifactual mimics in dermatologic microscopy. The overall model workflow, including preprocessing, architectural components, and inference logic, is illustrated in Figure 4.3.

All experiments were conducted using the Ultralytics (version: 8.2.70) implementation of the RT-DETR model, with the RT-DETR-L configuration pretrained on the COCO dataset used for initialization. To ensure reproducibility, we note that training was performed on a workstation equipped with an NVIDIA GeForce RTX 4090 24 GB (NVIDIA Corporation, Santa Clara, CA, USA). To obtain statistically robust performance estimates and address the sensitivity of single-split evaluation in moderate-sized clinical imaging datasets, all three architectures were trained under a five-fold stratified cross-validation protocol applied to the training and validation pool (2,286 images). Stratification was performed at the image level to preserve the proportion of annotated positive and negative images across all folds. In each iteration, four folds (approximately 1,829 images) were used for training and one fold (approximately 457 images) for validation. Model checkpoints were selected based on AP@0.50:0.95 on the fold-level validation partition. The independent held-out test set was excluded from all training and validation procedures and was used exclusively for final performance evaluation.

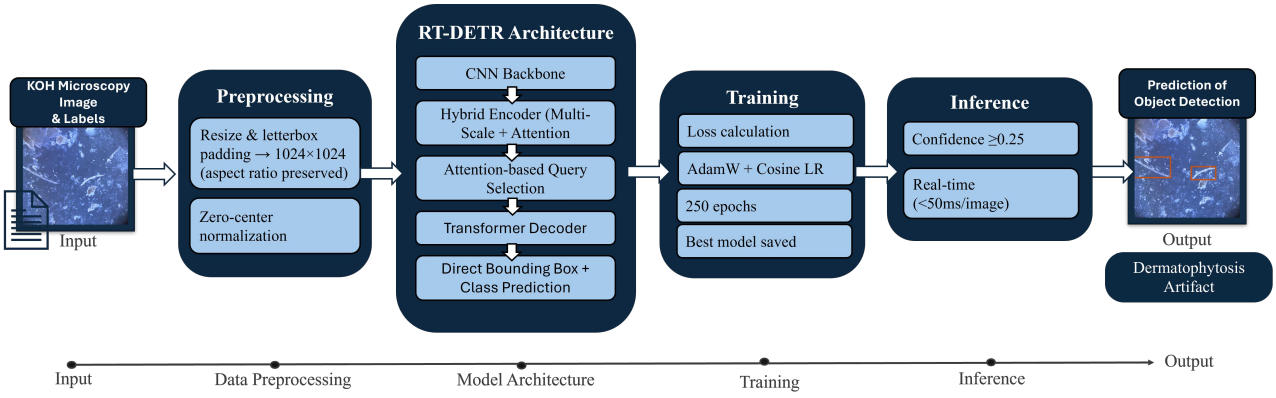


Figure 4.3: Overview of the fungal detection pipeline. KOH microscopy images are preprocessed and passed through the RT-DETR architecture, which combines a CNN backbone, hybrid encoder, attention-based query selection, and a transformer decoder to generate bounding box predictions for fungal elements and confounding artifacts.

Model training followed the hyperparameters and optimization strategy defined in the custom training pipeline. Specifically, the model was trained for up to 250 epochs using the AdamW optimizer, with an initial learning rate of 5×10^{-4}, cosine-annealed decay with warm-up, a batch size of 8, and mixed-precision training to reduce GPU memory consumption. Loss functions were weighted following standard practice (box = 7.5, classification = 1.0), and early stopping with a patience of 50 epochs was applied to prevent overfitting.

Given the morphological sensitivity of fungal hyphae and the need to preserve distinct artifact features, aggressive augmentations such as MixUp were intentionally disabled. Standard geometric variations were utilized, including horizontal flipping ($p = 0.2$), scaling ($\pm 20\%$), translation ($\pm 5\%$), and minor rotations ($\pm 2^{\circ}$). These choices ensured preservation of clinically relevant fungal morphology while still providing sufficient variability for robust learning.

Model selection was based on validation performance measured by AP@0.50:0.95, which reflects detection consistency across different intersection-over-union thresholds for all classes. During inference, predictions were filtered using a confidence threshold of 0.25. Due to its transformer-based matching mechanism, the model enables real-time prediction (less than 50 ms per image on the NVIDIA RTX 4090 GPU) and improves reliability when fungal structures appear in dense or overlapping formations. This end-to-end inference pipeline is summarized in Figure 4.3, which visualizes the progression from microscopy image preprocessing to final multiclass fungal region localization.

4.3.2.4 Baseline Model Evaluation

All three architectures were trained under the same five-fold stratified cross-validation protocol on the training and validation pool (2,286 images) using the identical two-class annotation scheme (fungal elements and artifacts). To facilitate direct architectural comparison, key training parameters were harmonized

across the three frameworks where applicable. Specifically, all models were trained for up to 250 epochs with an input resolution of $1{,}024 \times 1{,}024$ pixels, a batch size of 8, and COCO-pretrained weights as initialization. The initial learning rate was set to 5×10^{-4} with cosine annealing and early stopping (patience = 50 epochs). Architecture-specific components without direct equivalents across frameworks (e.g., the transformer query mechanism in RT-DETR, the distribution focal loss in YOLOv11, or the region proposal network in Faster R-CNN) were kept at their recommended default settings. Each fold-level model was then evaluated on the same independent held-out test set. To ensure a fair comparison, the same confidence threshold (0.25) was applied to all architectures, and detection metrics were computed using identical intersection-over-union criteria (AP@0.50 and AP@0.50:0.95). Inference speed was benchmarked under identical hardware conditions on the NVIDIA RTX 4090 GPU. This consistent evaluation protocol ensures that observed performance differences reflect architectural characteristics rather than differences in experimental conditions.

All reported metrics represent the mean and standard deviation of the five fold-level models evaluated on this independent test set. This design was adopted to quantify training stability rather than to perform ensemble inference. Each of the five independently trained models was evaluated on the same held-out test set solely to estimate the variance in performance attributable to training randomness and fold-specific data composition. The held-out test set was not used at any stage of model selection or hyperparameter tuning and was accessed only once per fold-level model for final performance reporting.

4.3.3 Results

Table 4.2 summarizes the object-level detection and image-level classification performance of all three architectures, reported as the mean and standard deviation across five models trained on different cross-validation folds and evaluated on the same independent held-out test set.

At the object level, RT-DETR achieved a mean precision of 0.921 ± 0.014, a mean recall of 0.831 ± 0.011, a mean F1-score of 0.874 ± 0.012, a mean AP@0.50 of $89.73 \pm 1.48\%$, a mean AP@0.50:0.95 of $76.85 \pm 1.22\%$, and a mean IoU of 0.838 ± 0.009 on the independent test set. YOLOv11 achieved a mean precision of 0.812 ± 0.016, a mean recall of 0.842 ± 0.012, a mean F1-score of 0.827 ± 0.013, a mean AP@0.50 of $81.34 \pm 1.52\%$, a mean AP@0.50:0.95 of $61.28 \pm 1.41\%$, and a mean IoU of 0.846 ± 0.008. Faster R-CNN achieved a mean precision of 0.333 ± 0.013, a mean recall of 0.800 ± 0.008, a mean F1-score of 0.470 ± 0.015, a mean AP@0.50 of $51.54 \pm 1.23\%$, a mean AP@0.50:0.95 of $27.47 \pm 1.01\%$, and a mean IoU of 0.750 ± 0.008.

Image-level classification performance was assessed by aggregating object-level detections per image. A case was classified as positive if at least one fungal element was detected above the confidence threshold (0.25). On the independent held-out test set, RT-DETR achieved a mean sensitivity of 0.989 ± 0.022, a mean specificity of 0.970 ± 0.005, a mean accuracy of 0.976 ± 0.008, a mean F1-score of 0.967 ± 0.012, and a mean of 0.2 ± 0.4 missed positive cases across the five evaluations. YOLOv11 achieved a mean sensitivity

Table 4.2: Object-level detection and image-level classification performance. All metrics are reported as mean ± standard deviation (SD) across five fold-level models evaluated on the independent held-out test set (254 images: 89 positive and 165 negative). A case was classified as positive when at least one fungal element was detected with a confidence score exceeding 0.25.

Level	Metric	RT-DETR	YOLOv11	Faster R-CNN
Object-Level	Precision	0.921 ± 0.014	0.812 ± 0.016	0.333 ± 0.013
	Recall	0.831 ± 0.011	0.842 ± 0.012	0.800 ± 0.008
	F1-score	0.874 ± 0.012	0.827 ± 0.013	0.470 ± 0.015
	AP@0.50 (%)	89.73 ± 1.48	81.34 ± 1.52	51.54 ± 1.23
	AP@0.50:0.95 (%)	76.85 ± 1.22	61.28 ± 1.41	27.47 ± 1.01
	Mean IoU	0.838 ± 0.009	0.846 ± 0.008	0.750 ± 0.008
Image-Level	Accuracy	0.976 ± 0.008	0.886 ± 0.012	0.765 ± 0.018
	Sensitivity	0.989 ± 0.022	0.767 ± 0.014	0.522 ± 0.021
	Specificity	0.970 ± 0.005	0.951 ± 0.008	0.896 ± 0.013
	F1-score	0.967 ± 0.012	0.826 ± 0.013	0.610 ± 0.019
	Missed Cases (FN)	0.2 ± 0.4	4.2 ± 1.1	8.6 ± 1.5

SD: standard deviation across five fold-level models evaluated on the independent test set. AP: average precision. IoU: intersection over union. FN: mean false negatives (missed positive cases).

of 0.767 ± 0.014, a mean specificity of 0.951 ± 0.008, a mean accuracy of 0.886 ± 0.012, a mean F1-score of 0.826 ± 0.013, and a mean of 4.2 ± 1.1 missed positive cases. Faster R-CNN achieved a mean sensitivity of 0.522 ± 0.021, a mean specificity of 0.896 ± 0.013, a mean accuracy of 0.765 ± 0.018, a mean F1-score of 0.610 ± 0.019, and a mean of 8.6 ± 1.5 missed positive cases.

4.3.4 Discussion

This study presents an object detection model for localizing fungal structures in KOH microscopy images of dermatophytosis. Previous deep learning work in this field has largely focused on image-level classification of infected versus non-infected samples or pixel-wise segmentation of fungal elements [66, 178, 179]. In contrast, our approach targets explicit object-level localization with bounding boxes, utilizing a multiclass detection strategy to distinguish true fungal elements from confounding artifacts, thereby aiming to provide interpretable and explainable visual cues on routine microscopy images.

Modern detection architectures have recently been applied to the automated analysis of fungal infections. Koo et al. employed a YOLOv4-based one-stage convolutional detector for hyphae detection in KOH microscopy images, demonstrating that automated bounding-box localization is feasible in clinical practice [180]. While their work provided an important proof of concept for CNN-based object detection, YOLOv4 relies on convolutional feature hierarchies and predefined anchor mechanisms, which represent a different design paradigm compared to transformer-based detectors. Moreover, their study focused on superficial mycoses under relatively controlled imaging conditions, with less visual complexity than is typically encountered in KOH preparations. In contrast, the present study evaluates the RT-DETR model, a transformer-based detector that performs end-to-end object detection, using attention-driven query selection and global contex-

tual reasoning. This architecture was evaluated in high-resolution KOH microscopy settings, where fungal hyphae appear small, low-contrast, and are often embedded within dense backgrounds containing keratin debris and preparation-related artifacts. By explicitly modeling mimics, such as fibers and air bubbles, as a distinct class during training, our model is designed to incorporate global contextual reasoning, which may help reduce false-positive detections in artifact-rich KOH microscopy settings. From a computational perspective, the model enables stable real-time inference in our experimental setting (below 50 ms per image), while maintaining a simplified detection pipeline without post-processing. These characteristics support not only efficient image-level analysis but also scalability to large microscopic fields, which may be advantageous for time-sensitive microscopic analysis in experimental settings. Taken together, while YOLOv4-based approaches have demonstrated the feasibility of automated hyphae detection, alternative detection frameworks that emphasize global contextual information and end-to-end prediction may offer complementary advantages in artifact-rich KOH microscopy settings.

The comparative evaluation of detection architectures provides additional insight into the suitability of transformer-based detectors for artifact-rich KOH microscopy. YOLOv11, a modern single-stage convolutional detector, achieved a mean recall of 0.842 ± 0.012 but a lower mean precision of 0.812 ± 0.016 (mean F1: 0.827 ± 0.013; mean AP@0.50: $81.34 \pm 1.52\%$), suggesting a tendency to generate false-positive detections in the presence of visually ambiguous artifacts. Faster R-CNN demonstrated the weakest overall performance, with a mean precision of 0.333 ± 0.013 and a mean AP@0.50 of $51.54 \pm 1.23\%$, indicating that the region-proposal mechanism frequently nominated background structures, including keratin debris and synthetic fibers, as candidate fungal regions across all five test-set evaluations. Its inference speed of 255.4 ms per image also represents a substantial computational disadvantage compared to RT-DETR (41.9 ms). The comparatively high false-positive rates observed for both CNN-based architectures may suggest that global contextual reasoning could be advantageous when discriminating morphologically similar structures in artifact-rich KOH microscopy. However, the present comparative evaluation assesses complete detection frameworks rather than isolating the contribution of individual architectural components. Therefore, the observed results should be interpreted as a comparative evaluation between detection architectures rather than as direct evidence for the superiority of the self-attention mechanism itself. These findings suggest that transformer-based end-to-end detection may offer practical advantages in artifact-rich microscopy settings, though validation on larger and more diverse datasets would be required to establish a definitive performance ranking across architectures.

Object-level localization is also closely aligned with the way clinicians read KOH slides. In daily practice, fungal elements must be identified among numerous mimics, including keratin debris, air bubbles, fibers, and staining artifacts, many of which can resemble hyphae and contribute to diagnostic variability [159]. Less-experienced observers in particular may confuse hyphae with these structures, a well-documented limitation of KOH microscopy that contributes to diagnostic variability [160]. Moreover, routine KOH examination often requires manual scanning of multiple large microscopic fields, which is time-consuming and may further increase reader fatigue and variability. By drawing bounding boxes around regions with a high probability of containing fungal elements while ignoring learned artifacts, the proposed model provides

visual cues. However, whether these cues improve diagnostic decision-making requires prospective reader studies. Such visual localization aligns with principles of explainable artificial intelligence, which emphasize transparency and human interpretability as key requirements for clinical deployment of AI systems [187]. In addition, the number and distribution of detected regions could in the future be used to approximate fungal burden or to monitor changes during treatment.

Quantitative analysis of the misclassified instances revealed that most false detections were associated with structures that closely resemble fungal hyphae, including synthetic fibers, scratch marks, overlapping cell borders, and sharp keratin edges. In addition, a small number of very faint or out-of-focus hyphal fragments with extremely low contrast against dense keratin backgrounds were occasionally missed. These cases represent the technical edge conditions encountered during the detection of subtle biological structures in heterogeneous KOH preparations. Overall, the quantitative results support the technical feasibility of the proposed approach. The low standard deviation of AP@0.50:0.95 across the five test-set evaluations ($76.85 \pm 1.22\%$) suggests that spatial localization quality is stable across the five independently trained models and that the results are not driven by a particularly favorable model initialization. The low mean number of missed positive cases on the independent test set (0.2 ± 0.4 missed positive cases) is an encouraging preliminary finding for screening-oriented applications; however, it should be interpreted within the constraints of the single-center retrospective design, and prospective validation is required before any inference about clinical reliability can be drawn. Evaluating AP across multiple IoU thresholds provides a more stringent measure of spatial agreement than a single-threshold metric and is consistent with current practice in object detection. The performance levels observed here are comparable to those reported for other microscopic detection tasks, including leukocyte detection, parasitic egg localization, and hyphae detection in KOH images [180, 184, 175, 185].

To qualitatively evaluate the model's performance, a comparative failure analysis was conducted across different detection scenarios in Figure 4.4. The top row, comprising Figure 4.4a–c, illustrates the expert-verified ground truth (GT) annotations, highlighting the high density of fungal structures. The middle row shows the model's independent inference; while Figures 4.4d and 4.4f demonstrate high-confidence detections, Figure 4.4e identifies "mimics," non-fungal artifacts that visually resemble hyphae, and labels them as artifacts to prevent diagnostic confusion.

The bottom row (Figure 4.4g–i) specifically illustrates instances of misclassification and the limitations of the detection framework. In Figure 4.4g, the model produces a false-positive detection by identifying a non-fungal structure as a hypha, which would lead to a fungus-negative field being incorrectly flagged as infected (dermatophytosis). Similarly, Figure 4.4h reveals a critical misclassification where the model identifies an ovoid, non-fungal structure as dermatophytosis, further demonstrating the system's tendency to overpredict in the presence of ambiguous biological morphology. Finally, Figure 4.4i confirms that certain synthetic fibers and sharp keratin edges continue to trigger false detections despite the multiclass training strategy. These instances represent technical bottlenecks where the high morphological similarity between debris and true hyphae limits the model's precision in complex KOH preparations.

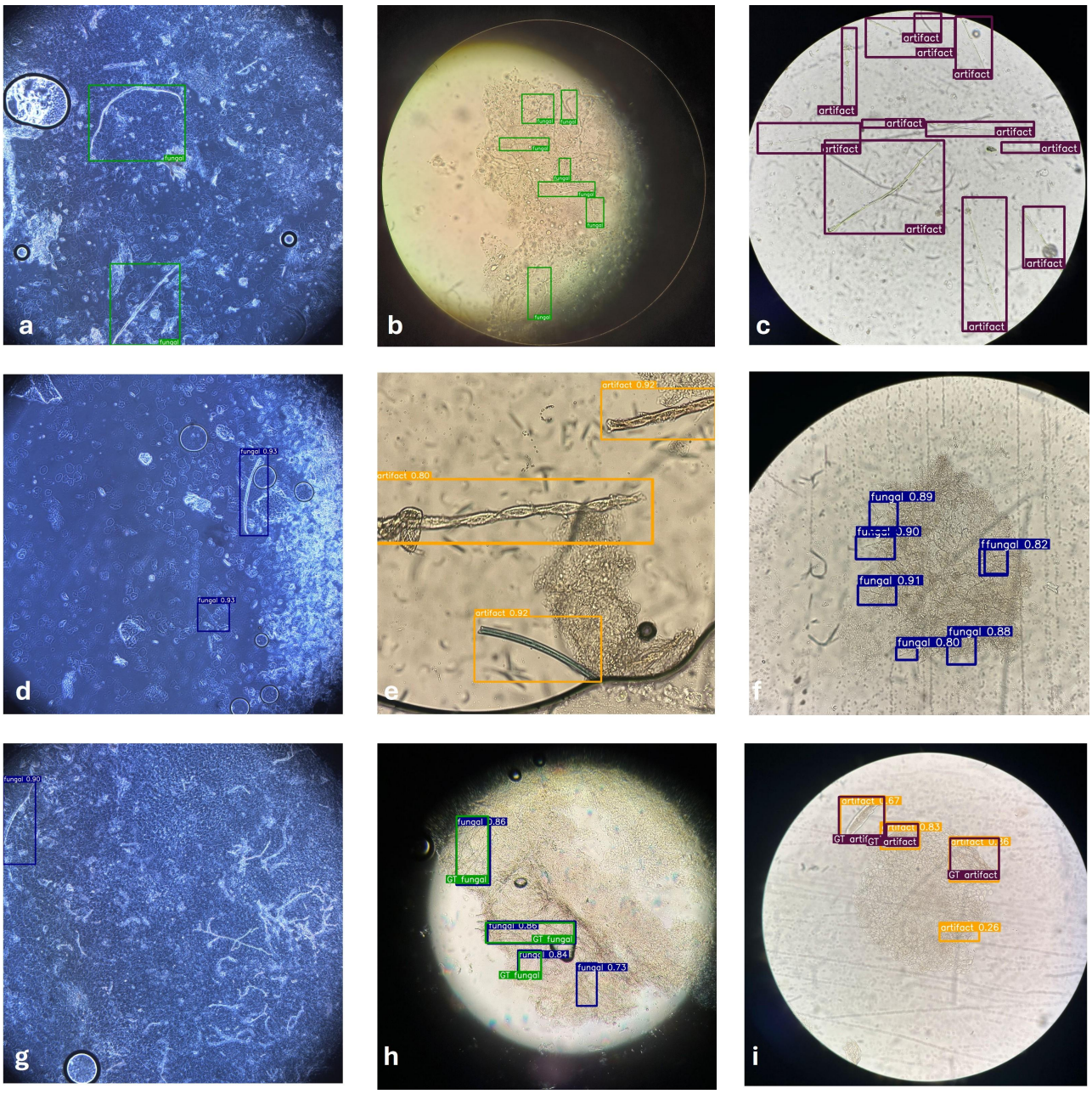


Figure 4.4: Qualitative examples of model predictions in KOH microscopy images. (a–c) Ground truth annotations showing fungal elements (green bounding boxes) and artifacts (purple bounding boxes). (d–f) Model predictions with confidence scores, including examples of correct discrimination between fungal structures (blue bounding boxes) and artifact mimics (yellow bounding boxes). (g–i) Overlay of ground truth and model predictions illustrating correct detections and representative misclassifications.

Several characteristics of the dataset and training strategy have likely contributed to these results. The images were acquired under routine conditions and deliberately included a wide range of KOH dissolution stages, from partially cleared preparations with abundant keratin to more transparent fields. This heterogeneity introduces considerable noise and artifacts, which makes the task more challenging for the model but also increases the dataset's relevance to real-world practice, where slides are rarely ideal. The use of high-resolution inputs and a transformer-based backbone allowed the network to combine local shape cues with broader contextual information from the surrounding tissue. This combination of fine-grained feature extraction and contextual modeling has also been shown to improve the detection of small, low-contrast structures in other microscopic imaging tasks [161].

This study has several limitations. The dataset originates from a single center, and its moderate size, while mitigated by COCO-based pretraining and five-fold cross-validation, nonetheless restricts generalizability; external multicenter validation is required to confirm performance across diverse clinical settings. Regarding class definitions, while the current model employs a multiclass strategy to distinguish fungal elements from artifacts, it treats all fungal species as a single category. Future extensions should therefore incorporate richer annotation schemes to differentiate specific fungal morphologies. A critical clinical consideration is the differential diagnosis of dermatophytosis. Nail unit malignancies, including subungual melanoma and squamous cell carcinoma, can clinically mimic dermatophytosis [188, 189, 190]. Since the proposed model is trained to detect fungal presence but not to exclude these mimics, relying solely on automated screening could obscure concurrent malignancies. Therefore, future work aims to integrate anomaly detection mechanisms to flag atypical patterns for expert review. Finally, prospective reader studies will be essential to translate these image-level performance gains into tangible improvements in clinical diagnostic accuracy and reporting efficiency.

In conclusion, this study demonstrates that a transformer-based detection framework, reinforced by a multiclass training strategy, can reliably identify fungal elements even within the complex, artifact-laden environment of routine KOH microscopy. By translating neural network predictions into interpretable bounding-box overlays, the system provides localized visual cues, which may support human-in-the-loop review in future clinical workflows. Ultimately, this localization capability may support the future integration of AI-based decision-support systems into clinical microscopy workflows, pending prospective validation, thereby potentially contributing to more consistent diagnostic assessment in dermatomycology.

4.4 Privacy-Preserving Local Language Models for Pemphigus Follow-Up

Chronic dermatologic diseases such as pemphigus require long-term follow-up, generating extensive longitudinal clinical documentation that is difficult to review comprehensively during routine visits and increasing clinician workload as well as the risk of missing critical historical information. We evaluated whether a locally deployed, privacy-preserving small language model (SLM) could retrieve structured clinical features and generate longitudinal summaries from long-term dermatology follow-up records. In this retrospective case series, 30 patients with pemphigus contributed 541 visit notes that were aggregated into full longitudinal records (89,336 words); 55 clinically relevant features were annotated by two expert dermatologists. The locally deployed SLM (Qwen3 4B Thinking 2507) was queried with each complete record to retrieve 55 features and generate one final report summary. Across feature retrieval tasks, mean accuracy was $82.25 \pm 0.17\%$. Dermatologists' ratings of AI-generated summaries were high for overall quality (8.23 for Evaluator 1 and 8.47 for Evaluator 2), clinical accuracy (7.93 and 8.20, respectively), and usefulness (8.50 and 8.47, respectively), with no statistically significant differences in mean ratings between evaluators. In a separate, independently rated pairwise comparison, AI-generated reports were preferred in 53.3% of evaluations. These findings suggest that privacy-preserving, locally deployed SLMs can generate clinician-facing longitudinal summaries that receive high expert ratings for overall quality, clinical accuracy, and usefulness when reviewed in the context of the patients' original longitudinal follow-up documentation.

4.4.1 Introduction

Chronic dermatologic diseases such as pemphigus necessitate prolonged, often lifelong follow-up, during which patients accumulate extensive clinical documentation spanning many years [191, 192]. These longitudinal records may include dozens of visits authored by different clinicians, each documenting disease activity, treatment adjustments, laboratory findings, and adverse events. In routine clinical practice, clinicians are expected to rapidly synthesize this historical information during follow-up visits to guide management decisions. However, comprehensive review of long and heterogeneous clinical narratives is frequently impractical in busy outpatient settings, particularly when patients are seen by different physicians over time. To manage this growing documentation burden, clinicians often rely on partial chart review or reuse of previous documentation through copying and pasting. While expedient, this practice introduces variability, reduces standardization, and increases the risk that critical historical details, such as prior exposure to rituximab, cumulative corticosteroid toxicity, or serious infections, are overlooked. In chronic autoimmune blistering diseases, where treatment decisions depend heavily on longitudinal disease behavior rather than isolated visits, incomplete recall of historical data may adversely affect patient safety and clinical outcomes.

Recent advances in artificial intelligence, particularly NLP and language models (LMs), have demonstrated substantial potential for clinical summarization, information extraction, and clinical decision support [193,

194, 195]. In dermatology, LM-based approaches have been explored for tasks ranging from disease classification to clinical documentation analysis, with growing interest in multimodal and domain-adapted models [77, 196, 197]. More broadly, LM-based systems have shown promise in generating clinical visit summaries, assisting with patient messaging, and supporting clinician communication workflows [193, 198].

Despite these advances, the application of generative LMs to real-world EHR data remains at an early stage, with clinical summarization relatively underexplored compared with other clinical tasks [199]. The use of LMs with longitudinal clinical records also raises important concerns regarding patient privacy, data governance, safety, and real-world deployability, particularly in healthcare systems with strict data-protection requirements [200, 201]. These challenges are especially relevant to longitudinal dermatology follow-up, where clinically meaningful interpretation may require synthesizing information across multiple years of care rather than from isolated encounters. Resource-efficient and domain-specific AI models have recently been proposed to address some of these challenges in dermatology [197]; however, systematic evaluations of locally deployed, privacy-preserving LMs using real-world longitudinal dermatology records remain limited.

In this study, we evaluate a fully local, privacy-preserving Qwen3-based SLM system designed to retrieve structured clinical features and generate longitudinal summaries directly from complete multi-year dermatology follow-up records. Using pemphigus as a representative chronic dermatologic disease, we assess whether such a system can support the automation of longitudinal data retrieval, while maintaining high accuracy.

4.4.2 Methods

4.4.2.1 Study Design and Dataset

The study was conducted in accordance with the Declaration of Helsinki and was approved by the Clinical Research Ethics Committee of Istanbul Research and Training Hospital (Approval No: 18, Date: 09 January 2026, Protocol No: 2011-KAEK-50). Given the retrospective design and use of de-identified clinical records, the requirement for informed consent was waived by the ethics committee. This retrospective case series included 30 patients diagnosed with pemphigus who underwent long-term dermatologic follow-up. All available visit notes for each patient were extracted from the hospital information system at Istanbul Research and Training Hospital (DermaBase) and aggregated into a single chronological longitudinal record, preserving the original narrative text without manual summarization. The cohort comprised a total of 541 visit notes, with a mean of 18.03 ± 13.13 visits per patient. The overview of the study workflow is presented in Figure 4.5.

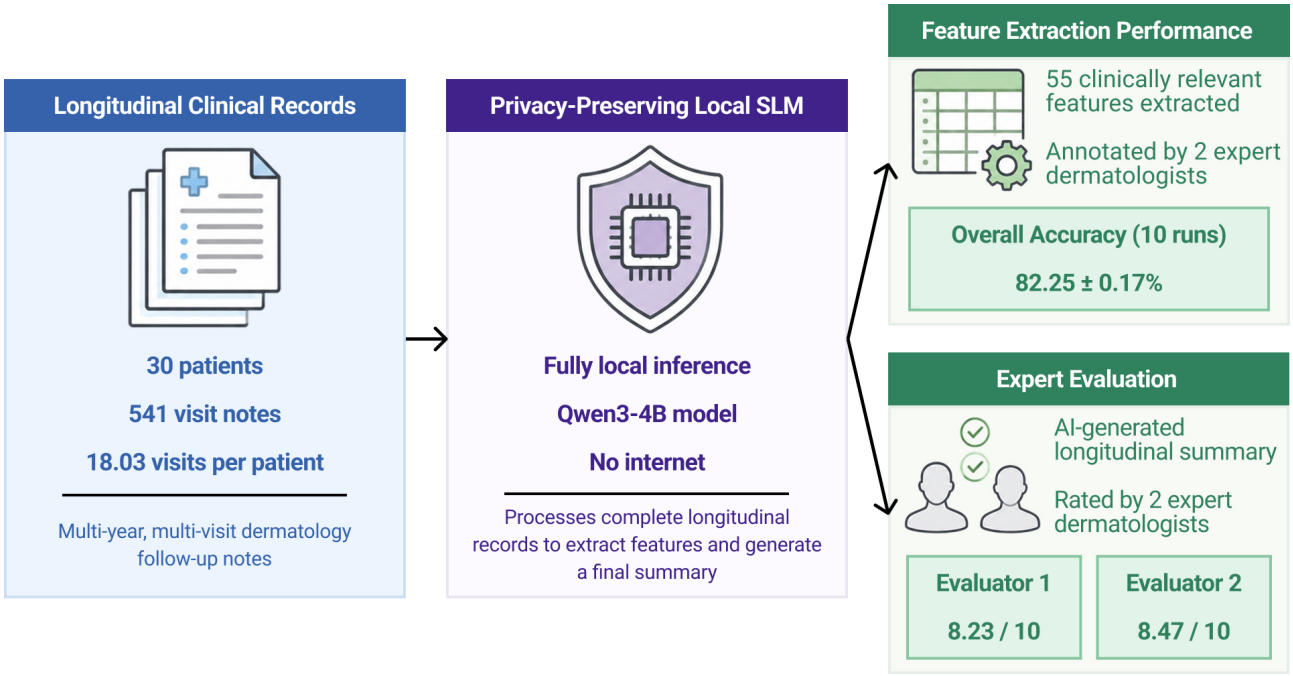


Figure 4.5: Overview of the study workflow: longitudinal pemphigus follow-up notes are aggregated into patient-level records and processed locally by a privacy-preserving SLM to retrieve structured clinical features and generate longitudinal summaries, which are then evaluated based on expert annotations.

4.4.2.2 Feature Schema and Ground-Truth Annotation

We defined a structured feature schema containing 55 features to capture clinically relevant longitudinal information from dermatology follow-up notes. These notes include diagnosis and phenotype, disease activity, laboratory findings (e.g., anti-desmoglein serology), systemic treatments (e.g., corticosteroids and steroid-sparing immunosuppressants), treatment dates and cycles (e.g., rituximab/IVIG), complications, and adverse events. The complete list of field names and their corresponding data types (categorical, numeric, or date) is provided in Table 4.3. In addition to structured fields, we evaluated the generation of a clinician-facing final report (*current_status*) summarizing the longitudinal course. Ground-truth annotations and completion of the summary reports were performed by two expert dermatologists who had long-term experience managing the included patients (A.E.K.A. and D.Y.).

Table 4.3: List of fields and their data types used for feature extraction.

Field name	Type
patient_id	text
visit_count_est	numeric
pemphigus_diagnosis	categorical
cyclophosphamide_ever	categorical
rituximab_ever	categorical

Field name	Type
rituximab_cycles_count	numeric
rituximab_first_date	date
rituximab_last_date	date
rituximab_response	categorical
ivig_ever	categorical
ivig_last_date	date
plasmapheresis_ever	categorical
plasmapheresis_last_date	date
renal_toxicity_ever	categorical
infusion_reaction_ever	categorical
dose_changes_documented	categorical
histology_pemphigus_compatible	categorical
report_id	text
pemphigus_subtype	categorical
cytopenia_ever	categorical
report_span_end_date	date
phenotype_mucosal	categorical
serious_infection_ever	categorical
diagnosis_confidence	categorical
anti_dsg1_positive_at_first_diagnosis	text
methotrexate_ever	categorical
steroid_complication_hypertension	categorical
longitudinal_course_clear	categorical
steroid_complication_myopathy	categorical
dif_pemphigus_compatible	categorical
anti_dsg3_positive_at_first_diagnosis	text
anti_dsg3_last_value	text
mmf_ever	categorical
steroid_complication_cataract	categorical
anti_dsg1_last_value	text
report_span_start_date	date
steroid_complication_avascular_necrosis	categorical
azathioprine_ever	categorical
phenotype_cutaneous	categorical
liver_toxicity_ever	categorical
steroid_complication_glaucoma	categorical

Field name	Type
steroid_complications_ever	categorical
medication_start_stop_dates_present	categorical
steroid_complication_hyperglycemia	categorical
immunosuppressant_stop_reason	categorical
anti_dsg3_last_date	date
current_systemic_treatment	categorical
anti_dsg1_last_date	date
current_prednisone_mg	numeric
last_flare_date	date
current_disease_activity	categorical
steroid_complication_osteoporosis_or_osteopenia	categorical
flare_count_est	numeric
current_steroid_duration_months	numeric
total_steroid_duration_months	numeric
current_status	text

4.4.2.3 Question Design and Inference Procedure

We designed the inference procedure in two sequential steps to balance fair model evaluation with clinically meaningful interpretation. We initially evaluated the Qwen3 model using the fixed main prompt (Appendix B.1) without additional domain-specific medical guidance to assess its intrinsic capability for longitudinal information extraction. During this initial prompt-development stage, we identified features for which interpretation depended on standardized clinical definitions, routine clinical conventions, or study-specific operational definitions. For these selected features, we subsequently added minimal explanatory information to the corresponding feature-specific notes (Appendix B.2). The main prompt itself remained unchanged throughout the study.

These explanatory notes were patient-independent and contained neither patient-specific information nor ground-truth answers. Their purpose was to reduce ambiguity in the interpretation of closely related medical concepts, particularly because Qwen3 is not specifically trained as a medical-domain model. Formal evaluation was then conducted using the fixed main prompt together with these predefined feature-specific notes.

For example, for the feature `steroid_complication_osteoporosis_or_osteopenia`, bone mineral density thresholds were explicitly defined: a T-score between -2.5 and -1.0 indicates osteopenia, whereas a T-score ≤ -2.5 indicates osteoporosis. These clarifications were introduced solely to reduce ambiguity and did not involve advanced prompting strategies or multi-step reasoning. Each feature was paired with a

feature-specific clinical question and an accompanying note specifying the expected answer format and interpretation rules (Appendix B.2). For the final report (*current_status*), a clinical template was provided to guide structured text generation. After the feature-specific notes were finalized, the complete prompt configuration, comprising the fixed main prompt, the feature-specific question, permitted output values, and any applicable explanatory note, was frozen and used unchanged across all ten repeated inference runs used to generate the reported performance metrics. The feature-specific notes were prespecified before formal evaluation, and the 30-patient evaluation cohort was not used to select or revise these definitions.

4.4.2.4 Model Deployment

Inference was performed using the Qwen3 4B Thinking 2507 [202] model deployed locally via LM Studio (version 0.3.22) on a workstation equipped with an RTX 4090 graphics card with 24 GB VRAM and 96 GB RAM. Internet access was disabled, and all data remained local throughout the study. The model was accessed via OpenAI-compatible local endpoints using LM Studio's local hosting feature. Predictions were repeated 10 times using an in-house Python script (version 3.12.3). Inference parameters included a temperature of 0.3, a context length of 90,000 tokens, and a timeout-based stopping criterion of 100 seconds. Key study details are summarized in Table 4.4.

Table 4.4: Summary of study and deployment statistics.

Category	Value
Patients	30
Total visit notes	541
Visits per patient	18.03 (range: 1–40)
Clinically relevant features	55
Inference requests (single-pass)	1,680
Repeats	10
Total model inferences	16,800
Model	Qwen3 4B Thinking 2507
Deployment	LM Studio v0.3.22
Hardware	RTX 4090 (24 GB VRAM), 96 GB RAM
Temperature	0.3
Context length	90,000 tokens
Timeout	100 s

4.4.2.5 Evaluation Metrics

Structured features were evaluated using exact-match accuracy based on expert annotations. Open-ended features were evaluated using a two-tiered approach. First, automated text similarity was assessed using BLEU and ROUGE metrics. Second, clinical relevance and usefulness were evaluated through expert assessment, followed by an independent pairwise preference comparison by two expert dermatologists. Two expert dermatologists (each with more than 20 years of clinical experience) evaluated 30 AI-generated patient reports. For each patient, the evaluator reviewed the AI-generated final report while consulting the

original longitudinal follow-up notes to determine whether the generated summary appropriately reflected the documented clinical course. The AI-generated report was rated for (i) overall quality, (ii) clinical accuracy, and (iii) clinical usefulness using 10-point Likert scales. A 10-point scale was selected to provide sufficient granularity for evaluating diverse, clinically meaningful error types that can occur in LM-generated reports. Reviewers were instructed to deduct points (typically 1–2 points per issue) for specific problems such as incorrect dates, incorrect clinical data, hallucinated statements, or clinically important missing information; smaller Likert ranges would have reduced sensitivity to these graded penalties and limited discrimination between near-correct and clearly incorrect outputs. In a separate pairwise preference assessment, the clinician-written and AI-generated reports were presented without revealing their origin, and evaluators selected the report they preferred (binary choice). All 30 patients were independently assessed by two evaluators, resulting in 60 total evaluations. Inter-rater reliability was assessed using the intraclass correlation coefficient (ICC) for continuous measures. Performance metrics from repeated model inference are presented as mean $\pm$ standard deviation (SD), whereas expert Likert-scale evaluations are reported as mean scores.

4.4.3 Results

The AI system accepted each patient's full longitudinal record as a single input, enabling feature retrieval from the entire follow-up history in one context. Specifically, the input consisted of the complete, chronologically aggregated free-text visit notes for a given patient. For each inference call, we queried the model using the complete record, resulting in 56 inference requests per patient (55 feature-specific requests and one report-generation request) and 1,680 requests across 30 patients. To assess output stability, we repeated the full inference procedure ten times under identical conditions, yielding 16,800 total model inferences. Inference speed was also favorable, with a time-to-first-token of approximately 0.05 seconds and a generation speed of roughly 130 tokens per second (approximately one word corresponds to 1.5 tokens). This performance allows rapid generation of longitudinal summaries, supporting rapid local inference in the experimental setting. Key performance metrics are summarized in Table 4.5 for feature-based metrics and Table 4.6 for report-level and expert-evaluation analyses.

Table 4.5: Summary of key performance results (mean $\pm$ SD across 10 repeated inference runs).

Metric	Value
Overall accuracy	$82.25 \pm 0.17\%$
Categorical feature accuracy	$86.34 \pm 0.13\%$
Numerical feature accuracy	$70.61 \pm 0.98\%$
Date-based feature accuracy	$80.56 \pm 0.59\%$
Final report BLEU	0.3602 ± 0.0029
Final report ROUGE-1	0.5195 ± 0.0051
Final report ROUGE-L	0.4662 ± 0.0055

Across all inference tasks, overall feature retrieval accuracy reached a mean of $82.25 \pm 0.17\%$ (range:

82.00–82.67%). Feature-level performance varied by data type. Categorical features achieved a mean exact-match accuracy of 86.34 ± 0.13% (range: 86.10–86.57%), while numerical features reached 70.61 ± 0.98% (range: 70.00–73.33%) and date-based features reached 80.56 ± 0.59% (range: 78.89–80.74%). Errors in numerical and date fields most commonly involved imprecise value extraction or partial temporal ambiguity rather than gross misclassification. For the open-ended final report, automated text similarity evaluation yielded a mean BLEU score of 0.3602 ± 0.0029, a ROUGE-1 score of 0.5195 ± 0.0051, and a ROUGE-L score of 0.4662 ± 0.0055. Ground-truth reports averaged 758.9 characters (105.1 words), whereas model-generated reports averaged 1,580.8 characters (217.7 words), corresponding to mean increases of 821.9 characters (108.3%) and 112.6 words (107.1%), respectively.

Table 4.6: Summary of report characteristics, expert evaluation of AI-generated reports, preference, and reliability analyses.

Category	Metric	Result
Report Length	Ground-truth length	758.9 characters (105.1 words)
	Model-generated length	1,580.8 characters (217.7 words)
	Absolute increase	+821.9 characters, +112.6 words
	Relative increase	+108.3% characters, +107.1% words
Expert Evaluation of AI-Generated Reports	Overall quality	8.23 vs 8.47 (Eval 1 vs Eval 2)
	Clinical accuracy	7.93 vs 8.20 (Eval 1 vs Eval 2)
	Clinical usefulness	8.50 vs 8.47 (Eval 1 vs Eval 2)
Pairwise AI-vs.-Clinician Preference	Overall preference	32/60: AI 53.3% vs Clinician 46.7%
	Evaluator 1 preference	AI 43.3% vs Clinician 56.7%
	Evaluator 2 preference	AI 63.3% vs Clinician 36.7%
	Total evaluations	60
Reliability & Correlation	Inter-rater reliability (ICC)	0.648–0.723
	Metric correlations (Pearson r)	0.92–0.97
	Preference vs quality correlation	0.70–0.81

In addition to automated metrics, the model-generated final reports were evaluated using a 10-point Likert-scale assessment by two expert dermatologists with more than 20 years of clinical experience. For each patient, evaluators reviewed the AI-generated report while consulting the original longitudinal follow-up documentation to verify the appropriateness and clinical accuracy of the generated summary. The AI-generated reports were rated for overall quality, clinical accuracy, and clinical usefulness. A total of 60 evaluations were completed. Quality scores were high across all metrics. Mean scores were as follows: overall quality 8.23 (Evaluator 1) vs 8.47 (Evaluator 2; $p = 0.147$); clinical accuracy 7.93 (Evaluator 1) vs 8.20 (Evaluator 2; $p = 0.133$); and clinical usefulness 8.50 (Evaluator 1) vs 8.47 (Evaluator 2; $p = 0.851$). Paired t-tests revealed no statistically significant differences between the two evaluators for any metric (all $p > 0.05$).

In a separate pairwise preference assessment (60 total evaluations), in which report origin was concealed from the evaluators, AI-generated reports were preferred in 53.3% of evaluations (32/60), compared to 46.7% (28/60) for clinician-generated reports. Preference patterns differed between evaluators: Evaluator

1 favored clinician reports (Ground Truth: 56.7% vs AI: 43.3%), whereas Evaluator 2 favored AI reports (Ground Truth: 36.7% vs AI: 63.3%). Reliability analysis showed moderate agreement between evaluators for continuous metrics (ICC = 0.648–0.723, all $p < 0.001$). Correlation analysis demonstrated strong positive correlations between quality metrics within each evaluation (Pearson r = 0.92–0.97, all $p < 0.001$). Preference was positively associated with quality scores (r = 0.70–0.81), suggesting that reports rated more highly were more often selected as preferred.

4.4.4 Discussion

This study demonstrates that a fully local, privacy-preserving SLM can retrieve clinically meaningful longitudinal information from complex dermatology follow-up records. By automating longitudinal data retrieval and generating clinician-facing reports, such systems have the potential to reduce clinician workload during routine follow-up visits [203, 204]. Importantly, our findings emphasize the role of LMs primarily as report-generation and summarization tools, rather than autonomous clinical decision-makers. The low standard deviations observed for type-wise structured extraction (categorical, numeric, and date fields) indicate high reproducibility across repeated inference runs under fixed decoding settings, even when mean performance differed by field type.

From a clinical perspective, the ability to rapidly surface key longitudinal information (such as prior biologic exposure, cumulative corticosteroid burden, flare frequency, and serious adverse events) may improve clinical decision-making and reduce cognitive load. This is particularly relevant in longitudinal dermatology care, where patients are often managed by multiple clinicians over extended periods.

A central finding of this study is that meaningful clinical benefit arises primarily from automated text generation for clinicians, rather than perfect extraction of every individual feature. The evaluated Qwen3 4B Thinking 2507 model appeared better suited to narrative synthesis than to strict structured extraction. While it missed some details during feature-level retrieval, it often integrated the same information more coherently when generating the final report. This behavior also highlights that different LMs have different capabilities for different clinical tasks [205]. No single small model can optimally handle all clinical NLP tasks; model selection should be driven by the intended workflow. In this study, our primary aim was longitudinal text generation, while also systematically assessing feature-based clinical data retrieval. Consistent with this, the Qwen3 model demonstrated strong report-generation performance, likely due to its stronger reasoning capabilities and ability to integrate long-context information.

Preliminary failure analysis conducted during prompt development indicated that some feature-level discrepancies arose when the model's default interpretation differed from a prespecified clinical convention or study-specific operational definition. For some complication fields, documented suspicion was considered sufficient evidence for a positive classification under our study definition, whereas the model initially classified the complication as absent because it had not been definitively confirmed. To standardize this interpretation, the corresponding feature-specific notes explicitly stated that documented suspicion was

1

GROUND TRUTH: Pemphigus type: Pemfigus Vulgaris; Current disease activity: Complete remission off therapy; Last eye exam date: NA; Last bone density test date: 2008-08-03; Osteoporosis treatment at visit if valid: NA; Last quantiferon date: NA; Inh prophylaxis date: NA; Last hepatitis serology date: NA; Last cbc date: NA; Last lft date: NA; Rituximab ever: No; Last rituximab date: NA; Prednisone use at the moment: None; Is there any safety & adverse events: Yes; Which treatment received and their dates: Prednol 2008 to 2010; Is there any side effect for any treatment: Cushingoid appearance (2008)

AI REPORT: Pemphigus type: Pemfigus vulgaris; Current disease activity: Remission (no new lesions reported since 2010, last clinical visit in 2014 with no symptoms); Last eye exam date: NA; Last bone density test date: 2008-08-03 (T-score: -1.4, normal); Osteoporosis treatment at visit: NA; Last quantiferon date: NA; Inh prophylaxis date: NA; Last hepatitis serology date: NA; Last cbc date: NA; Last lft date: NA; Rituximab ever: No; Last rituximab date: NA; Prednisone use at the moment: No; Is there any safety & adverse events: No reported adverse events; Treatments received and their dates: Prednisone (2008–2010, dose reduced from 16 mg/day to 4 mg/day); Is there any side effect for any treatment: No infections or steroid complications reported; Major surgery: None; Comorbid conditions: Cushingoid appearance (not confirmed as due to steroid use); Warning: Multiple entries on 2008-08-03 with identical dates and varying text content → possible copy-paste error; All entries from 2008–2014 show consistent clinical status with no new lesions, and the patient has been off medication since 2010 with no relapse reported.

2

GROUND TRUTH: Pemphigus type: Pemfigus herpetiformis; Current disease activity: Partial remission on treatment; Last eye exam date: NA; Last bone density test date: 2009-09-17; Osteoporosis treatment at visit if valid: NA; Last quantiferon date: NA; Inh prophylaxis date: NA; Last hepatitis serology date: NA; Last cbc date: 2009-12-03; Last lft date: NA; Rituximab ever: No; Last rituximab date: NA; Prednisone use at the moment: 32 mg/day; Is there any safety & adverse events: Yes; Which treatment received and their dates: Prednol 2006-2007 and 2009-07 to present; Is there any side effect for any treatment: Osteopenia (T score -2)

AI REPORT: Pemphigus type: Pemphigus herpetiformis; Current disease activity: Remission (no new lesions observed at last visit, stable clinical status with minimal erosions and widespread healed lesions); Last eye exam date: NA; Last bone density test date: 2009-09-17 (T-score: -2, Z-score: -1.9, osteopenia); Osteoporosis treatment at visit: NA; Last quantiferon date: NA; Inh prophylaxis date: NA; Last hepatitis serology date: NA; Last cbc date: NA; Last lft date: NA; Rituximab ever: No; Last rituximab date: NA; Prednisone use at the moment: Yes (32 mg/day); Is there any safety & adverse events: No reported safety events; Treatments received and their dates: Prednol 120 mg/day (started August 2006, discontinued one year prior, restarted 30.07.09, tapered to 32 mg/day by 12.03.09); Is there any side effect for any treatment: Steroid complications suspected including osteopenia (confirmed by bone density), no reported hyperglycemia, hypertension, cataract, glaucoma, myopathy, or avascular necrosis; Major surgery: None; Comorbid conditions: Anemia, hematuria; Warning: Duplicate entry detected in multiple visits on 2009-09-17, 2009-09-24, and 2009-12-03 with identical dates and similar text, suggesting possible copy-paste error; the clinical details remain consistent across entries, indicating likely repeated documentation of the same visit or follow-up.

Figure 4.6: Qualitative comparison between clinician-written ground-truth reports and AI-generated reports: Each example consists of two parts, with the clinician-written ground-truth report shown above and the AI-generated response shown below. The figure contains text-only excerpts annotated for qualitative analysis. Correct or clinically useful information is highlighted in green, explanatory content that does not align with the instruction is highlighted in yellow, and clear errors are highlighted in red. (1) Example of a generally good AI report: although disease activity is not detected correctly, the explanation is helpful, treatments and comorbidities are well captured, and the warning statements are useful for clinician review. (2) Example of a poor AI report: despite clear explanations and reasonable coverage of treatments and comorbidities, disease activity is not detected correctly, limiting overall report quality. Both AI-generated reports are longer than the ground-truth reports, as the AI model provides additional explanatory commentary. Clinicians prepared the ground-truth reports for automated metric calculation (BLEU and ROUGE), with primary comparisons conducted based on the original reports. This figure is included solely to support qualitative visualization and interpretation of model behavior.

sufficient evidence to classify the complication as present. These patient-independent clarifications were incorporated into the feature-specific notes without modifying the fixed main prompt. Once all feature-specific notes had been finalized, the complete prompt configuration was frozen, and no further prompt modifications were made during the ten repeated inference runs. These observations highlight the importance of explicit operational definitions when evaluating general-domain language models for structured clinical information extraction.

Crucially, this study was designed to mimic daily clinical practice as closely as possible. Clinician notes were provided to the model exactly as stored in the hospital information system (DermaBase), without preprocessing, cleaning, or human intervention. This evaluation allowed us to directly observe common challenges faced by LM-based clinical systems, most notably hallucination and misinformation generation. In contrast to many clinical information-extraction studies that evaluate LMs on pre-structured fields exported from hospital systems, our evaluation used raw longitudinal narrative documentation, which better reflects real-world clinical variability and incomplete recording practices. For instance, the model occasionally confused treatments, particularly regarding systemic corticosteroid (Prednol) use, likely due to duplicated or inconsistently documented entries across visits. The model also made incorrect causal assumptions, such as interpreting hepatosteatosis as evidence of drug-induced liver toxicity and labeling test dates as not available (NA) when they occurred in the distant past (Figure 4.6). While many of these issues could be addressed through more detailed prompts or standardized documentation, their occurrence highlights real-world risks when deploying LMs in unstructured clinical environments [206, 199, 207].

Feedback collected during the Likert-scale evaluation provided additional insight into the perceived quality of AI-generated reports. Most clinician comments focused on report structure, formatting, or the omission of preferred data elements, such as specific laboratory values or treatment timelines. Notably, these concerns were largely related to presentation rather than factual accuracy and could likely be addressed through relatively simple prompt engineering or template-based constraints.

From a technical standpoint, we deliberately evaluated model performance using a standard and accessible hardware configuration. All experiments were conducted on a computer equipped with a single NVIDIA RTX 4090 GPU, purchased for approximately USD 2,000, representing a setup that is feasible for many academic institutions and potentially for clinical research environments. Although the Qwen3 model supports a maximum context length of approximately 262,000 tokens, we limited the context length to 90,000 tokens to ensure compatibility with available GPU memory and system RAM.

In our dataset, the average number of visits per patient was 18.03, and the longest individual report contained approximately 7,975 words, requiring an estimated minimum context length of approximately 15,000 tokens to preserve full information content. For substantially longer longitudinal histories (e.g., more than 100 visits), larger context windows may be required to accommodate the full record; however, such cases were not directly evaluated in this study. While recent LM compression and engineering techniques have demonstrated promising reductions in VRAM requirements with minimal performance degradation [208], this study intentionally focused on uncompressed models to establish baseline hardware and performance

requirements. In this regard, our findings provide practical empirical guidance for researchers working in clinical natural language processing.

Our limitations should be acknowledged. First, pemphigus is a rare disease, and larger multicenter datasets will be required to validate these findings and improve the precision of performance estimates. Second, this was a single-center study, which may limit generalizability to institutions with different documentation styles or clinical workflows.

In conclusion, local LMs represent a promising, privacy-preserving approach for longitudinal data retrieval and report generation in dermatology. Future work should focus on prospective evaluation of clinician workload reduction, expansion to other chronic diseases, and integration into real-world clinical workflows. For future clinical decision-making studies, report-generation quality may be further improved by providing richer, better-grounded inputs (e.g., integrating PDF documents and structured summaries) and by integrating local models with retrieval-augmented generation (RAG) pipelines to reduce missing context and improve factuality. If local models can achieve consistently reliable longitudinal information synthesis across diverse settings, they may serve as a foundational component of privacy-preserving clinical decision support systems in future studies.

4.5 DermatoLlama: Resource-Efficient Dermatology Vision-Language Modeling

Vision-language models (VLMs) have enabled interpretable reasoning over multimodal data. However, privacy concerns in medicine require locally deployable systems, but the size of current VLMs restricts their use. Here, we present SCALEMED (Scalable Clinical Assistants and LEarning for MEDicine) to develop locally deployable high-performance small VLMs using synthetic data and the LoRA approach. We first applied SCALEMED to dermatology, developing DermatoLlama by fine-tuning Llama 3.2 on approximately 1.2 million synthetic samples, including 204,809 image–text question–answer pairs derived from 83,622 images, as well as text-only and multi-turn samples. We then demonstrated SCALEMED's scalability by fine-tuning the general medical model MedGemma 4B on the training split of 1.728 million image–text question–answer pairs generated from approximately 864,000 images. For dermoscopic image evaluation, DermatoLlama accuracy increased from 0.42 to 0.87, while text-generation performance improved substantially (BLEU-4: 0.06 to 0.66; ROUGE-L: 0.52 to 0.93). For clinical skin image evaluation, accuracy increased from 0.81 to 0.91, with corresponding improvements in BLEU-4 from 0.17 to 0.77 and ROUGE-L from 0.25 to 0.79. The increase in Tokenized F1-score from 0.33 to 0.47 in the analysis of general medical images further provided initial proof of concept for cross-domain adaptation. These results highlight SCALEMED's potential for creating efficient and explainable medical VLMs, while demonstrating improved performance on the evaluated dermatology benchmarks.

4.5.1 Introduction

Large language models are increasingly recognized as transformative tools in healthcare, enabling explainable analysis, structured diagnoses, and personalized treatment plans [209, 210, 211, 212]. They have also demonstrated promise in providing multimodal solutions, such as interpreting radiological images, augmenting clinical decision-making, and streamlining workflows [213, 214, 215]. Despite these advancements, widespread adoption of LLMs in medicine still faces significant barriers. These include the limited availability of high-quality, diverse datasets [87], stringent privacy regulations [216, 217], and high computational costs [93, 94] required for training and deployment of LLMs in clinical settings. Such constraints pose significant challenges for the adaptation of LLMs to real-world and resource-limited environments. Therefore, developing resource-efficient and privacy-preserving solutions allowing scalability to multiple fields is essential for effectively integrating LLMs into healthcare practice.

Building task-focused artificial intelligence assistant systems ("agents") and generating synthetic data, with or without ground truth, have emerged as powerful strategies for overcoming the high computational and extensive data demands of LLMs in healthcare [218, 219, 220]. By transferring knowledge from a large teacher model to a lighter, more efficient student model, "knowledge transfer" reduces the resources needed for training and inference. Collaborative AI assistants ("multi-agent systems") distribute complex tasks among specialized assistants that communicate and cooperate, enabling medical AI to process diverse

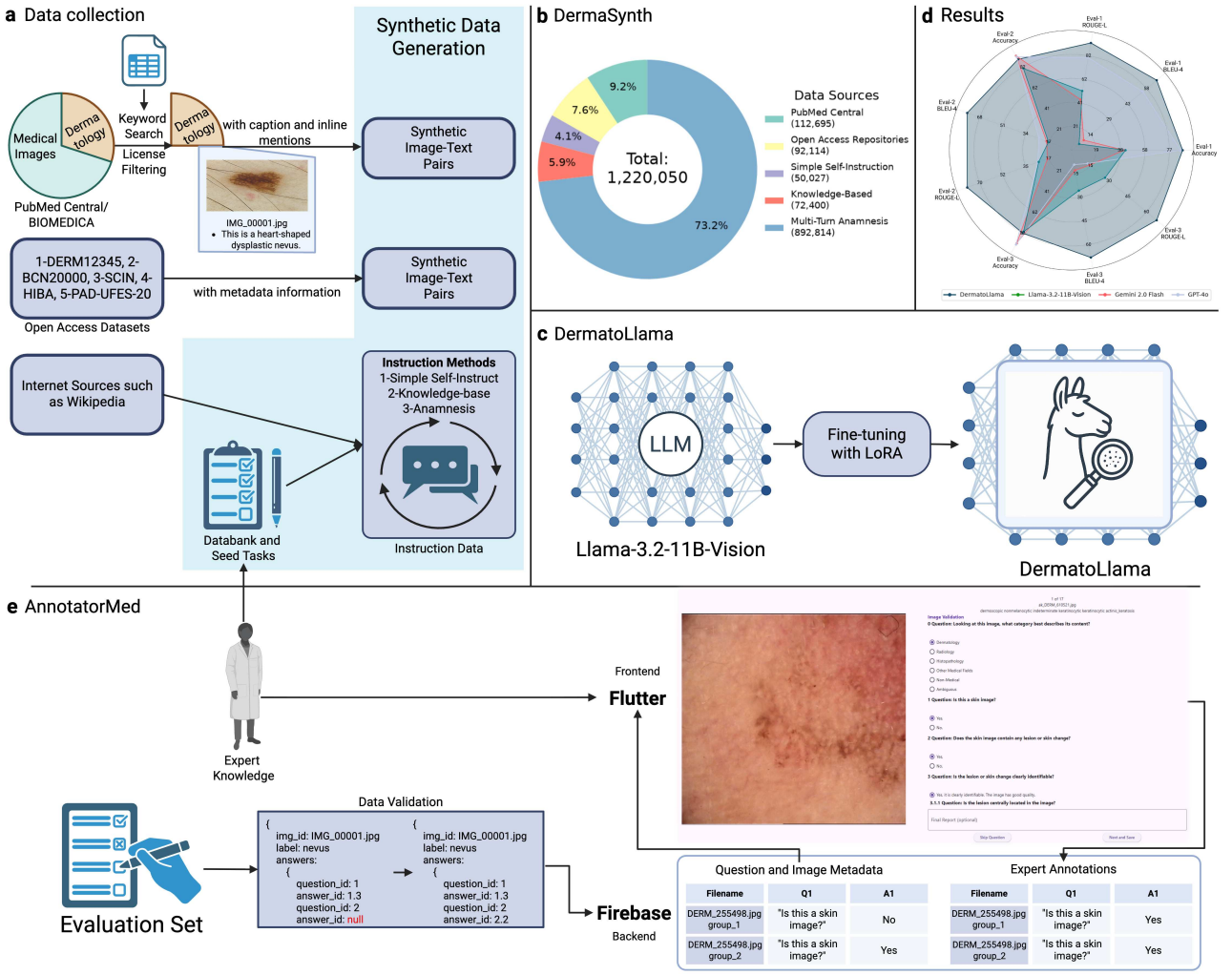


Figure 4.7: Overview of the SCALEMED framework. a, Data collection pipeline of DermaSynth, comprising 1.2 million data samples derived from PubMed, open-access datasets, and instruction-following methods. b, DermaSynth dataset size and data sources. c, DermatoLlama was fine-tuned from Llama-3.2-11B-Vision using the LoRA optimization method. d, DermatoLlama achieves accuracy comparable to GPT-4o and Gemini 2.0 Flash and surpasses all evaluated models in text generation across all evaluation sets. e, AnnotatorMed is a free and open-source tool for clinical data annotation.

inputs (e.g., images and text) while maintaining flexibility and scalability [221, 222, 223, 224]. Despite these promising methodologies, current evaluation practices and data-handling approaches rarely harness LLMs to their full potential. The field lacks a unified framework that incorporates knowledge transfer, multi-agent system designs, and synthetic data generation to address these real-world constraints [225].

Moreover, privacy concerns and data scarcity remain pressing issues. Many organizations are not allowed to use AI models hosted on external servers due to constraints on sharing confidential medical data. In addition, researchers require access to high-quality processed data to build sufficiently representative datasets that capture the full spectrum of patient variation. A solution addressing these issues could accelerate the deployment of LLM-driven healthcare tools, enabling robust performance across medical specialties while safeguarding patient confidentiality and optimizing resources for institutional usage.

To address these challenges in developing robust, privacy-preserving, and scalable AI systems for healthcare, we propose a data-centric framework, SCALEMED (Scalable Clinical Assistants and LEarning for MEDicine) (Figure 4.7 [226]). Unlike traditional approaches that rely on tightly controlled datasets and static evaluation formats, SCALEMED integrates synthetic data generation (Figure 4.7), leveraging open-access (OA) databases and PubMed (PM) literature to create a diverse, representative dataset of medical information, with a specific use case in dermatology. This dataset is enriched through flexible instruction methods, ranging from AI-generated clinical questions derived from high-quality references (self-instruct) to medical knowledge-based prompts and anamnesis-style narratives. This modular set of instruction-generation methods allows additional strategies to be integrated and exposes vision-language models (VLMs) to a wide range of clinical scenarios. SCALEMED also highlights critical licensing considerations, guiding researchers toward safe and compliant use of publicly available data. SCALEMED allows users to generate synthetic evaluation data by using state-of-the-art LLMs, which can be edited using our open-source annotation tool, AnnotatorMed. To enhance computational efficiency, SCALEMED can be integrated with techniques such as Low-Rank Adaptation (LoRA) and Quantized LoRA (QLoRA) during model fine-tuning. These techniques enable efficient adaptation of OA models, such as Llama and its vision variants, reducing the computational load without sacrificing model performance. SCALEMED further refines these models through knowledge transfer, while protecting patient privacy by enabling local data annotation and model development on low-cost computers. An iterative, modular design ensures the framework is adaptable, explainable, and scalable, facilitating multi-agent systems in which specialized agents can collaborate.

We first demonstrated the application of the SCALEMED framework in dermatology, a domain currently lacking sufficiently diverse, high-quality datasets essential for advanced clinical reasoning. We further extended SCALEMED beyond dermatology by introducing the clinical VQA corpus PMC-Clinical-VQA (864,000 clinical images; 1.728 million instruction-following (IF) QA pairs) and fine-tuning MedGemma 4B with LoRA. Using SCALEMED, we created DermaSynth (for dermatology) and PMC-Clinical-VQA (for general medicine), synthetic and LLM-assisted resources, comprising approximately 1.2 million and 1.728 million text/QA samples, respectively (Figure 4.7). PMC-Clinical-VQA is derived from PubMed Central figures and captions, and DermaSynth aggregates PubMed skin images and open-access skin images. We enriched both

resources via instruction-following procedures in which an LLM expands expert "seed tasks" into scalable supervision. This data-centric process supports scalable approximation of dermatology-oriented reasoning tasks and enabled the development of our VLM, DermatoLlama, for comprehensive text generation. A code-driven pipeline produces synthetic QA pairs, including image-grounded questions, for systematic assessment without exposing sensitive patient data. AnnotatorMed, our open-source annotation tool, preserves privacy by supporting local development on standard hardware, while its modular, iterative design enables cost-efficient deployments.

4.5.2 Methods

4.5.2.1 Data

SCALEMED focuses on data gathering, task structuring, and synthetic dataset creation. The foundation of our dataset is the BIOMEDICA dataset [147], a large-scale collection of over 24 million image-caption pairs from over 6 million articles on PubMed Central. To curate a dataset specifically for clinical applications, we implemented a filtering strategy on the BIOMEDICA corpus (Appendix C.4). We filtered the BIOMEDICA corpus, yielding a focused subset of 864,000 images, each accompanied by its original caption and inline mentions from the source article. For the dermatology dataset (DermaSynth, Section 3.3), we assembled a comprehensive dataset consisting of 112,695 image–text pairs from PubMed, 92,114 image–text pairs from open-access repositories providing limited metadata, and synthetic instruction-following data. The synthetic instruction-following data comprised 50,027 self-instruct prompts derived from 190 seed tasks (Appendix E.2.1), 72,400 knowledge-based tasks covering approximately 2,700 skin-condition entries generated from 40 seed tasks (Appendix E.2.2), and 9,963 multi-turn anamnesis interactions created from 137 disease-specific seed tasks based on 104 fixed questions (Appendices E.2.3, E.2.4, and E.2.5). This integrated approach streamlines the workflow, enabling researchers to progress seamlessly from raw data sources to a carefully curated synthetic dataset, which serves as the foundation for subsequent model training and evaluation. Details of the data generation process are shown in Figure 4.8 [227]. No human participants or personally identifiable information were involved in this study. The source images and associated textual information were obtained from publicly available open-access repositories and literature, while prompts, seed tasks, and derived synthetic data were generated as part of this study. Accordingly, the use of publicly available source materials did not require additional ethical approval. The dataset specifications are detailed in Table 3.3.

4.5.2.2 Evaluation Data

PMC-Clinical-VQA The PMC-Clinical-VQA dataset comprises approximately 1.728 million QA pairs, derived from 864,000 images with two QA pairs per image. The dataset was divided into training, validation, and test sets, accounting for approximately 98%, 1%, and 1% of the data, respectively. We evaluated both the base MedGemma 4B model and our final fine-tuned MedGemma 4B checkpoint on the held-out test set.

Dermatology Benchmark We first used the DermaBench-DERM12345 component for dermoscopic evaluation and then created two distinct evaluation benchmarks to comprehensively assess model performance.

1. Eval1 (DermaBench-DERM12345): The DermaBench-DERM12345 dataset used in this study for benchmarking AI models in dermoscopic image analysis is described in Section 5.2.

2. Eval2 (Clinical Image Multiple-Choice Question Answering (MCQA)): This benchmark comprises multiple-choice questions generated using an evaluation-style prompt (Appendix C.3) applied to the test split of the BIOMEDICA dataset. These questions assess comprehension and reasoning based on clinical images and their associated textual context, such as captions and inline mentions.

3. Eval3 (LLM-Generated Text-Only MCQA): Generated through a "Multiple-choice / balanced benchmark" pipeline, this benchmark consists of text-only multiple-choice questions covering various dermatology aspects, including diagnosis, treatment, and histopathology. Questions are balanced across difficulty levels (beginner, intermediate, expert), designed to assess the model's textual dermatologic knowledge base comprehensively.

4.5.2.3 Model Training and Evaluation

In this section, we discuss the model training and evaluation strategies used to ensure robust performance across various use cases, supporting flexibility for both vision-based and text-based tasks. Depending on specific application requirements and the medical domain, either vision-based or text-based LLMs can be employed. Multimodal models, which integrate vision, text, and other data types, offer further versatility for tasks requiring combined analysis of image and textual data. Text-based LLMs are particularly effective for structured data tasks, including medical record analysis, diagnostic reasoning, and clinical decision support. VLMs excel in image understanding and visual diagnostic applications. Choosing between these model types depends on data characteristics and specific medical domain needs, ensuring adaptability within diverse healthcare contexts. Different prompts were used to evaluate classification and text-generation performance. Additionally, single- and multi-select questions were used to calculate classification metrics, while all question types were used to calculate text-generation metrics.

Teacher-Student Paradigm and Model Training To balance performance and computational efficiency, we adopted a teacher-student approach. Initially, we used Gemini 2.0 Flash as a teacher model to generate synthetic supervision for two smaller multimodal student models, MedGemma 4B-it (developed by Google [205]) and Llama-3.2-11B-Vision in 16-bit precision (developed by Meta [228]), which were subsequently adapted using LoRA-based fine-tuning [93]. This approach reduces hardware requirements while retaining high-level reasoning capabilities. Fine-tuned MedGemma and DermatoLlama models were trained and evaluated using Python (v3.10.13), LlamaFactory (v0.9.2) [229], and DeepSpeed (v0.16.3). Training was performed on a low-cost workstation equipped with four NVIDIA RTX 4090 GPUs (96 GB combined VRAM) and 512 GB system RAM. For inference and benchmarking, we used a computer with a single RTX 4090 GPU (24 GB VRAM) and 96 GB RAM, confirming practical deployment feasibility on a single-GPU setup.

Low-Rank Adaptation (LoRA) Fine-Tuning Pretraining or full fine-tuning of LLMs involves updating all weights, a memory-intensive and computationally demanding process. In contrast, LoRA fine-tuning [93] updates only a small set of low-rank matrices.

For a pretrained linear layer with weight matrix $W_0 \in \mathbb{R}^{d \times k}$, where k and d denote the numbers of input and output features, respectively, and $r \ll \min(d,k)$ is the low-rank adaptation rank, LoRA learns a low-rank update:

$$W = W_0 + BA \tag{4.1}$$

where $A \in \mathbb{R}^{r \times k}$ and $B \in \mathbb{R}^{d \times r}$.

We employed LoRA to efficiently fine-tune the MedGemma 4B model on our dataset. The fine-tuning was configured with a LoRA rank (r) of 16 and an alpha (α) of 16. We used the `adamw_torch_fused` optimizer with a cosine learning rate scheduler and a learning rate of 4×10^{-4}. The model was trained with a per-device batch size of 24 and gradient accumulation steps of 4. We applied LoRA to DermatoLlama with a rank of 16, dropout rate of 0.05, and input/output cutoff of 2,048 for targeted modules (Appendix C.2).

Model Auditing To verify model improvements for MedGemma, the training was conducted for 6,500 steps, with checkpoints saved every 500 steps for intermediate evaluation. To provide a comprehensive assessment of the models' generative capabilities, we used a set of text-generation metrics (Tokenized F1, ROUGE-L, and BERTScore). For dermatology, we incrementally introduced samples from our synthetic dataset, DermaSynth, at scales of 10k, 50k, 100k, 200k, and the full dataset, assessing performance at each increment. The test split of DERM12345 was excluded from training. Model performance at each stage was evaluated using the evaluation datasets (Eval1, Eval2, Eval3), measuring accuracy, BLEU-4, ROUGE-1, and ROUGE-L scores to assess DermatoLlama variants and SOTA models.

4.5.2.4 License Issues

All synthetic data generated from PubMed articles and open-access datasets used in this study allow for commercial usage. However, synthetic data were generated using the OpenAI and Gemini APIs, each subject to its own usage restrictions. The VLM model, DermatoLlama, was fine-tuned using the Llama-3.2-11B-Vision model. DermaSynth and DermatoLlama are released under the CC BY-NC 4.0 license, prohibiting direct commercial use.

Researchers must independently verify compliance with all upstream licenses and implement appropriate safety measures before adopting these models in real-world patient-care scenarios.

4.5.3 Results

4.5.3.1 SCALEMED Framework Implementation

We implemented the SCALEMED framework (Figure 4.7) to generate a large-scale, multifaceted synthetic dataset for dermatology, DermaSynth, using three data-generation streams based on open-access image datasets (Figure 4.8), PubMed articles (Figure 4.8), and instruction-following data (Figure 4.8). DermaSynth comprises approximately 1.2 million synthetic samples (Figure 4.7), integrating visual question-answer (VQA) pairs derived from metadata-rich OA datasets on two common dermatologic image modalities (26,992 QA pairs derived from 13,496 clinical images and 65,122 QA pairs of 32,561 dermoscopic images) and the clinically relevant BIOMEDICA dataset (112,695 QA pairs derived from 37,565 images, no modality information) [147] based on PM articles. To construct the synthetic IF dataset, we initially designed seed tasks covering diverse clinical scenarios. IF synthetic data were then generated using a self-instruction framework comprising three categories: (1) simple IF QA (190 seed tasks (Appendix E.2.1), 50,027 synthetic samples), (2) knowledge-based QA (40 seed tasks (Appendix E.2.2), 72,400 synthetic samples), and (3) anamnesis-style case studies (137 seed tasks (Appendices E.2.3, E.2.4, and E.2.5), 892,814 text samples from 9,963 synthetic cases). Additionally, we created three evaluation sets to assess VLMs for dermatology: dermoscopic image analysis in multi-turn conversations (Eval1, 21,597 QA pairs) (Figure 4.7), clinical image analysis based on PubMed images (Eval2, 4,264 QA pairs), and general dermatology knowledge (Eval3, 1,998 QA pairs). This data-centric approach established comprehensive baseline datasets for training and evaluating a specialized dermatology VLM, DermatoLlama (Appendix E.1.1). To highlight the domain-agnostic applicability of the SCALEMED framework beyond dermatology, we also curated the PMC-Clinical-VQA dataset (864,000 clinical images; 1,728,000 instruction-following QA pairs) for general medicine using our LLM-assisted pipeline.

4.5.3.2 Model Performance on Benchmark Evaluations

We evaluated two fine-tuned models: MedGemma and DermatoLlama. MedGemma 4B (fine-tuned with LoRA on the PMC-Clinical-VQA corpus) was assessed on the test split of that corpus to provide cross-domain context within our evaluation pipeline. DermatoLlama (Llama-3.2-11B-Vision fine-tuned on DermaSynth) was evaluated against the base Llama-3.2-11B-Vision and SOTA models (GPT-4o, knowledge cutoff: October 2023; Gemini 2.0 Flash, knowledge cutoff: August 2024). The comparison was conducted across three distinct benchmarks (Eval1, Eval2, Eval3), each targeting different aspects of dermatologic reasoning and knowledge. Additionally, we performed targeted evaluations to assess the quality and effectiveness of each distinct synthetic data generation pipeline within DermaSynth (Figure 4.8). For this analysis, we trained separate model instances by progressively adding more data: 10k (OA), 50k (OA), 100k (OA), 200k (OA+PM), all training samples (OA+PM+IF). We then evaluated each trained instance on evaluation benchmarks (Eval1, Eval2, Eval3). This targeted approach validated the contribution of each data generation strategy. Detailed results comparing the DermatoLlama models trained on various dataset sizes with the SOTA models are presented in Table 4.7 and Figure 4.9 [230].

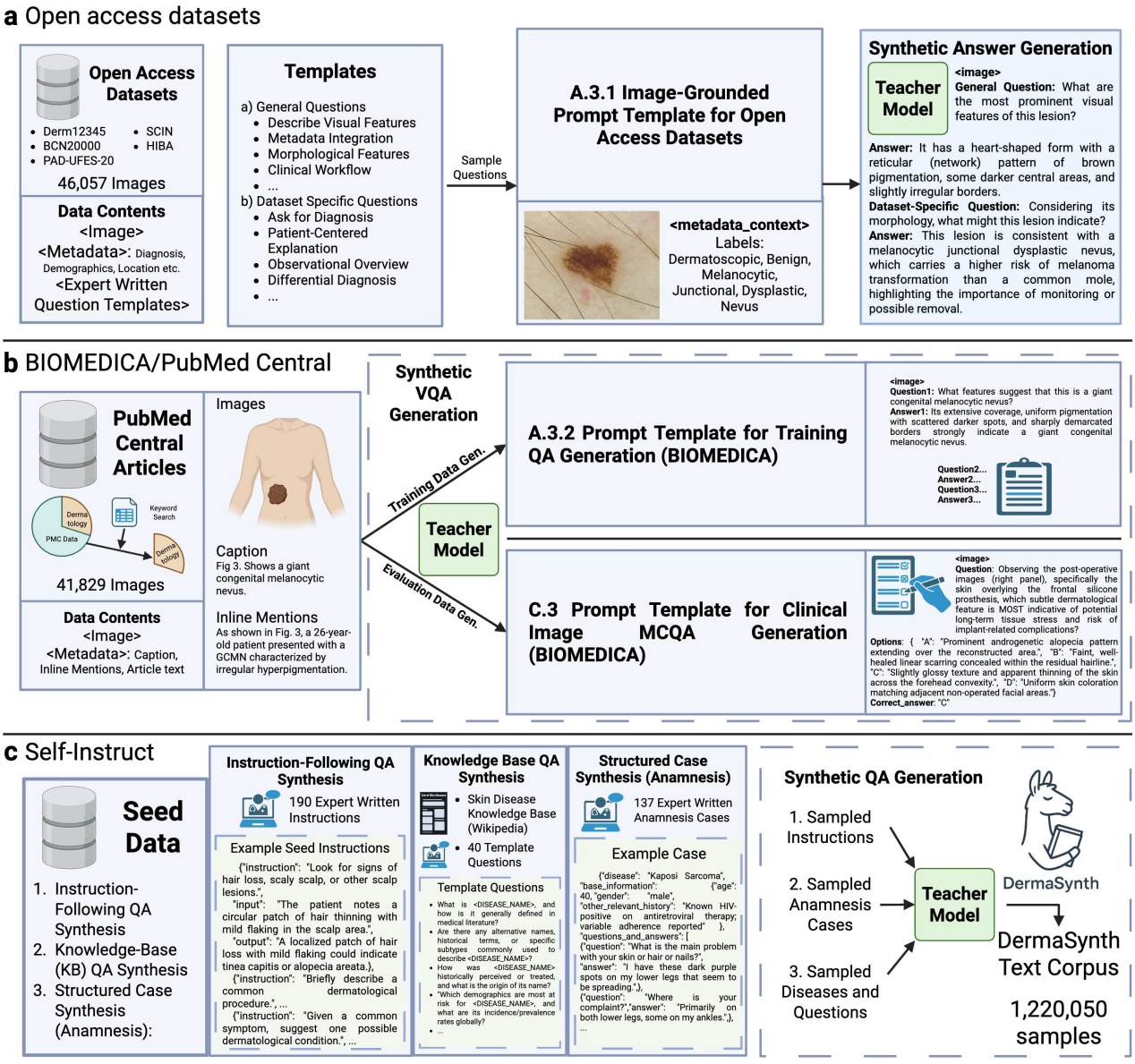


Figure 4.8: Overview of synthetic data generation workflow. a, 46,057 open-access dermatology images with metadata were used to generate metadata-informed, image-focused QA pairs. b, 41,829 open-access PubMed images with captions and inline mentions, comprising 37,565 images used to generate descriptive QA pairs for training and 4,264 images used to generate MCQAs for evaluation. c, 1,015,241 synthetic instruction-following samples were generated based on self-instruct, knowledge base, and anamnesis to cover clinical reasoning.

DermatoLlama Evaluation on Eval1: Multi-Turn Dermoscopic Image Analysis The DermatoLlama-Full model, trained on the full DermaSynth dataset excluding the DERM12345 test split, demonstrated improved performance compared with the base Llama model. The excluded test split was reserved for constructing the DERM12345 component of the DermaBench benchmark, referred to as Evaluation 1 (Eval1) in this chapter. DermatoLlama achieved an accuracy (Acc) of 0.87, along with text-generation scores of BLEU-4 = 0.66, ROUGE-1 = 0.93, and ROUGE-L = 0.93. In contrast, the base Llama model had considerably lower scores (Acc = 0.42, BLEU-4 = 0.06, ROUGE-1 = 0.52, ROUGE-L = 0.52). DermatoLlama also outperformed GPT-4o (Acc = 0.86, BLEU-4 = 0.55, ROUGE-1 = 0.81, ROUGE-L = 0.81) with higher accuracy and better text-generation performance, and surpassed Gemini 2.0 Flash (Acc = 0.41, BLEU-4 = 0.09, ROUGE-1 = 0.44, ROUGE-L = 0.44) across all metrics. Example reports of ground truth, DermatoLlama, and the base Llama model are presented in Figure 4.10 [231] and Appendix C.5. Other DermatoLlama models, trained on different sample sets of DermaSynth (OA, OA+PM, OA+PM+IF), also consistently outperformed the base Llama model, confirming that converting clinical data into structured descriptions provides high-fidelity supervision. The additional data obtained from PubMed and IF approaches demonstrate complementary benefits from combining diverse data sources.

DermatoLlama Evaluation on Clinical Multiple-Choice VQA (Eval2) The Eval2 benchmark tested the models' ability to answer MCQAs based primarily on clinical images and associated text from the BIOMEDICA dataset (Table 4.7) [147]. DermatoLlama-Full achieved an accuracy of 0.91, with BLEU-4 = 0.77 and ROUGE-L = 0.79. This model surpassed the base Llama model (Acc = 0.81, BLEU-4 = 0.17, ROUGE-L = 0.25). Although GPT-4o (Acc = 0.93) and Gemini 2.0 Flash (Acc = 0.94) demonstrated slightly higher classification accuracy, DermatoLlama outperformed both in BLEU-4 and ROUGE-L scores, indicating improved text-generation performance. Additionally, the performance of DermatoLlama-100k-OA (Acc = 0.62, BLEU-4 = 0.61, ROUGE-L = 0.63) and DermatoLlama-200k-OA+PM (Acc = 0.89, BLEU-4 = 0.62, ROUGE-L = 0.65) confirms the value of synthetic QA derived from PM sources. These results highlight that broader data exposure, as utilized in the full model, leads to improvements in both accuracy and text scores. These findings demonstrate that synthetic data derived from PM sources improve the model's ability to interpret dermatologic figures in the clinical literature. In addition, increased data diversity further enhances both diagnostic accuracy and language quality. This diversity was supported by incorporating PubMed-derived clinical figures and text, which capture a broad range of presentations, contexts, and documentation styles represented in the medical literature.

DermatoLlama Evaluation on Text-Only Dermatology Knowledge MCQAs (Eval3) To assess the transfer of textual dermatologic knowledge without images, we used the Eval3 benchmark, comprising text-only MCQAs (Table 4.7). DermatoLlama-Full achieved an accuracy of 0.83, BLEU-4 = 0.68, and ROUGE-L = 0.69, surpassing the base Llama model (Acc = 0.81, BLEU-4 = 0.25, ROUGE-L = 0.27) in both accuracy and text-generation metrics. While GPT-4o (Acc = 0.94) and Gemini 2.0 Flash (Acc = 0.93) performed better in classification accuracy, their text-generation quality remained low (BLEU-4 $\leq$ 0.12; ROUGE-L $\leq$ 0.15), highlighting DermatoLlama's explanation capabilities beneficial for clinicians. The improvement

after instruction-following training suggests that the synthetic self-instruction data strengthened domain-specific textual knowledge, effectively embedding domain-specific textual understanding into the language component of the VLM.

MedGemma 4B Evaluation on PMC-Clinical-VQA We fine-tuned MedGemma 4B with LoRA on the PMC-Clinical-VQA training split and evaluated it on the held-out test split using Tokenized F1, ROUGE-L, and BERTScore. The fine-tuned model consistently outperformed the base model across all metrics (Tokenized F1: 0.3330 $\rightarrow$ 0.4675; ROUGE-L: 0.2125 $\rightarrow$ 0.3501; BERTScore: 0.7858 $\rightarrow$ 0.8639), reflecting robust gains in long-form clinical VQA under a lightweight adaptation regime.

4.5.3.3 Resources Released

We make the following assets publicly available:

- **DermaSynth** – a dataset of approximately 1.2 million synthetic samples, including 204,809 clinical and dermoscopic image–text pairs, available in both Hugging Face Datasets and JSON formats (Figure 4.7).
- **DermatoLlama 1.0** – an 11B-parameter vision-language model with a context length of 2,048 tokens, fine-tuned using LoRA adapters on DermaSynth, a synthetic dermatology dataset, to process clinical text and images (Figure 4.7).
- **PMC-Clinical-VQA** – a large-scale, medical domain-general clinical visual question answering dataset curated from PubMed Central, consisting of 864,000 clinical images paired with 1,728,000 instruction-following QA examples. This dataset was constructed using our LLM-assisted pipeline to capture the broad range of biomedical visual presentations.
- **Fine-tuned MedGemma 4B** – a 4B-parameter multimodal medical model further adapted on PMC-Clinical-VQA across diverse medical image–text scenarios.
- **AnnotatorMed** – an open-source, locally deployable, and user-friendly web and mobile application. AnnotatorMed is a Flutter application with a Firebase backend for hierarchical data annotation, clinician testing, and optional logging of clinician feedback for RLHF-style workflows [214, 232], an approach to improve AI performance with relatively small amounts of feedback (Figure 4.7).
- **SCALEMED Framework** – end-to-end utilities (data scraping, synthetic generation, fine-tuning, evaluation) and a Colab notebook illustrating how to adapt the pipeline to any medical specialty (Figure 4.7).

Code is hosted on github.com/DermaVLM. The DermatoLlama model (trained on the full DermaSynth dataset, excluding the test split of DERM12345, which was reserved to create the DERM12345 component of the DermaBench benchmark) and the fine-tuned MedGemma 4B along with their checkpoints, training, and evaluation datasets, are available on Hugging Face (huggingface.co/DermaVLM). Additionally, the DermatoLlama model is directly accessible via the Hugging Face Space named DermatoLlama. Use of the

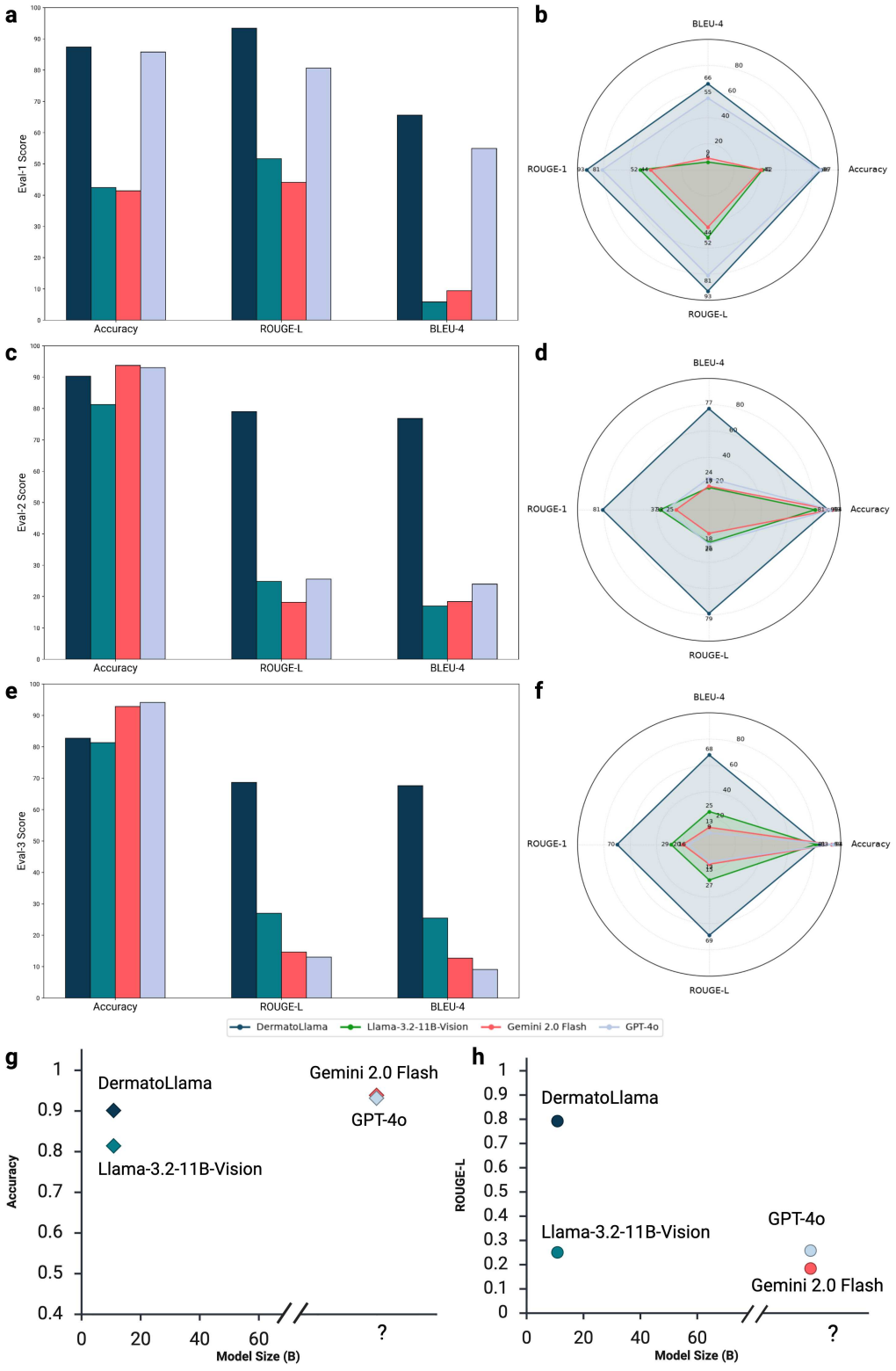


Figure 4.9: Comparison of results. (a, b) Eval1: classification and text-generation performance of DermatoLlama; (c, d) Eval2: classification and text-generation performance of DermatoLlama; (e, f) Eval3: classification and text-generation performance of DermatoLlama; (g, h) Eval2: scatter plots of accuracy and ROUGE-L versus parameter count illustrate the relationship between model size and performance, highlighting DermatoLlama's competitive performance despite its relatively small parameter count.

dataset, model, and code is restricted to non-commercial research, in accordance with the licensing terms documented in the repository.

4.5.4 Discussion

Deploying LLMs in resource-constrained organizations remains challenging due to limited data availability, stringent privacy regulations, and high costs of development and maintenance [233, 143, 234, 235, 236, 216]. We developed SCALEMED, a resource-efficient and privacy-preserving framework that addresses these constraints in medicine. SCALEMED is a modular, data-centric workflow for medical VLM development that integrates (i) synthetic data generation, (ii) an open-source, locally deployable data-curation panel (AnnotatorMed), and (iii) large-scale but affordable fine-tuning into a unified pipeline (Figure 4.7).

We demonstrated that: (i) synthetic data enriched with targeted clinical instructions substantially improves both classification and text-generation performance of Llama and MedGemma; (ii) performance generally improves as additional synthetic data sources, including open-access images, PubMed-derived contexts, and instruction-following methods, are introduced; (iii) the fine-tuned MedGemma 4B model shows substantial gains in image-grounded text-generation performance, as reflected by improvements in Tokenized F1, ROUGE-L, and BERTScore; and (iv) the full DermatoLlama variant demonstrated competitive overall performance against commercial, closed-source SOTA models (GPT-4o and Gemini 2.0 Flash), particularly on text-generation metrics. DermatoLlama surpasses the SOTA models on Eval1 (dermoscopic image analysis), which is based on DERM12345, a dataset published after the release of the SOTA models evaluated in this study. On Eval2 (clinical image analysis), containing images potentially present in the SOTA models' training datasets, DermatoLlama also achieves better text-generation metrics while attaining comparable accuracy. Furthermore, when assessed on questions generated by Gemini 2.0 Flash (Eval3), DermatoLlama still reaches an accuracy of 83%. Across all three evaluation sets, DermatoLlama achieved the highest scores for every text-generation metric, consistently outperforming GPT-4o, Gemini 2.0 Flash, and the base Llama model (Figure 4.9).

The SCALEMED workflow can reduce dependence on certain forms of manual multicenter data collection by enabling large-scale data generation within weeks. In our deployment, we assembled approximately 2 million image–text pairs and generated approximately 3 million samples using a limited amount of targeted expert input. Seed-design sessions and validation passes, together amounting to only tens of clinician-hours, were sufficient to bootstrap and quality-check the generators; the remaining work was automated and completed in 2 weeks. Attempting to assemble an equivalently broad, privacy-clean dataset from clinical sources would require sustained manual effort. Even at a conservative estimate of approximately 10 minutes per case, based on the average effort required to process one Eval1 image, curating approximately 3 million samples would require substantial human effort for consent checks, de-identification, extraction, and labeling, excluding the additional effort associated with Institutional Review Board (IRB) setup, data-use agreements, and multi-site coordination.

Table 4.7: Classification and text-generation performance of DermatoLlama with different variants and SOTA models, GPT-4o and Gemini 2.0 Flash. OA: Open-access images with QA pairs, PM: Open-access PubMed images with QA pairs, IF: Text-based instruction-following samples. The highest metric across all models per evaluation set is given in **bold**, and the second-highest value is underlined. The 10k, 50k, 100k, and 200k designations are nominal, rounded configuration labels rather than exact sample counts.

Data / Model	Acc	BLEU-4	ROUGE-1	ROUGE-L
Eval1 - 210 samples from test split of DERM12345				
Llama (base)	0.4244	0.0584	0.5201	0.5163
10k - OA	0.7635	0.3430	0.7986	0.7973
50k - OA	0.7874	0.3858	0.8197	0.8186
100k - OA	0.8146	0.3746	0.7768	0.7760
200k - OA + PM	0.8582	0.5304	0.9195	0.9187
Full - OA + PM + IF	**0.8742**	**0.6559**	**0.9349**	**0.9344**
Gemini 2.0 Flash	0.4140	0.0941	0.4441	0.4405
GPT-4o	0.8580	0.5495	0.8066	0.8065
Eval2 - PubMed Test Split				
Llama (base)	0.8102	0.1699	0.3693	0.2487
10k - OA	0.7085	0.2576	0.3365	0.3259
50k - OA	0.5229	0.5890	0.6458	0.6228
100k - OA	0.6203	0.6058	0.6572	0.6348
200k - OA + PM	0.8874	0.6226	0.6797	0.6542
Full - OA + PM + IF	0.9093	**0.7686**	**0.8056**	**0.7904**
Gemini 2.0 Flash	**0.9374**	0.1842	0.2482	0.1822
GPT-4o	0.9303	0.2399	0.3293	0.2561
Eval3 - LLM-generated questions				
Llama (base)	0.8133	0.2546	0.2893	0.2699
10k - OA	0.7387	0.6122	0.6290	0.6066
50k - OA	0.6597	0.6382	0.6642	0.6481
100k - OA	0.6491	0.6420	0.6689	0.6529
200k - OA + PM	0.7978	0.6616	0.6777	0.6715
Full - OA + PM + IF	0.8273	**0.6764**	**0.6973**	**0.6866**
Gemini 2.0 Flash	0.9284	0.1271	0.1962	0.1463
GPT-4o	**0.9414**	0.0910	0.1584	0.1304

Open-access dermatology images (OA) primarily enhanced the model's recognition accuracy in domain-specific tasks by expanding visual diversity. Accuracy improved from 0.424 to 0.787 and BLEU-4 from 0.058 to 0.386 (Eval1), indicating better visual-text alignment. However, OA alone led to performance decline under distribution shift, reducing accuracy to 0.523–0.708 for the 50k-OA and 10k-OA variants in Eval2 and 0.649–0.739 across the three OA-only variants (10k-OA, 50k-OA, and 100k-OA) in Eval3, suggesting over-specialization to image-centric cues. When PubMed dermatology papers (PM) were introduced, performance improved across all three evaluation sets, with accuracy increasing to 0.858 in Eval1, 0.887 in Eval2, and 0.798 in Eval3 for the 200k variant. Finally, instruction-following synthetic data (IF) generated from high-quality clinician seed tasks filled the reasoning and conversational gaps absent in OA and PM data. IF simulated contextual and multi-turn conversations, boosting text-generation quality from BLEU-4 0.530 to 0.656 (Eval1), 0.623 to 0.769 (Eval2), and 0.662 to 0.676 (Eval3). These improvements reflect how synthetic data introduced information about the "why" and "how" of clinicians' image interpretation, knowledge inaccessible from textbooks or public datasets, thereby enabling the model to reason across both images and narrative clinical logic. Among the DermatoLlama variants, the Full model provided the strongest combined classification and text-generation performance, following a three-source curriculum comprising OA data for visual alignment, PM data for semantic grounding, and IF data for structural and conversational diversity. This hierarchical design provided an initial cross-domain demonstration in a broader medical-image corpus. Fine-tuning MedGemma 4B on the PMC-Clinical-VQA training corpus, derived from 1.728 million PubMed-based synthetic QA pairs spanning multiple medical fields, substantially improved text-generation performance on the held-out evaluation set, as reflected by higher Tokenized F1, ROUGE-L, and BERTScore. Tokenized F1 improved from 0.3330 to 0.4675, ROUGE-L from 0.213 to 0.350, and BERTScore from 0.786 to 0.864, indicating improved token-level overlap and semantic similarity on the held-out PMC-Clinical-VQA evaluation set.

A critical insight from our dermatology use case is the importance of high-quality synthetic datasets. In settings where real-world data is scarce or cannot be easily shared due to privacy concerns, synthetic data can fill gaps by simulating a broad range of clinical presentations. Our experiments reinforced that the diversity and representativeness of synthetic samples are essential (Table 4.7). However, synthetic data should be complemented by clinically grounded resources, such as PM articles. Additionally, we found that seed tasks and databanks can be rapidly created with minimal yet strategic clinician input, streamlining the generation of high-quality synthetic data. This approach reduces the overall burden on domain experts while ensuring clinically relevant outputs.

Recent research in medical AI has focused on developing dermatology-focused foundation models [237, 212, 238] such as MONET [239], PanDerm [116], Derm1M [75], as well as pretrained models such as SkinGPT-4 [240]. These models primarily emphasize data annotation and short-context diagnostic tasks, falling short of addressing critical clinical needs such as comprehensive text generation. From a clinical perspective, there is increasing interest in the development and use of decision support systems that can handle longer contexts, including patient histories and detailed narratives [241, 149]. DermatoLlama, a fine-tuned 11-billion-parameter VLM, addresses this gap for a specific medical field (dermatology) by generating longer

sequences and detailed clinical reports while remaining practical to deploy. With a 2,048-token context window and training that favors concise structure, typical outputs remained at or below 150 words and were designed to provide concise differentials, next steps, and red-flag guidance without unnecessary verbosity. DermatoLlama runs efficiently on a single, low-cost NVIDIA RTX 4090 (approximately $2,000 initial setup). On a single RTX 4090, we observe approximately 15 tokens/second; assuming approximately 1.5 tokens per word, this corresponds to approximately 600 words/min, comfortably faster than typical silent reading of clinical prose (approximately 150–250 wpm). Thus, a 150-word report streams in approximately 15 seconds, which is faster than a clinician can typically read it end-to-end, suggesting that generation latency is unlikely to be the primary bottleneck. The combined cost of synthetic data generation, fine-tuning, and evaluation in this work was under $500, indicating that high-quality, domain-specific models can be developed and operated cost-effectively in academic and clinical settings.

To address the challenges of computational cost, privacy concerns, and scalability, we propose utilizing LoRA-based adapters to efficiently fine-tune LLMs for specialized medical models. Multiple condition-specific adapters fine-tuned with LoRA (e.g., for rare diseases) can be swapped in and out without duplicating the entire model. This model can also be deployed in cloud environments equipped with GPUs, such as Google Colab with NVIDIA A100 GPUs. A similar framework, LLaVA-Med [143], used eight A100 GPUs for training and reported a low serving cost, in addition to workforce requirements. This cost reduction could be substantial when considering the cumulative computational and deployment costs of multi-agent systems required across multiple departments and hospitals. We integrated LoRA adapters to efficiently fine-tune our model for specialized medical functions while preserving high-quality report generation and complex instruction-following abilities. Our framework also supports privacy-preserving local refinement using proprietary data, allowing institutions full control over data access. By leveraging synthetic data, knowledge transfer, and parameter-efficient fine-tuning with LoRA, we demonstrated that models of resource-accessible size can achieve clinically relevant performance without requiring massive architectures or full-model fine-tuning. Beyond diagnostic and treatment tasks, this model's high-quality text generation capabilities enable it to handle additional tasks, such as simplifying medical explanations for laypersons and providing supportive responses to stressed patients. This capability of the model underscores its broader applicability beyond clinical use cases and has the potential to reduce clinicians' workload by improving the generation and interpretability of clinical documentation.

We demonstrate with AnnotatorMed (described in detail in Section 5.1) how a purpose-built, free, open-source, and locally deployable tool can remove one of the largest bottlenecks in medical-AI development: secure, clinician-driven data curation. AnnotatorMed can run on local infrastructure (e.g., hospital servers) behind a firewall, ensuring sensitive images and reports remain securely within institutional storage while still enabling collaborative labeling projects. AnnotatorMed implements a hierarchical question workflow designed around user interface and user experience (UI/UX) principles, in which users are presented only with the options and subquestions relevant to their selections, while unselected and unrelated branches remain hidden. This design accommodates cases in which a single clinical finding branches into multiple differential diagnostic pathways (Figs. 5.1 and 5.2). This hierarchical workflow is particularly effective for

our dermatology use case, reducing screen clutter and increasing annotation speed. Eval1 would require 320 fields per image across 210 images, $210 \times 320 = 67{,}200$; by contrast, AnnotatorMed's dynamically rendered, skip-logic questionnaire reduced the realized interactions to 21,597 prompts, an approximate 68% reduction (Figure 4.7). Beyond labeling, AnnotatorMed can be used to capture RLHF annotations [214] or serve as a clinician testing platform, allowing experts to rank, correct, or critique model outputs (Fig. 5.3). Furthermore, the dermatology-specific hierarchical workflow can be adapted to other specialties with some domain-specific adjustments. By unifying multimodal input, flexible question logic, and secure on-premises deployment, AnnotatorMed provides a reusable architecture for high-quality data generation in clinical AI workflows.

This study has several limitations that future advances in AI may help address. Although synthetic data can increase dataset size, it may fail to capture subtle features, particularly those associated with rare or advanced cases [242]. Additionally, despite the low resource cost of training and deployment, clinical experts are still required to validate and annotate large volumes of generated data, especially for diagnostic models. While our trained DermatoLlama model contains 11 billion parameters, making it less resource-intensive than trillion-parameter alternatives, it may still lack the advanced reasoning capabilities required for more complex clinical tasks such as long multi-turn conversations for follow-ups. However, this reasoning capacity is likely to improve due to ongoing improvements in optimization methods, knowledge transfer, and incremental training. Furthermore, we emphasize that extensive safeguards, such as external clinical validation, bias audits across diverse populations, adversarial stress testing, and robust privacy controls, have not yet been implemented. These measures are crucial for safe clinical deployment, which demands formal regulatory review and rigorous prospective evaluation. Consequently, the current model remains a research prototype and is not yet suitable for guiding individual patient care.

Integrating SCALEMED with methods such as RLHF [232] and retrieval-augmented generation (RAG, a technique enabling AI systems to retrieve reliable medical information) [41] allows the development of AI systems that can mitigate AI hallucinations and improve contextual awareness [243]. These methods can be used to develop resource-efficient models by extending their knowledge bases without increasing parameter counts. In addition, establishing domain-ready baseline models empowers individual institutions to tailor AI solutions to their specific patient populations. Combining SCALEMED with complementary frameworks such as MONET [239], along with its automated concept annotation capability, can improve interpretability and clinical success. The hierarchical approaches used in Eval1 can similarly enhance AI assistant performance in medical applications by guiding researchers toward more effective strategies. An example would be a "hierarchy-of-thought" method, which we hypothetically name in analogy to existing chain-of-thought methods used in LLMs [244]. Our framework can seamlessly adapt to evolving clinical needs and diverse medical domains by providing a strong basis for developing advanced AI systems. It supports the development of multi-agent systems that distribute specialized tasks, such as image interpretation, textual reasoning, and patient triage, across smaller, dedicated submodels [221].

In conclusion, we introduce SCALEMED, a locally deployable framework for adaptable medical multimodal

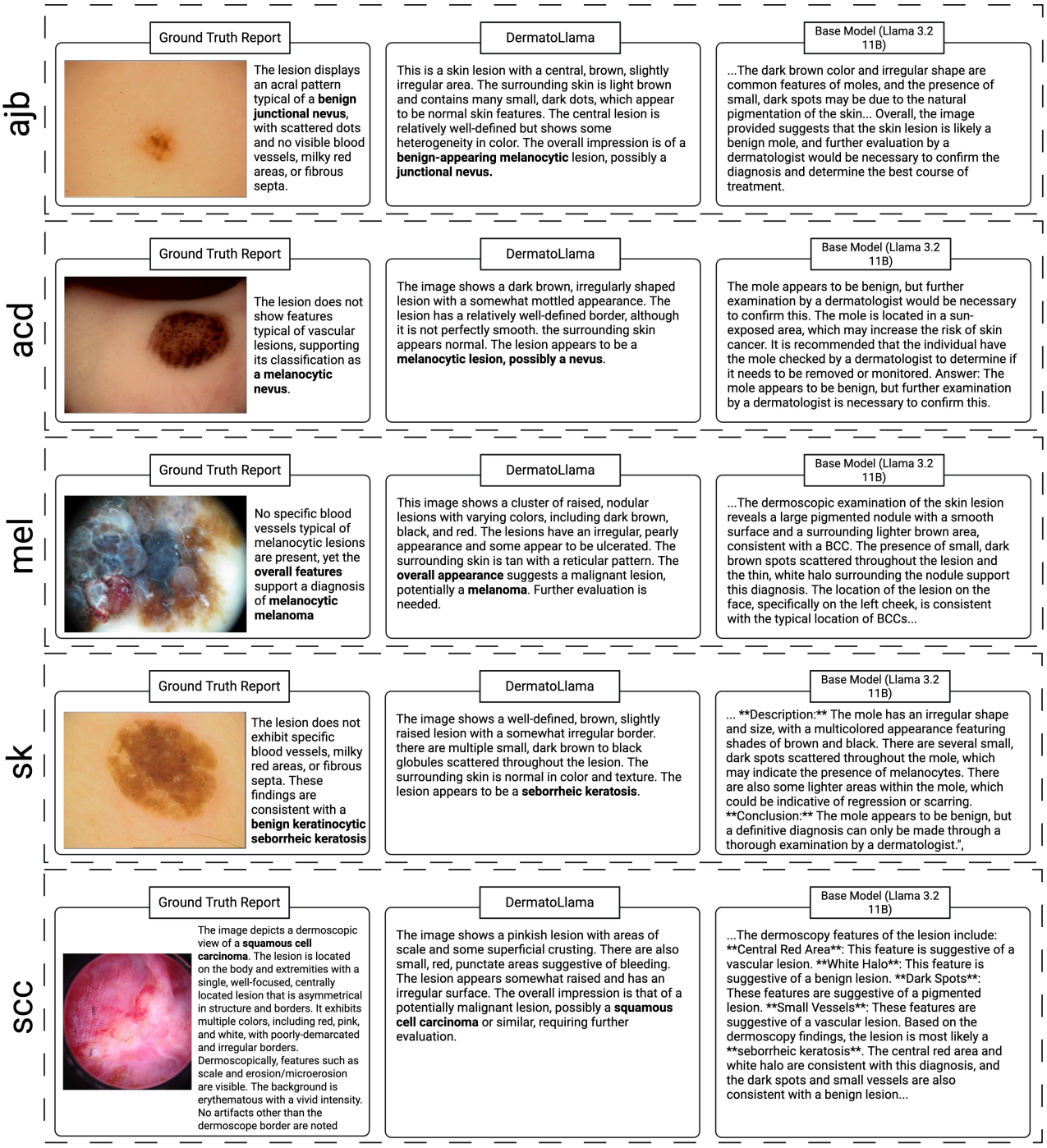


Figure 4.10: Example reporting results for dermoscopic cases: ajb (acral junctional banal nevus), acd (acral compound dysplastic nevus), mel (melanoma), sk (seborrheic keratosis), and scc (squamous cell carcinoma). For each case, the ground-truth report is shown on the left, the DermatoLlama output in the center, and the base Llama output on the right. DermatoLlama's descriptions align more closely with the reference reports than those generated by the base Llama model.

modeling. The system includes a domain-specialized VLM initially developed for dermatology (DermatoLlama), trained using synthetic QA pairs generated from dermatologic images (DermaSynth). We then demonstrate the generalizability and scalability of the framework by applying it to the general medical domain through fine-tuning MedGemma on the synthetic PMC-Clinical-VQA corpus, without requiring any architectural modifications. AnnotatorMed, a clinician-centric and domain-agnostic annotation tool, further supports expert-guided data curation across medical specialties. Together, these components establish SCALEMED as an efficient, lightweight platform, enabling privacy-preserving and cost-effective workflows in diverse clinical environments.

4.5.5 Data availability

The DermaSynth and PMC-Clinical-VQA datasets, along with the DermatoLlama and fine-tuned MedGemma models, are available through the DermaVLM organization on Hugging Face.

4.5.6 Code availability

Code is available through the DermaVLM GitHub organization.

Chapter 5

Benchmarks and Evaluations in Dermatology

This chapter begins with AnnotatorMed, a data annotation tool developed to support efficient clinical annotation workflows for dermatology AI (Section 5.1). It then introduces DermaBench, a clinician-annotated benchmark dataset for evaluating dermatology visual question answering and reasoning (Section 5.2). Finally, a hierarchical benchmarking framework based on the DERM12345 dataset is presented to systematically evaluate foundation models for fine-grained dermatology image classification (Section 5.3).

Parts of this chapter are reprinted and adapted from the following publications:

- (Section 5.2) **A. Yilmaz**, O. Erdem, E. Gokyayla, A. Acar, B. B. Dagtas, D. I. Erdil, G. Gencoglan, B. Temelkuran. "DermaBench: A Clinician-Annotated Benchmark Dataset for Dermatology Visual Question Answering and Reasoning," arXiv, https://doi.org/10.48550/arXiv.2601.14084
- (Section 5.3) F. Yuceyalcin*, **A. Yilmaz***, B. Temelkuran. "A Hierarchical Benchmark of Foundation Models for Dermatology," arXiv, https://doi.org/10.48550/arXiv.2601.12382

5.1 AnnotatorMed: Local Annotation Infrastructure for Benchmark Construction

5.1.1 Overview

Clinically meaningful benchmarking in dermatology AI requires more than collecting images and labels. Modern evaluation datasets, especially those designed for visual question answering, multimodal reasoning, and report generation, require structured annotations that capture image quality, modality, morphology, anatomical context, dermoscopic structures, diagnostic reasoning, and clinician confidence. These annotations are often hierarchical: one answer determines which additional questions should be asked. For example, a dermoscopic image may require a different set of follow-up questions than a clinical photograph, and a melanocytic lesion may require different dermoscopic criteria than a nonmelanocytic lesion. Standard flat annotation tools are therefore often inefficient for medical benchmarking workflows, as they may force annotators to navigate irrelevant fields, increase cognitive load, and reduce consistency.

In addition, medical data annotation has important privacy and governance constraints. Dermatology images and clinical records may contain sensitive or identifiable information, and many institutions cannot upload such data to externally hosted annotation platforms. For this reason, benchmarking infrastructure for medical AI should ideally support local or institution-controlled deployment, flexible database integration, and export formats compatible with downstream model evaluation. This need motivated the development of AnnotatorMed as a practical annotation interface for clinician-centered benchmark construction.

5.1.2 AnnotatorMed

AnnotatorMed was developed as a lightweight, locally deployable medical data annotation platform designed to support clinician-centered dataset construction and benchmarking workflows (Figure 5.1). The platform was implemented using Flutter (version 3.22.3), enabling deployment across both web and mobile interfaces. The system architecture allows local databases to be integrated for on-premises or institution-specific installations. This flexibility makes AnnotatorMed suitable for medical annotation settings in which data governance, institutional control, and privacy-preserving local deployment are essential.

A key feature of AnnotatorMed is its support for flexible hierarchical annotation. The interface is structured around a nested data model consisting of question, option, subquestion, and suboption levels, allowing complex clinical annotation logic to be represented within an intuitive workflow (Figure 5.2). This design is particularly relevant for dermatology and other medical domains, where one clinical observation may determine which additional questions are required. For example, selecting a specific lesion type, anatomical site, image modality, or image-quality feature can dynamically reveal only the relevant follow-up fields. By displaying only contextually necessary items, AnnotatorMed reduces visual clutter, unnecessary interactions, and annotator cognitive load.

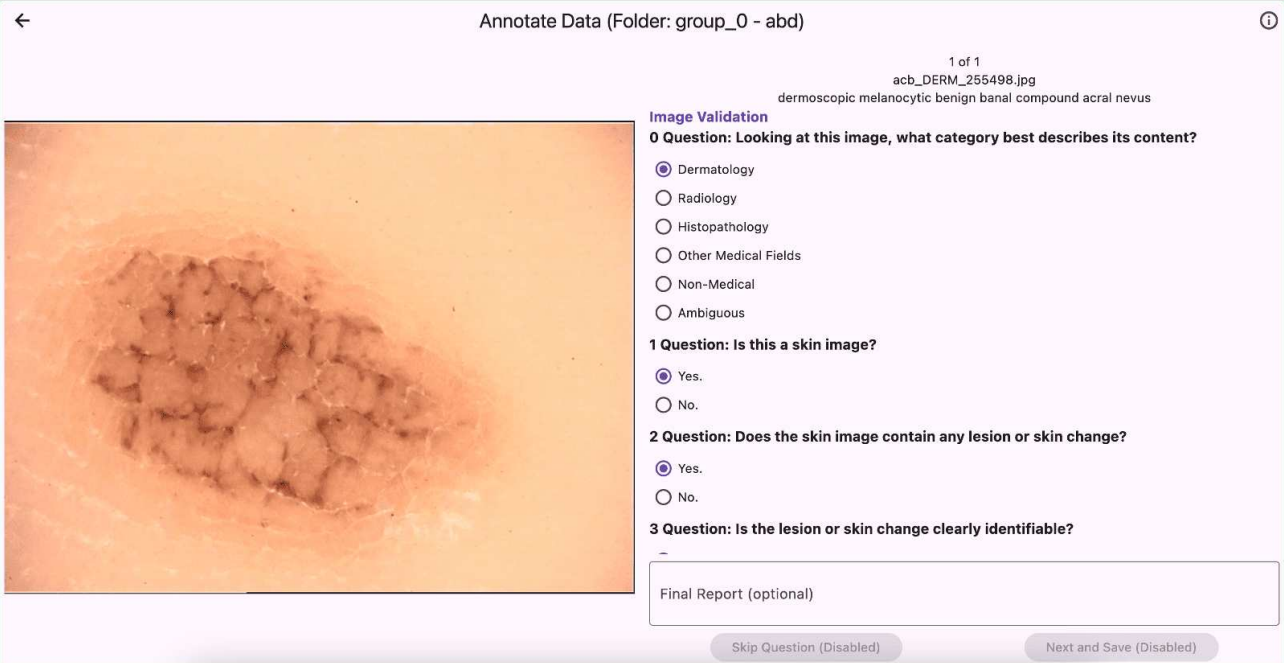


Figure 5.1: AnnotatorMed interface displaying hierarchical annotation questions on the right side, guiding systematic categorization of a dermatologic skin lesion image shown on the left. An optional text field in the bottom-right corner allows clinicians to add notes.

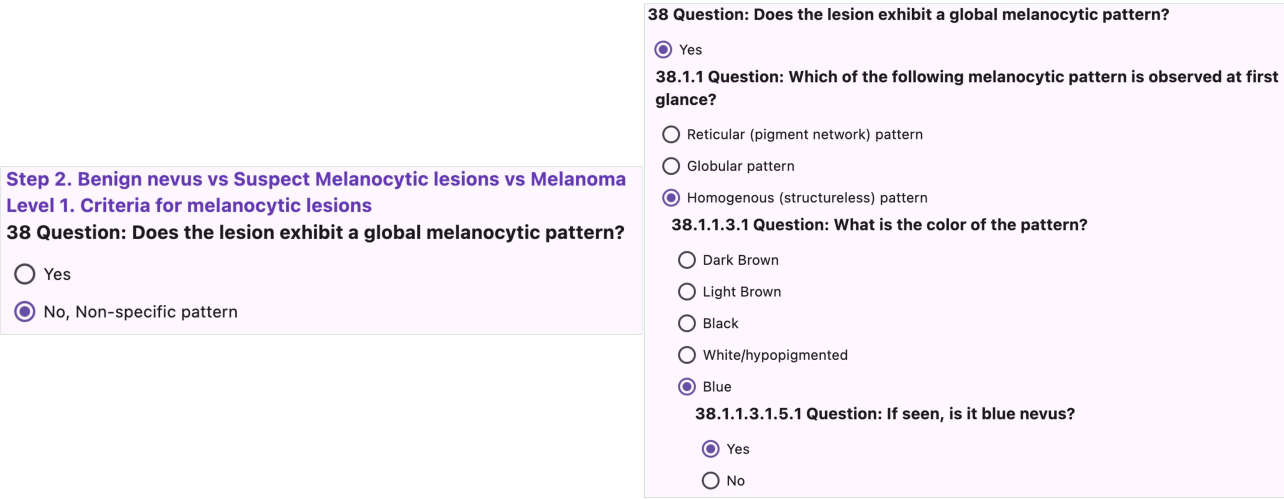


Figure 5.2: Dynamic hierarchical annotation interface in AnnotatorMed. Selecting “No” hides irrelevant follow-up questions, whereas selecting “Yes” reveals subsequent subquestions, streamlining the clinician workflow and reducing annotation time.

The platform supports single-select, multi-select, and open-ended question formats, allowing both structured categorical labels and descriptive clinical annotations to be collected within the same workflow. In addition to hierarchical annotation, the same component library can be used in a flat, non-hierarchical mode when simpler annotation tasks are required (Figure 5.3). This makes AnnotatorMed adaptable to a broad range of medical AI use cases, including dataset labeling, clinician testing, visual question-answering

benchmark construction, model output review, error analysis, and preference-based evaluation.

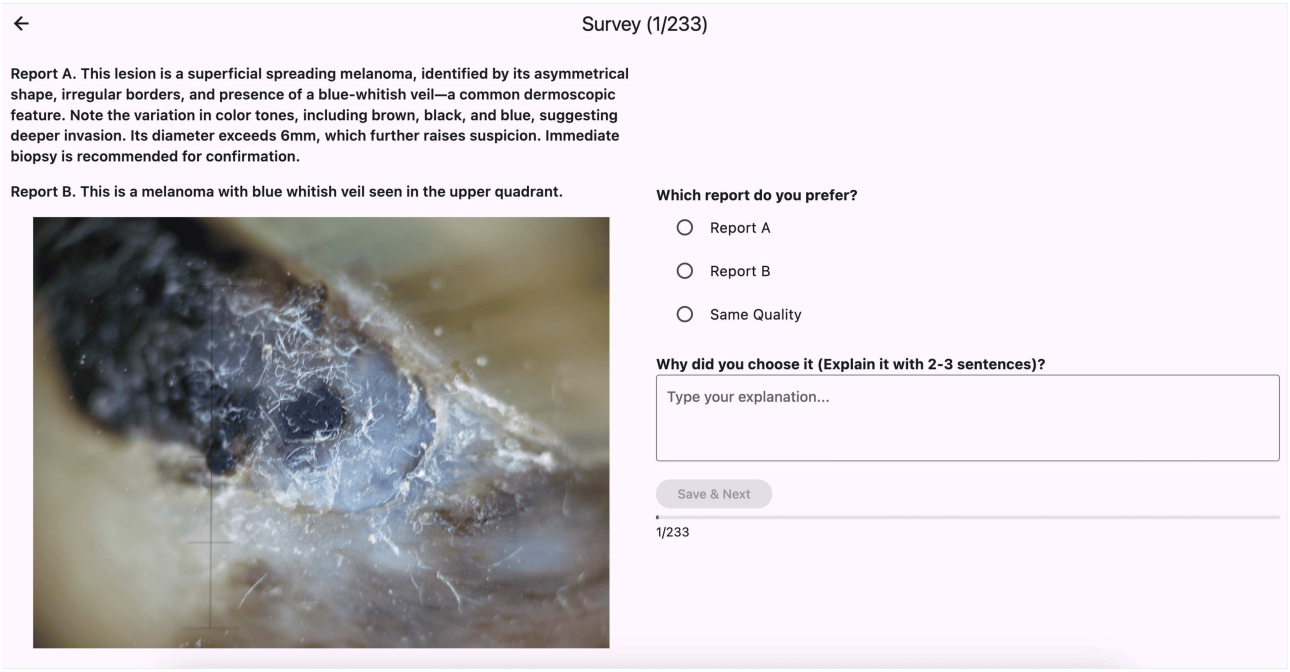


Figure 5.3: AnnotatorMed can be adapted to different evaluation needs, such as ranking model outputs or comparing alternative reports. This flexibility makes it suitable for clinician feedback collection, model output review, and reinforcement learning from human feedback (RLHF)-style workflows.

In this thesis, AnnotatorMed was used as part of the benchmarking and evaluation layer, particularly for clinician testing and DermaBench data annotation (Section 5.2). Its ability to support hierarchical clinical question structures made it suitable for generating dermatologist-authored visual question-answering annotations and for capturing complex dermatologic morphology in a scalable and reproducible manner. All annotation outputs can be exported in CSV or JSON format, facilitating integration with downstream model fine-tuning, evaluation, and benchmark analysis pipelines.

Overall, AnnotatorMed provides the annotation infrastructure required to connect clinical expertise with reproducible AI benchmarking. By supporting local deployment, hierarchical question logic, multiple response formats, and structured data export, it enables privacy-preserving and clinically flexible medical annotation workflows. This infrastructure forms the basis for the DermaBench benchmark described in the following section.

5.2 DermaBench: Clinician-Annotated Dermatology Visual Question Answering

Vision-language models (VLMs) are increasingly important in medical applications; however, their evaluation in dermatology remains limited by datasets that focus primarily on image-level classification tasks such as lesion recognition. While valuable for recognition, such datasets cannot assess the full visual understanding, language grounding, and clinical reasoning capabilities of multimodal models. Visual question answering (VQA) benchmarks are required to evaluate how models interpret dermatologic images, reason over fine-grained morphology, and generate clinically meaningful descriptions. We introduce DermaBench, a clinician-annotated dermatology VQA benchmark built on the DERM12345 and Diverse Dermatology Images datasets. DermaBench comprises 210 dermoscopic images from the DERM12345 dataset and a subset of 646 clinical images from the DDI dataset spanning Fitzpatrick skin types I–VI. Using a hierarchical annotation schema comprising 93 main questions for dermoscopic images, based on two-step pattern analysis, and 21 main questions for clinical images, expert dermatologists annotated each image for diagnosis, anatomic site, lesion morphology, distribution, surface features, color, and image quality, together with open-ended narrative descriptions and summaries, yielding 21,597 and 14,474 annotations for dermoscopic and clinical images, respectively. DermaBench is released as a metadata-only dataset to respect upstream licensing.

5.2.1 Background and Summary

Dermatology is one of the most visually driven specialties in medicine, relying heavily on pattern recognition across a wide spectrum of skin morphologies, colors, and anatomical sites. With the increasing adoption of teledermatology, mobile consultations, and electronic health records, vast quantities of dermatologic images are now captured daily in real-world settings. This surge of data offers an unprecedented opportunity for artificial intelligence to assist in diagnosis, triage, and education. However, despite these advances, the field lacks standardized, expert-annotated visual question answering (VQA) benchmark datasets that capture the true diversity of dermatologic conditions while maintaining clinically detailed labels and equitable representation of different skin tones.

Recent work in medical multimodal learning has produced a growing ecosystem of visual question answering (VQA) and vision-language datasets across medical fields such as radiology and pathology. Radiology datasets such as VQA-RAD comprise manually written questions but remain small (315 images, 3,515 QA pairs) and limited in scope, and their questions focus on modality or organ identification rather than clinical reasoning [245]. PathVQA expands scale but automatically derives QA pairs from textbook captions using rule-based NLP, producing shallow questions that are not aligned with real clinical tasks and involve no clinician input [246]. More recent multilingual resources such as WoundcareVQA collect naturally occurring text-image pairs but target consumer health wound triage rather than medically curated disease classification, and lack structured dermatologic annotation or expert-defined ground-truth reasoning [247]. Similarly, SLAKE provides semantic annotations and a medical knowledge graph, but focuses on radiology

and CT/MRI imaging modalities rather than clinical dermatology.

In dermatology, most publicly available datasets have been developed for image-level classification tasks, such as lesion diagnosis or malignancy detection. Widely used datasets and related collections have enabled substantial progress in supervised image classification, particularly for melanoma recognition and differential diagnosis [82, 83, 84, 115]. However, these datasets provide only a single diagnostic label per image and do not support evaluation of language grounding, fine-grained visual understanding, or clinical reasoning capabilities that are central to emerging vision-language models.

In contrast, dermatology-focused multimodal and visual question answering (VQA) datasets remain limited in trustworthiness, annotation depth, and clinical validity. DermaVQA [248] provides real-world consumer health queries paired with user-generated dermatology photographs and free-text responses; however, it primarily targets consultation-style response generation from clinical photographs rather than structured, expert-annotated evaluation across multiple dermatologic imaging modalities. MM-Skin seeks to scale dermatology VQA by extracting captions from textbooks and using large language models to synthesize approximately 27k QA pairs, but its supervision is synthetically generated rather than clinician-authored, and blends automatically extracted captions with LLM-written conversations instead of dermatologist-verified ground truth [74]. More recent dermatology datasets with richer medical labels, such as SkinCon and SkinCAP (including SkinCoT and SkinCaRe), provide concept annotations, captions, or chain-of-thought reasoning; however, they do not offer an integrated VLM benchmark with standardized question taxonomy, multi-expert consensus labeling, or consistent evaluation of reasoning and fairness across skin tones and imaging modalities [249, 250]. Collectively, existing dermatology multimodal datasets exhibit three recurrent gaps: (1) absence of dermatologist-curated ground-truth VQA annotations for clinical dermatology images, (2) reliance on internet-scraped or synthetically generated supervision, and (3) lack of benchmarks explicitly designed to evaluate visual understanding, language grounding, reasoning, and fairness in dermatology-focused VLMs.

To address these limitations, we introduce DermaBench, a clinician-curated, multi-source dermatology benchmark designed expressly for evaluating the visual understanding and reasoning capabilities of vision-language models. Unlike prior datasets, DermaBench provides dermatologist-authored question–answer pairs in a structured VQA format, ensuring medically reliable ground truth for each item. The benchmark is built on the DERM12345 dataset and the Diverse Dermatology Images dataset [79]. Importantly, every VQA item is written and validated by dermatologists, not synthesized by algorithms, overcoming the reliability limitations of DermaVQA and MM-Skin. To our knowledge, DermaBench is the first dermatology dataset to combine multimodal imagery, expert-verified VQA annotations, standardized ontology-driven labels, and diverse skin tone representation, enabling rigorous evaluation of perception, reasoning, and subgroup performance across skin tone groups. This design makes DermaBench a high-trust benchmark for dermatology-focused vision-language models, providing a foundation for reproducible, fair, and clinically aligned multimodal AI research. The dataset is publicly available at `https://doi.org/10.7910/DVN/Q4LBIW`.

5.2.2 Methods

We selected the DERM12345 and Diverse Dermatology Images datasets as the two primary components of DermaBench. These datasets were included to cover two complementary dermatology imaging settings: clinical photography and dermoscopy. The DDI dataset was selected because it is one of the few publicly available dermatology image collections designed explicitly for skin tone diversity and fairness research. It contains clinical photographs spanning Fitzpatrick skin types I–VI, with each image labeled by dermatologists for both diagnosis and Fitzpatrick phototype. Its structure, comprising 78 diagnostic categories across varied ethnicities, disease severities, and body sites, makes it a high-trust and clinically grounded resource for evaluating dermatology vision-language models in clinical image settings. This diversity enables DermaBench to represent lesions from both light- and dark-skinned patients more equitably and supports systematic fairness evaluation in downstream vision and vision-language models.

To complement this clinical image component, we also incorporated a dermoscopic subset from the DERM12345 test split. DERM12345 provides a hierarchical dermoscopic skin lesion taxonomy with 40 subclasses, making it suitable for evaluating fine-grained dermoscopic reasoning and skin lesion analysis. From the DERM12345 test split, we selected 210 dermoscopic images by sampling approximately five images from each of the 40 subclasses.

Together, the DDI and DERM12345 components allow DermaBench to evaluate both clinical dermatology image reasoning and dermoscopic skin lesion reasoning within a unified visual question-answering framework. The DDI component resulted in 14,474 image–question–answer pairs, while the DERM12345 dermoscopic component contributed 21,597 additional annotations. In total, DermaBench includes 856 images and 36,071 annotations across clinical and dermoscopic imaging modalities (Table 5.1). The overall development pipeline of DermaBench, including data collection, hierarchical question-set development, and expert annotation, is summarized in Figure 5.4.

5.2.2.1 Dataset Composition

DermaBench was developed as a benchmark focused on diagnostic diversity, demographic fairness, precise labeling, and multimodal dermatology reasoning. The dataset is based on two complementary dermatology image resources: the Diverse Dermatology Images dataset and the DERM12345 dataset. These two components were selected to represent different but clinically important dermatology imaging settings. The DDI component provides clinical dermatology photographs with Fitzpatrick skin tone annotations and broad diagnostic diversity, while the DERM12345 component provides dermoscopic images with a hierarchical skin lesion taxonomy suitable for fine-grained lesion analysis.

Each image in the DDI dataset was labeled using comprehensive hierarchical question sets (Appendix E.2.8) designed to capture global image characteristics, dermatologic morphology, distribution, surface features, color, image quality, and clinically meaningful narrative descriptions. The DDI schema consists of 21 main and 33 total questions, including single-select, multi-select, and open-ended formats (Table 5.2).

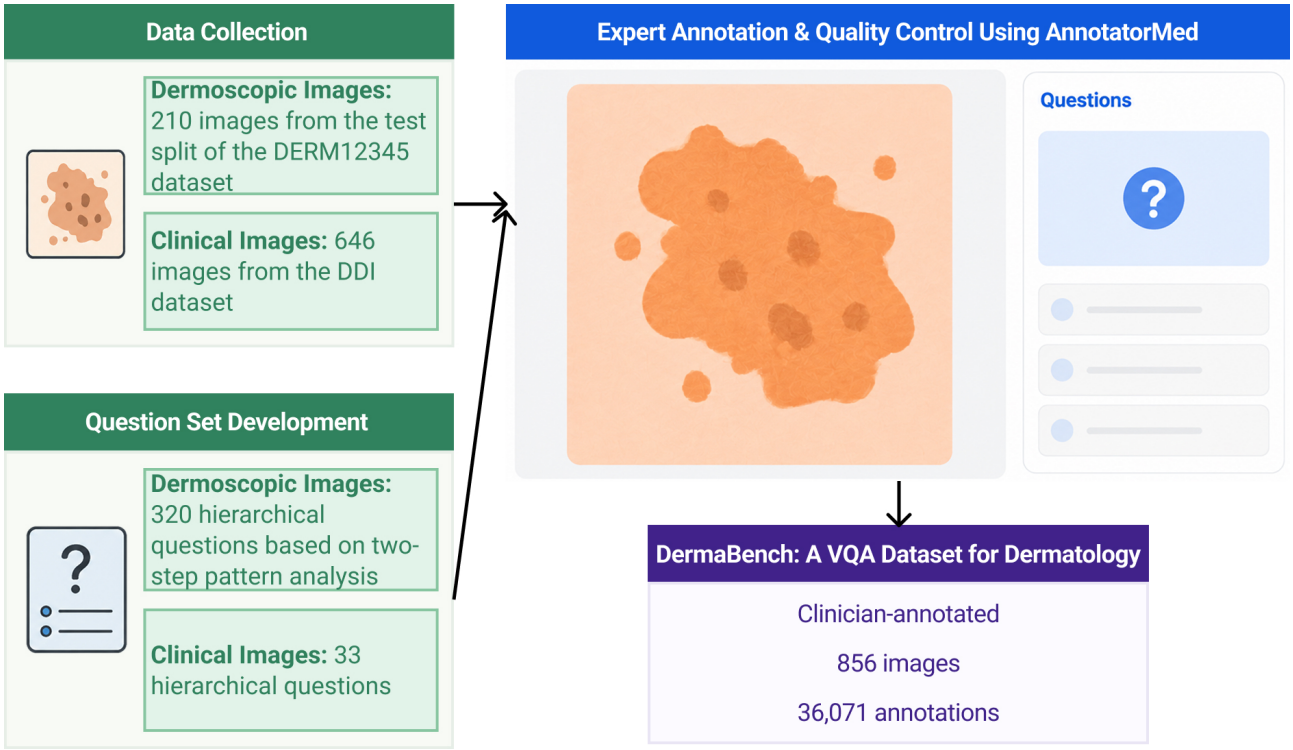


Figure 5.4: Dermoscopic images from DERM12345 and clinical images from DDI were annotated using expert-developed hierarchical question sets. Annotations were performed in AnnotatorMed, resulting in a clinician-annotated dermatology VQA benchmark comprising 856 images and 36,071 annotations.

Domain	Dataset	Modality	#Images	#QA Pairs	Details
Non-derm	VQA-RAD [245]	Radiology	315	3.5K	Open-ended
Non-derm	SLAKE [251]	Radiology	642	14K	Bilingual EN/ZH
Non-derm	PathVQA [246]	Pathology	5K	32K	Multi-type Qs
Derm	DermaVQA [248]	Clinical	3.4K	1.5K	Multilingual, Reddit + IIYI (online discussion forums)
Derm	MM-Skin [74]	Various[1]	∼11K	27K	Adds textbook image–text pairs
Derm	Lightweight Derm VQA [252]	Clinical	1,038	∼7.3K	7 QA pairs per image; low-resource focus
Derm	**DermaBench-DDI (Ours)**	Clinical	**646**	**14,474**	**Clinical VQA benchmark with skin tone diversity**
Derm	**DermaBench-DERM12345 (Ours)**	Dermoscopic	**210**	**21,597**	**Hierarchical dermoscopic VQA benchmark from the DERM12345 test split**

Table 5.1: Comparison of DermaBench with representative medical VQA corpora. Various[1]: clinical, dermoscopic, and pathology images.

Module / Step	Question ID
Image category and validation	0–4
Image quality and identifiability	5–9
Skin type assessment	10
Lesion number and anatomical location	11–12
Distribution pattern	13
Shape and configuration	14
Border features	15
Surface features	16
Lesion size estimation	17
Color assessment	18
Primary elementary lesion	19
Secondary elementary features	20
Final narrative report	21

Table 5.2: Hierarchy of the clinical image-analysis workflow used for the DDI component of DermaBench.

Expert clinicians designed hierarchical questions (93 main and 320 total questions, including single-select, multi-select, and open-ended formats) based on the two-step pattern analysis widely used in skin lesion diagnosis (Appendix E.2.6). They then annotated 210 images from the test split of the DERM12345 dataset (approximately five images per class across 40 classes) using these questions, creating a challenging VQA benchmark requiring nuanced clinical reasoning. In total, 93 main questions and 227 subquestions were developed and organized into five sections: intro, clinical, dermoscopy-basic, dermoscopy-melanocytic, and dermoscopy-nonmelanocytic (Table 5.3).

Module / Step / Level	**Question ID**
Image Validation	0–5
Clinical Image Analysis	6–22
Dermoscopic Image Analysis	23–77
Step 1. Basic lesion context	23–37
Step 2. Melanocytic vs. nonmelanocytic lesion	38–77
Level 1: Criteria for melanocytic lesion	38–44
Level 2: Specific blood vessels in melanocytic lesions	45–49
Level 3: Criteria for Dermatofibroma	50–54
Level 4: Criteria for Basal Cell Carcinoma	55–58
Level 5: Criteria for Seborrheic Keratosis and Solar Lentigo	59–61
Level 6: Criteria for vascular lesions (angioma, angiokeratoma, pyogenic granuloma, lymphangioma, Kaposi sarcoma)	62–64
Level 7: Criteria for keratinizing tumors (squamous cell carcinoma, keratoacanthoma, actinic keratosis)	65–68
Level 8: Specific blood vessels in nonmelanocytic lesions	69–71
Level 9: Criteria for structureless / featureless lesions	72–77
Step 3. Benign nevus vs. Suspicious melanocytic lesions vs. Melanoma	78–86
Level 1: Global pattern evaluation	78–81
Level 2: Local structure evaluation	82–86
Final Diagnosis Decision Questions	87–93

Table 5.3: Hierarchy of the image-analysis workflow used for the DERM12345 dermoscopic component of DermaBench.

The question sets were designed by expert dermatologists based on established dermatologic examination frameworks, dermoscopic pattern-analysis principles, and the morphology chapters of standard dermatology textbooks, ensuring clinical fidelity and alignment with real-world diagnostic workflows. For the clinical image component, the schema was designed to capture global image characteristics, lesion morphology, distribution, color, surface features, and narrative descriptions [253, 254, 255, 20]. For the dermoscopic component, the question set followed the two-step pattern analysis framework [26]. Step 1 evaluated the basic lesion context. Step 2 classified lesions as melanocytic or nonmelanocytic and was further subdivided into nine detailed diagnostic levels. The summarization section similarly followed a structured workflow: first, global and local lesion patterns were evaluated to classify the lesion as a benign

nevus, suspicious melanocytic lesion, or melanoma; second, the final diagnosis was determined using specific decision questions. Questions irrelevant to dermoscopic images, such as clinical-image-specific items, were omitted from the DERM12345 annotations. In addition, each section included two special questions requesting a summary report.

Schema development followed an iterative process: the full hierarchical question sets were reviewed, tested, and refined across three design rounds, with expert feedback incorporated at each stage to improve clarity, branching logic, and morphological completeness. To support consistent downstream annotation, consensus guidelines were prepared to define standardized interpretations for lesion morphology, distribution patterns, color terminology, surface characteristics, and dermoscopic structures. These guidelines were used throughout the annotation phase to reduce inter-rater variability and ensure uniform application of the hierarchical schemas. Collectively, these procedures strengthened the reliability, reproducibility, and clinical interpretability of the DermaBench annotations.

5.2.2.2 Annotation Framework: AnnotatorMed

All images were annotated using AnnotatorMed, an open-source medical data annotation interface developed within the SCALEMED framework (Section 5.1). AnnotatorMed is a web-based, modular tool that renders dynamic question hierarchies through a context-aware interface. Subquestions are automatically displayed only when their relevant parent options are selected, preventing screen clutter and reducing scrolling time. This UX design allows physicians to complete each annotation efficiently without navigating unnecessary question branches, enabling rapid yet comprehensive labeling during clinical workflows.

5.2.2.3 Consensus and Quality Control

A total of six dermatologists contributed to the DermaBench annotation workflow across multiple stages. Three dermatologists developed and refined the hierarchical question sets and annotation criteria used for the clinical and dermoscopic components. During annotation, image quality was also assessed by the dermatologists as an integral part of the annotation process. An initial quality-control review was subsequently performed jointly by one dermatologist and one engineer to identify residual clinical inconsistencies, annotation errors, and structural issues. In a second review stage, three dermatologists examined all annotations in a consensus setting and jointly resolved discrepancies in lesion morphology, distribution, diagnostic interpretation, Fitzpatrick skin type, and other clinically relevant annotation fields. Ten images from the DDI dataset with unresolved disagreements, insufficient visual quality, or ambiguous clinical content were excluded to ensure dataset reliability, resulting in a final dataset of 646 images from the original 656 images.

5.2.2.4 Annotation Efficiency and UX Performance

AnnotatorMed's dynamic rendering mechanism substantially improved annotation efficiency. By displaying subquestions only when their parent options were selected, the interface reduced visual clutter and the amount of content requiring scrolling by an estimated 50%, based on a comparison of the user interface

size. Dermatologists typically completed a full annotation in 3–5 minutes per image. The combination of structured single-select, multi-select, and open-ended questions and narrative fields allowed simultaneous quantitative and descriptive data capture, supporting downstream use in classification, retrieval, and multimodal VQA evaluation.

All data follow VQA-style structured schemas designed to support multimodal dermatology evaluation. For the DDI clinical image component, the structured annotation schema includes image path, modality tag, Fitzpatrick phototype, diagnostic and morphological labels, question category, answer text, annotator identifier, and additional metadata fields supporting fairness and reasoning analyses. For the DERM12345 dermoscopic component, the schema additionally captures hierarchical dermoscopic question categories, two-step pattern-analysis items, clinical and dermoscopic context, final diagnosis decision questions, and subclass-level information linked to the DERM12345 taxonomy.

The dataset is released for research use and is accompanied by documentation detailing the annotation protocol, consensus process, question hierarchy, and recommended evaluation workflows. This unified structure enables integration into multimodal training and evaluation pipelines while preserving the distinct strengths of each component: subgroup-aware clinical image evaluation across skin tone groups from DDI and fine-grained dermoscopic lesion reasoning from the DERM12345 dataset. Together, these components provide a reproducible and clinically grounded benchmark for dermatology-focused vision and vision-language models.

5.2.3 Data Records

DermaBench does not host or redistribute any dermatologic images. Instead, the dataset provides a structured metadata table on Harvard Dataverse at https://doi.org/10.7910/DVN/Q4LBIW [256] containing all annotation fields and image identifiers required to reproduce experiments or link to the original image sources [79, 115]. This design respects the licensing constraints of the upstream datasets while enabling standardized benchmarking. DermaBench was released under a Creative Commons Attribution-NonCommercial (CC BY-NC) license.

Researchers must obtain direct access to the images from the original repositories. For the Diverse Dermatology Images dataset, image files and access procedures are available at https://ddi-dataset.github.io/index.html#dataset. The DDI dataset is released under a restricted academic-use license: (i) permission is granted to view and use the dataset without charge for personal, non-commercial research purposes only, and (ii) users may not distribute, publish, or reproduce any portion of the dataset without prior written permission from the Stanford School of Medicine. DermaBench therefore provides only metadata for the DDI component and does not copy or redistribute DDI images in any form.

For the DERM12345 component, dermoscopic images and metadata are available through Harvard Dataverse at https://doi.org/10.7910/DVN/DAXZ7P. DERM12345 is released under a Creative Commons Attribution license, and researchers can access the original dermoscopic images directly from the repository.

In DermaBench, the DERM12345 component provides annotation metadata and image identifiers linked to the corresponding DERM12345 images, allowing reproducible evaluation while preserving the original dataset provenance.

The metadata is distributed as a tabular CSV-style dataset, with one row per image and annotation fields corresponding to the hierarchical question schema. Because DermaBench uses a conditional annotation workflow, many subquestion fields remain intentionally empty when a parent question was not triggered. For instance, subquestions related to the lesion location "scalp" appear only when the annotator selected "scalp" in Q12; otherwise those columns remain blank.

DermaBench does not include a predefined train-test split. Instead, a unified metadata file is provided to allow researchers to construct splits appropriate for classification, fairness auditing, VQA, or multimodal reasoning tasks.

Dataset Metadata The DermaBench metadata file includes the following fields:

- **file_ID:** Identifier linking to the corresponding DDI or DERM12345 image.
- **questionX:** The text of each question in the annotation schema.
- **answerX:** The dermatologist's answer to each question. Empty fields indicate questions not triggered by branching logic.
- **subquestion fields:** Hierarchical attributes (e.g., Q16.1, Q18.1). Present only when applicable.

The metadata schema supports both high-level tasks (disease classification, fairness assessment, distribution analysis) and fine-grained multimodal tasks (visual question answering, long-form reasoning, semantic parsing). Its hierarchical organization preserves the full richness of dermatologic morphology while maintaining compatibility with VQA-style architectures.

5.2.4 Technical Validation

Reliable evaluation of dermatology-focused AI systems requires benchmark datasets that are clinically trustworthy, demographically inclusive, and annotated with sufficient detail to support both diagnostic and reasoning-based tasks. The Diverse Dermatology Images dataset was originally curated to support fairness and subgroup evaluation in clinical image analysis, as it is a publicly accessible dermatology collection with expert-confirmed diagnoses and intentionally diverse representation across Fitzpatrick skin types. DermaBench builds upon this foundation by adding structured, multi-level visual question answering (VQA) annotations, enabling assessment not only of diagnostic performance but also of multimodal reasoning and fairness across skin tones.

To complement the clinical and fairness-oriented strengths of DDI, DermaBench also incorporates a dermoscopic component derived from DERM12345. DERM12345 is one of the most detailed and highly

subclassified dermoscopic datasets in the field, containing 40 skin lesion subclasses organized within a hierarchical taxonomy. This structure makes it particularly suitable for evaluating fine-grained dermoscopic reasoning, hierarchical lesion interpretation, and skin cancer-focused AI systems. By integrating DERM12345, DermaBench extends beyond clinical image reasoning and enables evaluation of dermatology-focused models across both clinical photographs and dermoscopic skin lesion images.

Clinically, AI has the potential to assist triage workflows by helping distinguish benign from malignant lesions, thereby reducing delays in dermatologic evaluation. However, dermatology datasets must represent both patient diversity and diagnostic complexity. Because algorithmic biases often arise from underrepresentation, the DDI dataset's intentional inclusion of Fitzpatrick I–VI skin tones provides a validated substrate for fairness auditing, which DermaBench directly inherits and extends through its annotation schema. In parallel, the hierarchical subclass structure of DERM12345 provides a complementary substrate for assessing whether AI systems can reason over fine-grained dermoscopic patterns and diagnostically similar skin lesion categories.

5.2.4.1 Validation of Source Datasets (DDI and DERM12345)

The images used in DermaBench originate from two clinically validated dermatology datasets: DDI and DERM12345. The DDI component underwent rigorous clinical and pathology-based validation at the time of the original dataset curation. As described in the DDI dataset paper [79], all images were retrospectively selected by reviewing pathology-confirmed cases from Stanford Clinics between 2010 and 2020. The original DDI dataset contained 656 images, of which 646 were retained for annotation. Each diagnosis was validated by both an expert dermatologist and a dermatopathologist using the corresponding biopsy report. Skin tone labels were assigned using in-person clinical assessments, demographic photographs, and review of the clinical images by two expert dermatologists. These procedures ensure high-fidelity diagnostic labels and accurate Fitzpatrick assignments, which form the basis for downstream fairness assessments in DermaBench.

The DERM12345 component provides the dermoscopic subset of DermaBench and adds fine-grained skin lesion coverage. As described in the DERM12345 dataset paper [132], the DERM12345 dataset contains 12,345 high-resolution dermoscopic images from 1,627 patients, collected between 2008 and 2021 from the Department of Dermatology and Venereology at Celal Bayar University (Manisa, Türkiye), Istinye University (Istanbul, Türkiye), and the University of Health Sciences Haydarpasa Numune Research and Training Hospital (Istanbul, Türkiye). Diagnostic confirmation in DERM12345 was based on consensus clinical diagnosis by two expert dermatologists, longitudinal follow-up, or histopathological confirmation where available. Importantly, all malignant lesions in DERM12345 were histopathologically confirmed. By incorporating the DERM12345 test-split subset, DermaBench extends its validation framework from clinical image reasoning and fairness assessment to hierarchical dermoscopic evaluation of fine-grained skin lesion categories.

5.2.4.2 Validation of Annotation Quality

DermaBench extends the DDI dataset through structured annotation of dermatologic morphology, image quality, distribution patterns, color hierarchy, and higher-level VQA tasks. All annotations were performed by expert dermatologists using a hierarchical schema implemented in AnnotatorMed and subsequently underwent dermatologist-led consensus review and quality control. The annotation workflow was designed to enhance consistency through controlled vocabularies and predefined branching logic, improve precision through specialist-performed labeling, and ensure reproducibility through explicit documentation of all question-and-answer fields.

Because the annotation interface displays subquestions only when parent attributes are selected, missing entries in sub-columns reflect intentional branching rather than incomplete annotation.

5.2.5 Usage Notes

Question ID 0 was included in both the DDI and DERM12345 components as a screening item to indicate images that should not be considered for downstream annotation or evaluation.

The DDI component of DermaBench was designed to be broadly compatible with the SkinCon [249] dermatologic concept taxonomy while remaining practical for VQA and LLM benchmarking. The concept labels are grounded in previously defined clinical terms [20], but they are not intended to be exhaustive, nor do they assume perfectly clear-cut boundaries between concepts. In practice, dermatologic descriptors may be annotated differently across datasets, including SkinCon, because attribute definitions, annotation guidelines, and clinical interpretation may vary. As a result of expanding the concept set from 48 to 197 concepts, partial differences in annotation granularity and alignment emerge between DermaBench and SkinCon, motivating the inclusion of explicit concept mappings. Core morphological attributes are aligned at the question level (Table 5.4), including primary lesion morphology (Q19), secondary morphology (Q20), surface features (Q16), color (Q18), and distribution (Q13). Given the potential for variability in definitions and annotations, we recommend providing SkinCon annotations alongside corresponding DermaBench columns (Table 5.4) when performing LLM benchmarking, enabling transparent comparison and more consistent cross-dataset analyses while preserving the VQA-oriented structure of DermaBench.

Table 5.4: Mapping between SkinCon concepts and their corresponding DermaBench question categories.

SkinCon Concept	Type	DermaBench Equivalent
Vesicle	Primary	Q19
Papule	Primary	Q19
Macule	Primary	Q19

Continued on next page

SkinCon Concept	Type	DermaBench Equivalent
Plaque	Primary	Q19
Pustule	Primary	Q19
Patch	Primary	Q19
Nodule	Primary	Q19
Ulcer	Secondary	Q20
Crust	Secondary	Q20
Erosion	Secondary	Q20
Excoriation	Secondary	Q20
Atrophy	Secondary	Q20
Exudate	Secondary	Q20
Purpura/Petechiae	Primary	Q19
Fissure	Secondary	Q20
Telangiectasia	Primary	Q19
Scale	Secondary	Q20
Scar	Secondary	Q20
Friable	Surface	Q16
Pedunculated	Surface	Q16
Exophytic/Fungating	Surface	Q16
Warty/Papillomatous	Surface	Q16
Dome-shaped	Surface	Q16
Flat-topped	Surface	Q16
Brown (Hyperpigmentation)	Color	Q18
Translucent	Color	Q18
White (Hypopigmentation)	Color	Q18
Purple	Color	Q18
Yellow	Color	Q18
Black	Color	Q18
Erythema	Color	Q18
Comedo	Primary	Q19
Lichenification	Secondary	Q20

Continued on next page

SkinCon Concept	Type	DermaBench Equivalent
Blue	Color	Q18
Umbilicated	Surface	Q16
Salmon	Color	Q18
Wheal	Primary	Q19
Gray	Color	Q18
Pigmented	Color	Q18
Cyst	Primary	Q19
Induration	Surface	Q16
Xerosis	Surface	Q16
Sclerosis	Secondary	Q20
Abscess	Primary	Q19
Bulla	Primary	Q19
Poikiloderma	Color	Q18
Acuminate	Surface	Q16
Burrow	Primary	Q19
Do not consider this image	Exclusion / Not answered	Q0

VQA Prompt for DermaBench This prompt is designed to benchmark vision–language models on DermaBench by standardizing how models interpret dermatologic images and respond to structured clinical questions. It ensures consistent use of dermatologic terminology and annotation rules by instructing the model to follow the definitions and guidelines provided in Appendix E.2.7. The prompt should be directly paired with the question definitions and answer options in Appendix E.2.8.

Main Prompt:

You are a board-certified dermatologist with expertise in dermatologic image interpretation. Carefully analyze the provided dermatologic image and answer the question below using only the visual information in the image and standard dermatologic knowledge.

Follow the instructions strictly based on the question type:

- If `QUESTION_TYPE` is **single-select**, select *one and only one* answer from the provided options and return the option text exactly as written.
- If `QUESTION_TYPE` is **multi-select**, select *all* options that apply based on the image. Return the selected option texts as a comma-separated list, using the exact wording.

- If `QUESTION_TYPE` is **open-ended**, write a concise, structured, and clinically meaningful report using standard dermatologic terminology. Do not speculate or include information not supported by the image.

Do not assume patient history, symptoms, or metadata unless explicitly provided. If the image cannot be evaluated for the given question, use the dataset-defined response indicating that the image should not be considered.

Question type: `{QUESTION_TYPE}`
Question: `{QUESTION_TEXT}`
Options: `{QUESTION_OPTIONS}`

5.2.6 Data availability

The DermaBench dataset is publicly available through Harvard Dataverse (https://doi.org/10.7910/DVN/Q4LBIW). The DERM12345 dataset is publicly available via Harvard Dataverse (https://doi.org/10.7910/DVN/DAXZ7P). The DDI dataset is available through the project website (https://ddi-dataset.github.io/index.html#dataset).

5.3 Hierarchical Benchmarking of Dermatology Foundation Models

Foundation models have transformed medical image analysis by providing robust feature representations that reduce the need for large-scale task-specific training. However, current benchmarks in dermatology often reduce the complex diagnostic taxonomy to flat, binary classification tasks, such as distinguishing melanoma from benign nevi. This oversimplification obscures a model's ability to perform fine-grained differential diagnoses, which is critical for clinical workflow integration. This study evaluates the utility of embeddings from 10 model families represented by 13 model variants, spanning general computer vision, general medical imaging, and dermatology-specific domains, for hierarchical skin lesion classification. Using the DERM12345 dataset, which comprises 40 lesion subclasses, we extracted embeddings from frozen encoders and trained lightweight adapter models using five-fold cross-validation. We introduce a hierarchical evaluation framework that assesses performance across four hierarchical levels using five label sets: subclass (40), main class (15), superclass (4), superclass (2), and malignancy (2). Our results reveal a granularity gap in model capabilities: MedImageInsight achieved the strongest performance across all reported hierarchy levels, including 97.20% weighted F1-score on binary malignancy classification and 59.98% on fine-grained 40-class subtype classification. MedSigLIP remained highly competitive at broader hierarchy levels, ranking second for malignancy (2), superclass (2), superclass (4), and main class (15), but ranked fourth on the subclass (40) task. Our findings suggest that while general medical foundation models are highly effective for high-level screening, specialized modeling strategies are necessary for the granular distinctions required in diagnostic support systems.

5.3.1 Introduction

Dermatologic diagnosis is an inherently hierarchical process [257]. When a clinician evaluates a skin lesion, the cognitive workflow proceeds from broad categorization (e.g., melanocytic versus nonmelanocytic) to risk assessment of malignancy, and finally to specific diagnosis. A robust diagnostic system must not only distinguish benign from malignant lesions but also differentiate between visually similar but biologically distinct entities, such as a compound nevus and a dermal nevus.

The automation of skin lesion analysis historically relied on convolutional neural networks trained end-to-end on specific datasets [258, 259]. While effective, this paradigm is resource-intensive and often yields models that generalize poorly across different clinical settings or imaging devices. Recently, foundation models have emerged as a powerful alternative [260, 261, 262, 263]. By pretraining on vast repositories of natural or medical images using self-supervised learning or image–text alignment, these models learn rich, transferable feature representations. In the general medical domain, models such as BiomedCLIP [264], MedImageInsight [265], and MedSigLIP [205] have demonstrated that leveraging diverse biomedical data can yield robust performance across various modalities, from radiology to histopathology [266, 267, 268]. Simultaneously, dermatology-specific foundation models, including PanDerm [116], SkinVL [74], and Derm

Foundation [269], have been developed to capture the unique visual semantics of skin diseases.

Despite the rapid increase in the number of foundation models in healthcare, a critical gap remains in their evaluation. Existing benchmarks typically focus on flat classification metrics or binary screening accuracy [270, 271, 268, 272]. Similarly, earlier benchmarks [259, 79] and recent few-shot evaluations [262] often reduce the diagnostic task to binary melanoma detection or a multiclass schema. Barata et al. [273] argued for the necessity of hierarchical diagnosis and Yan et al. [75] recently introduced hierarchical ontologies in the Derm1M dataset. However, a comparative benchmark of modern foundation model embeddings across the full diagnostic taxonomy is needed to understand their performance. Current evaluations fail to capture the granularity gap, in which a model may excel at high-level screening because of strong semantic priors yet struggle to capture the subtle intra-class visual variation required to distinguish biologically related but distinct entities.

In this study, we present a comprehensive hierarchical benchmark of foundation model embeddings for dermatology using the DERM12345 dataset [115]. We systematically evaluate ten model families represented by 13 encoder variants categorized into three domains: general computer vision models (DINOv2 [274], DINOv3 [275], CLIP [92], ResNet-50 [152]), general medical foundation models (MedSigLIP [205], BiomedCLIP [264], MedImageInsight [265]), and dermatology-specific models (MONET [239], PanDerm [116], Derm Foundation [269]). Instead of fine-tuning heavy backbones, we employ a standardized feature extraction pipeline followed by the training of lightweight adapters. This approach enables evaluation of the frozen representations without fine-tuning the foundation model backbones. Our evaluation strategy maps model predictions across five label sets: subclass (40), main class (15), superclass (4), superclass (2), and malignancy (2). This multi-level analysis reveals trade-offs between semantic understanding and visual discrimination, providing guidance for researchers selecting foundation models for specific clinical applications, from triage apps to expert-level decision support systems.

5.3.2 Methods

To evaluate the clinical utility of foundation model embeddings in dermatology, we designed a systematic two-phase experimental pipeline comprising data standardization, feature extraction, and hierarchical evaluation. The overall workflow, extending from raw image input to multi-level diagnostic prediction, is illustrated in Figure 5.5. This pipeline allows us to isolate the quality of the learned representations by keeping the foundation models frozen while training lightweight adapters. The following subsections detail the dataset characteristics, the selection of foundation models, the cross-validation protocol, and the hierarchical logic used to assess performance across the diagnostic taxonomy.

5.3.2.1 Study Design and Dataset

This study utilized the DERM12345 dataset [115], a large-scale collection of 12,345 dermoscopic images representing a diverse range of skin lesions collected from multiple clinical sources in Türkiye. The dataset is characterized by a granular taxonomic structure that classifies lesions into 40 distinct subclasses.

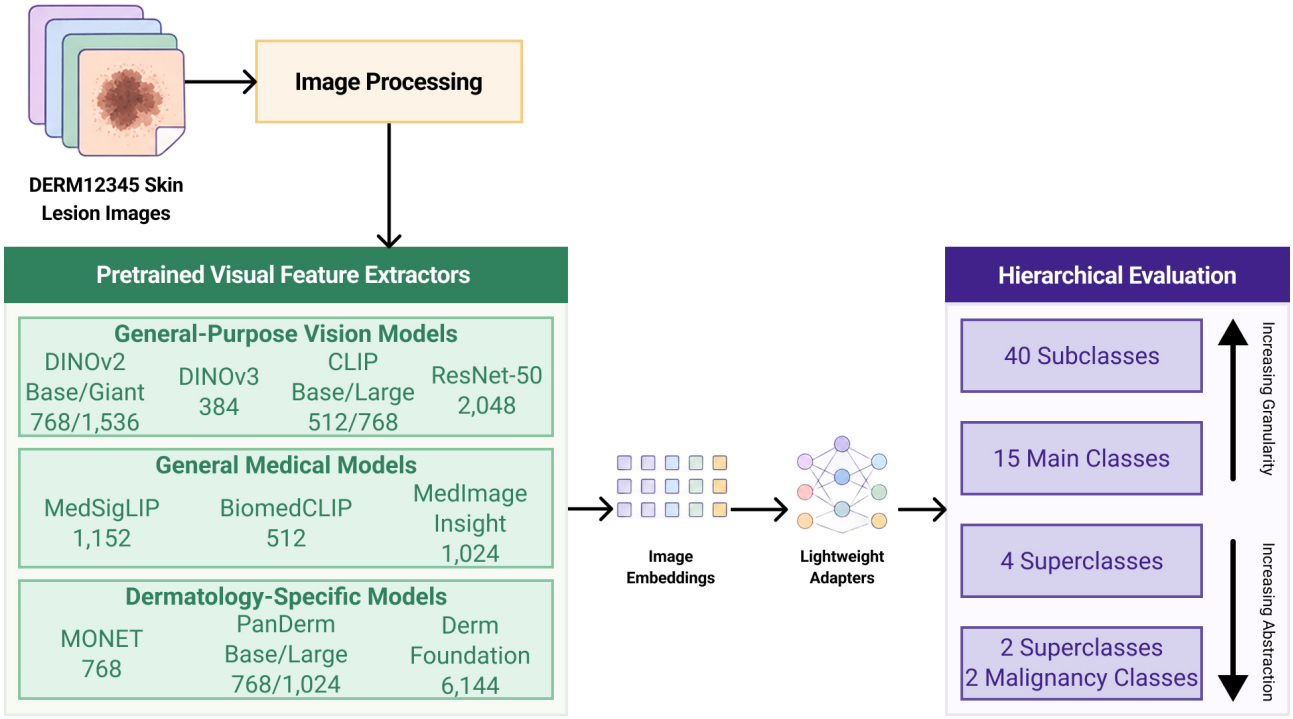


Figure 5.5: Experimental pipeline for benchmarking foundation models on the DERM12345 dataset [115]. High-resolution skin lesion images are preprocessed and passed through 13 evaluated model variants representing 10 foundation model families, used as frozen feature extractors and grouped into three domains: General Computer Vision (green), General Medical (purple), and Dermatology-Specific (yellow). Numbers in parentheses indicate the corresponding embedding dimensions. The extracted embeddings are used to train lightweight downstream classifiers, including K-Nearest Neighbors (KNN), Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), Multilayer Perceptron (MLP), and Extreme Gradient Boosting (XGBoost). Performance is evaluated hierarchically by mapping the predicted subclass labels from 40 subclasses to coarser diagnostic levels (main classes, superclasses, and binary malignancy), enabling assessment of model performance across the full clinical taxonomy.

The native taxonomy comprises 15 main classes and five superclasses; for the present benchmark, the indeterminate superclass was reassigned as described below, yielding four evaluation superclasses. The dataset includes a predefined split of 9,860 images for training and 2,485 images for testing.

5.3.2.2 Taxonomy Mapping

As described in the main text [115], the DERM12345 taxonomy was mapped to four hierarchical levels. For hierarchical evaluation, the 40 fine-grained subclasses were organized into progressively coarser diagnostic label sets: subclass (40), main class (15), superclass (4), superclass (2), and malignancy (2). The superclass (4) level comprised melanocytic benign, melanocytic malignant, nonmelanocytic benign, and nonmelanocytic malignant lesions. The superclass (2) level distinguished melanocytic from nonmelanocytic lesions, whereas malignancy (2) distinguished benign from malignant lesions. The complete subclass-to-parent mapping is provided in Table 5.5.

To facilitate the evaluation of malignancy detection, we mapped the dataset's native taxonomy to a malignancy (2) label map. The original dataset taxonomy includes an indeterminate category containing actinic keratosis cases. For the malignancy (2) task, this indeterminate category was assigned to the nonmalignant branch to create a two-class benign-versus-malignant task. This grouping was used solely for computational evaluation and should not be interpreted as classifying actinic keratosis as biologically benign. Thus, this mapping reduced the original five superclasses defined in the dataset to four superclasses for this benchmark: melanocytic benign, melanocytic malignant, nonmelanocytic benign, and nonmelanocytic malignant (Table 5.5).

Table 5.5: DERM12345 Hierarchical Taxonomy Map

Index	Subclass	Main Class	Superclass
0	acb	Banal Compound	Melanocytic Benign
1	acd	Dysplastic Compound	Melanocytic Benign
2	ajb	Banal Junctional	Melanocytic Benign
3	ajd	Dysplastic Junctional	Melanocytic Benign
4	ak	Keratinocytic (Indeterminate)	Nonmelanocytic Benign
5	alm	Melanoma	Melanocytic Malignant
6	angk	Vascular	Nonmelanocytic Benign
7	anm	Melanoma	Melanocytic Malignant
8	bcc	Keratinocytic	Nonmelanocytic Malignant
9	bd	Keratinocytic	Nonmelanocytic Malignant
10	bdb	Banal Dermal	Melanocytic Benign
11	cb	Banal Compound	Melanocytic Benign
12	ccb	Banal Compound	Melanocytic Benign
13	ccd	Dysplastic Compound	Melanocytic Benign

Index	Subclass	Main Class	Superclass
14	cd	Dysplastic Compound	Melanocytic Benign
15	ch	Keratinocytic	Nonmelanocytic Malignant
16	cjb	Banal Junctional	Melanocytic Benign
17	db	Banal Dermal	Melanocytic Benign
18	df	Fibro-histiocytic	Nonmelanocytic Benign
19	dfsp	Fibro-histiocytic	Nonmelanocytic Malignant
20	ha	Vascular	Nonmelanocytic Benign
21	isl	Lentigo	Melanocytic Benign
22	jb	Banal Junctional	Melanocytic Benign
23	jd	Dysplastic Junctional	Melanocytic Benign
24	ks	Vascular	Nonmelanocytic Malignant
25	la	Vascular	Nonmelanocytic Benign
26	lk	Keratinocytic	Nonmelanocytic Benign
27	lm	Melanoma	Melanocytic Malignant
28	lmm	Melanoma	Melanocytic Malignant
29	ls	Lentigo	Melanocytic Benign
30	mcb	Banal Compound	Melanocytic Benign
31	mel	Melanoma	Melanocytic Malignant
32	mpd	Keratinocytic	Nonmelanocytic Malignant
33	pg	Vascular	Nonmelanocytic Benign
34	rd	Dysplastic Recurrent	Melanocytic Benign
35	sa	Vascular	Nonmelanocytic Benign
36	scc	Keratinocytic	Nonmelanocytic Malignant
37	sk	Keratinocytic	Nonmelanocytic Benign
38	sl	Lentigo	Melanocytic Benign
39	srjd	Dysplastic Junctional	Melanocytic Benign

5.3.2.3 Foundation Models and Embedding Extraction

We selected a diverse suite of foundation models to evaluate the impact of domain-specific pretraining on dermatologic feature extraction. These models were categorized into three distinct domains. First, the General Vision category included DINOv2 (Base and Giant variants) [274], DINOv3 [275], and CLIP (Base and Large variants) [92]. These models were pretrained on large-scale natural image datasets and serve as a baseline for general visual feature extraction capabilities. Second, the General Medical category included MedSigLIP [205], BiomedCLIP [264], and MedImageInsight [265]. These models leverage large-scale medical image–text pairs or diverse radiological data, offering representations tuned to the

broader medical domain. Third, the Dermatology-specific category included MONET [239], PanDerm [116], and Derm Foundation [269]. These models were explicitly pretrained on dermatologic images, theoretically offering the most semantically relevant features for skin lesion analysis. The exact model variants, public checkpoint or weight sources, and embedding dimensions for each evaluated encoder are reported in Table 5.6.

Table 5.6: Foundation Model Specifications and Weight Sources

Model	Exact checkpoint / public weight source	Embedding dimension
DINOv2 (Base)	`facebook/dinov2-base` (ViT-B/14)	768
DINOv2 (Giant)	`facebook/dinov2-giant` (ViT-g/14)	1,536
DINOv3	`facebook/dinov3-vits16-pretrain-lvd1689m` (ViT-S/16)	384
CLIP (Base)	`openai/clip-vit-base-patch32` (ViT-B/32)	512
CLIP (Large)	`openai/clip-vit-large-patch14-336` (ViT-L/14, 336 px)	768
ResNet-50	`ResNet50_Weights.IMAGENET1K_V2`	2,048
MedSigLIP	`google/medsiglip-448`	1,152
BiomedCLIP	`microsoft/BiomedCLIP-PubMedBERT_256-vit_base_patch16_224` (ViT-B/16)	512
MedImageInsight	`lion-ai/MedImageInsight`, `2024.09.27/vision_model/medimageinsigt-v1.0.0.pt`	1,024
MONET	`suinleelab/monet` (CLIP ViT-L/14)	768
PanDerm (Base)	`panderm_bb_data6_checkpoint-499.pth` (ViT-B/16)	768
PanDerm (Large)	`panderm_ll_data6_checkpoint-499.pth` (ViT-L/16)	1,024
Derm Foundation	DERM12345 precomputed embeddings; Google Derm Foundation v1.0.0 (BiT-M ResNet101x3)	6,144

We implemented a standardized inference pipeline to extract feature embeddings from these models. Each model was wrapped in a unified interface that handled model-specific preprocessing requirements, such as resizing and normalization. We utilized the models as fixed feature extractors without fine-tuning their weights. For each image in the dataset, a forward pass was performed, and the embedding vector was extracted from the model's final hidden state, typically via mean pooling or the specific classification token output as dictated by the model architecture. This process resulted in a standardized feature matrix for the training and test sets for each foundation model.

The extracted embedding dimensions varied across the models (Table 5.6). DINOv3 had the smallest size at

384 dimensions, followed by CLIP (Base) and BiomedCLIP at 512 dimensions. DINOv2 (Base), CLIP (Large), MONET, and PanDerm (Base) produced 768-dimensional embeddings. MedImageInsight and PanDerm (Large) produced 1,024-dimensional embeddings, followed by MedSigLIP at 1,152 dimensions, DINOv2 (Giant) at 1,536 dimensions, and ResNet-50 at 2,048 dimensions. Derm Foundation produced the largest embeddings at 6,144 dimensions.

5.3.2.4 Adapter Training and Optimization

To assess the separability and utility of the extracted embeddings, we trained a set of lightweight adapter classifiers. We employed six distinct classification algorithms: K-Nearest Neighbors (KNN) [276], Logistic Regression (LR) [277], Support Vector Machines (SVM) [278], Random Forest (RF) [279], MLP [280], and XGBoost [281]. To account for differences in how individual classifiers interact with the embedding spaces of different foundation models, adapter selection was treated as part of the evaluation procedure. All six adapter classifiers (KNN, LR, SVM, RF, MLP, and XGBoost) were evaluated for each foundation model under two distinct training settings. In the primary fine-grained experiment, adapters were trained to predict the 40 subclass labels, and the resulting subclass predictions were subsequently mapped to the main class (15), superclass (4), superclass (2), and malignancy (2) levels for hierarchical evaluation. In the complementary coarse-level experiment, the same adapter families were trained directly on the superclass (4) labels, and their predictions were further mapped to the superclass (2) and malignancy (2) levels. Adapter selection and hyperparameter optimization were performed independently within each training setting using validation data. Detailed performance results for all evaluated adapters are reported in Appendix D.1. In the main text, we report only the validation-selected best-performing adapter for each foundation model at each hierarchy level. Adapter selection was therefore performed independently for each label set, and values across hierarchy levels for a given foundation model may originate from different adapter classifiers. Importantly, foundation models produce embeddings of vastly different dimensions, ranging from 384 in DINOv3 to 6,144 in Derm Foundation. A single classification head might favor some models while unfairly punishing others. To ensure a fair comparison, we treated adapter selection as part of the validation-stage model-selection procedure. For each foundation model, the adapters and their hyperparameters were selected using validation data, and the final reported result was obtained on the held-out test set. This approach reduces dependence on any single downstream classifier and provides an estimate of the best achievable performance from each foundation model's embeddings within the evaluated adapter set. It ensures we measure the potential of each foundation model under a consistent validation procedure, rather than penalizing it for being incompatible with a specific algorithm.

We adopted a five-fold stratified cross-validation strategy on the training set. For each fold, the training data were partitioned into training and validation subsets. Within each fold, we performed a hyperparameter search using GridSearchCV (Table 5.7) to optimize model-specific parameters, such as the number of neighbors for KNN, the regularization strength C for SVM and LR, and the tree depth for RF and XGBoost. The metric for optimization was weighted F1-score, chosen to account for the natural class imbalance present in the dataset. For each foundation model, the validation results were used to choose both the

Table 5.7: Hyperparameter Search Space

Classifier	Parameter	Search Space
KNN	`n_neighbors`	`[3, 5, 7]`
Logistic Regression	`C`	`[0.1, 1.0, 10]`
SVM	`C`	`[0.1, 1.0, 10]`
	`gamma`	`['scale']`
Random Forest	`n_estimators`	`[100, 200]`
	`max_depth`	`[10, 20]`
MLP	`hidden_layer_sizes`	`[(50,), (100,)]`
	`alpha`	`[0.0001, 0.001]`
XGBoost	`n_estimators`	`[100, 200]`
	`max_depth`	`[3, 5]`

adapters and the adapter-specific hyperparameters before final test-set evaluation. The held-out test set was not used for adapter selection or hyperparameter tuning.

5.3.2.5 Hyperparameter Optimization

All adapters were trained using five-fold stratified cross-validation with GridSearchCV. Table 5.7 lists the exact hyperparameter search spaces used for each classifier.

5.3.2.6 Hierarchical Performance Evaluation

We implemented a hierarchical evaluation framework to assess the trained adapters. The primary benchmark trained adapters on the fine-grained subclass (40) labels. During inference on the held-out test set, each adapter assigned a predicted subclass label to each test image. Each predicted subclass label was then mapped to its corresponding parent category at the coarser hierarchy levels, including main class (15), superclass (4), superclass (2), and malignancy (2). Reference subclass labels were mapped using the same hierarchy. The complete subclass-to-parent mapping used for these calculations is provided in Table 5.5. Thus, a coarse-level prediction was considered correct when the predicted subclass and reference subclass belonged to the same parent category, even if the specific subclass prediction was incorrect. Summed parent-level probabilities were used only for the one-vs-rest ROC visualizations and were not used to calculate the classification metrics reported in the performance tables.

We report weighted F1-score as the primary metric for the hierarchical analysis. Given the natural class imbalance in the dataset, where common nevi outnumber rare malignancies, weighted F1-score summarizes frequency-weighted classification performance across classes. To ensure that performance on rare classes is not obscured, we additionally report balanced accuracy in the Appendix tables.

For C_l classes at hierarchy level l, weighted F1-score was calculated as

$$\mathrm{F1}_{\mathrm{weighted}} = \sum_{g=1}^{C_l} \frac{n_g}{N} \left(\frac{2\,\mathrm{Precision}_g\,\mathrm{Recall}_g}{\mathrm{Precision}_g + \mathrm{Recall}_g} \right) \tag{5.1}$$

where n_g is the number of test samples in class g, and N is the total number of test samples.

As a complementary coarse-level analysis, we also trained the same adapter families directly on the superclass (4) labels: melanocytic benign, melanocytic malignant, nonmelanocytic benign, and nonmelanocytic malignant. For this experiment, performance on the directly trained superclass (4) task was reported alongside performance at coarser hierarchical levels, obtained by mapping the predicted four-class labels to the superclass (2) and malignancy (2) levels. The same weighted F1-score and balanced accuracy definitions were used for this complementary analysis, with C_l corresponding to the number of classes at the evaluated hierarchy level. This analysis was reported separately from the 40-subclass-trained benchmark because direct coarse-label training and fine-grained hierarchical mapping evaluate different use cases.

5.3.3 Results

5.3.3.1 Hierarchical Performance Benchmark

We quantified the classification performance of the foundation models across the five levels of the clinical taxonomy. Tables 5.8 and 5.9 present held-out test performance using weighted F1-score, reporting the performance of the validation-selected best-performing adapter for each foundation model at each hierarchical level. Adapter selection was performed independently for each label set; therefore, values across hierarchy levels may originate from different adapter classifiers. Balanced accuracy was additionally reported as a complementary measure, and detailed results are provided in Appendix Tables D.1–D.12.

Our hierarchical analysis exposed a divergence in model capabilities relative to diagnostic granularity. MedImageInsight achieved the highest weighted F1-scores across all reported hierarchy levels in Table 5.8, including malignancy (2) (97.20%), superclass (2) (95.62%), superclass (4) (93.89%), main class (15) (64.89%), and subclass (40) (59.98%). MedSigLIP remained a consistently strong general medical model, ranking second on the broader malignancy (2), superclass (2), superclass (4), and main class (15) label sets in Table 5.8, with weighted F1-scores of 96.49%, 94.37%, 91.94%, and 62.34%, respectively. Figure 5.6 complements these results by showing the weighted F1-scores achieved across all evaluated adapter classifiers using bar charts.

After establishing performance under fine-label training, we next examined whether these patterns persisted as the diagnostic task shifted from fine-grained category recognition to broader superclass discrimination (Table 5.9). Table 5.9 provides a complementary coarse-label experiment to address whether model behavior at broad hierarchy levels changes when adapters are trained on coarse labels rather than only on the 40 fine-grained subclasses. In this experiment, the same frozen embeddings and adapter families were

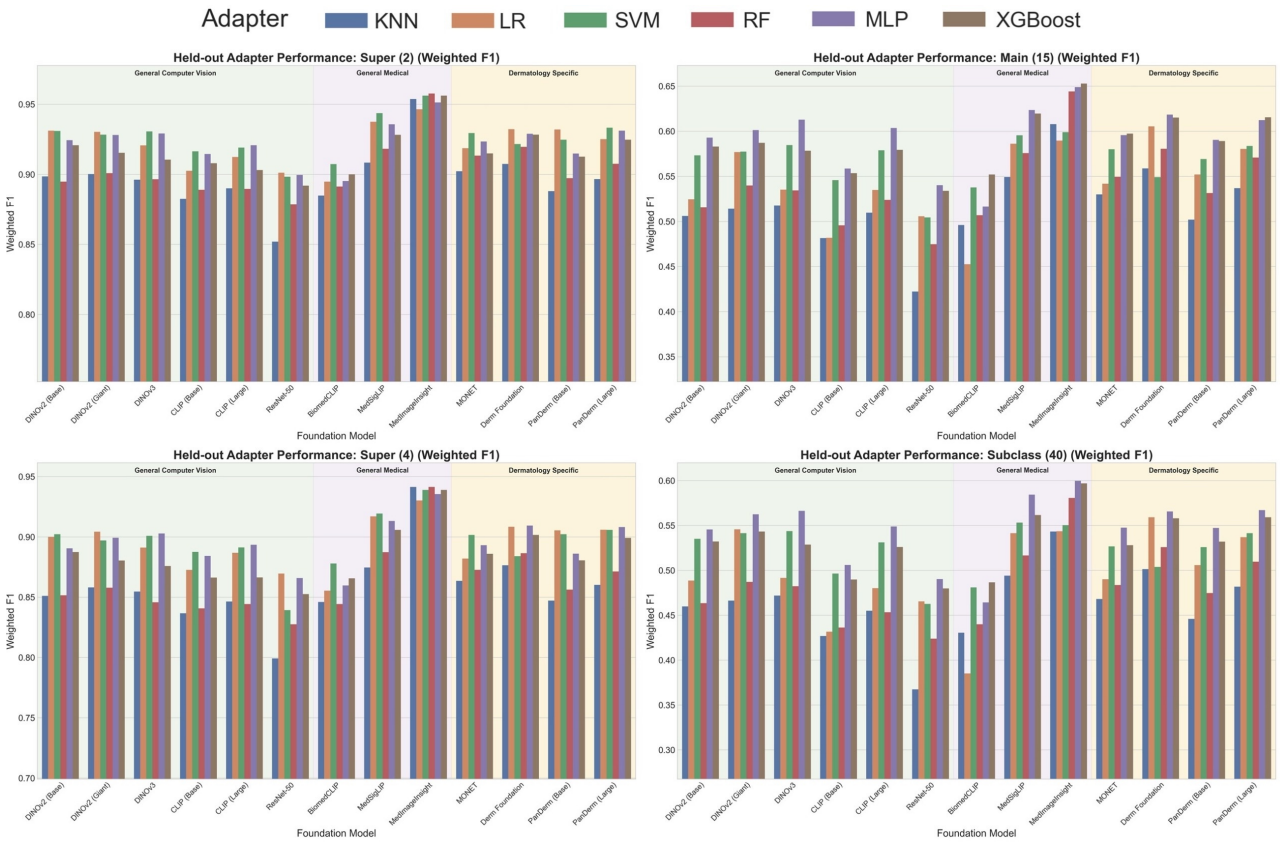


Figure 5.6: Hierarchical performance benchmark across four diagnostic label sets for fine-grained analysis: superclass (2), superclass (4), main class (15), and subclass (40). Bar charts show the weighted F1-scores achieved by the six adapter classifiers (KNN, LR, SVM, RF, MLP, XGBoost) for each foundation model. Foundation models are grouped by domain using background colors: general computer vision models are shown in green, general medical models in purple, and dermatology-specific models in yellow. MedImageInsight achieves the highest performance across the hierarchy, while MedSigLIP remains among the strongest general medical models, particularly at broader hierarchy levels.

Table 5.8: Fine-grained Foundation Model Leaderboard for the 40-subclass-trained hierarchical experiment. Weighted F1-score (%) achieved by the best-performing adapter, selected independently for each foundation model and evaluated hierarchy level. All adapters were trained to predict the 40 fine-grained subclasses, and their predictions were evaluated hierarchically at progressively coarser levels: malignancy (2) (benign vs. malignant); superclass (2) (melanocytic vs. nonmelanocytic); superclass (4) (melanocytic benign, melanocytic malignant, nonmelanocytic benign, and nonmelanocytic malignant); main class (15) (15 main diagnostic classes); and subclass (40) (40 fine-grained lesion subtypes). Models are sorted by their performance at the finest level of granularity, subclass (40). **Bold** values indicate the best performance, and *italic* values indicate the second-best performance. Malig., malignancy; SC, superclass; MC, main class; Sub., subclass.

Foundation Model	Domain	Malig. (2)	SC (2)	SC (4)	MC (15)	Sub. (40)
MedImageInsight	General Medical	**97.20%**	**95.62%**	**93.89%**	**64.89%**	**59.98%**
Derm Foundation	Dermatology	95.88%	93.22%	90.84%	61.84%	*56.57%*
DINOv2 (Giant)	General Vision	95.51%	92.83%	89.70%	60.12%	56.24%
MedSigLIP	General Medical	*96.49%*	*94.37%*	*91.94%*	*62.34%*	55.32%
DINOv3	General Vision	96.09%	93.06%	90.08%	58.47%	54.38%
PanDerm (Large)	Dermatology	96.25%	93.33%	90.58%	58.36%	54.14%
DINOv2 (Base)	General Vision	96.28%	93.10%	90.22%	57.32%	53.52%
CLIP (Large)	General Vision	96.08%	91.91%	89.12%	57.89%	53.11%
MONET	Dermatology	96.03%	92.95%	90.16%	58.01%	52.66%
PanDerm (Base)	Dermatology	96.13%	92.47%	90.22%	59.03%	52.59%
CLIP (Base)	General Vision	95.64%	91.64%	88.76%	54.57%	49.64%
ResNet-50	General Vision	95.27%	90.11%	86.95%	54.02%	49.03%
BiomedCLIP	General Medical	95.23%	90.73%	87.80%	53.77%	48.09%

Table 5.9: Coarse-level Foundation Model Leaderboard for the superclass (4)-trained hierarchical experiment. Weighted F1-score (%) achieved by the best-performing adapter, selected independently for each foundation model and evaluated hierarchy level. All adapters were trained to predict the four superclass labels, and their predictions were evaluated hierarchically at progressively coarser levels: malignancy (2) (binary benign vs. malignant); superclass (2) (melanocytic vs. nonmelanocytic); and superclass (4) (melanocytic benign, melanocytic malignant, nonmelanocytic benign, and nonmelanocytic malignant). Models are sorted by their performance on the superclass (4) task. **Bold** values indicate the best performance, and *italic* values indicate the second-best performance.

Foundation Model	Domain	Malignancy (2)	Superclass (2)	Superclass (4)
MedImageInsight	General Medical	**97.53%**	**96.28%**	**94.72%**
MedSigLIP	General Medical	96.93%	93.62%	*91.69%*
Derm Foundation	Dermatology	96.14%	*93.94%*	91.38%
DINOv3	General Vision	*97.10%*	93.33%	91.18%
PanDerm (Large)	Dermatology	96.61%	93.46%	91.06%
PanDerm (Base)	Dermatology	96.66%	93.43%	90.96%
DINOv2 (Base)	General Vision	96.43%	93.30%	90.43%
MONET	Dermatology	96.28%	92.69%	89.91%
DINOv2 (Giant)	General Vision	96.12%	92.80%	89.75%
CLIP (Large)	General Vision	96.70%	91.90%	89.55%
CLIP (Base)	General Vision	96.37%	91.27%	88.64%
ResNet-50	General Vision	95.43%	90.36%	87.35%
BiomedCLIP	General Medical	95.60%	90.43%	87.21%

trained using the superclass (4) labels, and the predicted four-class labels were subsequently mapped to the superclass (2) and malignancy (2) levels. Direct coarse-label training generally improved performance at broader hierarchy levels than the 40-subclass-derived evaluation. MedImageInsight remained the best-performing model, reaching 97.53% on malignancy (2), 96.28% on superclass (2), and 94.72% on superclass (4). MedSigLIP was also highly competitive, achieving the second-best result on superclass (4) (91.69%) and remaining close to the leading group on superclass (2) (93.62%), while Derm Foundation achieved the second-best superclass (2) result (93.94%) and DINOv3 achieved the second-best malignancy (2) score (97.10%).

5.3.3.2 Qualitative Analysis of Embedding Spaces

To qualitatively assess the semantic structure of the learned representations, we visualized the embedding spaces using t-distributed Stochastic Neighbor Embedding (t-SNE). Figure 5.7 presents a multi-panel comparison of six representative foundation models: MedSigLIP, DINOv3, MONET, Derm Foundation, PanDerm, and MedImageInsight. The projections are colored according to the 15 main classes of the taxonomy. Comprehensive t-SNE visualizations for all evaluated models are provided in Appendix Figures D.1–D.3.

5.3.3.3 Fine-Grained Error Analysis

To characterize the error patterns underlying this gap, we analyzed the specific failure modes of MedImageInsight. We selected this model for detailed inspection because it represents the most illustrative case

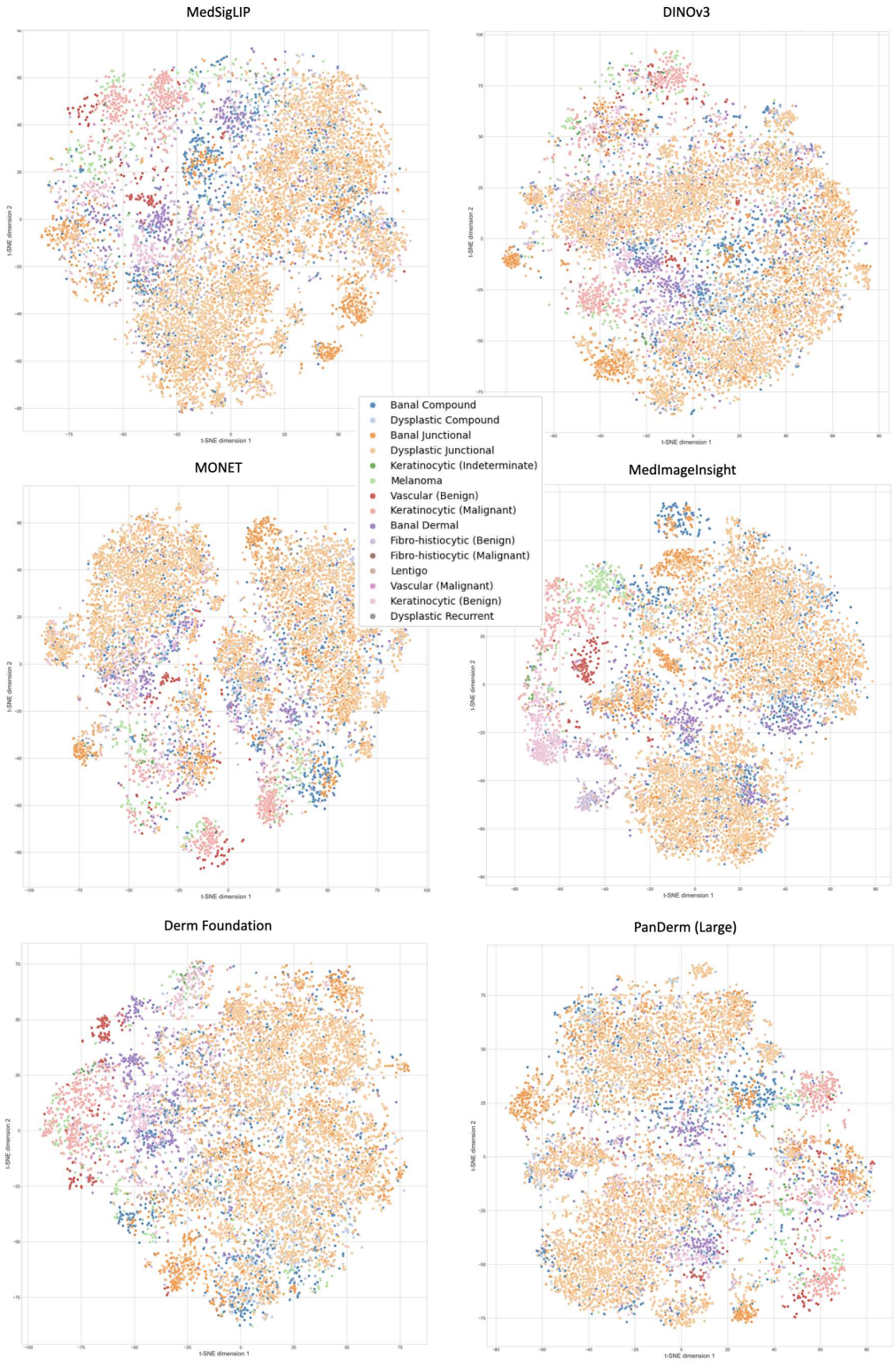


Figure 5.7: Projections are colored by the 15 main classes of the DERM12345 taxonomy. The panels display projections for MedSigLIP, DINOv3, MONET, MedImageInsight, Derm Foundation, and PanDerm (Large). Each point represents a skin lesion. The projections illustrate the two-dimensional t-SNE organization of the embeddings, showing the separation of distinct peripheral classes (e.g., Vascular) and the dense clustering of the melanocytic region.

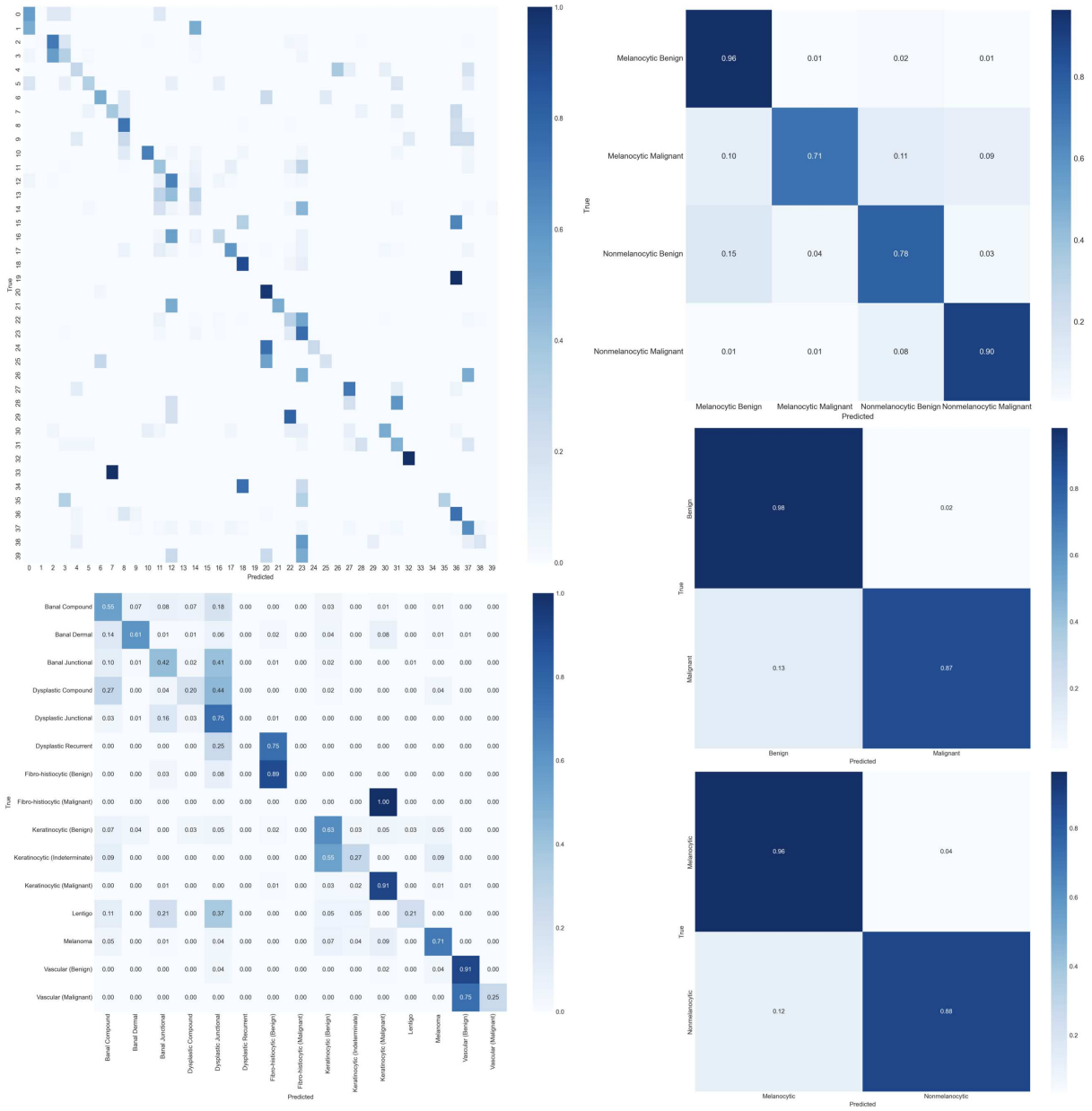


Figure 5.8: Multi-level confusion matrix analysis for MedImageInsight (MLP Adapter). Left column: Fine-grained confusion matrices at the subclass (top-left) and main class (bottom-left) levels, revealing substantial semantic overlap between visually similar lesion categories. In the subclass confusion matrix, the row and column indices (0–39) correspond to the subclass indices defined in Table 5.5, which provides the mapping from each subclass index to its subclass, main class, and superclass. The main class matrix illustrates substantial overlap between related melanocytic classes, with frequent confusion between dysplastic and banal nevi. Right column: Aggregated confusion matrices at progressively coarser levels (superclass (4), malignancy, and superclass (2)), demonstrating increasingly distinct block-diagonal patterns and improved discrimination of broad diagnostic categories. This progression highlights the degradation in performance from coarse-grained to fine-grained classification within the hierarchical taxonomy. The color intensity in each confusion matrix cell represents the normalized proportion of samples for a given true class that are predicted as the corresponding class. Darker blue indicates a higher proportion of samples, while lighter blue indicates a lower proportion. Cells along the main diagonal represent correct classifications, whereas darker cells outside the diagonal indicate more frequent misclassification between particular classes. The color bar shows the corresponding normalized values, ranging from 0 to 1.

of the granularity gap: while it achieved the highest overall performance on the malignancy (2) task (97.20%), it suffered a sharp decline in the fine-grained subclass task relative to its coarse-level performance. For qualitative error analysis, we examined the MLP adapter shown in Figure 5.8. The left panels (subclass and main class) reveal substantial classification confusion among visually similar categories. The main class (15) matrix (bottom-left) illustrates substantial overlap between related melanocytic classes, with 27.93% of Dysplastic Compound Nevi misclassified as Banal Compound Nevi. In contrast, the right panels (malignancy and superclass levels) exhibit increasingly pronounced block-diagonal structures, indicating improved performance and clearer separation at broader hierarchical levels. The model correctly classified 98.00% of benign lesions at the malignancy level. Detailed confusion matrices for all models on the main class (15) task are shown in Appendix Figures D.4–D.6. Hierarchical ROC curves are presented in Appendix Figure D.7. Although both are melanocytic nevi, dysplastic nevi are associated with increased melanoma risk at the patient level, whereas banal compound nevi generally lack the same risk association. The model's inability to separate them, despite its high binary accuracy, underscores that its embeddings cluster nevi together semantically but fail to resolve the intra-class visual continuum required for differential diagnosis.

5.3.4 Discussion

A key finding of this benchmark is the marked divergence in model performance across the diagnostic hierarchy, a phenomenon we term the granularity gap. MedImageInsight, a model pretrained on large-scale medical images with associated text and labels, achieved the best performance across the hierarchy. This may suggest that the model has encoded higher-level semantic concepts that help distinguish cancer from non-cancer based on linguistic patterns present in the training data. However, the reduction in absolute performance from the broad malignancy (2) task to the fine-grained subclass (40) task indicates that fine-grained visual discrimination remains substantially more difficult than coarse screening. This difficulty is illustrated by the confusion matrix analysis, which revealed a high error rate in distinguishing Dysplastic Compound Nevi from Banal Compound Nevi. Clinically, this is a critical distinction, as dysplastic nevi serve as potential markers for melanoma risk. The model's difficulty in resolving these visually similar yet biologically distinct entities suggests that its embedding space is better aligned with broad semantic categories than with the subtle textural features required for differential diagnosis.

It is important to contextualize these findings within our specific hierarchical training strategy. In the primary 40-subclass benchmark, unlike standard benchmarks that might train separate models for each task (e.g., a dedicated binary classifier), we trained adapters on the fine-grained 40-class labels. For each image, the adapter's predicted subclass label was mapped to its corresponding parent category at each coarser taxonomic level, including main class, superclass, and malignancy; the reference subclass label was mapped using the same hierarchy. Subclass probabilities were not aggregated to determine the classification predictions reported in the performance tables. A correct binary prediction therefore occurred when the predicted subclass and reference subclass belonged to the same malignancy category, even if the specific subclass prediction was incorrect. Consequently, the high binary weighted F1-scores achieved

by models such as MedImageInsight and MedSigLIP (97.20% and 96.49%, respectively) indicate that their predicted subclass labels frequently remained within the correct malignancy branch of the taxonomy.

Our results also challenge the assumption that domain specificity is the sole determinant of performance. MedImageInsight and MedSigLIP, both general medical foundation models, performed strongly across the hierarchy, while the dermatology-specific Derm Foundation model achieved the second-best result on the finest subclass (40) task. The top-three ranking for subclass (40) spans three model domains: MedImageInsight (general medical), Derm Foundation (dermatology-specific), and DINOv2 (Giant) (general vision). This mixed ranking may reflect differences in representation scale, pretraining diversity, and compatibility between the embedding space and a downstream adapter. MedSigLIP was particularly consistent at broader hierarchy levels, ranking second in Table 5.8 for malignancy (2), superclass (2), superclass (4), and main class (15), and fourth for subclass (40). This pattern suggests that large-scale general medical pretraining can yield representations that remain highly competitive in dermatology, especially for broader diagnostic groupings, while dermatology-specific pretraining may still be valuable for fine-grained lesion discrimination. The performance of DINOv2 (Giant), which ranked third on the subclass (40) task, further supports the transferability of large-scale natural-image representations to dermatologic classification when paired with a suitable adapter. Overall, the results indicate that researchers cannot simply assume a specialized model is inherently superior; thorough verification against robust generalist baselines remains essential.

The embedding visualizations highlight a key challenge in dermatologic modeling: substantial overlap among closely related melanocytic lesion classes. In all six projections, the core melanocytic classes, including Banal Compound, Dysplastic Compound, and Dysplastic Junctional nevi, formed a dense, overlapping central cluster. This suggests that these foundation models, regardless of domain pretraining, do not show clear visual separation for the most clinically challenging cases without supervised adaptation. Clinically, these distinctions often depend on subtle combinations of dermoscopic features rather than a single dominant visual cue. Relevant features include pigment-network atypia, lesion asymmetry, irregular dots and globules, streaks, color variation, regression structures, and blue-white veil. The overlap observed in the t-SNE projections may therefore indicate that the embedding spaces do not cleanly separate these fine-grained feature combinations before task-specific supervision. However, because t-SNE is a two-dimensional projection, this visualization cannot prove that any specific dermoscopic feature is absent from the original high-dimensional embeddings; these features should be interpreted as plausible clinical explanations for the observed overlap rather than direct measurements of model encoding.

Distinct patterns also emerged in the peripheral clusters. MedImageInsight showed the clearest separation for visually distinct lesion groups, including Vascular and Keratinocytic classes, which formed compact clusters away from the central melanocytic region. This suggests stronger separation of visually distinct biological groups within the projected embedding space. In contrast, MedSigLIP and DINOv3 showed more diffuse and less compact cluster structures. While less visually distinct in 2D projection, this dispersed structure may preserve richer intra-class variance, potentially aiding the fine-grained discrimination of

visually similar subtypes during the supervised adaptation phase. Across all foundation models, banal and dysplastic nevi formed dense, overlapping clusters with no clear visual separation in the t-SNE projections. The comparative adapter analysis further confirms this difficulty. Overall, supervised classifiers provided substantially better fine-grained discrimination than would be suggested by the visual overlap observed in the two-dimensional t-SNE projections, highlighting that t-SNE separation should not be interpreted as a direct measure of high-dimensional class separability.

We acknowledge several limitations in this study. First, the DERM12345 dataset was collected exclusively from multiple clinical sources in Türkiye. Fitzpatrick skin type was not available in the dataset metadata, and aggregate skin type percentages were not reported in the source publication; consequently, skin-type-stratified performance could not be evaluated. Collection from several centers in Türkiye also does not establish global demographic representativeness, potentially affecting the generalizability of the embeddings to diverse populations. Second, our evaluation is restricted to dermoscopic images. While this modality offers high diagnostic precision, the performance of these foundation models on clinical photography remains to be verified. Future work should extend this hierarchical benchmarking framework to multicenter datasets with explicit demographic and skin type annotation, as well as diverse imaging modalities, to ensure broader clinical applicability.

In conclusion, this benchmark demonstrates that the optimal choice of foundation model in dermatology depends strongly on the diagnostic granularity of the target task. The choice of vision encoder should therefore be guided by the specific clinical use case. By evaluating embeddings across the full hierarchy, from binary malignancy to 40 distinct subclasses, we provide a roadmap for selecting the appropriate visual encoder for diverse dermatologic applications, ensuring that AI tools align with the nuanced reality of clinical diagnosis.

5.3.5 Data availability

The DERM12345 dataset used in this study is publicly available through Harvard Dataverse (https://doi.org/10.7910/DVN/DAXZ7P).

Chapter 6

Conclusion

This thesis set out to develop a practical, explainable, resource-efficient, and privacy-aware artificial intelligence ecosystem for dermatology. The central argument of the thesis is that clinically useful dermatology AI cannot be reduced to a single model, dataset, or benchmark. Instead, it requires a connected pipeline that begins with high-quality and clinically meaningful data, expands into multimodal supervision, develops models appropriate for different dermatologic workflows, and evaluates those systems using benchmarks that reflect diagnostic hierarchy, reasoning, interpretability, and deployment constraints.

The work presented in this thesis therefore addresses dermatology AI across three main layers. First, the data layer focused on the development of structured and scalable datasets, including a hierarchical dermoscopic dataset for skin lesion analysis and a synthetic multimodal image–text dataset for broader dermatology reasoning. Second, the model layer explored a progression of AI methods, from conventional convolutional neural networks to object detection, local language models, and multimodal vision-language models. Third, the evaluation layer focused on benchmarking strategies that go beyond conventional classification accuracy, including hierarchical foundation-model evaluation and clinician-annotated visual question-answering benchmarks.

Together, these contributions demonstrate that the future of dermatology AI will depend not only on stronger models, but also on better data structures, more interpretable outputs, more realistic evaluation, and deployment strategies that respect the privacy and operational constraints of healthcare environments.

6.1 Conclusions and Limitations

Chapter 3: Datasets

The dataset contributions in this thesis address two major needs in dermatology AI: structured dermoscopic data for skin lesion and skin cancer analysis, and multimodal image–text data for broader dermatologic reasoning.

1. Section 3.2—DERM12345

DERM12345 provides a large hierarchical dermoscopic dataset with 40 subclasses, enabling fine-grained classification and hierarchical evaluation. This is important because many previous dermatology AI datasets focus on broad diagnostic categories or binary malignancy classification, whereas clinical dermatology often requires more detailed subclass-level reasoning.

However, DERM12345 also has important limitations. The dataset is retrospective and therefore shaped by the patient populations, imaging devices, and clinical practices of the contributing centers. Although it includes a detailed diagnostic hierarchy, some rare subclasses remain underrepresented because uncommon dermatologic entities are difficult to collect at scale. In addition, detailed skin-tone or demographic information is not available for all images, limiting the ability to perform comprehensive subgroup fairness analysis. Future expansions of hierarchical dermoscopic datasets should therefore aim to include broader demographic metadata, more diverse acquisition settings, and larger numbers of rare lesions.

2. Section 3.3—DermaSynth

DermaSynth addresses a different but equally important limitation: the lack of large-scale image–text datasets in dermatology. By generating synthetic image–text pairs from open-access dermatology datasets and PubMed-derived resources, DermaSynth provides scalable multimodal supervision for dermatology vision-language model development. This is particularly valuable because dermatology images often lack formal reports, unlike radiology or pathology. Nevertheless, synthetic data generation depends on the quality of source metadata and the reliability of teacher model outputs. Although metadata-informed prompt construction and explicit image-focused instructions may reduce unsupported generation, synthetic descriptions may still contain errors, simplified reasoning, or incomplete medical context. Therefore, synthetic datasets should be treated as a scalable training resource rather than a substitute for clinician-validated data.

A further limitation is that dermatology is closely connected with other medical domains, including histopathology, surgery, rheumatology, infectious diseases, allergy, oncology, and internal medicine. DermaSynth is primarily dermatology-focused, but future multimodal datasets should include richer medical context across related specialties. This would allow models to reason more effectively in cases where skin findings are part of systemic disease, treatment complications, surgical decision-making, or histopathologic confirmation.

Chapter 4: AI

1. Section 4.2—Classification Models

The model-development chapter began with CNN-based baseline models trained on DERM12345. These models established a reference point for conventional image-level dermoscopic classification. The results demonstrated that fine-grained 40-subclass classification is substantially more challenging than broad or binary classification. This supports one of the core arguments of the thesis: dermatology AI should not be evaluated only on simplified diagnostic tasks. However, CNN models remain limited because they generally provide black-box outputs in the form of class probabilities. They do not naturally explain which dermatologic structures support a prediction or why a lesion belongs to a particular category. More advanced foundation-model evaluations using the DERM12345 test split are therefore discussed in the benchmarking chapter.

2. Section 4.3—Object Detection Models

To move beyond image-level classification, this thesis explored object detection for KOH microscopy in dermatophytosis. This study demonstrated that object detection can provide more interpretable outputs by localizing fungal hyphae and explicitly distinguishing them from artifacts. This approach is particularly appropriate for dermatologic microscopy, where the clinical task is not simply to classify an image but to identify diagnostically relevant structures within the field. However, the study remains limited by the diversity of microscopy settings. Future validation should include images from different microscopes, laboratories, staining conditions, sample preparation protocols, and clinical centers. In addition, automated or semi-automated microscopy systems could further improve scalability by enabling high-throughput image acquisition and analysis.

3. Sections 4.4 and 4.5—Language Models

The pemphigus language-model study demonstrated the feasibility of local, privacy-preserving longitudinal clinical data retrieval. This is important because chronic dermatologic diseases generate extensive clinical documentation that is difficult to review comprehensively during routine care. A locally deployed model can reduce dependence on cloud-based inference and may help protect patient privacy. However, the model used in this study was a general-purpose open-source language model rather than a dermatology-specific or medically pretrained model. This may lead to basic interpretation errors, inconsistent handling of clinical definitions, and the need for significant prompt refinement or post-processing. Although local deployment improves privacy, it does not remove the need for governance, auditability, reliability testing, bias assessment, and clinician oversight.

The multimodal language-model studies represent a further step toward advanced dermatology AI. The focused bacterial skin disease model showed that compact vision-language models can be adapted to specific dermatologic tasks using synthetic supervision [151]. DermatoLlama then extended this idea to a broader dermatology setting using large-scale synthetic multimodal data. These studies show the potential of resource-efficient multimodal models for report generation, visual question answering, and dermatology-specific reasoning. However, DermatoLlama was fine-tuned using parameter-efficient methods rather than trained through full pretraining on a large dermatology

corpus. Pretraining or continued pretraining on larger, higher-quality dermatology image–text corpora could substantially improve performance, robustness, and domain-specific reasoning.

Chapter 5: Benchmarks

The benchmarking contributions in this thesis were designed to evaluate dermatology AI systems more meaningfully than standard classification benchmarks. DERM12345 provides a hierarchical test split for fine-grained dermoscopic evaluation, while the foundation-model benchmark demonstrates that models performing strongly on broad malignancy classification may not necessarily perform best on subclass-level diagnosis. This supports the need for evaluation across multiple levels of dermatologic granularity.

Section 5.2—DermaBench

DermaBench extends the evaluation framework toward multimodal reasoning by providing visual question-answering annotations for both clinical and dermoscopic dermatology images. However, DermaBench also has limitations. The clinical and dermoscopic images are derived from open-source datasets and therefore do not include full patient histories, laboratory results, treatment information, or longitudinal clinical context. This limits the ability to evaluate models on complete real-world dermatology reasoning. In addition, although DermaBench includes structured morphology and image-analysis questions, future versions should include more conversational, reasoning-oriented, and clinically contextual questions. For example, future benchmarks should test whether models can answer questions such as why a lesion is suspicious, what features support a diagnosis, which differential diagnoses should be considered, and how uncertainty should be communicated.

More detailed localized annotations are also needed. Current explainability approaches often show heatmaps or attention regions, but clinically meaningful dermatology AI should ideally identify specific features, such as a blue-whitish veil in a particular region of a lesion, irregular pigment network, regression structures, follicular pigmentation, or vascular patterns. Future benchmarks should therefore include concept annotations, localization labels, segmentation masks, and region-level reasoning tasks, especially for dermoscopic skin cancer analysis.

Finally, all evaluations in this thesis remain primarily offline and retrospective. Real-world deployment would require prospective testing, clinical workflow integration, clinician training, safety monitoring, calibration analysis, and assessment of effects on patient outcomes. Future benchmarks should therefore evaluate not only answer correctness, but also uncertainty, refusal behavior, calibration, consistency across patient groups, safety, robustness, and clinical usefulness.

6.2 Future Work

Chapter 3: Datasets

1. Expanding Datasets for Specialized AI Models

Future work should expand dermatology datasets in both depth and breadth. For specialized AI systems, larger datasets are needed for rare dermatologic diseases, uncommon lesion subtypes, and clinically challenging presentations. Many common dermatologic diseases are underrepresented in published datasets because they are well known to clinicians and less frequently reported in the literature. As a result, open-source data often overrepresents rare, severe, or publication-worthy cases, while common outpatient dermatology remains underrepresented. Future dataset-building efforts should therefore include routine clinical images, common inflammatory diseases, infectious dermatoses, hair and nail disorders, pediatric dermatology, and treatment-related skin findings.

2. Expanding Datasets for Multimodal Language Models

For multimodal language models, future datasets should include more conversational image–text data, visual question-answering examples, differential diagnosis discussions, patient history, and EHR-like longitudinal context. Dermatology models should learn not only to classify images but also to reason over clinical information, explain morphology, summarize patient histories, and produce concise reports.

3. Large-Scale Text and Image Datasets for Dermatology

Large-scale text and image datasets derived from open-access literature, PubMed Central, clinical atlases, and institutional datasets could support the next generation of dermatology foundation models. However, such datasets should include careful provenance tracking, licensing review, expert validation, and safeguards against hallucinated or low-quality synthetic supervision.

4. Improved Synthetic Datasets

Synthetic data will remain important because it is not feasible to collect sufficiently diverse real-world datasets for every dermatologic disease and every clinical scenario. Future work should explore improved synthetic data generation through stronger teacher models, multimodal self-instruction, expert-guided seed tasks, and quality-control pipelines.

5. Synthetic Image Datasets for Dermatology

Synthetic image generation may also become increasingly useful for augmenting rare diseases, underrepresented skin tones, and difficult visual patterns [117]. However, synthetic images must be validated carefully to ensure that they are clinically realistic and do not introduce misleading morphology or bias.

6. Concept and Localized Annotations for Dermoscopic Images

Future datasets should become more detailed, standardized, and clinically realistic. For dermoscopic skin cancer analysis, concept-level and localized annotations are especially important. Datasets should contain not only whether a model predicts melanoma, but also whether it identifies the relevant dermoscopic features and localizes them correctly. This requires moving beyond image-level labels

toward region-level annotations, bounding boxes, and segmentation masks for clinically meaningful structures. For example, a clinically meaningful dataset should be able to evaluate whether a model identifies and segments a blue-whitish veil in the upper-left region of a lesion, localizes atypical vessels in a specific area, or distinguishes regression structures, irregular network, dots, globules, and ulceration within the lesion. Such localized and segmented concept annotations would enable more interpretable evaluation by linking diagnostic predictions to visible dermatologic evidence.

Chapter 4: AI

1. Hybrid AI Systems Integrated with Language Models

A major future direction is the integration of specialized dermatology AI models with language models. For example, a skin cancer classifier, a dermoscopic feature detector, or a fungal object detector could be combined with a language model that explains the result in clinically meaningful terms. Such hybrid systems could provide both model performance and interpretability. Instead of only predicting a diagnosis, the system could state which visual structures were detected, where they were located, and how they support the final interpretation.

This approach could help bridge the gap between black-box classification and clinically useful explanation. In dermoscopy, future models should be able to identify and localize features such as blue-whitish veil, irregular network, dots and globules, regression structures, atypical vessels, follicular pigmentation, and ulceration. In microscopy, future models could estimate fungal burden or quantify treatment response. In chronic diseases, language models could summarize treatment trajectories and detect missing safety information.

2. Large Dermatology Language and Vision-Language Models

Future work should continue the development of dermatology-specific language and vision-language models. DermatoLlama represents an early step in this direction, but further progress will likely require larger and more diverse datasets, better pretraining strategies, improved clinical grounding, and more rigorous evaluation. Dermatology-specific models could be developed for multiple subfields, including pigmented lesions, inflammatory dermatoses, infectious diseases, hair disorders, nail disorders, pediatric dermatology, dermatopathology, and autoimmune blistering diseases.

A broader dermatology language model could support text summarization, long-term follow-up, clinical decision support, patient education, data extraction, and research analysis. Compared with general medical models, dermatology-specialized models may reduce hallucinations and misinformation by learning domain-specific terminology, morphology, disease patterns, and clinical workflows. Future work could develop such models directly or compress larger dermatology models into smaller resource-efficient versions suitable for local deployment.

3. Agentic and Modular Dermatology AI Systems

Future dermatology AI systems may benefit from modular or agentic architectures [282, 283]. Instead of relying on a single general-purpose model, different specialized models or agents could be developed for different tasks. For example, one agent could evaluate dermoscopic structures,

another could summarize patient history, another could retrieve guideline information, another could assess uncertainty, and another could act as an LLM-as-a-judge to critique or validate the output. Such systems could coordinate multiple specialized components to improve reliability and interpretability. A skin cancer agent, trichoscopy agent, dermatopathology agent, inflammatory disease agent, and safety-checking agent could each contribute to a final decision-support workflow. However, these systems require careful design, transparent orchestration, robust evaluation, and strong governance. They should be developed as clinician-assistive tools rather than autonomous diagnostic systems.

Chapter 5: Benchmarks

1. More Detailed and Standardized Benchmarks

For general dermatologic disease analysis, future benchmarks should include common diseases, rare diseases, clinical photographs, patient history, follow-up data, and reasoning questions. Benchmarks should move beyond multiple-choice question answering and include conversational assessment, report generation, uncertainty handling, safety responses, and refusal behavior. They should also test consistency across skin tones, anatomical sites, imaging devices, disease categories, and patient groups.

Finally, benchmarking should be tied to real clinical usefulness. Dermatology AI systems should be evaluated not only for producing plausible text but also for supporting clinical decisions, reducing workload, improving documentation, increasing diagnostic consistency, and avoiding unsafe recommendations.

6.3 Final Remarks

This thesis demonstrates that trustworthy dermatology AI requires an ecosystem rather than a standalone algorithm. High-granularity datasets make clinically meaningful tasks possible. Synthetic supervision helps scale multimodal learning. Domain-adapted models make deployment more practical. Object detection provides more interpretable visual evidence. Local language models make privacy-preserving clinical retrieval feasible. Expert-designed benchmarks make evaluation more realistic.

Overall, the work presented in this thesis provides several key steps toward the wider adoption of explainable, resource-efficient, privacy-preserving, and multimodal artificial intelligence systems in dermatology. The most important lesson is that clinical usefulness emerges from the alignment of data, models, evaluation, and deployment with the realities of dermatologic practice.

Bibliography

[1] Chante Karimkhani, Robert P Dellavalle, Luc E Coffeng, et al. “Global Skin Disease Morbidity and Mortality: An Update From the Global Burden of Disease Study 2013”. In: *JAMA Dermatology* 153.5 (2017), pp. 406–412. DOI: 10.1001/jamadermatol.2016.5538.

[2] Roderick J Hay, Nicole E Johns, Hywel C Williams, et al. “The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions”. In: *Journal of Investigative Dermatology* 134.6 (2014), pp. 1527–1534. DOI: 10.1038/jid.2013.446.

[3] C Flohr and R Hay. “Putting the burden of skin diseases on the global map”. In: *British Journal of Dermatology* 184.2 (2021), pp. 189–190. DOI: 10.1111/bjd.19704.

[4] Andrew Y Finlay. “The burden of skin disease: quality of life, economic aspects and social issues”. In: *Clinical Medicine* 9.6 (2009), pp. 592–594. DOI: 10.7861/clinmedicine.9-6-592.

[5] MR Laughter, Mayra BC Maymone, Soudeh Mashayekhi, et al. “The global burden of atopic dermatitis: lessons from the Global Burden of Disease Study 1990–2017”. In: *British Journal of Dermatology* 184.2 (2021), pp. 304–309. DOI: 10.1111/bjd.19580.

[6] AM Layton, D Thiboutot, and J Tan. “Reviewing the global burden of acne: how could we improve care to reduce the burden?” In: *British Journal of Dermatology* 184.2 (2021), pp. 219–225. DOI: 10.1111/bjd.19477.

[7] IYK Iskandar, R Parisi, CEM Griffiths, DM Ashcroft, and Global Psoriasis Atlas. “Systematic review examining changes over time and variation in the incidence and prevalence of psoriasis by age and gender”. In: *British Journal of Dermatology* 184.2 (2021), pp. 243–258. DOI: 10.1111/bjd.19169.

[8] Margarida Gonçalo, A Gimenéz-Arnau, M Al-Ahmad, et al. “The global burden of chronic urticaria for the patient and society”. In: *British Journal of Dermatology* 184.2 (2021), pp. 226–236. DOI: 10.1111/bjd.19561.

[9] T Loney, MS Paulo, A Modenese, et al. “Global evidence on occupational sun exposure and keratinocyte cancers: a systematic review”. In: *British Journal of Dermatology* 184.2 (2021), pp. 208–218. DOI: 10.1111/bjd.19152.

[10] V Cox, LC Fuller, D Engelman, A Steer, and RJ Hay. “Estimating the global burden of scabies: what else do we need?” In: *British Journal of Dermatology* 184.2 (2021), pp. 237–242. DOI: 10.1111/bjd.19170.

[11] Divya Seth, Khatiya Cheldize, Danielle Brown, and Esther E Freeman. “Global burden of skin disease: inequities and innovations”. In: *Current Dermatology Reports* 6.3 (2017), pp. 204–210. DOI: 10.1007/s13671-017-0192-7.

[12] RJ Hay, M Augustin, CEM Griffiths, et al. “The global challenge for skin health”. In: *British Journal of Dermatology* 172.6 (2015), pp. 1469–1472. DOI: 10.1111/bjd.13854.

[13] Hyeon Ki Jeong, Christine Park, Ricardo Henao, and Meenal Kheterpal. “Deep learning in dermatology: a systematic review of current approaches, outcomes, and limitations”. In: *JID Innovations* 3.1 (2023), p. 100150. DOI: 10.1016/j.xjidi.2022.100150.

[14] Shern Ping Choy, Byung Jin Kim, Alexandra Paolino, et al. “Systematic review of deep learning image analyses for the diagnosis and monitoring of skin disease”. In: *npj Digital Medicine* 6.1 (2023), p. 180. DOI: 10.1038/s41746-023-00914-8.

[15] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. “Large language models in medicine”. In: *Nature Medicine* 29.8 (2023), pp. 1930–1940. DOI: 10.1038/s41591-023-02448-8.

[16] William J Nahm, Nayyab Sohail, Joshua Burshtein, Mohamad Goldust, and Maria Tsoukas. “Artificial intelligence in dermatology: a comprehensive review of approved applications, clinical implementation, and future directions”. In: *International Journal of Dermatology* 64.9 (2025), pp. 1568–1583. DOI: 10.1111/ijd.17847.

[17] Hensin Tsao, Jeannette M Olazagasti, Kelly M Cordoro, et al. “Early detection of melanoma: reviewing the ABCDEs”. In: *Journal of the American Academy of Dermatology* 72.4 (2015), pp. 717–723. DOI: 10.1016/j.jaad.2015.01.025.

[18] NWJ Smeets, DIM Kuijpers, Patricia Nelemans, et al. “Mohs’ micrographic surgery for treatment of basal cell carcinoma of the face–results of a retrospective study and review of the literature”. In: *British Journal of Dermatology* 151.1 (2004), pp. 141–147. DOI: 10.1111/j.1365-2133.2004.06047.x.

[19] Arif M Aslam and Anand N Patel. “Facial cutaneous squamous cell carcinoma”. In: *BMJ* 352 (2016), p. i1513. DOI: 10.1136/bmj.i1513.

[20] Jean L. Bolognia, Julie V. Schaffer, and Lorenzo Cerroni, eds. *Dermatology*. 5th ed. Elsevier, 2024. ISBN: 9780702082252.

[21] Ashfaq A. Marghoob, Ralph P. Braun, and Natalia Jaimes, eds. *Atlas of Dermoscopy*. 3rd ed. CRC Press, 2022. ISBN: 9781138595989. DOI: 10.1201/9780429487873.

[22] Harald Kittler, H Pehamberger, K Wolff, and MJTIO Binder. “Diagnostic accuracy of dermoscopy”. In: *The Lancet Oncology* 3.3 (2002), pp. 159–165. DOI: 10.1016/s1470-2045(02)00679-4.

[23] Giuseppe Argenziano, C Catricalà, M Ardigo, et al. “Seven-point checklist of dermoscopy revisited”. In: *British Journal of Dermatology* 164.4 (2011), pp. 785–790. DOI: 10.1111/j.1365-2133.2010.10194.x.

[24] Franz Nachbar, Wilhelm Stolz, Tanja Merkle, et al. “The ABCD rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions”. In: *Journal of the American Academy of Dermatology* 30.4 (1994), pp. 551–559. DOI: 10.1016/S0190-9622(94)70061-3.

[25] ME Vestergaard, PHPM Macaskill, PE Holt, and SW Menzies. “Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting”. In: *British Journal of Dermatology* 159.3 (2008), pp. 669–676. DOI: 10.1111/j.1365-2133.2008.08713.x.

[26] Giuseppe Argenziano, H Peter Soyer, Sergio Chimenti, et al. “Dermoscopy of pigmented skin lesions: results of a consensus meeting via the Internet”. In: *Journal of the American Academy of Dermatology* 48.5 (2003), pp. 679–693. DOI: 10.1067/mjd.2003.281.

[27] Rayan Krishnan, Pranav Rajpurkar, and Eric J Topol. “Self-supervised learning in medicine and healthcare”. In: *Nature Biomedical Engineering* 6.12 (2022), pp. 1346–1352. DOI: 10.1038/s41551-022-00914-1.

[28] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. “A survey of transfer learning”. In: *Journal of Big Data* 3.1 (2016), p. 9. DOI: 10.1186/s40537-016-0043-6.

[29] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, et al. “Foundation models for generalist medical artificial intelligence”. In: *Nature* 616.7956 (2023), pp. 259–265. DOI: 10.1038/s41586-023-05881-4.

[30] Wasif Khan, Seowung Leem, Kyle B See, Joshua K Wong, Shaoting Zhang, and Ruogu Fang. “A Comprehensive Survey of Foundation Models in Medicine”. In: *IEEE Reviews in Biomedical Engineering* 19 (2026), pp. 283–304. DOI: 10.1109/rbme.2025.3531360.

[31] Syed Muhammad Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, and Muhammad Khurram Khan. "Medical Image Analysis using convolutional neural networks: a review". In: *Journal of Medical Systems* 42.11 (2018), p. 226. DOI: 10.1007/s10916-018-1088-1.

[32] Kai Han, Yunhe Wang, Hanting Chen, et al. "A survey on vision transformer". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.1 (2022), pp. 87–110. DOI: 10.1109/TPAMI.2022.3152247.

[33] Yian Zhao, Wenyu Lv, Shangliang Xu, et al. "DETRs Beat YOLOs on Real-time Object Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 16965–16974. DOI: 10.1109/cvpr52733.2024.01605.

[34] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. "End-to-End Object Detection with Transformers". In: *Computer Vision - ECCV 2020*. 2020, pp. 213–229. DOI: 10.1007/978-3-030-58452-8_13.

[35] Rana Gursoy, Abdurrahim Yilmaz, Baris Kizilyaprak, et al. "Artifact-Aware Fungal Detection in Dermatophytosis: A Transformer-Based Approach for KOH Microscopy". In: *Bioengineering* 13.5 (2026), p. 591. DOI: 10.3390/bioengineering13050591.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[37] Larry R. Medsker and Lakhmi C. Jain, eds. *Recurrent Neural Networks: Design and Applications*. CRC Press, 2000. ISBN: 9780849371813.

[38] Alex Graves. "Long short-term memory". In: *Supervised sequence labelling with recurrent neural networks*. Springer, 2012, pp. 37–45. DOI: 10.1007/978-3-642-24797-2_4.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. "Attention is all you need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 6000–6010. URL: https://dl.acm.org/doi/10.5555/3295222.3295349.

[40] Shengyu Zhang, Linfeng Dong, Xiaoya Li, et al. "Instruction tuning for large language models: A survey". In: *ACM Computing Surveys* 58.169 (7 2026), pp. 1–36. DOI: 10.1145/3777411.

[41] Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. "Retrieval-augmented generation for knowledge-intensive NLP tasks". In: *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 793. 2020, pp. 9459–9474. URL: https://dl.acm.org/doi/abs/10.5555/3495724.3496517.

[42] Wallace H Clark Jr, David E Elder, DuPont Guerry IV, et al. "Model predicting survival in stage I melanoma based on tumor progression". In: *JNCI: Journal of the National Cancer Institute* 81.24 (1989), pp. 1893–1904. DOI: 10.1093/jnci/81.24.1893.

[43] T Schindewolf, Wilhelm Stolz, Rene Albert, Wolfgang Abmayr, and Harry Harms. "Classification of melanocytic lesions with color and texture analysis using digital image processing." In: *Analytical and Quantitative Cytology and Histology* 15.1 (1993), pp. 1–11. URL: https://pubmed.ncbi.nlm.nih.gov/8471104/.

[44] Michael Binder, Heinz Kittler, A Seeber, A Steiner, H Pehamberger, and K Wolff. "Epiluminescence microscopy-based classification of pigmented skin lesions using computerized image analysis and an artificial neural network". In: *Melanoma Research* 8.3 (1998), pp. 261–266. DOI: 10.1097/00008390-199806000-00009.

[45] Barbara Rosado, Scott Menzies, Alexandra Harbauer, et al. "Accuracy of computer diagnosis of melanoma: a quantitative meta-analysis". In: *Archives of Dermatology* 139.3 (2003), pp. 361–367. DOI: 10.1001/archderm.139.3.361.

[46] Marco Burroni, Rosamaria Corona, Giordana Dell'Eva, et al. "Melanoma computer-aided diagnosis: reliability and feasibility study". In: *Clinical Cancer Research* 10.6 (2004), pp. 1881–1886. DOI: 10.1158/1078-0432.ccr-03-0039.

[47] Ammara Masood and Adel Ali Al-Jumaily. "Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms". In: *International Journal of Biomedical Imaging* 2013.1 (2013), p. 323268. DOI: 10.1155/2013/323268.

[48] Abder-Rahman H Ali, Jingpeng Li, and Guang Yang. "Automating the ABCD rule for melanoma detection: a survey". In: *IEEE Access* 8 (2020), pp. 83333–83346. DOI: 10.1109/access.2020.2991034.

[49] Andre Esteva, Brett Kuprel, Roberto A Novoa, et al. "Dermatologist-level classification of skin cancer with deep neural networks". In: *Nature* 542.7639 (2017), pp. 115–118. DOI: 10.1038/nature21056.

[50] Noel Codella, Junjie Cai, Mani Abedini, Rahil Garnavi, Alan Halpern, and John R Smith. "Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images". In: *Machine Learning in Medical Imaging*. Springer International Publishing, 2015, pp. 118–126. DOI: 10.1007/978-3-319-24888-2_15.

[51] David Gutman, Noel CF Codella, Emre Celebi, et al. *Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC)*. 2016. arXiv: 1605.01397.

[52] Noel CF Codella, David Gutman, M Emre Celebi, et al. "Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)". In: *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE. 2018, pp. 168–172. DOI: 10.1109/isbi.2018.8363547.

[53] Noel Codella, Veronica Rotemberg, Philipp Tschandl, et al. *Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)*. 2019. arXiv: 1902.03368.

[54] Nimra Shehzadi, Arshia Rehman, Saeeda Naz, Saba Rehman, and Fahmi Khalifa. "Machine learning and deep learning based psoriasis recognition system: evaluation, management, prognosis—where we are and the way to the future". In: *Artificial Intelligence Review* 58.9 (2025), p. 263. DOI: 10.1007/s10462-025-11195-w.

[55] Leo Huang, Wai Hoh Tang, Rahman Attar, et al. "Remote assessment of eczema severity via AI-powered skin image analytics: a systematic review". In: *Artificial Intelligence in Medicine* 156 (2024), p. 102968. DOI: 10.1016/j.artmed.2024.102968.

[56] Fang Cao, Yujie Yang, Cui Guo, Hui Zhang, Qianying Yu, and Jing Guo. "Advancements in artificial intelligence for atopic dermatitis: diagnosis, treatment, and patient management". In: *Annals of Medicine* 57.1 (2025), p. 2484665. DOI: 10.1080/07853890.2025.2484665.

[57] Neha Yadav, Sk Md Alfayeed, Aditya Khamparia, Babita Pandey, Dang NH Thanh, and Sagar Pande. "HSV model-based segmentation driven facial acne detection using deep learning". In: *Expert Systems* 39.3 (2022), e12760. DOI: 10.1111/exsy.12760.

[58] Ghasem Rahmatpour Rokni, Nasim Gholizadeh, Mahsa Babaei, and Kinnor Das. "Artificial intelligence in infectious skin disease". In: *Dermatological Reviews* 5.3 (2024), e241. DOI: 10.1002/der2.241.

[59] Wenchao Gao, Meirong Li, Rong Wu, et al. "The design and application of an automated microscope developed based on deep learning for fungal detection in dermatology". In: *Mycoses* 64.3 (2021), pp. 245–251. DOI: 10.1111/myc.13209.

[60] Ali Mansour Abdelmula, Omid Mirzaei, Emrah Güler, and Kaya Süer. "Assessment of deep learning models for cutaneous Leishmania parasite diagnosis using microscopic images". In: *Diagnostics* 14.1 (2023), p. 12. DOI: 10.3390/diagnostics14010012.

[61] Gustavo Blanco, Agma JM Traina, Caetano Traina Jr, et al. "A superpixel-driven deep learning approach for the analysis of dermatological wounds". In: *Computer Methods and Programs in Biomedicine* 183 (2020), p. 105079. DOI: 10.1016/j.cmpb.2019.105079.

[62] Ana Maria Malciu, Mihai Lupu, and Vlad Mihai Voiculescu. "Artificial intelligence-based approaches to reflectance confocal microscopy image analysis in dermatology". In: *Journal of Clinical Medicine* 11.2 (2022), p. 429. DOI: 10.3390/jcm11020429.

[63] Xuan Liu, Nadiya Chuchvara, Yuwei Liu, and Babar Rao. "Real-time deep learning assisted skin layer delineation in dermal optical coherence tomography". In: *OSA Continuum* 4.7 (2021), pp. 2008–2023. DOI: 10.1364/osac.426962.

[64] Joanna Czajkowska, Pawel Badura, Szymon Korzekwa, Anna Płatkowska-Szczerek, and Monika Słowińska. "Deep learning-based high-frequency ultrasound skin image classification with multicriteria model evaluation". In: *Sensors* 21.17 (2021), p. 5846. DOI: 10.3390/s21175846.

[65] Zhengshuai Jiang, Xiaming Gu, Dongdong Chen, Min Zhang, and Congcong Xu. "Deep learning-assisted multispectral imaging for early screening of skin diseases". In: *Photodiagnosis and Photodynamic Therapy* 48 (2024), p. 104292. DOI: 10.1016/j.pdpdt.2024.104292.

[66] Abdurrahim Yilmaz, Fatih Göktay, Rahmetullah Varol, Gulsum Gencoglan, and Huseyin Uvet. "Deep convolutional neural networks for onychomycosis detection using microscopic images with KOH examination". In: *Mycoses* 65.12 (2022), pp. 1119–1126. DOI: 10.1111/myc.13498.

[67] Fatima N Mirza, Zaim Haq, Parsa Abdi, Michael J Diaz, and Tiffany J Libby. "Artificial Intelligence for Mohs and Dermatologic Surgery: A Systematic Review and Meta-Analysis". In: *Dermatologic Surgery* 50.9 (2024), pp. 799–806. DOI: 10.1097/dss.0000000000004297.

[68] Aditya K Gupta, Iordanka A Ivanova, and Helen J Renaud. "How good is artificial intelligence (AI) at solving hairy problems? A review of AI applications in hair restoration and hair disorders". In: *Dermatologic Therapy* 34.2 (2021), e14811. DOI: 10.1111/dth.14811.

[69] Solam Lee, Jong Won Lee, Sung Jay Choe, et al. "Clinically applicable deep learning framework for measurement of the extent of hair loss in patients with alopecia areata". In: *JAMA Dermatology* 156.9 (2020), pp. 1018–1020. DOI: 10.1001/jamadermatol.2020.2188.

[70] Hui Wang and Xin Liu. "YOLO-OHFD: A YOLO-Based Oriented Hair Follicle Detection Method for Robotic Hair Transplantation". In: *Applied Sciences* 15.6 (2025), p. 3208. DOI: 10.3390/app15063208.

[71] Vishal Gaurav, Chander Grover, Mehul Tyagi, and Suman Saurabh. "Artificial intelligence in diagnosis and management of nail disorders: a narrative review". In: *Indian Dermatology Online Journal* 16.1 (2025), pp. 40–49. DOI: 10.4103/idoj.idoj_460_24.

[72] Pat Vatiwutipong, Sirawich Vachmanus, Thanapon Noraset, and Suppawong Tuarob. "Artificial intelligence in cosmetic dermatology: a systematic literature review". In: *IEEE Access* 11 (2023), pp. 71407–71425. DOI: 10.1109/access.2023.3295001.

[73] Amirata Ghorbani, Vivek Natarajan, David Coz, and Yuan Liu. "DermGAN: Synthetic generation of clinical skin images with pathology". In: *Machine Learning for Health (ML4H) at NeurIPS 2019*. PMLR. 2020, pp. 155–170. URL: https://proceedings.mlr.press/v116/ghorbani20a/ghorbani20a.pdf.

[74] Wenqi Zeng, Yuqi Sun, Chenxi Ma, Weimin Tan, and Bo Yan. "MM-Skin: Enhancing dermatology vision-language model with an image-text dataset derived from textbooks". In: *Proceedings of the 33rd ACM International Conference on Multimedia*. 2025, pp. 3769–3778. DOI: 10.1145/3746027.3755187.

[75] Siyuan Yan, Ming Hu, Yiwen Jiang, et al. "Derm1M: A Million-Scale Vision-Language Dataset Aligned with Clinical Ontology Knowledge for Dermatology". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025, pp. 12681–12690. DOI: 10.1109/iccv51701.2025.01178.

[76] Maxine Joly-Chevrier, Anne Xuan-Lan Nguyen, Michael Lesko-Krleza, and Philippe Lefrançois. "Performance of ChatGPT on a practice dermatology board certification examination". In: *Journal of Cutaneous Medicine and Surgery* 27.4 (2023), pp. 409–411. DOI: 10.1177/12034754231188437.

[77] Alessia Paganelli, Marco Spadafora, Cristian Navarrete-Dechent, Stefania Guida, Giovanni Pellacani, and Caterina Longo. "Natural language processing in dermatology: A systematic literature review and state of the art". In: *Journal of the European Academy of Dermatology and Venereology* 38.12 (2024), pp. 2225–2234. DOI: 10.1111/jdv.20286.

[78] Polat Goktas and Andrzej Grzybowski. "Assessing the impact of ChatGPT in dermatology: a comprehensive rapid review". In: *Journal of Clinical Medicine* 13.19 (2024), p. 5909. DOI: 10.3390/jcm13195909.

[79] Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, et al. "Disparities in dermatology AI performance on a diverse, curated clinical image set". In: *Science Advances* 8.32 (2022), eabq6147. DOI: 10.1126/sciadv.abq6147.

[80] M. Emre Celebi, Teresa Mendonca, and Jorge S. Marques. "PH2: A public database for the analysis of dermoscopic images". In: *Dermoscopy Image Analysis*. CRC Press, 2015. DOI: 10.1201/b19107-17.

[81] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. "Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets". In: *IEEE Journal of Biomedical and Health Informatics* 23.2 (2019), pp. 538–546. DOI: 10.1109/JBHI.2018.2824327.

[82] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions". In: *Scientific Data* 5 (2018), p. 180161. DOI: 10.1038/sdata.2018.161.

[83] Carlos Hernández-Pérez, Marc Combalia, Sebastian Podlipnik, et al. "BCN20000: Dermoscopic lesions in the wild". In: *Scientific Data* 11.1 (2024), p. 641. DOI: 10.1038/s41597-024-03387-w.

[84] International Skin Imaging Collaboration. *ISIC Archive*. URL: https://www.isic-archive.com/.

[85] Andre GC Pacheco, Gustavo R Lima, Amanda S Salomao, et al. "PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones". In: *Data in Brief* 32 (2020), p. 106221. DOI: 10.1016/j.dib.2020.106221.

[86] Matthew Groh, Caleb Harris, Luis Soenksen, et al. "Evaluating deep neural networks trained on clinical images in dermatology with the Fitzpatrick 17k dataset". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2021, pp. 1820–1828. DOI: 10.1109/cvprw53098.2021.00201.

[87] Abbi Ward, Jimmy Li, Julie Wang, et al. "Creating an Empirical Dermatology Dataset Through Crowdsourcing With Web Search Advertisements". In: *JAMA Network Open* 7.11 (2024), e2446615. DOI: 10.1001/jamanetworkopen.2024.46615.

[88] Melanie Champendal, Henning Müller, John O Prior, and Cláudia Sá Dos Reis. "A scoping review of interpretability and explainability concerning artificial intelligence methods in medical imaging". In: *European Journal of Radiology* 169 (2023), p. 111159. DOI: 10.1016/j.ejrad.2023.111159.

[89] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. "Grad-CAM: visual explanations from deep networks via gradient-based localization". In: *International Journal of Computer Vision* 128.2 (2020), pp. 336–359. DOI: 10.1007/s11263-019-01228-7.

[90] Biraja Ghoshal, Allan Tucker, Bal Sanghera, and Wai Lup Wong. "Estimating uncertainty in deep learning for reporting confidence to clinicians in medical image segmentation and diseases detection". In: *Computational Intelligence* 37.2 (2021), pp. 701–734. DOI: 10.1111/coin.12411.

[91] Tyler J Loftus, Benjamin Shickel, Matthew M Ruppert, et al. "Uncertainty-aware deep learning in healthcare: a scoping review". In: *PLOS Digital Health* 1.8 (2022), e0000085. DOI: 10.1371/journal.pdig.0000085.

[92] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. "Learning transferable visual models from natural language supervision". In: *International Conference on Machine Learning*. PmLR. 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a/radford21a.pdf.

[93] Edward J Hu, Yelong Shen, Phillip Wallis, et al. "LoRA: Low-rank adaptation of large language models." In: *ICLR*. Vol. 1. 2. 2022, p. 3. URL: https://openreview.net/pdf?id=nZeVKeeFYf9.

[94] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. "QLORA: efficient finetuning of quantized LLMs". In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. 441. 2023, pp. 10088–10115. URL: https://dl.acm.org/doi/10.5555/3666122.3666563.

[95] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. *BERTScore: Evaluating text generation with BERT*. ICLR 2020. 2019. arXiv: 1904.09675.

[96] Philipp Tschandl, Christoph Rinner, Zoe Apalla, et al. "Human–computer collaboration for skin cancer recognition". In: *Nature Medicine* 26.8 (2020), pp. 1229–1234. DOI: 10.1038/s41591-020-0942-0.

[97] Giuseppe Argenziano, H Peter Soyer, Vincenzo De Giorgio, et al. *Interactive atlas of dermoscopy*. EDRA Medical Publishing & New Media, 2000. ISBN: 9788886457309.

[98] *Dermofit Library: A cognitive prosthesis to aid focal skin lesion diagnosis*. URL: https://homepages.inf.ed.ac.uk/rbf/DERMOFIT/.

[99] Ioannis Giotis, Nynke Molders, Sander Land, Michael Biehl, Marcel F Jonkman, and Nicolai Petkov. "MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images". In: *Expert Systems with Applications* 42.19 (2015), pp. 6578–6585. DOI: 10.1016/j.eswa.2015.04.034.

[100] Seung Seog Han, Myoung Shin Kim, Woohyung Lim, Gyeong Hun Park, Ilwoo Park, and Sung Eun Chang. "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm". In: *Journal of Investigative Dermatology* 138.7 (2018), pp. 1529–1538. DOI: 10.1016/j.jid.2018.01.028.

[101] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. "A benchmark for automatic visual classification of clinical skin disease images". In: *European Conference on Computer Vision*. Springer International Publishing, 2016, pp. 206–222. DOI: 10.1007/978-3-319-46466-4_13.

[102] Jufeng Yang, Xiaoxiao Sun, Jie Liang, and Paul L Rosin. "Clinical skin lesion diagnosis using representations inspired by dermatologist criteria". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 1258–1266. DOI: 10.1109/cvpr.2018.00137.

[103] Jufeng Yang, Xioapeng Wu, Jie Liang, et al. "Self-paced balance learning for clinical skin disease recognition". In: *IEEE Transactions on Neural Networks and Learning Systems* 31.8 (2020), pp. 2832–2846. DOI: 10.1109/tnnls.2019.2917524.

[104] *Atlas dermatologico*. URL: https://atlasdermatologico.com.br/.

[105] Jehad Amin AlKattash. *Dermaamin*. URL: https://www.dermaamin.com/site/.

[106] Nicholas R Kurtansky, Brian M D'Alessandro, Maura C Gillis, et al. "The SLICE-3D dataset: 400,000 skin lesion image crops extracted from 3D TBP for skin cancer detection". In: *Scientific Data* 11.1 (2024), p. 884. DOI: 10.1038/s41597-024-03743-w.

[107] *The Cancer Genome Atlas*. URL: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.

[108] *DermNet NZ Dataset*. URL: https://www.dermnetnz.org/.

[109] Jordan Yap, William Yolland, and Philipp Tschandl. "Multimodal skin lesion classification using deep learning". In: *Experimental Dermatology* 27.11 (2018), pp. 1261–1267. DOI: 10.1111/exd.13777.

[110] Teresa Mendonça, Pedro M Ferreira, Jorge S Marques, André RS Marcal, and Jorge Rozeira. "PH2 - A dermoscopic image database for research and benchmarking". In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2013, pp. 5437–5440. DOI: 10.1109/EMBC.2013.6610779.

[111] Veronica Rotemberg, Nicholas Kurtansky, Brigid Betz-Stablein, et al. "A patient-centric dataset of images and metadata for identifying melanomas using clinical context". In: *Scientific Data* 8.1 (2021), p. 34. DOI: 10.1038/s41597-021-00815-z.

[112] Michael A Marchetti, Emily A Cowen, Nicholas R Kurtansky, et al. "Prospective validation of dermoscopy-based open-source artificial intelligence for melanoma diagnosis (PROVE-AI study)". In: *npj Digital Medicine* 6.1 (2023), p. 127. DOI: 10.1038/s41746-023-00872-1.

[113] María Agustina Ricci Lara, María Victoria Rodríguez Kowalczuk, Maite Lisa Eliceche, et al. "A dataset of skin lesion images collected in Argentina for the evaluation of AI tools in this population". In: *Scientific Data* 10.1 (2023), p. 712. DOI: 10.1038/s41597-023-02630-0.

[114] Philipp Tschandl, Bengu Nisa Akay, Cliff Rosendahl, et al. "MILK10k: A Hierarchical Multimodal Imaging-Learning Toolkit for Diagnosing Pigmented and Nonpigmented Skin Cancer and its Simulators". In: *Journal of Investigative Dermatology* 146.2 (2026), 357–364.e7. DOI: 10.1016/j.jid.2025.06.1594.

[115] Abdurrahim Yilmaz, Sirin Pekcan Yasar, Gulsum Gencoglan, and Burak Temelkuran. "DERM12345: A Large, Multisource Dermatoscopic Skin Lesion Dataset with 40 Subclasses". In: *Scientific Data* 11.1 (2024), p. 1302. DOI: 10.1038/s41597-024-04104-3.

[116] Siyuan Yan, Zhen Yu, Clare Primiero, et al. "A multimodal vision foundation model for clinical dermatology". In: *Nature Medicine* 31.8 (2025), pp. 2691–2702. DOI: 10.1038/s41591-025-03747-y.

[117] Jiajun Sun, Zhen Yu, Siyuan Yan, Jason J Ong, Zongyuan Ge, and Lei Zhang. "Controllable Skin Synthesis via Lesion-Focused Vector Autoregression Model". In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Vol. LNCS 15975. Springer Nature Switzerland, 2025, pp. 129–139. DOI: 10.1007/978-3-032-05325-1_13.

[118] Michael A Marchetti, Noel CF Codella, Stephen W Dusza, et al. "Results of the 2016 International Skin Imaging Collaboration International Symposium on Biomedical Imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images". In: *Journal of the American Academy of Dermatology* 78.2 (2018), 270–277.e1. DOI: 10.1016/j.jaad.2017.08.016.

[119] Memorial Sloan Kettering Cancer Center. *PROVe-AI [Data set]*. ISIC Archive, 2022. DOI: 10.34970/576276.

[120] Gursoy Dogan. "Melanocytic nevi in 2783 children and adolescents in Turkey". In: *Pediatric Dermatology* 24.5 (2007), pp. 489–494. DOI: 10.1111/j.1525-1470.2007.00500.x.

[121] Yiğit Özer TIFTIKCIOĞLU, Önder Karaaslan, Hasan Mete Aksoy, Berna Aksoy, and Uğur Koçer. "Basal cell carcinoma in Turkey". In: *The Journal of Dermatology* 33.2 (2006), pp. 91–95. DOI: 10.1111/j.1346-8138.2006.00019.x.

[122] Zeynep Karaca Ural, Leyla Baykal Selcuk, Deniz Aksu Arica, Savas Yayli, and Sevgi Bahadir. "Prevalence and Dermoscopic Patterns of Acral Melanocytic Nevi in Turkey". In: *Acta Dermatovenerologica Croatica* 29.2 (2021), pp. 94–101. URL: https://pubmed.ncbi.nlm.nih.gov/34477075/.

[123] Abdurrahim Yilmaz. *Created in BioRender.* 2024. URL: https://biorender.com/g99y437.

[124] Alon Scope, Konstantinos Liopyris, Jochen Weber, et al. "International Skin Imaging Collaboration-Designated Diagnoses (ISIC-DX): Consensus terminology for lesion diagnostic labeling". In: *Journal of the European Academy of Dermatology and Venereology* 39.1 (2025), pp. 117–125. DOI: 10.1111/jdv.20055.

[125] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, et al. "Datasheets for datasets". In: *Communications of the ACM* 64.12 (2021), pp. 86–92. DOI: 10.1145/3458723.

[126] Roxana Daneshjou, Catarina Barata, Brigid Betz-Stablein, et al. "Checklist for evaluation of image-based AI reports in dermatology: CLEAR Derm consensus guidelines from the International Skin Imaging Collaboration artificial intelligence working group". In: *JAMA Dermatology* 158.1 (2022), pp. 90–96. DOI: 10.1001/jamadermatol.2021.4915.

[127] Dana Baigrie and Laura S Tanner. "Dysplastic nevi". In: *StatPearls [Internet]*. StatPearls Publishing, 2022. URL: https://www.ncbi.nlm.nih.gov/books/NBK482210/.

[128] VA Kinsler, P O'hare, N Bulstrode, et al. "Melanoma in congenital melanocytic naevi". In: *British Journal of Dermatology* 176.5 (2017), pp. 1131–1143. DOI: 10.1111/bjd.15301.

[129] Luise Ribeiro Daltro, Lygia Bertalha Yaegashi, Rodrigo Abdalah Freitas, Bruno de Carvalho Fantini, and Cacilda da Silva Souza. "Atypical cellular blue nevus or malignant blue nevus?" In: *Anais Brasileiros de Dermatologia* 92.1 (2017), pp. 110–112. DOI: 10.1590/abd1806-4841.20174502.

[130] Abdurrahim Yilmaz. *Created in BioRender.* 2024. URL: https://biorender.com/z11p239.

[131] Abdurrahim Yilmaz. *Created in BioRender.* 2024. URL: https://biorender.com/z45e991.

[132] Abdurrahim Yilmaz, Sirin Pekcan Yasar, Gulsum Gencoglan, and Burak Temelkuran. *DERM12345: A Large, Multisource Dermatoscopic Skin Lesion Dataset with 40 Subclasses*. Harvard Dataverse, 2024. DOI: 10.7910/DVN/DAXZ7P.

[133] Tiffany W Cheng, Madeline C Ahern, and Alessio Giubellino. "The spectrum of Spitz melanocytic lesions: From morphologic diagnosis to molecular classification". In: *Frontiers in Oncology* 12 (2022), p. 889223. DOI: 10.3389/fonc.2022.889223.

[134] Catarina Barata and Jorge S Marques. "Deep learning for skin cancer diagnosis with hierarchical architectures". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE. 2019, pp. 841–845. DOI: 10.1109/isbi.2019.8759561.

[135] Cristina Carrera and Ashfaq A Marghoob. "Discriminating nevi from melanomas: clues and pitfalls". In: *Dermatologic Clinics* 34.4 (2016), pp. 395–409. DOI: 10.1016/j.det.2016.05.003.

[136] Kee Suck Suh, Jong Bin Park, Joon Hee Kim, et al. "Dysplastic nevus: Clinical features and usefulness of dermoscopy." In: *The Journal of Dermatology* 46.2 (2019), e76–e77. DOI: 10.1111/1346-8138.14583.

[137] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. *Instruction tuning with GPT-4*. 2023. arXiv: 2304.03277.

[138] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. "Visual instruction tuning". In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. 1516. 2023, pp. 34892–34916. URL: https://dl.acm.org/doi/10.5555/3666122.3667638.

[139] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. "Improved baselines with visual instruction tuning". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 26286–26296. DOI: 10.1109/cvpr52733.2024.02484.

[140] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. "PMC-LLaMA: toward building open-source language models for medicine". In: *Journal of the American Medical Informatics Association* 31.9 (2024), pp. 1833–1843. DOI: 10.1093/jamia/ocae045.

[141] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, et al. *MedAlpaca–An Open-Source Collection of Medical Conversational AI Models and Training Data*. 2023. arXiv: 2304.08247.

[142] Chang Shu, Baian Chen, Fangyu Liu, Zihao Fu, Ehsan Shareghi, and Nigel Collier. *Visual Med-Alpaca: A Parameter-Efficient Biomedical LLM with Visual Capabilities*. 2023. URL: https://cambridgeltl.github.io/visual-med-alpaca/.

[143] Chunyuan Li, Cliff Wong, Sheng Zhang, et al. "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day". In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. 1240. 2023, pp. 28541–28564. URL: https://dl.acm.org/doi/10.5555/3666122.3667362.

[144] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, et al. *Stanford Alpaca: An Instruction-following LLaMA model*. 2023. URL: https://github.com/tatsu-lab/stanford_alpaca.

[145] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, et al. "Self-instruct: Aligning language models with self-generated instructions". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 13484–13508. DOI: 10.18653/v1/2023.acl-long.754.

[146] Abdurrahim Yilmaz. *DermaSynth: Rich Synthetic Image-Text Pairs Using Open Access Dermatology Datasets*. 2025. DOI: 10.7910/DVN/QWQEMM.

[147] Alejandro Lozano, Min Woo Sun, James Burgess, et al. "BIOMEDICA: An Open Biomedical Image-Caption Archive, Dataset, and Vision-Language Models Derived from Scientific Literature". In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 19724–19735. DOI: 10.1109/cvpr52734.2025.01837.

[148] *Wikipedia - List of skin conditions*. URL: https://en.wikipedia.org/wiki/List_of_skin_conditions.

[149] Mor Zarfati, Girish N Nadkarni, Benjamin S Glicksberg, et al. "Exploring the Role of Large Language Models in Melanoma: A Systematic Review". In: *Journal of Clinical Medicine* 13.23 (2024), p. 7480. DOI: 10.3390/jcm13237480.

[150] Xu Cao, Wenqian Ye, Kenny Moise, and Megan Coffee. *MpoxVLM: A vision-language model for diagnosing skin lesions from mpox virus infection*. 2024. arXiv: 2411.10888.

[151] Shiwan Zhang, Abdurrahim Yilmaz, Gulsum Gencoglan, and Burak Temelkuran. "Fine-Tuning a Small Vision Language Model Using Synthetic Data for Explaining Bacterial Skin Disease Images". In: *Diagnostics* 16.4 (2026), p. 603. DOI: 10.3390/diagnostics16040603.

[152] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[153] François Chollet. "Xception: Deep learning with depthwise separable convolutions". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 1800–1807. DOI: 10.1109/cvpr.2017.195.

[154] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. "Inception-v4, Inception-ResNet and the impact of residual connections on learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017. DOI: 10.1609/aaai.v31i1.11231.

[155] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009, pp. 248–255. DOI: 10.1109/cvpr.2009.5206848.

[156] Archana Singal and Deepshikha Khanna. "Onychomycosis: Diagnosis and management". In: *Indian Journal of Dermatology, Venereology and Leprology* 77 (2011), pp. 659–672. DOI: 10.4103/0378-6323.86475.

[157] AK Gupta, N Stec, RC Summerbell, et al. "Onychomycosis: a review". In: *Journal of the European Academy of Dermatology and Venereology* 34.9 (2020), pp. 1972–1990. DOI: 10.1111/jdv.16394.

[158] Jacob Oren Levitt, Barrie H Levitt, Arash Akhavan, and Howard Yanofsky. "The sensitivity and specificity of potassium hydroxide smear and fungal culture relative to clinical assessment in the evaluation of tinea pedis: a pooled analysis". In: *Dermatology Research and Practice* 2010.1 (2010), p. 764843. DOI: 10.1155/2010/764843.

[159] Veronica Velasquez-Agudelo and Jaiberth Antonio Cardona-Arias. "Meta-analysis of the utility of culture, biopsy, and direct KOH examination for the diagnosis of onychomycosis." In: *BMC Infectious Diseases* 17.1 (2017), p. 166. DOI: 10.1186/s12879-017-2258-3.

[160] Sophie Soyeon Lim, Jungyoon Ohn, and Je-Ho Mun. "Diagnosis of onychomycosis: from conventional techniques and dermoscopy to artificial intelligence". In: *Frontiers in Medicine* 8 (2021), p. 637216. DOI: 10.3389/fmed.2021.637216.

[161] Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, et al. "HoVer-Net: simultaneous segmentation and classification of nuclei in multi-tissue histology images". In: *Medical Image Analysis* 58 (2019), p. 101563. DOI: 10.1016/j.media.2019.101563.

[162] Dhanya Bibin, Madhu S Nair, and P Punitha. "Malaria parasite detection from peripheral blood smear images using deep belief networks". In: *IEEE Access* 5 (2017), pp. 9099–9108. DOI: 10.1109/access.2017.2705642.

[163] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, et al. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42 (2017), pp. 60–88. DOI: 10.1016/j.media.2017.07.005.

[164] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. "AI in health and medicine". In: *Nature Medicine* 28.1 (2022), pp. 31–38. DOI: 10.1038/s41591-021-01614-0.

[165] Dinggang Shen, Guorong Wu, and Heung-Il Suk. "Deep learning in medical image analysis". In: *Annual Review of Biomedical Engineering* 19 (2017), pp. 221–248. DOI: 10.1146/annurev-bioeng-071516-044442.

[166] Clare McGenity, Emily L Clarke, Charlotte Jennings, et al. "Artificial intelligence in digital pathology: a systematic review and meta-analysis of diagnostic test accuracy". In: *npj Digital Medicine* 7.1 (2024), p. 114. DOI: 10.1038/s41746-024-01106-8.

[167] Paras Lakhani and Baskaran Sundaram. “Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks”. In: *Radiology* 284.2 (2017), pp. 574–582. DOI: 10.1148/radiol.2017162326.

[168] Jeroen Van Der Laak, Geert Litjens, and Francesco Ciompi. “Deep learning in histopathology: the path to the clinic”. In: *Nature Medicine* 27.5 (2021), pp. 775–784. DOI: 10.1038/s41591-021-01343-4.

[169] Titus J Brinker, Achim Hekler, Alexander H Enk, et al. “Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task”. In: *European Journal of Cancer* 113 (2019), pp. 47–54. DOI: 10.1016/j.ejca.2019.04.001.

[170] Konstantinos Liopyris, Stamatios Gregoriou, Julia Dias, and Alexandros J Stratigos. “Artificial intelligence in dermatology: challenges and perspectives”. In: *Dermatology and Therapy* 12.12 (2022), pp. 2637–2651. DOI: 10.1007/s13555-022-00833-8.

[171] Chongming Zhang, Yaxuan Zhang, Zhanyuan Chang, and Chuanjiang Li. “Sperm YOLOv8E-TrackEVD: a novel approach for sperm detection and tracking”. In: *Sensors* 24.11 (2024), p. 3493. DOI: 10.3390/s24113493.

[172] Rani Oomman Panicker, Kaushik S Kalmady, Jeny Rajan, and MK Sabu. “Automatic detection of tuberculosis bacilli from microscopic sputum smear images using deep learning methods”. In: *Biocybernetics and Biomedical Engineering* 38.3 (2018), pp. 691–699. DOI: 10.1016/j.bbe.2018.05.007.

[173] Sivaramakrishnan Rajaraman, Sameer K Antani, Mahdieh Poostchi, et al. “Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images”. In: *PeerJ* 6 (2018), e4568. DOI: 10.7717/peerj.4568.

[174] Khalid Hamed S Allehaibi, Lukito Edi Nugroho, Lutfan Lazuardi, Anton Satria Prabuwono, Teddy Mantoro, et al. “Segmentation and classification of cervical cells using deep learning”. In: *IEEE Access* 7 (2019), pp. 116925–116941. DOI: 10.1109/access.2019.2936017.

[175] Yifei Chen, Chenyan Zhang, Ben Chen, et al. “Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases”. In: *Computers in Biology and Medicine* 170 (2024), p. 107917. DOI: 10.1016/j.compbiomed.2024.107917.

[176] Md Mamunur Rahaman, Chen Li, Yudong Yao, et al. “DeepCervix: A deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques”. In: *Computers in Biology and Medicine* 136 (2021), p. 104649. DOI: 10.1016/j.compbiomed.2021.104649.

[177] Peter Naylor, Marick Laé, Fabien Reyal, and Thomas Walter. “Segmentation of nuclei in histopathology images by deep regression of the distance map”. In: *IEEE Transactions on Medical Imaging* 38.2 (2018), pp. 448–459. DOI: 10.1109/TMI.2018.2865709.

[178] Philipp Jansen, Adelaida Creosteanu, Viktor Matyas, et al. “Deep learning assisted diagnosis of onychomycosis on whole-slide images”. In: *Journal of Fungi* 8.9 (2022), p. 912. DOI: 10.3390/jof8090912.

[179] Florence Decroos, Sebastian Springenberg, Marc PÄPPER, et al. “A deep learning approach for histopathological diagnosis of onychomycosis: not inferior to analogue diagnosis by histopathologists”. In: *Acta Dermato-Venereologica* 101.8 (2021). DOI: 10.2340/00015555-3893.

[180] Taehan Koo, Moon Hwan Kim, and Mihn-Sook Jue. “Automated detection of superficial fungal infections from microscopic images through a regional convolutional neural network”. In: *PLOS One* 16.8 (2021), e0256290. DOI: 10.1371/journal.pone.0256290.

[181] Nidhal Jegham, Chan Young Koh, Marwan Abdelatti, and Abdeltawab Hendawi. *YOLO Evolution: A Comprehensive Benchmark and Architectural Review of YOLOv12, YOLO11, and Their Previous Versions*. 2024. arXiv: 2411.00201.

[182] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: Towards real-time object detection with region proposal networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2016), pp. 1137–1149. DOI: 10.1109/TPAMI.2016.2577031.

[183] Richa Garg and S Gupta. "Mimickers of dermatophytes on KOH mount". In: *IP Indian J Clin Exp Dermatol* 6.1 (2020), pp. 98–101. DOI: 10.18231/j.ijced.2020.021.

[184] Wenbin Xu, Qiang Zhai, Jizhong Liu, Xingyu Xu, and Jing Hua. "A lightweight deep-learning model for parasite egg detection in microscopy images". In: *Parasites & Vectors* 17.1 (2024), p. 454. DOI: 10.1186/s13071-024-06503-2.

[185] Suprim Nakarmi, Sanam Pudasaini, Safal Thapaliya, et al. "Deep-learning assisted detection and quantification of (oo)cysts of Giardia and Cryptosporidium on smartphone microscopy images". In: *Machine Learning for Biomedical Imaging* 2 (2024), pp. 956–976. DOI: 10.59275/j.melba.2024-a333.

[186] Tongxue Zhou, Su Ruan, and Stéphane Canu. "A review: Deep learning for medical image segmentation using multi-modality fusion". In: *Array* 3–4 (2019), p. 100004. DOI: 10.1016/j.array.2019.100004.

[187] Erico Tjoa and Cuntai Guan. "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.11 (2021), pp. 4793–4813. DOI: 10.1109/TNNLS.2020.3027314.

[188] Ryan R Riahi, Philip R Cohen, and Leonard H Goldberg. "Subungual nail bed melanoma masquerading as tinea ungium". In: *Dermatology Online Journal* 24.6 (2018). DOI: 10.5070/D3246040694.

[189] Kaya L Curtis, Snehal P Amin, Nathaniel J Jellinek, et al. "Diagnosis and management of nail unit squamous cell carcinoma: a clinical review by an expert panel". In: *JAAD Reviews* 4 (2025), pp. 187–194. DOI: 10.1016/j.jdrv.2024.12.012.

[190] Federico Venturi, Elisabetta Magnaterra, Biagio Scotti, et al. "Squamous Cell Carcinoma of the Nail Unit: A Comprehensive Review of Clinical Features, Diagnostic Workflow, Management Strategies and Therapeutic Options". In: *Diagnostics* 15.18 (2025), p. 2378. DOI: 10.3390/diagnostics15182378.

[191] Michael Kasperkiewicz, Christoph T. Ellebrecht, Hayato Takahashi, et al. "Pemphigus". In: *Nature Reviews Disease Primers* 3.1 (2017), p. 17026. DOI: 10.1038/nrdp.2017.26.

[192] Enno Schmidt, Michael Kasperkiewicz, and Pascal Joly. "Pemphigus". In: *The Lancet* 394.10201 (2019), pp. 882–894. DOI: 10.1016/S0140-6736(19)31778-7.

[193] Dave Van Veen, Cara Van Uden, Louis Blankemeier, et al. "Adapted large language models can outperform medical experts in clinical text summarization". In: *Nature Medicine* 30.4 (2024), pp. 1134–1142. DOI: 10.1038/s41591-024-02855-5.

[194] Honghan Wu, Minhong Wang, Jinge Wu, et al. "A survey on clinical natural language processing in the United Kingdom from 2007 to 2022". In: *npj Digital Medicine* 5.1 (2022), p. 186. DOI: 10.1038/s41746-022-00730-6.

[195] Yoav Goldberg. "A primer on neural network models for natural language processing". In: *Journal of Artificial Intelligence Research* 57 (2016), pp. 345–420. DOI: 10.1613/jair.4992.

[196] Ghazal Azarfar, Sara Naimimohasses, Sirisha Rambhatla, et al. “Responsible adoption of multimodal artificial intelligence in health care: promises and challenges”. In: *The Lancet Digital Health* 7.12 (2025), p. 100917. DOI: 10.1016/j.landig.2025.100917.

[197] Abdurrahim Yilmaz, Furkan Yuceyalcin, Rahmetullah Varol, et al. *Resource-efficient medical vision language model for dermatology via a synthetic data generation framework*. medRxiv preprint. 2025. DOI: 10.1101/2025.05.17.25327785.

[198] Charlotte MHHT Bootsma-Robroeks, Jessica D Workum, Stephanie CE Schuit, et al. “AI-generated draft replies to patient messages: exploring effects of implementation”. In: *Frontiers in Digital Health* 7 (2025), p. 1588143. DOI: 10.3389/fdgth.2025.1588143.

[199] Xinsong Du, Zhengyang Zhou, Yifei Wang, et al. “Testing and evaluation of generative large language models in electronic health record applications: a systematic review”. In: *Journal of the American Medical Informatics Association* 33.3 (2026), pp. 743–753. DOI: 10.1093/jamia/ocaf233.

[200] Oliver J Bear Don’t Walk IV, Harry Reyes Nieva, Sandra Soo-Jin Lee, and Noémie Elhadad. “A scoping review of ethics considerations in clinical natural language processing”. In: *JAMIA Open* 5.2 (2022), ooac039. DOI: 10.1093/jamiaopen/ooac039.

[201] Dillon Obika, Christopher Kelly, Nicola Ding, et al. “Safety principles for medical summarization using generative AI”. In: *Nature Medicine* 30.12 (2024), pp. 3417–3419. DOI: 10.1038/s41591-024-03313-y.

[202] An Yang, Anfeng Li, Baosong Yang, et al. *Qwen3 Technical Report*. 2025. arXiv: 2505.09388.

[203] Christine Dymek, Bryan Kim, Genevieve B Melton, Thomas H Payne, Hardeep Singh, and Chun-Ju Hsiao. “Building the evidence–base to reduce electronic health record–related clinician burden”. In: *Journal of the American Medical Informatics Association* 28.5 (2021), pp. 1057–1061. DOI: 10.1093/jamia/ocaa238.

[204] Tejal K Gandhi, David Classen, Christine A Sinsky, et al. “How can artificial intelligence decrease cognitive and work burden for front line practitioners?” In: *JAMIA Open* 6.3 (2023), ooad079. DOI: 10.1093/jamiaopen/ooad079.

[205] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, et al. *MedGemma Technical Report*. 2025. arXiv: 2507.05201.

[206] Savyasachi V. Shah. “Accuracy, Consistency, and Hallucination of Large Language Models When Analyzing Unstructured Clinical Notes in Electronic Medical Records”. In: *JAMA Network Open* 7.8 (2024), e2425953. DOI: 10.1001/jamanetworkopen.2024.25953.

[207] Yubin Kim, Hyewon Jeong, Shan Chen, et al. *Medical Hallucinations in Foundation Models and Their Impact on Healthcare*. 2025. arXiv: 2503.05777.

[208] Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. “A survey on model compression for large language models”. In: *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 1556–1577. DOI: 10.1162/tacl_a_00704.

[209] Karan Singhal, Tao Tu, Juraj Gottweis, et al. “Toward expert-level medical question answering with large language models”. In: *Nature Medicine* 31.3 (2025), pp. 943–950. DOI: 10.1038/s41591-024-03423-7.

[210] Sílvia Mamede and Henk G Schmidt. “Making large language models into reliable physician assistants”. In: *Nature Medicine* 31.4 (2025), pp. 1071–1072. DOI: 10.1038/s41591-025-03606-w.

[211] Ethan Goh, Robert J Gallo, Eric Strong, et al. “GPT-4 assistance for improvement of physician performance on patient care tasks: a randomized controlled trial”. In: *Nature Medicine* 31.4 (2025), pp. 1233–1238. DOI: 10.1038/s41591-024-03456-y.

[212] Xiaohong Liu, Hao Liu, Guoxing Yang, et al. “A generalist medical language model for disease diagnosis assistance”. In: *Nature Medicine* 31.3 (2025), pp. 932–942. DOI: 10.1038/s41591-024-03416-6.

[213] Senliang Lu, Yehang Chen, Yuan Chen, et al. “General lightweight framework for vision foundation model supporting multi-task and multi-center medical image analysis”. In: *Nature Communications* 16.1 (2025), p. 2097. DOI: 10.1038/s41467-025-57427-z.

[214] Ryutaro Tanno, David GT Barrett, Andrew Sellergren, et al. “Collaboration between clinicians and vision–language models in radiology report generation”. In: *Nature Medicine* 31.2 (2025), pp. 599–608. DOI: 10.1038/s41591-024-03302-1.

[215] Yixing Jiang, Jesutofunmi A Omiye, Cyril Zakka, et al. *Evaluating general vision-language models for clinical medicine*. medRxiv preprint. 2024. DOI: 10.1101/2024.04.12.24305744.

[216] Branden Lee, Shivam Patel, Crystal Favorito, Sara Sandri, Maria Rain Jennings, and Tinglong Dai. “Development and commercialization pathways of AI medical devices in the United States: implications for safety and regulatory oversight”. In: *NEJM AI* 2.7 (2025), AIra2500061. DOI: 10.1056/aira2500061.

[217] Alexander Ziller, Tamara T Mueller, Simon Stieger, et al. “Reconciling privacy and accuracy in AI for medical imaging”. In: *Nature Machine Intelligence* 6.7 (2024), pp. 764–774. DOI: 10.1038/s42256-024-00858-y.

[218] Dani Kiyasseh, Aaron Cohen, Chengsheng Jiang, and Nicholas Altieri. “A framework for evaluating clinical artificial intelligence systems without ground-truth annotations”. In: *Nature Communications* 15.1 (2024), p. 1808. DOI: 10.1038/s41467-024-46000-9.

[219] James Zou and Eric J Topol. “The rise of agentic AI teammates in medicine”. In: *The Lancet* 405.10477 (2025), p. 457. DOI: 10.1016/s0140-6736(25)00202-8.

[220] Thierry Hanser, Ernst Ahlberg, Alexander Amberg, et al. “Data-driven federated learning in drug discovery with knowledge distillation”. In: *Nature Machine Intelligence* 7.3 (2025), pp. 423–436. DOI: 10.1038/s42256-025-00991-2.

[221] Xiaolan Chen, Jiayang Xiang, Shanfu Lu, Yexin Liu, Mingguang He, and Danli Shi. “Evaluating large language models and agents in healthcare: key challenges in clinical applications”. In: *Intelligent Medicine* 5.2 (2025), pp. 151–163. DOI: 10.1016/j.imed.2025.03.002.

[222] Binxu Li, Tiankai Yan, Yuanting Pan, et al. *MMedAgent: Learning to use medical tools with multi-modal agent*. 2024. arXiv: 2407.02483.

[223] Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. *MedRAX: Medical Reasoning Agent for Chest X-ray*. 2025. arXiv: 2502.02673.

[224] Ziyue Wang, Junde Wu, Chang Han Low, and Yueming Jin. *MedAgent-Pro: Towards multi-modal evidence-based medical diagnosis via reasoning agentic workflow*. 2025. arXiv: 2503.18968.

[225] Harsha Nori, Yin Tat Lee, Sheng Zhang, et al. *Can generalist foundation models outcompete special-purpose tuning? case study in medicine*. 2023. arXiv: 2311.16452.

[226] Abdurrahim Yilmaz. *Created in BioRender.* 2026. URL: https://biorender.com/g21q4l5.

[227] Abdurrahim Yilmaz. *Created in BioRender.* 2026. URL: https://biorender.com/umlcapf.

[228] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. *The Llama 3 Herd of Models*. 2024. arXiv: 2407.21783.

[229] Yaowei Zheng, Richong Zhang, Junhao Zhang, et al. *LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models*. Bangkok, Thailand, 2024. arXiv: 2403.13372.

[230] Abdurrahim Yilmaz. *Created in BioRender.* 2026. URL: https://biorender.com/ru9no8o.

[231] Abdurrahim Yilmaz. *Created in BioRender.* 2026. URL: https://biorender.com/82r169e.

[232] Long Ouyang, Jeffrey Wu, Xu Jiang, et al. "Training language models to follow instructions with human feedback". In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. 2011. 2022, pp. 27730–27744. URL: https://dl.acm.org/doi/10.5555/3600270.3602281.

[233] Tao Tu, Mike Schaekermann, Anil Palepu, et al. "Towards conversational diagnostic artificial intelligence". In: *Nature* 642.8067 (2025), pp. 442–450. DOI: 10.1038/s41586-025-08866-7.

[234] Katrina Cirone, Mohamed Akrout, Latif Abid, and Amanda Oakley. "Assessing the utility of multimodal large language models (GPT-4 vision and large language and vision assistant) in identifying melanoma across different skin tones". In: *JMIR Dermatology* 7 (2024), e55508. DOI: 10.2196/55508.

[235] Shreya Johri, Jaehwan Jeong, Benjamin A Tran, et al. "An evaluation framework for clinical use of large language models in patient interaction tasks". In: *Nature Medicine* 31.1 (2025), pp. 77–86. DOI: 10.1038/s41591-024-03328-5.

[236] Gongbo Zhang, Zihan Xu, Qiao Jin, et al. "Leveraging long context in retrieval augmented language models for medical question answering". In: *npj Digital Medicine* 8.1 (2025), p. 239. DOI: 10.1038/s41746-025-01651-w.

[237] Atilla P Kiraly, Sebastien Baur, Kenneth Philbrick, et al. *Health AI Developer Foundations*. 2024. arXiv: 2411.15128.

[238] Tao Tu, Shekoofeh Azizi, Danny Driess, et al. "Towards generalist biomedical AI". In: *NEJM AI* 1.3 (2024), Aloa2300138. DOI: 10.1056/aioa2300138.

[239] Chanwoo Kim, Soham U Gadgil, Alex J DeGrave, et al. "Transparent medical image AI via an image–text foundation model grounded in medical literature". In: *Nature Medicine* 30.4 (2024), pp. 1154–1165. DOI: 10.1038/s41591-024-02887-x.

[240] Juexiao Zhou, Xiaonan He, Liyuan Sun, et al. "Pre-trained multimodal large language model enhances dermatological diagnosis using SkinGPT-4". In: *Nature Communications* 15.1 (2024), p. 5649. DOI: 10.1038/s41467-024-50043-3.

[241] Haiwen Gui, Jesutofunmi A Omiye, Crystal T Chang, and Roxana Daneshjou. "The promises and perils of foundation models in dermatology". In: *Journal of Investigative Dermatology* 144.7 (2024), pp. 1440–1448. DOI: 10.1016/j.jid.2023.12.019.

[242] Mihai Nadăş, Laura Dioşan, and Andreea Tomescu. "Synthetic Data Generation Using Large Language Models: Advances in Text and Code". In: *IEEE Access* 13 (2025), pp. 134615–134633. DOI: 10.1109/ACCESS.2025.3589503.

[243] Crystal T Chang, Hodan Farah, Haiwen Gui, et al. "Red teaming ChatGPT in medicine to yield real-world insights on model behavior". In: *npj Digital Medicine* 8.1 (2025), p. 149. DOI: 10.1038/s41746-025-01542-0.

[244] Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. "Chain-of-thought prompting elicits reasoning in large language models". In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. 1800. 2022, pp. 24824–24837. URL: https://dl.acm.org/doi/10.5555/3600270.3602070.

[245] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. "A dataset of clinically generated visual questions and answers about radiology images". In: *Scientific Data* 5.1 (2018), p. 180251. DOI: 10.1038/sdata.2018.251.

[246] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. *PathVQA: 30000+ questions for medical visual question answering*. 2020. arXiv: 2003.10286.

[247] Wen-wai Yim, Asma Ben Abacha, Robert Doerning, et al. "WoundcareVQA: A multilingual visual question answering benchmark dataset for wound care". In: *Journal of Biomedical Informatics* 170 (2025), p. 104888. DOI: 10.1016/j.jbi.2025.104888.

[248] Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. "DermaVQA: A multilingual visual question answering dataset for dermatology". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 209–219. DOI: 10.1007/978-3-031-72086-4_20.

[249] Roxana Daneshjou, Mert Yuksekgonul, Zhuo Ran Cai, Roberto Novoa, and James Y Zou. "SkinCon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis". In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. 1320. 2022, pp. 18157–18167. URL: https://dl.acm.org/doi/10.5555/3600270.3601590.

[250] Yuhao Shen, Liyuan Sun, Yan Xu, et al. *SkinCaRe: A Multimodal Dermatology Dataset Annotated with Medical Caption and Chain-of-Thought Reasoning*. 2024. arXiv: 2405.18004.

[251] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. "SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering". In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 1650–1654. DOI: 10.1109/isbi48211.2021.9434010.

[252] Abtahi Noor, Raiyan Habib Mahe, Azwad Aziz, et al. "Lightweight visual question answering (VQA) model for skin disease detection". In: *2025 7th International Conference on Electrical Information and Communication Technology (EICT)*. IEEE. 2025, pp. 1–6. DOI: 10.1109/EICT68394.2025.11355641.

[253] Gerd Plewig, Lars E. French, Thomas Ruzicka, Roland Kaufmann, and Michael Hertl, eds. *Braun-Falco's Dermatology*. 4th ed. Springer, 2022. ISBN: 9783662637081.

[254] William D. James, Dirk M. Elston, James R. Treat, and Misha Rosenbach. *Andrews' Diseases of the Skin: Clinical Dermatology*. 13th ed. Elsevier, 2019. ISBN: 9780323547536.

[255] Sewon Kang, Masayuki Amagai, Anna L. Bruckner, et al., eds. *Fitzpatrick's Dermatology*. 9th ed. McGraw-Hill Education, 2019. ISBN: 9780071837798.

[256] Abdurrahim Yilmaz, Ozan Erdem, Ece Gokyayla, et al. *DermaBench: A Clinician-Annotated Benchmark Dataset for Dermatology Visual Question Answering and Reasoning*. 2026. DOI: 10.7910/DVN/Q4LBIW.

[257] Ralph P. Braun, Harold S. Rabinovitz, Margeret Oliviero, Alfred W. Kopf, and Jean H. Saurat. "Pattern analysis: a two-step procedure for the dermoscopic diagnosis of melanoma". In: *Clinics in Dermatology* 20.3 (2002), pp. 236–239. DOI: 10.1016/S0738-081X(02)00216-X.

[258] Manu Goyal, Thomas Knackstedt, Shaofeng Yan, and Saeed Hassanpour. "Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities". In: *Comput. Biol. Med.* 127 (2020), p. 104065. DOI: 10.1016/j.compbiomed.2020.104065.

[259] Md Ashraful Alam Milton. *Automated Skin Lesion Classification Using Ensemble of Deep Neural Networks in ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection Challenge*. 2019. arXiv: 1901.10802.

[260] Magdalini Paschali, Zhihong Chen, Louis Blankemeier, et al. "Foundation Models in Radiology: What, How, Why, and Why Not". In: *Radiology* 314.2 (2025), e240597. DOI: 10.1148/radiol.240597.

[261] Fan Liu, Tianshu Zhang, Wenwen Dai, et al. "Few-shot adaptation of multi-modal foundation models: a survey". In: *Artificial Intelligence Review* 57.10 (2024), p. 268. DOI: 10.1007/s10462-024-10915-y.

[262] Stefano Woerner and Christian F. Baumgartner. "Navigating Data Scarcity Using Foundation Models: A Benchmark of Few-Shot and Zero-Shot Learning Approaches in Medical Imaging". In: *Foundation Models for General Medical AI. MedAGI 2024*. Springer Nature Switzerland, 2024, pp. 30–39. DOI: 10.1007/978-3-031-73471-7_4.

[263] Xue Li, Jameson Merkow, Noel C. F. Codella, et al. "From Embeddings to Accuracy: Comparing Foundation Models for Radiographic Classification". In: *Journal of Imaging Informatics in Medicine* (2025). DOI: 10.1007/s10278-025-01747-5.

[264] Sheng Zhang, Yanbo Xu, Naoto Usuyama, et al. "A multimodal biomedical foundation model trained from fifteen million image–text pairs". In: *NEJM AI* 2.1 (2025), AIoa2400640. DOI: 10.1056/aioa2400640.

[265] Noel C. F. Codella, Ying Jin, Shrey Jain, et al. *MedImageInsight: An Open-Source Embedding Model for General Domain Medical Imaging*. 2024. arXiv: 2410.06542.

[266] Peter Neidlinger, Omar S. M. El Nahhas, Hannah Sophie Muti, et al. "Benchmarking foundation models as feature extractors for weakly supervised computational pathology". In: *Nature Biomedical Engineering* 10.6 (2025), pp. 1113–1123. DOI: 10.1038/s41551-025-01516-3.

[267] Jaeung Lee, Jeewoo Lim, Keunho Byeon, and Jin Tae Kwak. "Benchmarking pathology foundation models: Adaptation strategies and scenarios". In: *Computers in Biology and Medicine* 190 (2025), p. 110031. DOI: 10.1016/j.compbiomed.2025.110031.

[268] Dequan Wang, Xiaosong Wang, Lilong Wang, et al. "A Real-world Dataset and Benchmark For Foundation Model Adaptation in Medical Image Classification". In: *Scientific Data* 10.1 (2023), p. 574. DOI: 10.1038/s41597-023-02460-0.

[269] Google. *Derm Foundation Model*. 2024. URL: https://developers.google.com/health-ai-developer-foundations/derm-foundation.

[270] Sonnet Xu, Haiwen Gui, Veronica Rotemberg, Tongzhou Wang, Yiqun T Chen, and Roxana Daneshjou. *A framework for evaluating the efficacy of foundation embedding models in healthcare*. medRxiv preprint. 2024. DOI: 10.1101/2024.04.17.24305983.

[271] Ruinan Jin, Zikang Xu, Yuan Zhong, et al. "FairMedFM: fairness benchmarking for medical imaging foundation models". In: *Proceedings of the 38th International Conference on Neural Information Processing Systems*. 3535. 2024, pp. 111318–111357. URL: https://dl.acm.org/doi/10.5555/3737916.3741451.

[272] Joana Huix, Adithya Ganeshan, Johan Haslum, Magnus Soderberg, Christos Matsoukas, and Kevin Smith. "Are Natural Domain Foundation Models Useful for Medical Image Classification?" In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 7619–7628. DOI: 10.1109/WACV57701.2024.00746.

[273] Catarina Barata, Jorge S. Marques, and M. Emre Celebi. "Deep Attention Model for the Hierarchical Diagnosis of Skin Lesions". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2019, pp. 2757–2765. DOI: 10.1109/CVPRW.2019.00334.

[274] Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. "DINOv2: Learning Robust Visual Features without Supervision". In: *Transactions on Machine Learning Research* (2024). URL: https://openreview.net/pdf?id=a68SUt6zFt.

[275] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, et al. "DINOv3". In: *Transactions on Machine Learning Research* (2025). URL: https://openreview.net/pdf?id=2NlGyqNjns.

[276] Thomas Cover and Peter Hart. “Nearest neighbor pattern classification”. In: *IEEE Transactions on Information Theory* 13.1 (1967), pp. 21–27. DOI: 10.1109/tit.1967.1053964.

[277] David R Cox. “The regression analysis of binary sequences”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 20.2 (1958), pp. 215–232. DOI: 10.1111/j.2517-6161.1958.tb00292.x.

[278] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine Learning* 20.3 (1995), pp. 273–297. DOI: 10.1007/BF00994018.

[279] Leo Breiman. “Random Forests”. In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[280] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors”. In: *Nature* 323.6088 (1986), pp. 533–536. DOI: 10.1038/323533a0.

[281] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. Association for Computing Machinery, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.

[282] Jinghan Ru, Siyuan Yan, Yuguo Yin, Yuexian Zou, and Zongyuan Ge. *DermoGPT: Open Weights and Open Data for Morphology-Grounded Dermatological Reasoning MLLMs*. 2026. arXiv: 2601.01868.

[283] Yize Liu, Siyuan Yan, Ming Hu, et al. *DermAgent: A Self-Reflective Agentic System for Dermatological Image Analysis with Multi-Tool Reasoning and Traceable Decision-Making*. 2026. arXiv: 2605.14403.

Appendices

The appendices provide supplementary material related to the studies presented throughout this thesis. These materials include extended methodological details, additional results, and supporting analyses that complement the main chapters without interrupting the flow of the thesis. The appendices are organized according to the main research components of the thesis, including synthetic dataset generation, local language models for longitudinal follow-up, multimodal dermatology model development, and foundation model benchmarking.

- **Appendix A — DermaSynth — Section 3.3**
  This appendix includes supplementary information related to the DermaSynth dataset, including details of the synthetic image–text generation pipeline, prompt design, data sources, and filtering steps.

- **Appendix B — Follow-up Study of Pemphigus Disease — Section 4.4**
  This appendix provides additional material related to the privacy-preserving local language model study for longitudinal pemphigus follow-up. It includes supplementary details on the feature schema, prompting strategy, evaluation procedure, and additional analyses of model-generated clinical summaries.

- **Appendix C — DermatoLlama — Section 4.5**
  This appendix provides supplementary material related to DermatoLlama and the SCALEMED framework, including additional methodological details, model outputs, and supporting analyses for the resource-efficient dermatology vision-language model described in the thesis.

- **Appendix D — Benchmarking of Foundation Models in Dermatology — Section 5.3**
  This appendix provides supplementary material related to the hierarchical benchmark of foundation models for dermatology. It includes additional details on the evaluated models and supplementary results.

- **Appendix E — Supplementary Resources**
  This appendix provides supplementary resources developed throughout this thesis and deposited in the accompanying Mendeley Data repository. The materials support the development and reproducibility of DermaSynth, DermatoLlama, and DermaBench. They include instruction libraries and seed tasks for instruction-following, knowledge-based, and anamnesis-based question–answer synthesis; disease lists; structured anamnesis templates; expert-curated benchmark question sets; DermaBench annotation guidelines; hierarchical annotation resources for the DERM12345 and DDI datasets; example model outputs; and a video demonstration of DermatoLlama. The DermaBench

resources include a DERM12345 hierarchical dermoscopic question set with 93 root questions and 320 total questions, as well as DDI annotation guidelines and question lists used for clinical skin image annotation. All supplementary materials are publicly available through Mendeley Data under the CC BY-NC 4.0 license.

Appendix A

DermaSynth Appendix

This appendix provides supplementary material related to "DermaSynth: Synthetic Image–Text Supervision for Dermatology," which is described in Section 3.3.

A.1 Prompts

A.1.1 Question Generation Prompt

We used the following prompt to generate the question set by giving metadata to ChatGPT o1 pro:

Instructions

Write questions to generate synthetic data using large language models.

For generalizability, we need to create more prompt samples so that, while using the API to generate the synthetic dataset, we can randomly sample from any prompt example.

Let's continue enhancing the dataset_name dataset prompts. Generate more prompt types, starting from 1. You can use the current ones if you'd like.

The only things available to us are images, labels, and metadata. We cannot assume anything about the images, and we will not annotate anything.

The prompt should be centered on the image, not the label.

Note that we do not always have to give the label and ask a question. We can create prompt types that ask for the label, metadata, etc., depending on the dataset.

This simulates a dermatologist using the chatbot and inferring from an image by asking questions such as "What is shown in this image?" or "What type of ... is shown in this image?"

Dataset Context

```
{metadata_context}
```

(e.g., Anatomical Site - Sex - Skin Type)

A.1.2 API Prompt Instruction

During synthetic data generation, we use the following instruction prompt to focus on visible lesion characteristics and not reveal or rely on metadata:

Instructions

You are a dermatologist examining an image of a skin lesion or dermatologic condition.

You do not have direct access to the patient's metadata. Although contextual information is provided below, do not reference or rely on it in your answer.

Base your answer solely on the visible characteristics of the lesion or dermatologic condition shown in the image.

Provide your answer in exactly the format specified below.

Context

`{metadata_context}`

(e.g., Label: Actinic keratosis - Age: 70.0 - Anatomical Site: head/neck - Sex: female)

Question

`{prompt_text}`

Response Format

`{answer}`

A.2 Example Questions

All questions are accessible in the code repositories.

Example General Questions

- What are the most prominent visual features in this lesion or disease?
- Based on the image, what might be the diagnosis?
- What conditions might present similarly to the lesion or disease in this image?
- How would you rank the most likely alternative diagnoses for this lesion or disease?
- How would you explain the lesion or disease in this image to a worried patient?
- What language would you use to describe this lesion or disease in a patient consultation?
- What follow-up procedures might be necessary for this lesion or disease?
- What further evaluations are essential for diagnosing or treating this lesion or disease?
- Develop a question–answer set for teaching purposes about this lesion or disease.
- Write a concise Q&A pair for explaining the lesion or disease in this image.

Example Dataset-Specific Questions (for the DERM12345 dataset [115])

- Based on the appearance of this lesion/skin disease, what condition might it represent?
- Considering its morphology, what might this lesion or disease indicate?
- What does the lesion or disease's distribution of pigmentation suggest?
- Could this lesion/skin disease represent a malignant condition? Why or why not?
- What dermatologic disorders might resemble the lesion/skin disease in this image?

A.3 Data Generation Prompts

Carefully designed prompt templates were crucial for generating high-quality synthetic data for DermaSynth. For each data category, we developed tailored prompts optimized to enhance the accuracy, coherence, and clinical relevance of the generated content.

A.3.1 Image-Grounded Prompt Template for Open-Access Datasets

For open-access datasets containing dermatologic images with metadata, metadata were included as contextual input during prompt construction; however, the answer-generation instruction explicitly prohibited the model from referencing or relying on these metadata and directed it to focus on visible image characteristics:

Instructions

You are a dermatologist evaluating an image of a skin lesion or dermatologic condition.

You do not have direct access to the patient's clinical metadata. Although contextual information is provided below for internal reference, you must not mention, reference, or rely on this information when formulating your response.

Base your answer exclusively on the visible characteristics of the lesion or dermatologic condition shown in the image.

Provide your response using exactly the format specified below.

Context for Internal Reference Only

```
{metadata_context}
```

Question

```
{prompt_text}
```

Response Format

```
{"question": "<question>", "answer": "<answer>"}
```

Output only the JSON object. Do not include any additional explanation, commentary, or text outside the specified JSON structure.

A.3.2 Prompt Template for Training QA Generation (BIOMEDICA)

For the BIOMEDICA dataset, we used the following prompt template for training QA generation:

Instructions

You are an AI assistant tasked with generating question–answer pairs about a medical image from a research paper.

Generate exactly three question–answer pairs concerning the visible characteristics shown in the image. Although the figure caption and image context are provided to clarify the image type and medical setting, both the questions and answers must focus exclusively on features that can be directly observed in the image.

The image context provided below contains all mentions and references to the image throughout the research paper, concatenated into a single passage. It may therefore contain duplicate or overlapping information. Use this context only to understand the general medical setting of the image and not as a source of non-visible information for the generated questions or answers.

Context

Figure Caption: `{caption}`

Image Context: `{image_context}`

Response Format

Return the response using exactly the following JSON structure:

```
{
  "qa_pairs": [
    {
      "question_1": "<first_question>",
      "answer_1": "<first_answer>"
    },
    {
      "question_2": "<second_question>",
      "answer_2": "<second_answer>"
    },
    {
      "question_3": "<third_question>",
      "answer_3": "<third_answer>"
```

```
    }
  ]
}
```

Output only the JSON object. Do not include any additional commentary, explanation, or text outside the specified JSON structure.

A.3.3 Prompt Template for Instruction-Following QA Synthesis

For generating diverse instruction-following data, we used a template that emphasized clinical relevance and diversity:

You are a board-certified dermatologist with expertise in clinical dermatology, dermatopathology, and procedural dermatology. Your task is to generate unique, factually accurate dermatology instructions that reflect realistic clinical scenarios.

Instruction requirements:

- Each instruction must be medically accurate and consistent with current dermatologic practice.
- Generate diverse instructions covering a range of tasks, including diagnosis, treatment planning, patient education, procedural techniques, dermoscopic interpretation, and differential diagnosis.
- Refer to specific dermatologic conditions (e.g., actinic keratosis, squamous cell carcinoma (SCC), melanoma, psoriasis, atopic dermatitis, hidradenitis suppurativa) rather than generic skin disorders.
- Include scenarios involving special populations, such as pediatric, geriatric, immunocompromised, and skin-of-color populations.
- Include both common and uncommon or rare dermatologic conditions.

Formatting requirements:

- Prefix each instruction with “Instruction N:”, where N denotes the instruction number.
- Immediately following each instruction, include the line “Is classification task: Yes/No”.
- Classification tasks should require selection from predefined options or categories.
- Non-classification tasks should require detailed analysis, explanation, interpretation, or planning.
- Each instruction should consist of 1–3 sentences and should be clear, concise, and specific.

Examples:

```
Instruction 1: {task['instruction']}
Is classification task: {'Yes' if task.get('is_classification', False)
 ↪  else 'No'}

Instruction 2: {task['instruction']}
```

```
Is classification task: {'Yes' if task.get('is_classification', False)
↪  else 'No'}
```

Generate `{num_instructions}` new dermatology instructions. The generated instructions must be unique with respect to both the provided examples and one another. Ensure that each instruction is factually accurate, clinically relevant, and explicitly identified as either a classification or non-classification task.

A.3.4 Prompt Template for Knowledge-Base QA Synthesis

This template was designed for broad synthetic dermatology knowledge generation rather than strict source-grounded question answering. Disease-specific knowledge-base content was supplied as contextual seed information, while the model was permitted to supplement incomplete context using its general dermatologic knowledge.

For generating knowledge from a comprehensive database of dermatologic diseases:

You are a dermatology expert specializing in the diagnosis and management of skin diseases and conditions. Your task is to answer questions concerning `{disease_name}`.

Context information:

The following information is provided for background knowledge only and must not be explicitly referenced in the response:

`{disease_context}`

Task:

`{task_title}`

Task description:

`{task_description}`

Question:

`{question}`

Provide an authoritative answer from the perspective of a dermatology expert. Use a direct, confident, and professional clinical style. Do not refer to the provided context, source material, or the process used to formulate the answer. Avoid phrases such as “based on the information”, “according to the context”, or similar wording that implies reliance on externally supplied material.

Respond as an expert with established knowledge of `{disease_name}`. If the provided context does not contain sufficient information to address all aspects of the question, supplement the response with appropriate general dermatologic knowledge.

The response should be:

- factually accurate and clinically sound;
- clearly structured, with appropriate headings and organization where relevant;

- comprehensive, addressing all clinically relevant aspects of the question; and
- practical and applicable to clinical practice.

Answer:

```
{answer}
```

A.3.5 Prompt Template for Structured Case Synthesis (Anamnesis)

For generating comprehensive dermatologic case studies with standardized structure:

You are a board-certified dermatologist tasked with creating realistic dermatology case studies for medical student training. Generate a new and unique dermatology case that follows exactly the same structure and format as the examples provided below. Ensure that the case is medically accurate, clinically realistic, and educational.

Requirements:

- Follow the exact format of the example cases, preserving all sections and their order.
- Generate a dermatologic condition that is different from all conditions represented in the examples.
- Include all of the following top-level fields: case_id, disease, base_information, final_diagnosis, questions_and_answers, treatment_of_choice, and any_important_additional_information.
- The questions_and_answers field must contain exactly the same sections, subsections, and questions as those used in the example cases.
- For each question, include the following fields:
  - `section`: the major category, such as `Main Complaint` or `History of Present Complaint`.
  - `subsection`: the corresponding subcategory, when applicable, such as `Location` or `Onset`.
  - `question`: the exact wording of the question used in the examples.
  - `answer`: the patient's response.
  - `answer_status`: one of `answered`, `not_applicable`, or `unanswered`.
- If `answer_status` is `not_applicable` or `unanswered`, include a `reason_explanation` field explaining why.
- Ensure that `base_information` includes age, gender, and other relevant medical or dermatologic history.
- Use `subsection` as the nested key for question subcategories, and ensure that the corresponding parser and schema use this key consistently.

- Ensure that all clinical information, including history, examination findings, diagnosis, and treatment, is medically accurate and internally consistent.

Example cases:

[Example cases in JSON format]

Generate one new and unique dermatology case that follows exactly the same structure as the examples. The `questions_and_answers` section must preserve the same section organization, subsections, and exact question wording, while providing new, medically accurate answers appropriate to the generated case.

The case must represent a dermatologic condition not included in the examples while maintaining the same JSON schema and organizational structure.

Output only the newly generated case as valid JSON. Do not include markdown, commentary, explanations, or any additional text outside the JSON object.

A.4 Wikipedia API Scraping

We developed an asynchronous scraper to construct a comprehensive dermatologic knowledge base from Wikipedia. Leveraging Python's asyncio and aiohttp libraries, we systematically queried the Wikipedia API to retrieve an exhaustive list of skin conditions from the "List_of_skin_conditions" page. For each listed condition, we extracted relevant page content, structured it into clear subsections (e.g., overview, symptoms, treatments), and excluded irrelevant sections such as references. Section headers were retained to ensure organized data representation. This process resulted in approximately 2,700 skin-condition entries. To adhere to Wikipedia's API rate limits, we implemented request throttling with 1-second delays between calls. The final knowledge base was saved in JSON format, enabling efficient retrieval for instruction-following data generation.

Appendix B

Follow-up Study of Pemphigus Disease

This appendix provides supplementary material related to "Privacy-Preserving Local Language Models for Pemphigus Follow-Up," which is described in Section 4.4.

The fixed main prompt and all feature-specific questions, permitted output values, and explanatory notes used in the final evaluation are provided below. Each clinical feature was queried separately using the same fixed main prompt.

B.1 Main Prompt

> You are an expert dermatologist tasked with summarizing long-term clinical notes for patients with pemphigus. Based on the provided input type, options, and notes, please complete the request by generating output in horizontal CSV format that incorporates data from all patient visits.

B.2 Clinical Features

1. `report_id`, Open text
2. `patient_id`, Open text (unique patient identifier)
3. `report_span_start_date`, Date (YYYY-MM-DD) | NA
4. `report_span_end_date`, Date (YYYY-MM-DD) | NA
5. `visit_count_est`, Integer | NA
6. `pemphigus_diagnosis`, Yes | No | NA
7. `pemphigus_subtype`, Pemphigus Vulgaris | Pemphigus Foliaceus | Paraneoplastic Pemphigus (PNP) | IgA pemphigus | Pemphigus Herpetiformis | Pemphigus Vegetans | NA
8. `diagnosis_confidence`, High | Low
    a. 1) "High" if at least 2 of the following are concordant: histology, direct immunofluorescence, anti-

DSG1/3 values; 2) “Low” if 1 or none are concordant.

9. `phenotype_mucosal`, Yes | No | NA

10. `phenotype_cutaneous`, Yes | No | NA

11. `current_disease_activity`, Active | Partial remission on treatment | Partial remission off treatment | Minimal treatment | Complete remission on treatment | Complete remission off treatment | NA

    a. You need to decide `current_disease_activity` based on that information:

        i. Complete remission off therapy: No new lesions for at least 2 months without receiving any treatment.

        ii. Complete remission on therapy: No new lesions while receiving minimal therapy.

        iii. Minimal therapy: Prednisone 10 mg/day (or equivalent) for at least 2 months and/or minimal adjuvant therapy.

        iv. Partial remission off therapy: Development of lesions that heal spontaneously within 1 week, occurring over a 2-month period without treatment.

        v. Partial remission on therapy: Development of lesions that heal within 1 week while receiving minimal therapy.

        vi. Active (relapse/flare): While the disease is under control, the appearance of 3 or more new lesions per month that do not heal spontaneously within 1 week.

12. `last_flare_date`, Date (YYYY-MM-DD) | NA

13. `flare_count_est`, Integer | NA

    a. The initial presentation of the disease should not be counted, unless the patient received the initial diagnosis elsewhere and then came here.

14. `histology_pemphigus_compatible`, Yes | No | NA

15. `dif_pemphigus_compatible`, Yes | No | NA

16. `anti_dsg1_positive_at_first_diagnosis`, Numeric | Positive | Negative | NA

17. `anti_dsg3_positive_at_first_diagnosis`, Numeric | Positive | Negative | NA

18. `anti_dsg1_last_value`, Numeric | Positive | Negative | NA

19. `anti_dsg1_last_date`, Date (YYYY-MM-DD) | NA
20. `anti_dsg3_last_value`, Numeric | Positive | Negative | NA
21. `anti_dsg3_last_date`, Date (YYYY-MM-DD) | NA
22. `current_systemic_treatment`, None | Steroid | Azathioprine | MMF | MTX | Cyclophosphamide | Rituximab | Combination | NA
23. `current_prednisone_mg`, Numeric | NA
    a. If the patient is not receiving treatment, enter 0.
24. `current_steroid_duration_months`, Numeric | NA
25. `total_steroid_duration_months`, Numeric | NA
26. `azathioprine_ever`, Yes | No
27. `mmf_ever`, Yes | No
28. `methotrexate_ever`, Yes | No
29. `cyclophosphamide_ever`, Yes | No
30. `immunosuppressant_stop_reason`, Ineffective | Adverse event | Remission | Other | NA
31. `rituximab_ever`, Yes | No
32. `rituximab_cycles_count`, Integer
33. `rituximab_first_date`, Date (YYYY-MM-DD) | NA
34. `rituximab_last_date`, Date (YYYY-MM-DD) | NA
35. `rituximab_response`, Complete | Partial | None | NA
36. `ivig_ever`, Yes | No
37. `ivig_last_date`, Date (YYYY-MM-DD) | NA
38. `plasmapheresis_ever`, Yes | No
39. `plasmapheresis_last_date`, Date (YYYY-MM-DD) | NA

40. `serious_infection_ever`, Yes | No | NA
    a. Infections requiring hospitalization and intravenous treatment (treatment through a vein).
41. `cytopenia_ever`, Yes | No | NA
    a. Complete Blood Count (CBC): If available and there is no toxicity, write “No”; if there is no data, write “NA”.
42. `liver_toxicity_ever`, Yes | No | NA
    a. Biochemistry: If available and there is no toxicity, write “No”; if there is no data, write “NA”.
    b. Hepatosteatosis (fatty liver) does not indicate liver toxicity.
43. `renal_toxicity_ever`, Yes | No | NA
    a. Biochemistry: If available and there is no toxicity, write “No”; if there is no data, write “NA”.
44. `infusion_reaction_ever`, Yes | No | NA
45. `steroid_complications_ever`, Yes | No | NA
46. `steroid_complication_hyperglycemia`, Yes | No | NA
    a. If there is suspicion of steroid complications, you can mark the criterion as “Yes”.
47. `steroid_complication_hypertension`, Yes | No | NA
    a. If there is suspicion of steroid complications, you can mark the criterion as “Yes”.
48. `steroid_complication_osteoporosis_or_osteopenia`, Osteoporosis | Osteopenia | No | NA
    a. Osteopenia: $-2.5 <$ T-score < -1.0; osteoporosis: T-score ≤ -2.5.
49. `steroid_complication_cataract`, Yes | No | NA
    a. If there is suspicion of steroid complications, you can mark the criterion as “Yes”.
50. `steroid_complication_glaucoma`, Yes | No | NA
    a. If there is suspicion of steroid complications, you can mark the criterion as “Yes”.
51. `steroid_complication_myopathy`, Yes | No | NA
    a. If there is suspicion of steroid complications, you can mark the criterion as “Yes”.

52. `steroid_complication_avascular_necrosis`, Yes | No | NA
    a. If there is suspicion of steroid complications, you can mark the criterion as “Yes”.
53. `medication_start_stop_dates_present`, Yes | Partial | No | NA
54. `dose_changes_documented`, Yes | No | NA
55. `longitudinal_course_clear`, Yes | No | NA
56. `current_status`, Open text
    a. You need to write a plain-text report based on that information (a single paragraph, no bullet list). Include: pemphigus type; current disease activity; last eye exam date; last bone density test date; osteoporosis treatment at visit (if valid); last QuantiFERON date; isoniazid (INH) prophylaxis date (if given); last hepatitis serology date; last CBC date; last liver function tests (LFTs) date; rituximab ever (if yes, last rituximab date); current prednisone use; any safety/adverse events; treatments received and dates; side effects (immunosuppressant-related infections; steroid complications: hyperglycemia, hypertension, osteoporosis, cataract, glaucoma, myopathy, avascular necrosis, other; rituximab complications: infusion reaction, cytopenia, serious infection, other). Also include major surgery (cardiac surgery, organ transplantation, etc.) and comorbid conditions. If there is suspicion of steroid complications, mention it in the summary.
    b. If repeated entries occur but the date stays the same while text changes (suggesting multiple admissions), warn: (1) if both date and text are the same $\rightarrow$ duplicate entry; (2) if date is the same but text is different $\rightarrow$ possible date error.
    c. The treatment summary should indicate the start year and end year of treatment. If unclear, state that it is unclear.

Appendix C

DermatoLlama

This appendix provides supplementary material related to "DermatoLlama: Resource-Efficient Dermatology Vision-Language Modeling," which is described in Section 4.5.

C.1 PubMed Scraping Tool

PubMed Central (PMC) provides open-access scientific articles frequently accompanied by figures and supplementary materials.

We first implemented a multi-threaded scraping approach to extract data from PMC's website and servers, enabling efficient and flexible retrieval for various dermatologic conditions. For each keyword (e.g., "dermatology," "skin lesion," "cutaneous disorders"), we generated a text file listing relevant PMC IDs. Each PMC ID was then used to download associated HTML files. Leveraging Beautiful Soup, we systematically extracted article texts, figure captions, figure mentions, and corresponding figures directly from PMC's website at `https://pmc.ncbi.nlm.nih.gov/articles/PMC_ID/`. Alternatively, articles could be retrieved from PubMed's FTP server; however, the FTP files often contained numerous large and unrelated supplementary files, necessitating an additional filtering step to isolate relevant article content. Through this process, we obtained a comprehensive and domain-specific corpus of dermatologic images and accompanying textual data licensed for research purposes.

To improve data quality, rigorous filtering was performed on the figures sourced from PMC. Given the substantial volume of initial data, we adopted a multi-step approach combining automated techniques with manual validation:

1. **Feature Extraction**: Image features were extracted using the `tf_efficientnetv2_b0.in1k` model from the `timm` library, facilitating systematic image analysis.
2. **Dimensionality Reduction**: Principal Component Analysis (PCA) was applied to the extracted image embeddings, reducing dimensionality to 50 principal components. This procedure was iterated twice to enhance cluster separability, employing batch processing optimized for machines with 96 GB of RAM.
3. **Clustering**: Reduced features underwent k-means clustering (k = 20) to categorize visually similar images, effectively grouping them into dermatologic themes or disease-related clusters.

4. **Manual Inspection**: A final manual review was conducted by domain experts, focusing on clusters identified as borderline or ambiguous. This meticulous validation step ensured the removal of irrelevant or incorrectly labeled content, significantly enhancing dataset accuracy.

By systematically integrating automated filtering, advanced feature extraction, clustering methodologies, and expert manual curation, we successfully constructed a refined corpus suitable for synthetic data generation. Nevertheless, we recommend utilizing the BIOMEDICA dataset for future research [147].

C.2 Training Parameters of DermatoLlama for LoRA

- Input cutoff: 2,048
- Output cutoff: 2,048
- Learning rate: 5×10^{-5}
- Epochs: 5
- Max. gradient norm: 1.0
- Compute type: bf16
- Batch size: 2
- Gradient accumulation: 8
- LR scheduler: linear
- Optimizer: AdamW
- LoRA Configuration:
- Rank: 16
- Alpha: 16
- Dropout: 0.05
- DeepSpeed Configuration:
- Stage: 2
- Enable offload: false
- LoRA target modules:
  - `mlp.up_proj`
  - `cross_attn.v_proj`

- self_attn.k_proj
- cross_attn.q_proj
- self_attn.v_proj
- self_attn.o_proj
- self_attn.q_proj
- cross_attn.o_proj
- cross_attn.k_proj
- mlp.gate_proj
- mlp.down_proj

C.3 Prompt Template for Clinical Image MCQA Generation (BIOMEDICA)

For the BIOMEDICA dataset, we used the following prompt template for clinical image MCQA generation:

Instructions

You are an AI assistant tasked with generating a challenging multiple-choice evaluation question based on a dermatologic image from a research paper.

Generate exactly one highly challenging multiple-choice question with four answer options (A, B, C, and D). The question should require detailed interpretation of dermatologic features, patterns, or diagnostic characteristics visible in the image and should be suitable for benchmarking advanced medical image understanding.

The generated question should:

- Focus on specific visual characteristics that require dermatologic expertise to identify.
- Assess the recognition of subtle or diagnostically relevant visual features.
- Include plausible distractors that require fine-grained discrimination between related dermatologic findings or diagnoses.
- Where appropriate, require integration of visual findings with relevant dermatologic knowledge.
- Be sufficiently challenging for evaluating advanced medical image understanding capabilities.

Context

- **Figure caption:** `{caption}`.
- **Image context:** `{image_context}`.

Response Format

Return the response using exactly the following JSON structure:

```
{
  "qa_pair": {
    "question": "<difficult_dermatology_question>",
    "options": {
      "A": "<option_A>",
      "B": "<option_B>",
```

```
      "C": "<option_C>",
      "D": "<option_D>"
    },
    "correct_answer": "<A, B, C, or D>"
  }
}
```

Output only the JSON object and do not include any additional commentary, explanation, or text.

C.4 BIOMEDICA Dataset Filtering

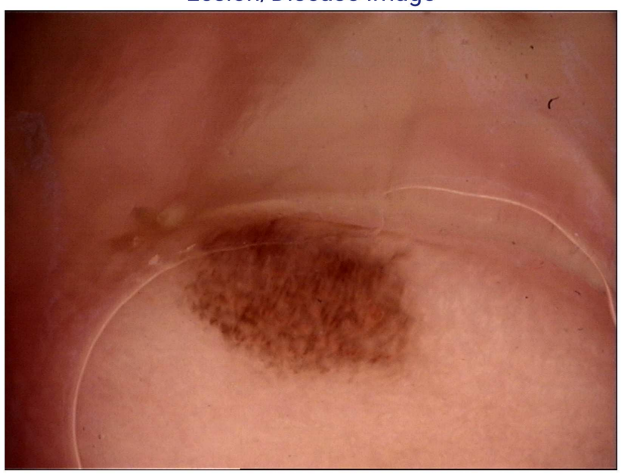
To efficiently extract dermatology-specific content from the BIOMEDICA dataset, we implemented a parallelized processing pipeline designed specifically for handling large-scale WebDataset formats. Utilizing Python's ProcessPoolExecutor, our filter class concurrently processes `.tar` archives retrieved from the Hugging Face repository, significantly boosting throughput by enabling simultaneous handling of multiple `.tar` archives. The filtering logic identifies dermatology-relevant images by inspecting key metadata fields (image_primary_label, image_secondary_label, and image_panel_subtype) for predefined clinical imaging terms. For each relevant sample, both the image and its corresponding metadata, including captions and contextual information, were preserved. The resulting dataset was systematically structured into separate directories for images and metadata, facilitating streamlined and efficient integration into our downstream synthetic data generation pipeline.

C.5 DermatoLlama Example Reports

The following figures illustrate example reports generated by DermatoLlama for diverse dermatologic cases from the test dataset. Each example includes a dermoscopic image, the ground truth diagnosis, the clinical question, and the corresponding report generated by DermatoLlama.

Dermatological Case Analysis: Test Set DERM_105042

Lesion/Disease Image



Ground Truth:

Benign Melanocytic Banal Acral

Question:

You are a dermatologist tasked with writing a brief report based on this image. If possible, please provide a diagnosis. If you are unsure, please state that you are unsure.

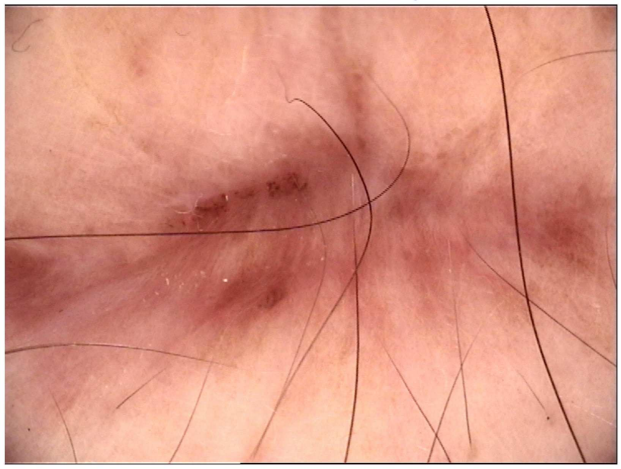
DermatoLlama-Generated Report:

This is a pigmented lesion with a somewhat irregular border and variable pigmentation. There is a central darker area with a surrounding lighter area. Given the irregular border and variable pigmentation, it is concerning for a melanocytic lesion. Further evaluation is needed to determine the diagnosis. I recommend a biopsy to rule out malignancy.

Figure C.1: Example of DermatoLlama-generated report using the test split from the DERM12345 dataset. The input image (left) is accompanied by the ground truth diagnosis, the clinical question, and the corresponding report generated by DermatoLlama (right).

Dermatological Case Analysis: Test Set DERM_110622

Lesion/Disease Image



Ground Truth:

Benign Melanocytic Dysplastic Recurrent

Question:

You are a dermatologist tasked with writing a brief report based on this image. If possible, please provide a diagnosis. If you are unsure, please state that you are unsure.

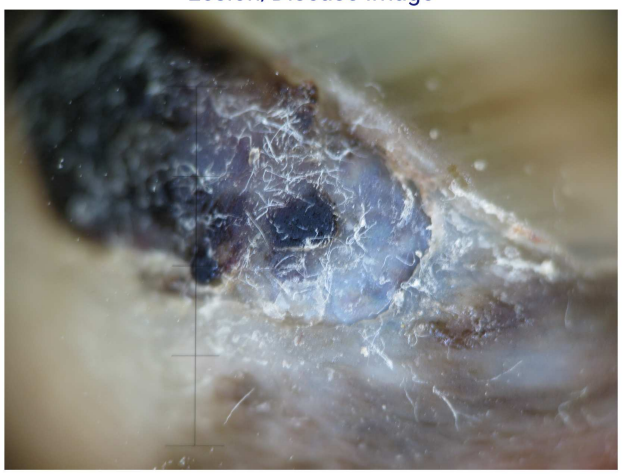
DermatoLlama-Generated Report:

This is a dermoscopic image showing a skin lesion with a light brown to reddish-brown color. There are some darker brown areas present, as well as a few small white specks. The lesion appears to be relatively well-demarcated from the surrounding skin. There are some hairs crossing the image. Based on these features, the lesion appears to be a melanocytic nevus, likely benign. However, without clinical context, I cannot be certain. I recommend monitoring for any changes.

Figure C.2: Example of DermatoLlama-generated report using the test split from the DERM12345 dataset. The input image (left) is accompanied by the ground truth diagnosis, the clinical question, and the corresponding report generated by DermatoLlama (right).

Dermatological Case Analysis: Test Set DERM_127766

Lesion/Disease Image



Ground Truth:

Malignant Melanocytic Melanoma Acral_nodular

Question:

You are a dermatologist tasked with writing a brief report based on this image. If possible, please provide a diagnosis. If you are unsure, please state that you are unsure.

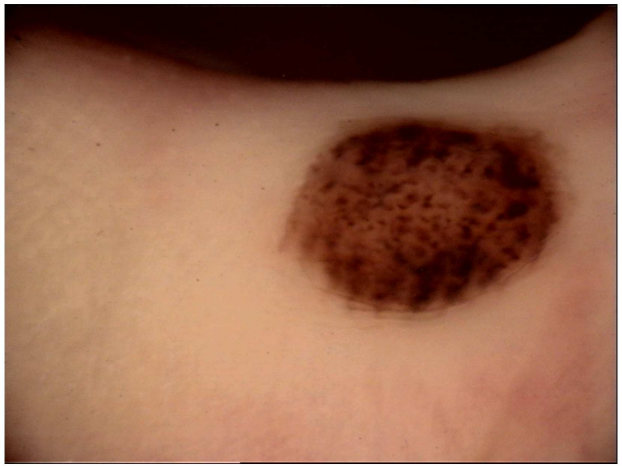
DermatoLlama-Generated Report:

The image shows a dark, irregular lesion with a central black area and surrounding lighter, bluish-gray pigmentation. There are also some fine, white, thread-like structures present on the surface. Given these features, a melanocytic lesion, possibly a melanoma, should be considered. However, the presence of the white structures makes a diagnosis difficult. Further examination and/or biopsy is recommended.

Figure C.3: Example of DermatoLlama-generated report using the test split from the DERM12345 dataset. The input image (left) is accompanied by the ground truth diagnosis, the clinical question, and the corresponding report generated by DermatoLlama (right).

Dermatological Case Analysis: Test Set DERM_134644

Lesion/Disease Image



Ground Truth:

Benign Melanocytic Dysplastic Acral

Question:

You are a dermatologist tasked with writing a brief report based on this image. If possible, please provide a diagnosis. If you are unsure, please state that you are unsure.

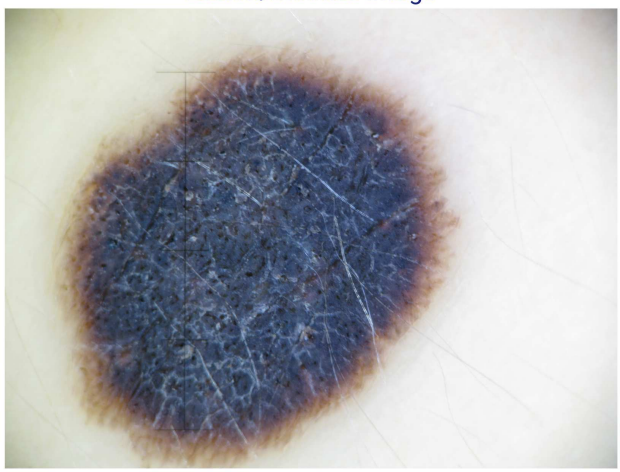
DermatoLlama-Generated Report:

The image shows a brown, oval-shaped lesion with a relatively uniform color and a somewhat reticular pattern. There are some darker and lighter areas within the lesion. The borders appear mostly regular but slightly irregular in places. Based on the image, a melanocytic nevus is a possible diagnosis. Further evaluation may be warranted to rule out other possibilities.

Figure C.4: Example of DermatoLlama-generated report using the test split from the DERM12345 dataset. The input image (left) is accompanied by the ground truth diagnosis, the clinical question, and the corresponding report generated by DermatoLlama (right).

Dermatological Case Analysis: Test Set DERM_137346

Lesion/Disease Image



Ground Truth:

Benign Melanocytic Dysplastic Spitz_reed

Question:

You are a dermatologist tasked with writing a brief report based on this image. If possible, please provide a diagnosis. If you are unsure, please state that you are unsure.

DermatoLlama-Generated Report:

This is a well-defined, dark brown to black lesion with a somewhat irregular border. The pigment is mostly uniform, but there are some subtle variations in color and texture. There are also some fine hairs visible within the lesion. Based on these characteristics, a diagnosis of a melanocytic nevus is likely, but a biopsy may be warranted to rule out malignancy.

Figure C.5: Example of DermatoLlama-generated report using the test split from the DERM12345 dataset. The input image (left) is accompanied by the ground truth diagnosis, the clinical question, and the corresponding report generated by DermatoLlama (right).

Appendix D

Benchmarking of Foundation Models in Dermatology

This appendix provides supplementary material related to "Hierarchical Benchmarking of Dermatology Foundation Models," which is described in Section 5.3.

D.1 Comprehensive Performance Tables

D.1.1 40-Subclass-Trained Hierarchical Experiment

All adapters in this experiment were trained to predict the 40 fine-grained subclasses. Their predictions were evaluated at the subclass (40), main class (15), superclass (4), superclass (2), and malignancy (2) levels. Tables D.1–D.10 report two complementary results for each task: Panel A shows five-fold validation performance as mean ± standard deviation, and Panel B shows the corresponding held-out test performance. Adapter selection was performed independently for each evaluated hierarchy level based on mean five-fold validation weighted F1-score before final held-out test evaluation.

Table D.1: Hierarchical evaluation of the 40-subclass-trained model: weighted F1-score for malignancy (2) classification (%).

Panel A. Five-fold validation (mean ± SD)

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	97.4 ± 0.3	97.9 ± 0.2	97.9 ± 0.3	96.3 ± 0.4	98.6 ± 0.2	97.2 ± 0.3
DINOv2 (Giant)	97.6 ± 0.2	98.6 ± 0.1	98.1 ± 0.1	97.1 ± 0.3	98.6 ± 0.2	97.8 ± 0.1
DINOv3	98.1 ± 0.2	97.9 ± 0.2	98.1 ± 0.2	97.3 ± 0.3	99.0 ± 0.2	97.6 ± 0.2
CLIP (Base)	97.1 ± 0.2	97.6 ± 0.4	97.7 ± 0.2	96.4 ± 0.2	98.3 ± 0.3	97.0 ± 0.2
CLIP (Large)	97.8 ± 0.4	97.9 ± 0.3	98.1 ± 0.2	97.2 ± 0.2	98.6 ± 0.2	97.2 ± 0.2
ResNet-50	95.1 ± 0.3	97.1 ± 0.3	96.8 ± 0.3	95.5 ± 0.3	95.8 ± 0.3	96.4 ± 0.4
BiomedCLIP	97.3 ± 0.1	95.6 ± 0.4	96.7 ± 0.3	96.9 ± 0.3	98.1 ± 0.3	96.9 ± 0.2
MedSigLIP	98.4 ± 0.2	98.4 ± 0.3	98.4 ± 0.2	97.9 ± 0.2	99.1 ± 0.3	98.0 ± 0.3
MedImageInsight	98.5 ± 0.2	98.7 ± 0.2	98.7 ± 0.1	98.4 ± 0.2	98.9 ± 0.3	98.6 ± 0.2
MONET	98.3 ± 0.2	98.0 ± 0.2	98.2 ± 0.2	97.5 ± 0.1	98.9 ± 0.2	97.5 ± 0.4
Derm Foundation	97.6 ± 0.2	98.6 ± 0.1	98.5 ± 0.2	97.7 ± 0.2	96.8 ± 0.6	98.0 ± 0.2
PanDerm (Base)	98.0 ± 0.2	98.1 ± 0.3	98.2 ± 0.1	97.5 ± 0.2	98.8 ± 0.2	97.8 ± 0.2
PanDerm (Large)	98.3 ± 0.3	98.5 ± 0.2	98.5 ± 0.4	98.0 ± 0.1	99.1 ± 0.1	98.0 ± 0.3

Panel B. Held-out test

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	94.47	95.59	95.38	95.09	96.28	95.37
DINOv2 (Giant)	94.62	95.88	95.56	95.31	95.51	95.10
DINOv3	95.41	95.15	95.76	94.72	96.09	95.08
CLIP (Base)	94.28	94.90	94.87	94.34	95.64	94.86
CLIP (Large)	95.19	95.27	95.54	94.96	96.08	95.65
ResNet-50	92.45	95.27	95.19	93.71	92.95	94.47
BiomedCLIP	94.53	93.51	94.59	94.47	95.23	95.08
MedSigLIP	95.72	96.42	96.20	96.17	96.49	96.70
MedImageInsight	97.86	96.86	97.09	97.23	97.20	96.96
MONET	95.12	94.59	95.25	95.57	96.03	95.72
Derm Foundation	95.00	95.88	96.35	95.39	93.72	95.53
PanDerm (Base)	94.90	95.87	95.25	95.00	96.13	95.58
PanDerm (Large)	95.13	96.41	96.26	95.52	96.25	96.41

Table D.2: Hierarchical evaluation of the 40-subclass-trained model: weighted F1-score for superclass (2) classification (%).

Panel A. Five-fold validation (mean ± SD)

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	93.4 ± 0.1	95.3 ± 0.2	95.6 ± 0.4	92.5 ± 0.3	96.2 ± 0.3	94.1 ± 0.3
DINOv2 (Giant)	94.0 ± 0.2	96.3 ± 0.3	96.3 ± 0.4	93.4 ± 0.3	96.7 ± 0.2	95.2 ± 0.3
DINOv3	94.7 ± 0.3	95.1 ± 0.5	95.8 ± 0.6	93.6 ± 0.5	96.7 ± 0.4	94.7 ± 0.6
CLIP (Base)	92.9 ± 0.2	93.6 ± 0.5	94.5 ± 0.4	91.8 ± 0.2	95.4 ± 0.2	93.6 ± 0.6
CLIP (Large)	94.5 ± 0.4	94.9 ± 0.4	95.7 ± 0.5	92.6 ± 0.5	96.5 ± 0.5	94.5 ± 0.7
ResNet-50	89.6 ± 0.5	93.3 ± 0.2	93.3 ± 0.6	90.9 ± 0.4	93.1 ± 0.4	93.0 ± 0.5
BiomedCLIP	93.4 ± 0.2	91.0 ± 0.5	92.5 ± 0.2	92.7 ± 0.4	94.8 ± 0.7	93.6 ± 0.5
MedSigLIP	95.7 ± 0.4	96.0 ± 0.6	96.3 ± 0.6	94.9 ± 0.7	97.3 ± 0.4	96.0 ± 0.4
MedImageInsight	97.2 ± 0.3	97.6 ± 0.3	97.4 ± 0.2	96.9 ± 0.4	97.8 ± 0.3	97.4 ± 0.2
MONET	95.0 ± 0.3	95.7 ± 0.5	95.8 ± 0.4	94.2 ± 0.3	96.9 ± 0.3	94.9 ± 0.4
Derm Foundation	94.9 ± 0.4	96.5 ± 0.3	96.4 ± 0.2	94.7 ± 0.6	95.4 ± 0.5	95.7 ± 0.5
PanDerm (Base)	94.2 ± 0.2	95.7 ± 0.3	95.8 ± 0.5	93.4 ± 0.6	96.7 ± 0.3	95.0 ± 0.4
PanDerm (Large)	95.3 ± 0.4	96.6 ± 0.4	96.5 ± 0.2	94.7 ± 0.3	97.3 ± 0.2	95.9 ± 0.4

Panel B. Held-out test

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	89.85	93.12	92.44	89.48	93.10	92.07
DINOv2 (Giant)	90.02	93.03	92.81	90.08	92.83	91.54
DINOv3	89.62	92.06	92.92	89.65	93.06	91.05
CLIP (Base)	88.25	90.25	91.45	88.90	91.64	90.79
CLIP (Large)	89.00	91.24	92.07	88.96	91.91	90.31
ResNet-50	85.20	90.11	89.96	87.86	89.83	89.20
BiomedCLIP	88.49	89.47	89.53	89.13	90.73	89.99
MedSigLIP	90.84	93.76	93.58	91.83	94.37	92.82
MedImageInsight	95.38	94.64	95.13	95.76	95.62	95.62
MONET	90.22	91.86	92.34	91.34	92.95	91.49
Derm Foundation	90.74	93.22	92.89	91.96	92.16	92.83
PanDerm (Base)	88.80	93.21	91.49	89.72	92.47	91.26
PanDerm (Large)	89.66	92.51	93.12	90.75	93.33	92.47

Table D.3: Hierarchical evaluation of the 40-subclass-trained model: weighted F1-score for superclass (4) classification (%).

Panel A. Five-fold validation (mean ± SD)

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	91.7 ± 0.2	93.9 ± 0.1	94.1 ± 0.4	89.9 ± 0.4	95.2 ± 0.4	92.2 ± 0.4
DINOv2 (Giant)	92.4 ± 0.4	95.5 ± 0.4	95.0 ± 0.3	91.4 ± 0.5	95.9 ± 0.4	93.8 ± 0.3
DINOv3	93.3 ± 0.4	93.8 ± 0.3	94.5 ± 0.7	91.5 ± 0.5	96.0 ± 0.4	93.0 ± 0.6
CLIP (Base)	90.8 ± 0.2	91.9 ± 0.4	93.0 ± 0.4	89.2 ± 0.2	94.3 ± 0.3	91.3 ± 0.7
CLIP (Large)	93.0 ± 0.5	93.5 ± 0.3	94.3 ± 0.3	90.6 ± 0.6	95.5 ± 0.5	92.5 ± 0.3
ResNet-50	86.9 ± 0.5	91.6 ± 0.4	91.2 ± 0.8	87.7 ± 0.5	90.1 ± 0.2	90.4 ± 0.6
BiomedCLIP	91.5 ± 0.4	88.4 ± 0.3	90.4 ± 0.4	90.3 ± 0.2	93.3 ± 0.7	91.3 ± 0.5
MedSigLIP	94.5 ± 0.4	95.0 ± 0.6	95.3 ± 0.6	93.3 ± 0.7	96.6 ± 0.4	94.6 ± 0.6
MedImageInsight	96.2 ± 0.2	96.6 ± 0.3	96.5 ± 0.2	95.7 ± 0.4	97.1 ± 0.3	96.4 ± 0.1
MONET	93.8 ± 0.3	94.3 ± 0.5	94.6 ± 0.4	92.4 ± 0.3	96.2 ± 0.3	93.2 ± 0.5
Derm Foundation	93.3 ± 0.3	95.6 ± 0.3	95.3 ± 0.3	93.2 ± 0.5	93.6 ± 0.5	94.4 ± 0.5
PanDerm (Base)	92.9 ± 0.3	94.4 ± 0.3	94.6 ± 0.5	91.7 ± 0.5	95.9 ± 0.2	93.3 ± 0.3
PanDerm (Large)	94.1 ± 0.4	95.6 ± 0.5	95.5 ± 0.4	93.0 ± 0.3	96.8 ± 0.2	94.5 ± 0.4

Panel B. Held-out test

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	85.11	90.00	89.05	85.15	90.22	88.75
DINOv2 (Giant)	85.81	90.42	89.93	85.79	89.70	88.04
DINOv3	85.47	89.11	90.28	84.58	90.08	87.58
CLIP (Base)	83.67	87.26	88.42	84.08	88.76	86.62
CLIP (Large)	84.64	88.68	89.34	84.43	89.12	86.64
ResNet-50	79.91	86.95	86.59	82.76	83.92	85.25
BiomedCLIP	84.60	85.53	85.97	84.43	87.80	86.56
MedSigLIP	87.46	91.70	91.31	88.73	91.94	90.58
MedImageInsight	94.14	93.02	93.55	94.14	93.89	93.89
MONET	86.36	88.20	89.31	87.26	90.16	88.59
Derm Foundation	87.64	90.84	90.94	88.65	88.40	90.16
PanDerm (Base)	84.72	90.54	88.60	85.62	90.22	88.06
PanDerm (Large)	86.02	90.59	90.82	87.12	90.58	89.92

Table D.4: Hierarchical evaluation of the 40-subclass-trained model: weighted F1-score for main class (15) classification (%).

Panel A. Five-fold validation (mean ± SD)

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	63.4 ± 0.5	64.0 ± 0.8	69.9 ± 0.6	60.1 ± 0.7	69.9 ± 0.9	66.0 ± 0.6
DINOv2 (Giant)	64.2 ± 0.9	69.5 ± 0.7	72.1 ± 0.5	62.9 ± 0.6	70.4 ± 0.5	68.1 ± 0.6
DINOv3	66.8 ± 0.9	62.6 ± 0.6	69.9 ± 0.8	63.7 ± 0.2	71.9 ± 0.5	67.4 ± 0.8
CLIP (Base)	61.4 ± 0.8	59.3 ± 1.4	66.5 ± 0.6	58.0 ± 0.4	67.9 ± 0.4	63.3 ± 0.4
CLIP (Large)	65.4 ± 1.5	63.9 ± 0.8	70.2 ± 0.4	61.0 ± 0.7	70.6 ± 0.4	66.0 ± 0.9
ResNet-50	53.3 ± 0.8	59.4 ± 0.2	62.3 ± 1.1	53.1 ± 0.7	58.8 ± 0.8	60.5 ± 0.8
BiomedCLIP	61.3 ± 1.2	52.5 ± 0.5	60.9 ± 0.7	60.2 ± 0.7	64.4 ± 0.9	63.0 ± 0.6
MedSigLIP	68.7 ± 1.2	67.0 ± 0.9	72.6 ± 1.0	66.5 ± 1.0	72.6 ± 0.8	69.9 ± 1.4
MedImageInsight	70.3 ± 0.7	67.9 ± 0.9	74.0 ± 0.7	70.0 ± 0.6	68.4 ± 1.0	72.6 ± 0.9
MONET	67.4 ± 1.0	64.4 ± 0.7	71.4 ± 0.8	64.3 ± 0.6	71.4 ± 0.5	68.1 ± 1.1
Derm Foundation	66.0 ± 0.8	71.3 ± 1.0	72.7 ± 1.1	65.7 ± 1.0	65.6 ± 1.0	69.9 ± 1.2
PanDerm (Base)	66.0 ± 0.7	65.6 ± 0.2	70.6 ± 0.5	63.7 ± 0.6	70.5 ± 0.6	67.3 ± 0.4
PanDerm (Large)	68.1 ± 0.9	68.8 ± 0.9	72.4 ± 0.7	67.1 ± 0.5	72.8 ± 0.8	69.8 ± 0.7

Panel B. Held-out test

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	50.61	52.45	59.28	51.56	57.32	58.30
DINOv2 (Giant)	51.41	57.69	60.12	53.98	57.73	58.71
DINOv3	51.76	53.52	61.29	53.43	58.47	57.84
CLIP (Base)	48.15	48.19	55.86	49.57	54.57	55.35
CLIP (Large)	50.97	53.49	60.36	52.39	57.89	57.94
ResNet-50	42.23	50.58	54.02	47.48	50.43	53.39
BiomedCLIP	49.59	45.25	51.65	50.69	53.77	55.20
MedSigLIP	54.92	58.62	62.34	57.57	59.55	61.96
MedImageInsight	60.78	58.96	64.89	64.41	59.89	65.28
MONET	53.00	54.18	59.56	54.93	58.01	59.73
Derm Foundation	55.88	60.54	61.84	58.04	54.90	61.51
PanDerm (Base)	50.20	55.21	59.03	53.15	56.91	58.91
PanDerm (Large)	53.68	58.03	61.23	57.08	58.36	61.54

Table D.5: Hierarchical evaluation of the 40-subclass-trained model: weighted F1-score for subclass (40) classification (%).

Panel A. Five-fold validation (mean ± SD)

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	60.3 ± 0.4	62.1 ± 1.1	67.0 ± 0.8	56.0 ± 0.6	68.3 ± 1.1	62.4 ± 0.6
DINOv2 (Giant)	61.2 ± 0.8	67.8 ± 0.6	69.8 ± 0.7	58.9 ± 0.3	68.9 ± 0.5	64.6 ± 0.7
DINOv3	64.2 ± 0.8	60.4 ± 0.7	67.4 ± 0.8	59.4 ± 0.4	70.6 ± 0.3	64.3 ± 0.9
CLIP (Base)	57.2 ± 0.9	56.4 ± 1.6	63.4 ± 0.6	53.0 ± 0.4	65.8 ± 0.4	59.2 ± 0.6
CLIP (Large)	61.8 ± 1.1	61.7 ± 0.9	67.5 ± 0.4	56.2 ± 0.7	68.8 ± 0.4	62.1 ± 0.6
ResNet-50	49.6 ± 0.8	57.0 ± 0.5	59.1 ± 0.9	48.9 ± 0.7	56.5 ± 0.8	56.1 ± 0.6
BiomedCLIP	57.3 ± 1.4	48.0 ± 0.7	57.6 ± 0.6	54.9 ± 0.6	62.0 ± 0.8	58.7 ± 0.6
MedSigLIP	65.4 ± 1.1	65.0 ± 0.9	70.0 ± 1.0	62.9 ± 1.0	71.1 ± 0.9	66.6 ± 1.2
MedImageInsight	67.6 ± 0.8	66.0 ± 1.0	71.6 ± 1.0	66.6 ± 1.0	67.0 ± 1.1	69.5 ± 1.0
MONET	64.5 ± 1.1	62.3 ± 0.9	68.7 ± 0.9	59.5 ± 0.8	70.0 ± 0.6	64.6 ± 1.1
Derm Foundation	62.5 ± 0.8	69.4 ± 1.0	70.2 ± 1.0	62.4 ± 0.8	63.9 ± 0.9	66.4 ± 1.1
PanDerm (Base)	62.7 ± 0.8	63.7 ± 0.3	67.9 ± 0.5	59.3 ± 0.5	68.8 ± 0.5	63.8 ± 0.7
PanDerm (Large)	65.0 ± 0.7	67.2 ± 0.8	70.2 ± 0.6	62.8 ± 0.4	71.3 ± 0.8	66.5 ± 0.7

Panel B. Held-out test

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	45.98	48.85	54.56	46.34	53.52	53.21
DINOv2 (Giant)	46.62	54.57	56.24	48.72	54.14	54.32
DINOv3	47.19	49.16	56.63	48.23	54.38	52.86
CLIP (Base)	42.68	43.17	50.60	43.63	49.64	48.98
CLIP (Large)	45.50	48.01	54.88	45.33	53.11	52.60
ResNet-50	36.75	46.54	49.03	42.38	46.26	47.99
BiomedCLIP	43.05	38.52	46.44	43.99	48.09	48.67
MedSigLIP	49.40	54.15	58.43	51.66	55.32	56.16
MedImageInsight	54.33	54.35	59.98	58.06	55.04	59.69
MONET	46.80	49.01	54.76	48.36	52.66	52.80
Derm Foundation	50.14	55.91	56.57	52.59	50.38	55.80
PanDerm (Base)	44.60	50.58	54.72	47.48	52.59	53.19
PanDerm (Large)	48.18	53.69	56.70	50.96	54.14	55.91

Table D.6: Hierarchical evaluation of the 40-subclass-trained model: balanced accuracy for malignancy (2) classification (%).

Panel A. Five-fold validation (mean ± SD)

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	90.7 ± 1.8	94.0 ± 0.5	93.6 ± 0.8	89.3 ± 1.2	94.7 ± 0.6	89.6 ± 1.2
DINOv2 (Giant)	92.1 ± 0.7	95.9 ± 0.4	94.4 ± 0.9	91.6 ± 1.1	95.3 ± 1.0	91.7 ± 0.9
DINOv3	93.1 ± 0.6	94.7 ± 0.8	94.4 ± 1.1	91.2 ± 0.9	96.5 ± 0.6	90.8 ± 0.7
CLIP (Base)	90.9 ± 0.9	93.8 ± 1.2	93.3 ± 1.0	89.0 ± 1.3	94.2 ± 1.2	88.3 ± 0.7
CLIP (Large)	92.9 ± 0.8	94.0 ± 1.2	93.8 ± 1.3	90.1 ± 1.3	94.8 ± 0.9	89.9 ± 0.6
ResNet-50	81.5 ± 0.9	90.8 ± 1.3	89.3 ± 0.9	87.8 ± 1.6	82.6 ± 1.2	87.1 ± 1.5
BiomedCLIP	90.1 ± 0.9	89.9 ± 2.2	90.2 ± 1.5	90.5 ± 0.7	93.9 ± 1.1	89.4 ± 1.1
MedSigLIP	94.8 ± 0.6	95.9 ± 1.0	95.7 ± 0.5	93.5 ± 0.6	96.8 ± 1.2	92.8 ± 1.1
MedImageInsight	94.7 ± 0.6	96.5 ± 0.6	96.0 ± 0.8	95.4 ± 1.2	97.5 ± 0.5	95.4 ± 0.6
MONET	94.1 ± 0.5	94.6 ± 0.9	94.6 ± 1.1	92.1 ± 0.5	96.3 ± 0.6	90.9 ± 1.7
Derm Foundation	92.0 ± 0.7	96.4 ± 0.5	95.3 ± 0.6	92.9 ± 0.7	94.9 ± 0.5	93.3 ± 0.8
PanDerm (Base)	93.7 ± 0.8	94.7 ± 1.3	94.3 ± 0.4	92.3 ± 0.8	95.7 ± 1.1	91.2 ± 0.6
PanDerm (Large)	94.0 ± 0.3	95.9 ± 1.2	95.3 ± 1.3	92.9 ± 0.6	97.0 ± 0.6	92.3 ± 1.0

Panel B. Held-out test

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	83.33	88.56	86.06	85.88	87.88	84.42
DINOv2 (Giant)	83.41	88.73	86.17	86.93	86.69	84.08
DINOv3	84.63	89.55	88.29	85.28	87.77	82.98
CLIP (Base)	82.84	88.30	85.56	83.42	86.40	82.84
CLIP (Large)	85.76	89.81	89.80	84.71	89.22	87.31
ResNet-50	72.97	84.54	85.76	82.30	74.52	82.06
BiomedCLIP	83.18	86.85	86.83	84.40	86.70	84.24
MedSigLIP	87.36	90.89	88.75	88.91	89.65	89.41
MedImageInsight	93.59	92.07	92.58	93.58	93.93	89.57
MONET	87.17	86.11	88.52	88.72	89.92	85.91
Derm Foundation	84.74	89.46	90.85	86.97	88.62	86.33
PanDerm (Base)	85.58	90.37	87.07	86.91	88.70	85.09
PanDerm (Large)	86.63	90.52	89.33	87.60	88.96	88.51

Table D.7: Hierarchical evaluation of the 40-subclass-trained model: balanced accuracy for superclass (2) classification (%).

Panel A. Five-fold validation (mean ± SD)

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	83.9 ± 0.4	91.6 ± 0.8	91.2 ± 1.3	83.2 ± 0.9	92.3 ± 0.9	86.2 ± 1.3
DINOv2 (Giant)	86.2 ± 0.8	93.1 ± 1.2	92.9 ± 1.3	87.1 ± 0.6	93.6 ± 0.9	89.3 ± 0.9
DINOv3	87.0 ± 0.6	91.9 ± 1.0	91.7 ± 1.0	85.4 ± 0.8	93.4 ± 1.2	87.7 ± 0.8
CLIP (Base)	83.5 ± 0.8	89.4 ± 1.0	88.8 ± 1.2	82.2 ± 0.3	91.2 ± 0.6	85.3 ± 1.2
CLIP (Large)	86.5 ± 1.1	90.6 ± 0.7	91.3 ± 1.0	84.6 ± 0.6	92.8 ± 1.4	87.5 ± 1.6
ResNet-50	76.2 ± 1.2	87.6 ± 1.0	85.8 ± 1.3	80.1 ± 1.4	84.6 ± 1.1	83.7 ± 1.2
BiomedCLIP	84.7 ± 0.6	85.2 ± 1.3	85.9 ± 0.4	85.4 ± 0.9	90.4 ± 1.2	86.1 ± 0.7
MedSigLIP	89.4 ± 0.9	92.8 ± 1.1	93.0 ± 1.4	89.2 ± 1.5	94.9 ± 1.0	91.6 ± 1.0
MedImageInsight	94.1 ± 0.7	95.4 ± 0.8	95.0 ± 0.6	95.1 ± 0.8	97.1 ± 0.4	95.2 ± 0.9
MONET	88.3 ± 0.5	92.5 ± 0.9	91.7 ± 1.2	88.6 ± 0.6	93.7 ± 0.7	88.5 ± 0.9
Derm Foundation	88.1 ± 0.5	93.3 ± 1.0	93.1 ± 0.5	91.3 ± 0.7	93.3 ± 0.8	91.1 ± 0.7
PanDerm (Base)	86.7 ± 0.7	92.4 ± 1.1	92.0 ± 1.1	85.4 ± 1.1	93.8 ± 0.7	88.4 ± 0.6
PanDerm (Large)	88.9 ± 0.9	94.0 ± 0.8	92.8 ± 0.8	87.6 ± 0.4	94.8 ± 0.8	90.6 ± 1.0

Panel B. Held-out test

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	76.86	87.39	84.72	77.00	86.06	82.07
DINOv2 (Giant)	78.24	87.54	86.38	80.84	85.68	80.83
DINOv3	75.83	86.49	86.45	78.09	86.03	80.01
CLIP (Base)	74.04	83.06	81.67	76.41	83.69	78.07
CLIP (Large)	74.65	84.53	85.29	78.79	84.07	79.32
ResNet-50	67.49	80.87	78.79	74.72	78.02	75.36
BiomedCLIP	74.77	82.32	78.99	77.73	83.28	77.74
MedSigLIP	78.98	88.52	87.38	83.51	89.33	84.26
MedImageInsight	90.48	90.32	92.18	94.45	94.57	92.59
MONET	78.28	86.05	84.75	82.59	86.36	80.99
Derm Foundation	79.61	86.85	86.53	85.62	86.96	85.79
PanDerm (Base)	75.47	88.16	83.89	78.43	85.04	81.04
PanDerm (Large)	77.12	86.59	86.78	80.30	87.32	84.24

Table D.8: Hierarchical evaluation of the 40-subclass-trained model: balanced accuracy for superclass (4) classification (%).

Panel A. Five-fold validation (mean ± SD)

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	76.0 ± 1.5	85.9 ± 1.1	84.3 ± 1.7	69.9 ± 2.0	85.9 ± 0.8	74.8 ± 2.3
DINOv2 (Giant)	77.8 ± 1.5	89.4 ± 2.2	86.5 ± 1.7	75.3 ± 2.0	88.4 ± 2.2	80.0 ± 1.8
DINOv3	80.2 ± 1.4	87.0 ± 0.8	85.6 ± 1.8	73.7 ± 1.2	89.5 ± 1.6	78.1 ± 1.1
CLIP (Base)	73.4 ± 1.0	83.2 ± 1.1	81.5 ± 1.2	68.0 ± 1.1	85.1 ± 1.2	72.1 ± 1.5
CLIP (Large)	79.8 ± 1.4	84.5 ± 0.9	84.1 ± 1.2	73.1 ± 2.1	86.5 ± 1.2	75.9 ± 0.8
ResNet-50	61.7 ± 1.7	79.1 ± 1.8	75.1 ± 2.0	64.9 ± 2.3	67.4 ± 0.8	70.4 ± 1.6
BiomedCLIP	75.0 ± 1.8	75.0 ± 2.2	75.2 ± 1.8	72.0 ± 0.6	82.8 ± 1.8	73.3 ± 1.3
MedSigLIP	83.0 ± 1.6	88.3 ± 1.6	88.3 ± 1.6	79.2 ± 2.0	90.3 ± 1.6	82.8 ± 2.4
MedImageInsight	88.5 ± 1.0	92.1 ± 0.8	90.7 ± 0.9	88.4 ± 1.8	93.7 ± 0.9	89.3 ± 1.2
MONET	82.4 ± 1.0	86.6 ± 1.6	86.0 ± 2.1	77.4 ± 1.4	89.9 ± 1.0	78.3 ± 1.8
Derm Foundation	79.8 ± 0.7	89.5 ± 1.1	87.3 ± 0.6	80.9 ± 1.0	84.4 ± 0.6	82.7 ± 0.9
PanDerm (Base)	80.2 ± 1.2	86.9 ± 1.2	86.3 ± 0.7	75.9 ± 0.8	89.5 ± 1.0	78.5 ± 1.0
PanDerm (Large)	82.9 ± 0.8	89.8 ± 1.1	87.6 ± 1.8	79.2 ± 0.6	91.5 ± 1.1	82.3 ± 1.5

Panel B. Held-out test

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	56.24	75.06	68.59	55.58	70.47	66.05
DINOv2 (Giant)	57.51	74.88	71.01	58.39	69.23	62.96
DINOv3	57.09	75.79	74.53	54.61	70.02	63.02
CLIP (Base)	55.57	71.16	67.62	53.94	67.16	59.11
CLIP (Large)	56.55	74.14	73.49	55.79	68.45	61.22
ResNet-50	42.96	64.08	62.26	50.39	52.09	56.97
BiomedCLIP	55.01	72.02	63.98	56.00	67.19	60.20
MedSigLIP	62.29	78.45	74.46	65.33	75.57	72.76
MedImageInsight	83.80	83.52	83.72	85.82	85.93	80.77
MONET	61.64	70.22	71.23	62.62	72.60	65.28
Derm Foundation	62.23	75.02	75.91	66.87	68.88	70.42
PanDerm (Base)	59.02	77.57	69.15	59.72	72.98	65.61
PanDerm (Large)	62.59	76.77	74.74	63.14	73.02	69.94

Table D.9: Hierarchical evaluation of the 40-subclass-trained model: balanced accuracy for main class (15) classification (%).

Panel A. Five-fold validation (mean ± SD)

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	50.4 ± 2.9	68.4 ± 4.7	61.9 ± 3.6	41.3 ± 2.9	62.3 ± 2.3	44.5 ± 3.0
DINOv2 (Giant)	53.3 ± 3.6	72.8 ± 2.8	67.3 ± 2.3	46.3 ± 4.2	68.3 ± 4.4	52.0 ± 2.0
DINOv3	57.5 ± 4.3	70.8 ± 3.2	66.4 ± 2.3	48.0 ± 2.8	70.3 ± 4.7	49.6 ± 4.2
CLIP (Base)	49.6 ± 4.9	62.0 ± 4.4	58.1 ± 3.0	40.1 ± 2.9	63.0 ± 2.4	43.5 ± 4.4
CLIP (Large)	58.5 ± 4.6	68.9 ± 4.1	64.6 ± 2.6	43.7 ± 3.2	66.5 ± 2.8	48.9 ± 2.5
ResNet-50	36.6 ± 2.8	55.7 ± 3.1	47.7 ± 3.9	34.3 ± 2.7	43.8 ± 3.0	41.2 ± 2.6
BiomedCLIP	47.2 ± 2.8	54.8 ± 2.7	52.1 ± 2.5	43.0 ± 2.7	58.1 ± 2.4	44.9 ± 2.3
MedSigLIP	61.6 ± 6.0	71.3 ± 3.4	66.7 ± 6.3	53.6 ± 4.0	72.3 ± 3.6	55.5 ± 2.7
MedImageInsight	67.4 ± 3.1	74.0 ± 4.4	68.0 ± 0.7	58.3 ± 3.0	69.8 ± 3.0	60.7 ± 4.2
MONET	57.3 ± 4.0	67.6 ± 1.2	63.9 ± 3.4	48.0 ± 2.5	68.6 ± 3.8	50.9 ± 3.1
Derm Foundation	53.5 ± 4.4	68.2 ± 3.1	65.5 ± 1.7	54.5 ± 1.6	58.6 ± 2.8	56.9 ± 1.6
PanDerm (Base)	57.9 ± 2.9	69.1 ± 2.2	62.4 ± 5.4	46.5 ± 3.8	69.2 ± 2.2	50.3 ± 2.9
PanDerm (Large)	63.5 ± 3.3	74.4 ± 3.6	67.9 ± 2.0	52.9 ± 4.4	74.4 ± 2.7	54.8 ± 4.3

Panel B. Held-out test

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	24.46	44.79	37.43	23.52	37.08	31.60
DINOv2 (Giant)	32.53	52.28	42.69	25.59	38.23	31.49
DINOv3	24.86	51.72	39.97	24.20	38.80	29.29
CLIP (Base)	22.06	42.96	32.38	21.97	34.04	26.08
CLIP (Large)	24.79	41.73	42.42	24.24	35.47	29.68
ResNet-50	15.52	29.86	27.75	19.55	25.41	24.34
BiomedCLIP	25.33	37.69	31.57	24.23	37.02	29.18
MedSigLIP	29.78	54.56	41.87	34.02	43.34	37.61
MedImageInsight	48.41	52.75	48.87	45.69	49.84	46.81
MONET	34.23	45.83	36.45	28.72	37.58	32.20
Derm Foundation	31.06	48.24	48.06	35.35	37.66	37.36
PanDerm (Base)	24.74	43.33	37.69	25.29	37.59	31.58
PanDerm (Large)	29.03	45.40	40.62	30.05	39.59	34.85

Table D.10: Hierarchical evaluation of the 40-subclass-trained model: balanced accuracy for subclass (40) classification (%).

Panel A. Five-fold validation (mean ± SD)

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	48.7 ± 2.4	67.2 ± 1.6	59.8 ± 4.3	40.8 ± 2.2	59.1 ± 2.1	39.2 ± 2.9
DINOv2 (Giant)	51.2 ± 2.8	71.9 ± 3.6	62.9 ± 3.1	44.5 ± 2.2	61.6 ± 2.3	43.7 ± 2.5
DINOv3	59.4 ± 3.0	71.0 ± 3.3	64.9 ± 3.0	46.8 ± 2.7	69.4 ± 2.1	44.8 ± 2.0
CLIP (Base)	48.6 ± 4.4	61.8 ± 1.1	56.6 ± 1.4	39.0 ± 2.5	60.4 ± 1.8	37.8 ± 2.8
CLIP (Large)	58.3 ± 3.5	67.4 ± 3.6	60.3 ± 2.7	44.5 ± 2.1	63.5 ± 3.6	43.0 ± 2.9
ResNet-50	36.4 ± 2.9	56.9 ± 0.9	46.1 ± 2.9	34.6 ± 1.6	36.9 ± 0.5	37.4 ± 3.1
BiomedCLIP	47.6 ± 4.0	52.2 ± 2.3	50.3 ± 3.5	39.6 ± 2.6	56.4 ± 3.0	37.3 ± 2.1
MedSigLIP	62.2 ± 5.6	70.8 ± 4.8	65.0 ± 5.1	53.6 ± 2.0	71.1 ± 4.6	51.1 ± 2.4
MedImageInsight	66.7 ± 3.0	75.7 ± 4.0	68.5 ± 2.5	57.5 ± 3.1	70.6 ± 3.0	56.7 ± 3.0
MONET	61.2 ± 3.8	68.6 ± 1.0	63.2 ± 3.8	47.6 ± 1.6	69.7 ± 1.6	47.8 ± 1.9
Derm Foundation	50.8 ± 3.3	69.9 ± 2.1	63.4 ± 3.7	50.7 ± 2.5	55.2 ± 2.0	52.1 ± 2.8
PanDerm (Base)	58.0 ± 3.3	70.0 ± 3.6	61.5 ± 2.6	48.6 ± 1.5	68.0 ± 1.6	46.1 ± 1.7
PanDerm (Large)	62.6 ± 3.0	74.7 ± 3.1	67.3 ± 2.4	53.5 ± 1.9	73.7 ± 2.6	53.2 ± 5.6

Panel B. Held-out test

Model	KNN	LR	MLP	RF	SVM	XGBoost
DINOv2 (Base)	21.19	36.79	28.72	18.01	27.00	23.69
DINOv2 (Giant)	23.64	39.21	33.75	19.24	30.44	22.22
DINOv3	20.45	40.95	33.09	17.78	29.07	19.31
CLIP (Base)	16.00	30.26	24.58	15.97	24.33	16.42
CLIP (Large)	21.27	31.73	29.37	15.04	26.69	21.31
ResNet-50	9.63	24.83	22.07	13.29	16.69	17.64
BiomedCLIP	17.73	29.53	22.76	15.16	24.76	18.75
MedSigLIP	25.30	40.90	34.85	23.53	33.44	25.99
MedImageInsight	34.19	43.56	37.67	31.02	36.42	34.47
MONET	24.39	35.13	29.83	19.23	28.13	21.20
Derm Foundation	23.77	36.67	34.28	25.01	25.67	25.96
PanDerm (Base)	22.86	35.65	31.81	21.24	30.29	24.47
PanDerm (Large)	24.87	38.17	33.25	23.32	32.52	25.13

D.1.2 Four-Superclass-Trained Model Evaluation

All adapters in this experiment were trained to predict the four superclass labels. Their predictions were evaluated at the superclass (4), superclass (2), and malignancy (2) levels. Tables D.11 and D.12 report held-out test performance using weighted F1-score and balanced accuracy, respectively. Adapter selection was performed independently for each evaluated hierarchy level based on mean five-fold validation weighted F1-score before final held-out test evaluation.

Table D.11: Four-superclass-trained model evaluation: weighted F1-score on the held-out test set (%).

Model	Malignancy (2)	Superclass (2)	Superclass (4)
DINOv2 (Base)	96.43	93.30	90.43
DINOv2 (Giant)	96.12	92.80	89.75
DINOv3	97.10	93.33	91.18
CLIP (Base)	96.37	91.27	88.64
CLIP (Large)	96.70	91.90	89.55
ResNet-50	95.43	90.36	87.35
BiomedCLIP	95.60	90.43	87.21
MedSigLIP	96.93	93.62	91.69
MedImageInsight	97.53	96.28	94.72
MONET	96.28	92.69	89.91
Derm Foundation	96.14	93.94	91.38
PanDerm (Base)	96.66	93.43	90.96
PanDerm (Large)	96.61	93.46	91.06

Table D.12: Four-superclass-trained model evaluation: balanced accuracy on the held-out test set (%).

Model	Malignancy (2)	Superclass (2)	Superclass (4)
DINOv2 (Base)	87.06	84.06	70.61
DINOv2 (Giant)	86.69	83.14	68.70
DINOv3	89.28	84.89	73.24
CLIP (Base)	87.39	79.27	67.61
CLIP (Large)	89.41	81.36	69.40
ResNet-50	85.73	79.45	65.47
BiomedCLIP	86.19	79.20	65.03
MedSigLIP	89.73	84.97	74.50
MedImageInsight	92.47	93.53	84.67
MONET	88.98	83.07	71.18
Derm Foundation	88.89	88.02	74.63
PanDerm (Base)	89.20	84.75	73.33
PanDerm (Large)	90.09	85.68	74.86

D.2 Extended Visualizations

D.2.1 Embedding Space Analysis (t-SNE)

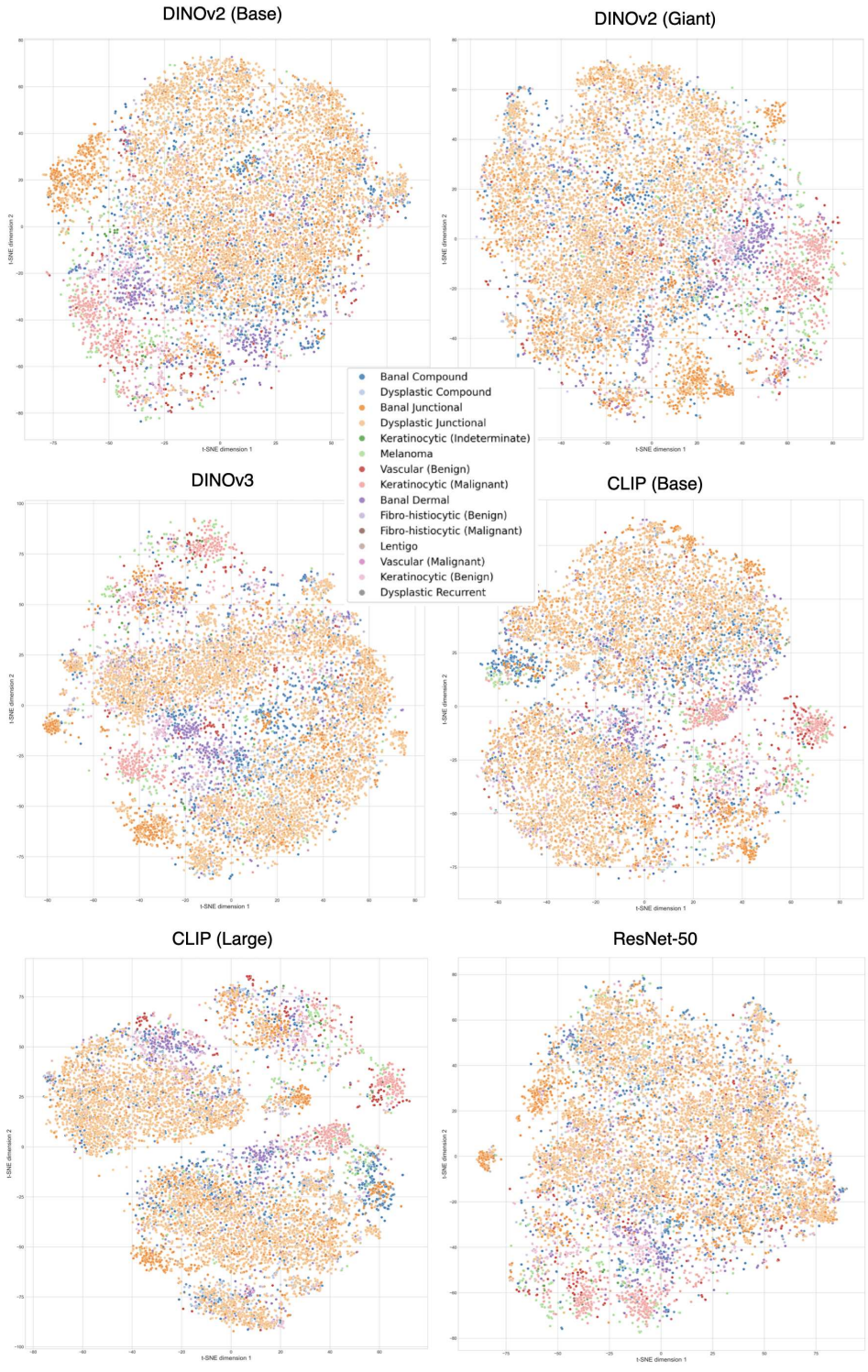


Figure D.1: t-SNE projections of embeddings for general vision models (DINOv2 (Base), DINOv2 (Giant), DINOv3, CLIP (Base), CLIP (Large), ResNet-50). Points are colored by main class.

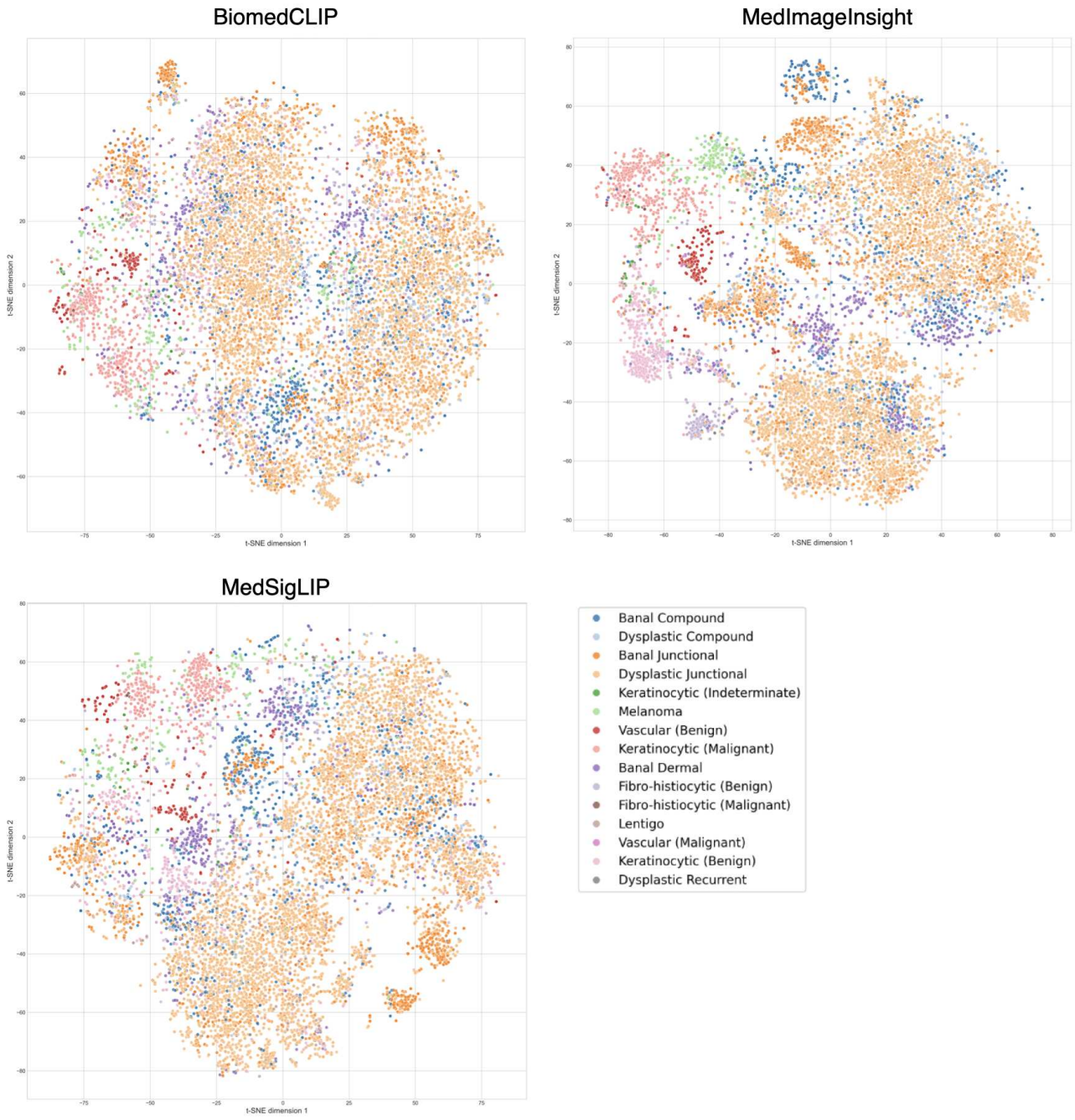


Figure D.2: t-SNE projections for general medical models (BiomedCLIP, MedImageInsight, MedSigLIP).

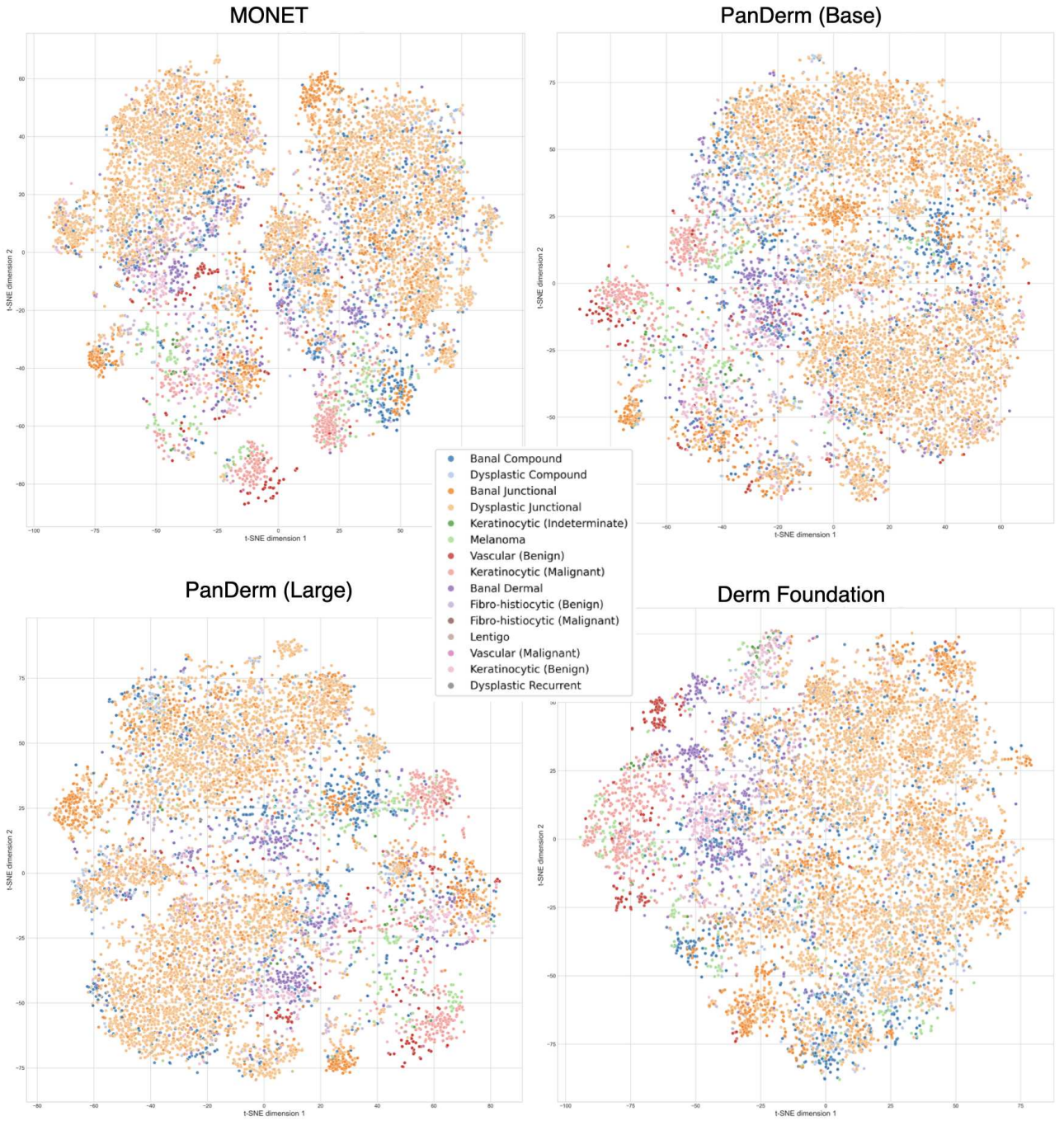


Figure D.3: t-SNE projections for dermatology-specific models (MONET, PanDerm (Base), PanDerm (Large), Derm Foundation).

D.2.2 Fine-Grained Confusion Matrices

The following figures show normalized confusion matrices at the main class (15) evaluation level, derived from the predictions of the 40-subclass-trained adapters. For clarity, only results from the MLP adapter are visualized for each foundation model.

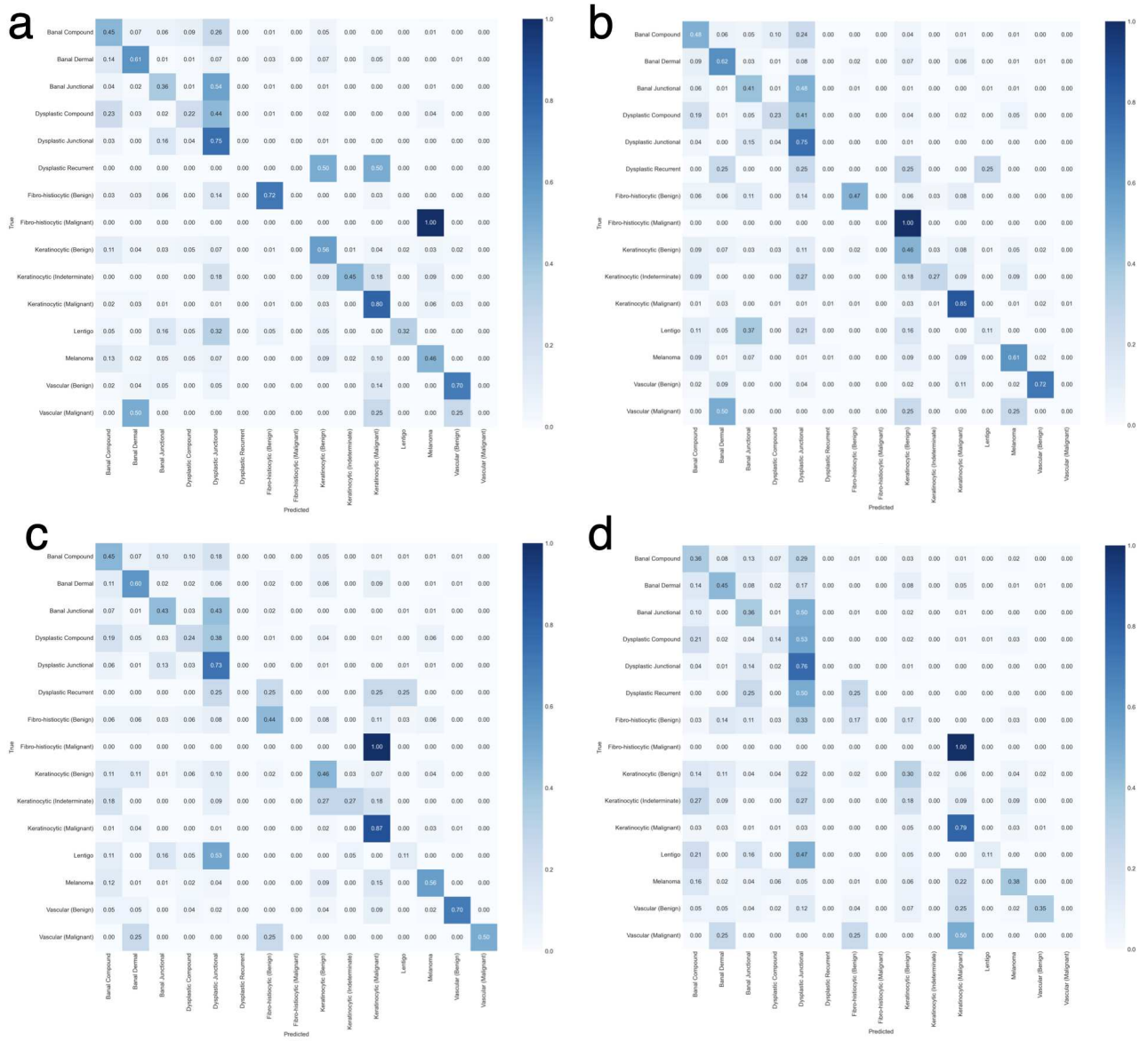


Figure D.4: Confusion matrices for general vision models. Normalized confusion matrices evaluated on the main class (15) task. The panels show results for (a) DINOv2 (Giant), (b) DINOv3, (c) CLIP (Large), and (d) ResNet-50.

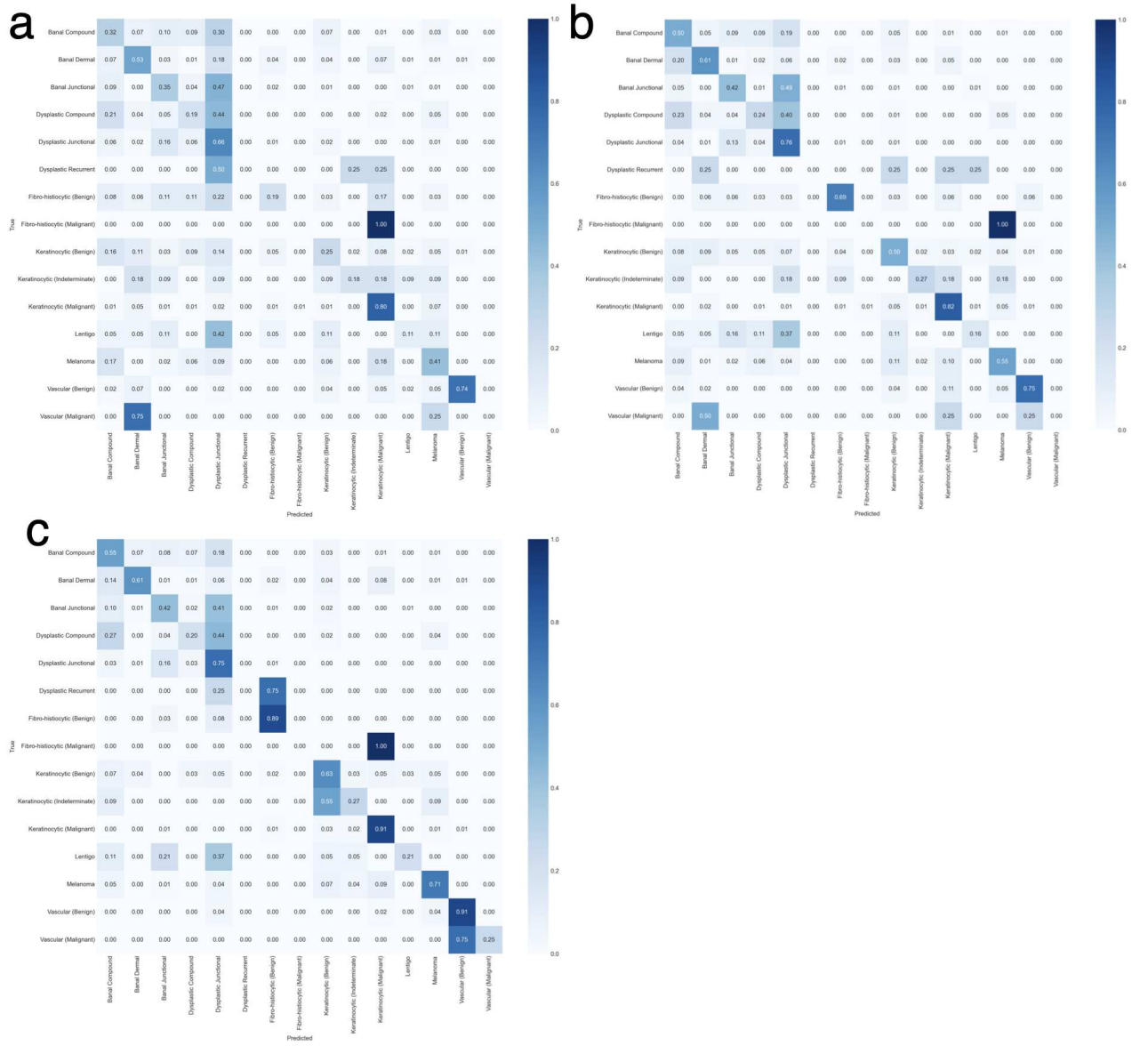


Figure D.5: Confusion matrices for general medical models. Normalized confusion matrices evaluated on the main class (15) task. The panels show results for (a) BiomedCLIP, (b) MedSigLIP, and (c) MedImageInsight.

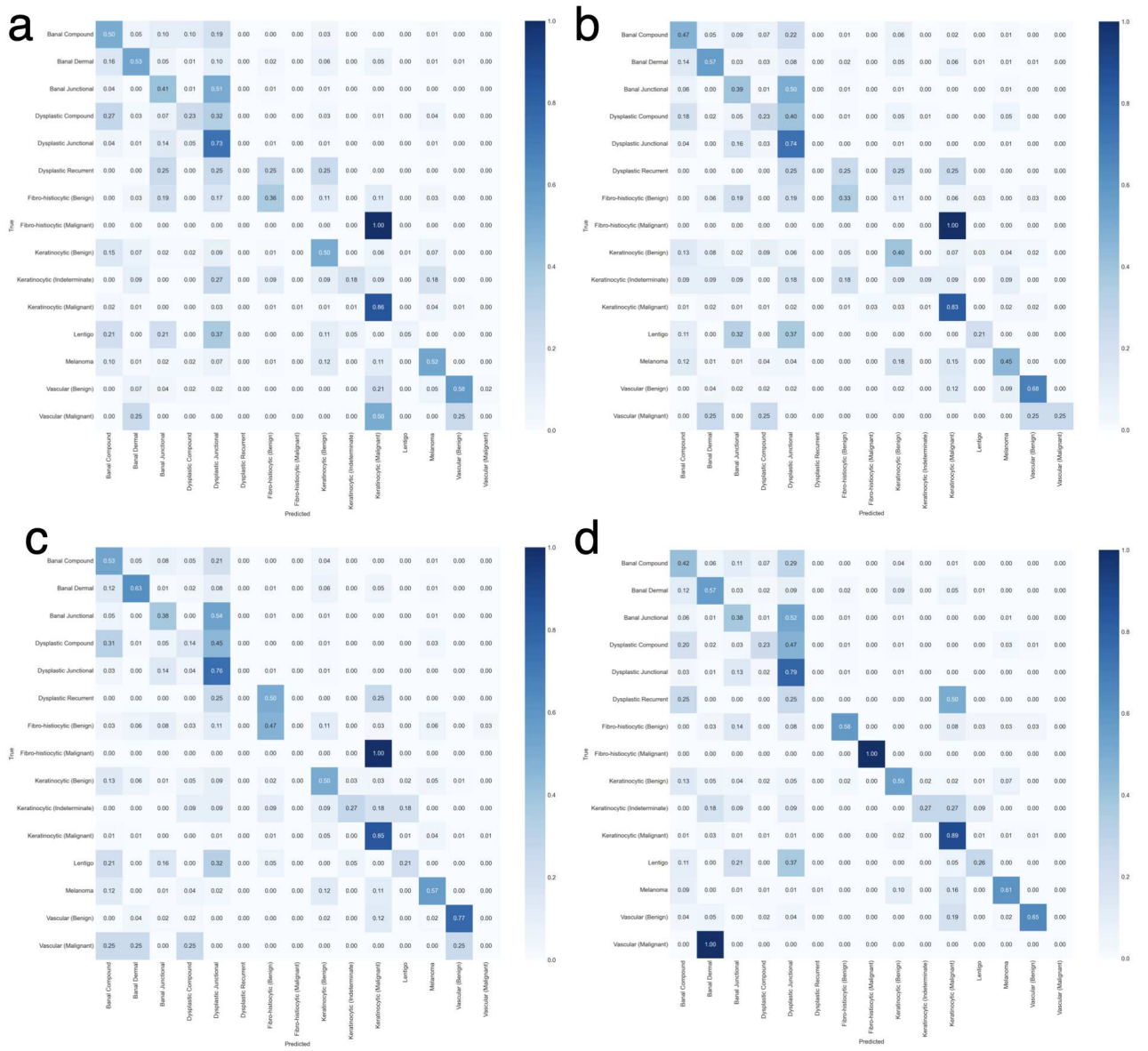


Figure D.6: Confusion matrices for dermatology-specific models. Normalized confusion matrices evaluated on the main class (15) task. The panels show results for (a) MONET, (b) PanDerm (Base), (c) PanDerm (Large), and (d) Derm Foundation.

D.2.3 Hierarchical ROC Curves

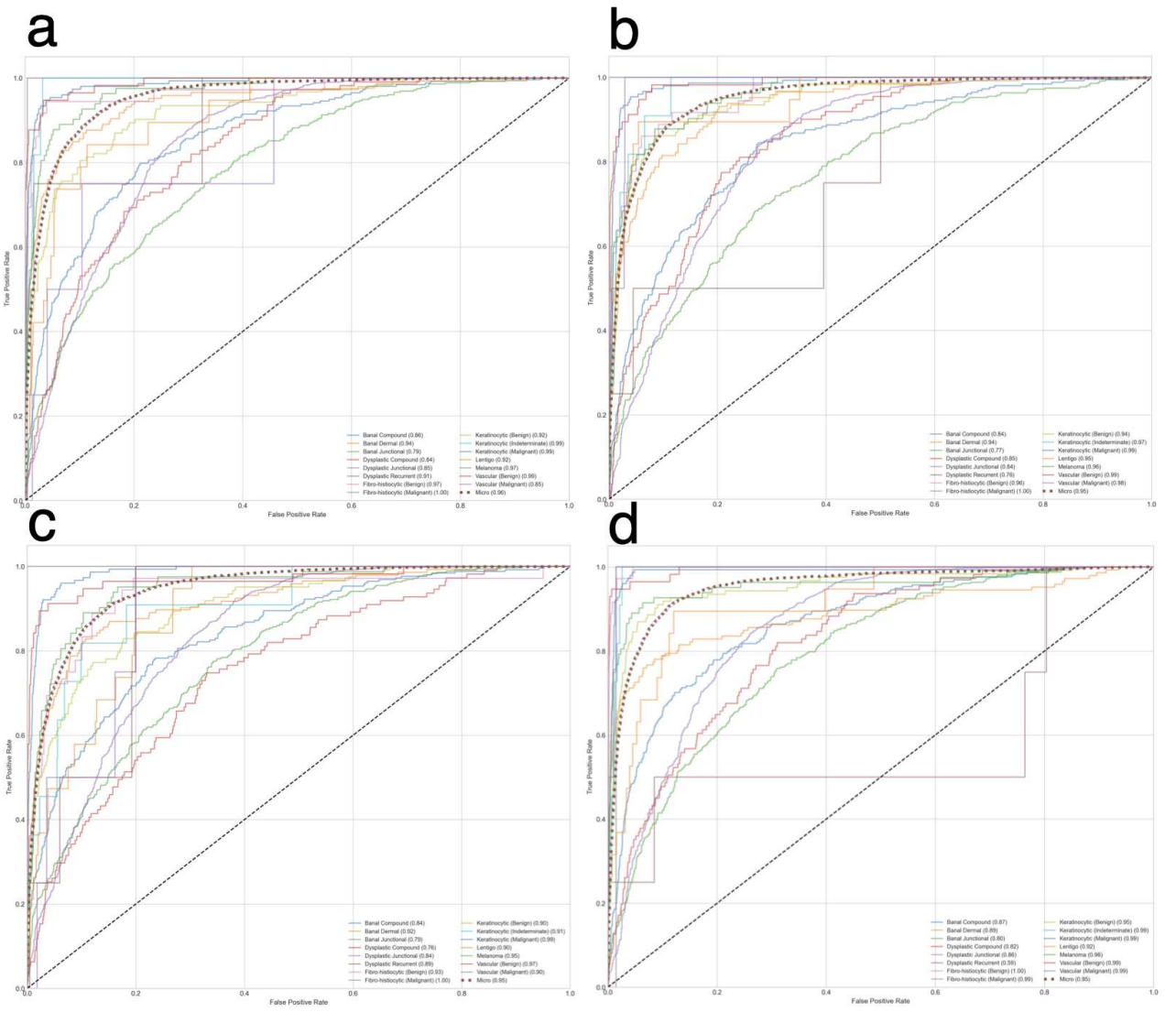


Figure D.7: One-vs-rest ROC curves at the main class (15) hierarchy level for the 40-subclass-trained experiment. The four panels show one-vs-rest ROC curves for (a) MedSigLIP, (b) Derm Foundation, (c) MONET, and (d) MedImageInsight.

Appendix E

Supplementary Resources

This appendix summarizes the supplementary resources developed and used throughout this thesis. The complete files are deposited in Mendeley Data and are available through the following persistent identifier:

DOI: 10.17632/h3wky2skbp.1

The repository contains the seed tasks, structured prompt templates, benchmark questions, hierarchical visualizations, demonstration materials, and example model outputs used in the DermaSynth, DermaBench, and DermatoLlama studies.

E.1 Supplementary Movie

E.1.1 Supplementary Movie 1: DermatoLlama Demonstration

A video demonstration of DermatoLlama in operation, including example image-based interactions and model-generated responses, is provided.

E.2 Supplementary Data

E.2.1 Supplementary Data 1: Seed Tasks for Instruction-Following Question Answering Synthesis

A collection of 190 seed tasks used for instruction-following question–answer synthesis.

E.2.2 Supplementary Data 2: Seed Tasks for Knowledge-Based Question Answering Synthesis

A collection of 40 seed tasks used for knowledge-based question–answer synthesis.

E.2.3 Supplementary Data 3: Seed Tasks for Structured Case Synthesis (Anamnesis)

A collection of 137 seed tasks used for structured clinical case synthesis based on anamnesis.

E.2.4 Supplementary Data 4: Disease List for Structured Case Synthesis

A list of 137 disease names used with the seed tasks for structured clinical case synthesis.

E.2.5 Supplementary Data 5: Anamnesis Question Template

A set of 104 predefined questions forming the anamnesis template used for structured clinical case synthesis.

E.2.6 Supplementary Data 6: DERM12345 Dataset Annotation Question List for DermaBench

An expert-curated hierarchical dermoscopic VQA question set comprising 93 root questions and 320 questions in total, including single-select, multiple-select, and open-ended formats. This resource was used for annotation of the DERM12345 component of DermaBench (Section 5.2).

E.2.7 Supplementary Data 7: DDI Dataset Annotation Guidelines for DermaBench

This resource contains the annotation guidelines used during the construction of DermaBench for the DDI dataset. This document describes the annotation workflow, question definitions, response criteria, and instructions provided to annotators to ensure consistent and standardized data annotation.

E.2.8 Supplementary Data 8: DDI Dataset Annotation Question List for DermaBench

An expert-curated hierarchical clinical skin image annotation question set comprising 21 main questions and 33 questions in total, including single-select, multiple-select, and open-ended formats. This resource was used for annotation of the DDI component of DermaBench (Section 5.2).