DOCTORAL SCHOOL
UNIVERSITÀ DEGLI STUDI DELL' INSUBRIA

Department of **Theoretical and Applied Sciences (DiSTA)**

Ph. D. program in **Computer Science and Mathematics of Computation**, **XXXVIII** cycle

Advancing Automated Skin Cancer Detection through Transformer Architectures

Candidate: **Mirco Gallazzi**
Registration number: **730379**

Supervisor: **Prof.ssa Silvia Corchs**
Co-Supervisor: **Prof Ignazio Gallo**

Coordinator: **Prof.ssa Barbara Carminati**

Academic Year 2025/2026

Abstract

Skin cancer is among the most rapidly increasing malignancies worldwide, and early detection remains the most effective strategy to reduce mortality. Melanoma, though representing a minority of cases, accounts for the majority of skin cancer–related deaths, underscoring the importance of timely and accurate diagnosis. In this context, artificial intelligence (AI) and computer vision have become essential tools to support dermatologists in their decision-making processes. Two fundamental and closely related computer vision tasks—classification, which identifies the diagnostic category of a lesion, and segmentation, which delineates its spatial extent—form the methodological foundation of this doctoral research.

Despite significant advances, existing computer-aided diagnostic systems still face critical limitations in generalization, interpretability, and clinical applicability. Even with the availability of large annotated repositories such as the International Skin Imaging Collaboration (ISIC) archive, which includes the widely adopted HAM10000 dataset, most models struggle to maintain performance under domain shifts (e.g., changes in acquisition devices, imaging conditions, patient populations, or clinical settings) or across heterogeneous clinical data. To address these issues, this thesis explores Transformer-based architectures and investigates three progressive research stages: dataset unification, segmentation analysis, and sequential task learning.

First, a Large Dataset was assembled by merging and harmonizing several public dermatological datasets to increase data diversity and mitigate class imbalance. This unified dataset enabled a comprehensive evaluation of Swin Transformer–based models for skin lesion classification, showing improved accuracy and robustness compared to single-source training. Building upon these results, the research then examined the role of segmentation in enhancing classification performance, demonstrating that segmentation alone does not guarantee improvement, yet provides a meaningful structural prior that can guide representation learning.

From these insights, a Sequential Learning framework was developed to explicitly model the interaction between segmentation and classification. A modular Sequential Swin Transformer architecture was designed to investigate how the order of learning—performing segmentation before classification or vice versa—affects performance and interpretability. A standardized evaluation protocol was defined and implemented in this thesis, using HAM10000 as a reference training dataset, and an independent external test set was used for external validation to ensure reproducibility and fair comparison. Experiments revealed that performing segmentation first yields superior generalization and more structured latent representations. Explainability analyses, utilizing gradient-based visualization and t-distributed stochastic neighbor embedding, further confirmed that segmentation-first learning enhances spatial coherence and interpretability.

Beyond algorithmic contributions, this research also emphasizes the importance of data realism and clinical collaboration. In collaboration with dermatologists from the "Ospedale di Circolo e Fondazione Macchi" in Varese, a new dataset named SKINPAN was constructed, comprising over 10,000 high-resolution panoramic dermatological images annotated by expert dermatologists. Designed to reflect real clinical conditions and lesion distribution, SKINPAN bridges the gap between curated dermoscopic datasets and real-world clinical imagery.

Overall, this doctoral work demonstrates that progress in AI-assisted dermatology depends not solely on model complexity but on the synergy between data quality, task sequencing, and rigorous evaluation. The proposed framework establishes a reproducible foundation for integrating structural and semantic information, fostering both robustness and interpretability. Through methodological

innovation and interdisciplinary collaboration, this research contributes to the advancement of trustworthy, context-aware AI systems for clinical dermatology.

Index

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Definition
ABCD	Asymmetry, Border, Color, Diameter (clinical rule for nevus assessment)
AI	Artificial Intelligence
AKIEC	Actinic Keratosis / Intraepithelial Carcinoma
ALBEF	Align-Before-Fuse (multimodal architecture)
AMFAM	Adversarial Multimodal Fusion with Attention
AUC	Area Under the ROC Curve
BAT	Boundary-Aware Transformer
BCC	Basal Cell Carcinoma
BCN	BCN20000 (dermoscopic dataset)
Bi-ConvLSTM	Bidirectional Convolutional Long Short-Term Memory
BKL	Benign Keratosis-like Lesions
CA-Net	Contextual Attention Network (segmentation)
CCTM	Convolutional Compact Transformer Model
CNN	Convolutional Neural Network
COCO	Common Objects in Context (benchmark and evaluation suite)
CS	Classification→Segmentation
SC	Segmentation→Classification
DF	Dermatofibroma
Dice	Dice Similarity Coefficient
DL	Deep Learning
DSCATNet	Dual-Scale Cross-Attention Transformer Network
F1	F1-score (macro or weighted, as specified)
FC	Full Count (total number of images in the dataset)
FAT-Net	Feature-Adaptive Transformer Network
GAN	Generative Adversarial Network
Grad-CAM	Gradient-weighted Class Activation Mapping
GPU	Graphics Processing Unit
GS-TransUNet	Gaussian Splatting Transformer U-Net
HAM	HAM10000 (dermoscopic dataset)

Abbreviation	Definition
HAMt	HAM test set (external validation subset)
IoU (Jaccard)	Intersection over Union (Jaccard index)
ISIC	International Skin Imaging Collaboration (archive and challenge series)
KvasirC	Kvasir Classification dataset
KvasirS	Kvasir Segmentation dataset
LD	Large Dataset (merged multi-source dataset proposed in this thesis)
mAP	Mean Average Precision
MEL	Melanoma
mIoU	Mean Intersection over Union
MSK	Memorial Sloan Kettering (dataset series)
MTL	Multi-Task Learning
NV	Melanocytic Nevi
PH2	PH2 (dermoscopic dataset with segmentation masks)
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SAM	Segment Anything Model
SEER	Surveillance, Epidemiology, and End Results Program
SKINPAN	Panoramic Dermatology Image Dataset introduced in this thesis
SOTA	State of the Art
SST	Sequential Swin Transformer (proposed in this thesis)
SVM	Support Vector Machine
Swin	Swin Transformer
SwinV2	Swin Transformer V2
t-SNE	t-distributed Stochastic Neighbor Embedding
TA	Test Accuracy
TP	Test Precision
TR	Test Recall
TF1	Test F1-score
TAb	Test Accuracy binary
TAm	Test Accuracy multiclass
TPm	Precision
TRm	Recall
TF1m	F1-score (multiclass)
TESL-Net	Transformer-Enhanced Segmentation Learning Network
TransUNet	Transformer-based U-Net architecture
U-Net	Encoder–decoder CNN for biomedical segmentation
UDA	Unsupervised Domain Adaptation (also dataset name in ISIC/MSK)
VASC	Vascular Lesions
ViT	Vision Transformer
WHO	World Health Organization

Chapter 1

Introduction

Skin cancer is one of the most common malignancies worldwide, with its global incidence rising steadily over the past few decades. Among its various types, **melanoma** stands out due to its particularly aggressive nature and high metastatic potential. According to the Global Cancer Observatory (GLOBOCAN), there were over 320,000 new melanoma cases and more than 57,000 deaths globally in 2020 [3]. More recently, data from the World Health Organization (WHO) report that more than 1.5 million new cases of skin cancer were diagnosed globally in 2022, with melanoma alone responsible for over 60,000 deaths annually [4]. Although melanoma accounts for a small fraction of all skin cancer cases, it is responsible for the majority of skin cancer-related deaths, underscoring the critical need for early detection and intervention.

The prognosis of melanoma is highly stage-dependent. Data from the American Cancer Society and SEER show that when diagnosed at an early stage (I, II), the five-year survival rate exceeds **95–99%**, whereas in advanced or metastatic stages (III, IV), survival drops dramatically below **30–35%** [5, 6]. This gap highlights the decisive impact of timely detection. Prognosis and therapeutic decisions are guided by the AJCC 8^{th} edition staging system, which incorporates parameters such as Breslow thickness, ulceration, and sentinel lymph node status [7].

From a prevention perspective, excessive exposure to ultraviolet (UV) radiation remains the main modifiable risk factor. Both natural sunlight and indoor tanning devices are classified as *Group 1 carcinogens* by the WHO and IARC [8]. Additional risk factors include fair skin, multiple nevi, immunosuppression, and genetic predispositions such as mutations in *BRAF*, *NRAS*, and *CDKN2A* [9, 10].

In terms of diagnosis, dermatologists often rely on *dermoscopic imaging* to improve accuracy. A dermatoscope is a non-invasive tool that magnifies skin lesions and reveals subsurface structures not visible to the naked eye. Meta-analyses confirm that dermoscopy increases diagnostic accuracy by 5–30% compared to naked-eye examination alone [11, 12]. Despite its widespread use, diagnostic performance remains dependent on clinician expertise and the type of lesion.

Over the past decade, the therapeutic landscape has undergone profound changes. For advanced melanoma, *immunotherapy* and *targeted therapy* have significantly improved survival rates [13, 14]. Five-year survival in stage IV melanoma has more than doubled compared to the pre-immunotherapy era [15]. Nevertheless, these treatments remain costly and are not universally effective, reinforcing the need for early detection and reliable diagnostic tools.

These clinical considerations naturally highlight the increasing importance of technological sup-

port systems in dermatology. Traditional diagnostic workflows, even when aided by dermoscopy, remain subject to inter-observer variability and are highly dependent on physician expertise and training [16]. This has created a strong demand for automated methods that can provide consistent, reproducible, and scalable diagnostic support.

In recent years, the field of **artificial intelligence (AI)** has emerged as a key enabler of such solutions. By learning discriminative representations directly from medical images, AI systems can assist clinicians in screening, triage, and risk stratification tasks [17]. Dermatology has been one of the most active areas of this research, with automated skin lesion analysis positioned as a promising tool for early melanoma detection [18, 19]. Several studies have demonstrated that AI models can match or even exceed the diagnostic accuracy of expert dermatologists, and, when combined with human expertise, further improve clinical decision-making [20, 21]. However, models still show limited generalization across external datasets, suffer from under-representation of darker skin tones and rare conditions [22], and may exploit spurious correlations from acquisition artifacts (e.g., ruler markings, color calibration charts, ink annotations, or background patterns) [23].

From a computer science perspective, automated skin lesion analysis is a prototypical problem in **computer vision**, where algorithms are designed to extract, represent, and classify visual patterns. Early approaches relied on handcrafted features inspired by clinical heuristics such as the ABCD rule (Asymmetry, Border, Color, Diameter), or on generic descriptors like color histograms, edge detectors, and texture measures [24]. While these techniques provided initial insights, they were limited in capturing the high intra-class variability and subtle morphological differences of skin lesions, and their performance was often sensitive to acquisition conditions.

The introduction of **machine learning** brought more adaptable classifiers, such as support vector machines and random forests, which improved performance over purely rule-based methods. However, they remained critically dependent on the quality of manually engineered features. The paradigm shifted dramatically with the emergence of **Deep Learning (DL)**, particularly convolutional neural networks (CNNs), which enabled end-to-end learning of hierarchical feature representations directly from raw image data [24, 25]. CNNs proved highly effective in capturing both low-level visual attributes have proven highly effective in capturing both low-level visual attributes and high-level semantic patterns, thereby scaling natural image benchmarks, such as ImageNet [25], quickly translated to the medical domain, where CNNs achieved state-of-the-art performance in classification, segmentation, and detection tasks [17, 18].

Nonetheless, CNNs are inherently constrained by the locality of their receptive fields, limiting their ability to model long-range dependencies and global context. In dermatology, this is particularly relevant, as lesions must often be interpreted not only through fine-grained structures but also in relation to broader contextual patterns. To overcome these limitations, **Transformer-based architectures** have recently been introduced into computer vision [26]. By leveraging self-attention mechanisms, Transformers can capture both fine-grained details and global contextual cues, offering a powerful alternative to CNNs. These models have rapidly established themselves as a new state of the art across medical imaging domains, including dermatology.

Within this context, this doctoral work is positioned at the intersection of medical imaging and computer vision. It investigates how deep learning methodologies—and, in particular, Transformer-based architectures—can be adapted and extended to tackle the challenges of skin lesion analysis, with the ultimate goal of enhancing diagnostic support in real-world clinical scenarios.

1.0.1 Public Awareness and Motivation Toward AI-Assisted Skin Cancer Prevention

To contextualize the societal relevance of this research, an exploratory survey was conducted among 137 participants to assess their awareness of skin cancer prevention, familiarity with artificial intelligence (AI), and openness toward the use of AI-assisted diagnostic tools in dermatology. Participants were recruited through online dissemination, and inclusion was voluntary; no clinical expertise was required. The sample was intended to provide a broad, non-clinical perspective on public attitudes. The anonymous questionnaire, distributed through online forms, included both quantitative (Likert scale and multiple-choice) and qualitative questions. The objective was not to achieve statistical inference, but rather to capture general attitudes and identify motivational trends supporting the integration of AI into preventive dermatology.

Awareness and prevention. Participants showed a strong interest in skin cancer prevention and early detection: over 75% of respondents rated their interest as 4 or 5 on a 5-point scale (Figure 1.1). Similarly, 71% considered the problem highly important, suggesting that the topic resonates with public health concerns. Notably, 50.4% of respondents reported having already undergone a professional skin screening, and about one third declared a family history of skin lesions or melanoma, indicating personal exposure to the issue.

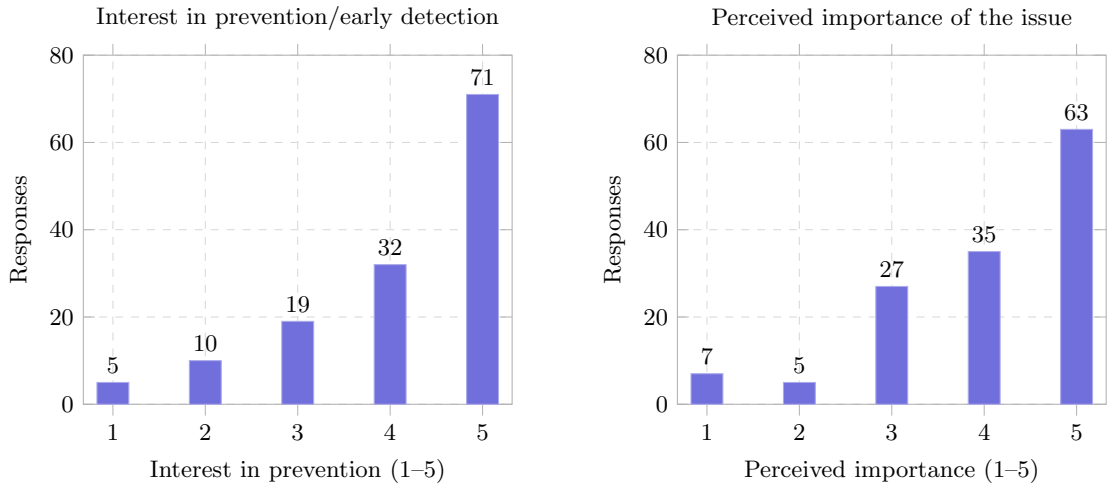


Figure 1.1: Public awareness: interest in prevention and perceived importance of early detection of skin lesions.

Knowledge and perception of AI. When asked about their familiarity with artificial intelligence, 65% of participants reported only a moderate understanding, yet 70% had already used AI-based tools (e.g., ChatGPT or automated assistants) for personal purposes (Figure 1.2). Despite limited technical knowledge, a generally positive perception emerged: 65% of respondents expressed trust in AI technologies (scores ≥3 on a 5-point scale), and over 80% stated that they would like to better understand how AI can support health-related tasks.

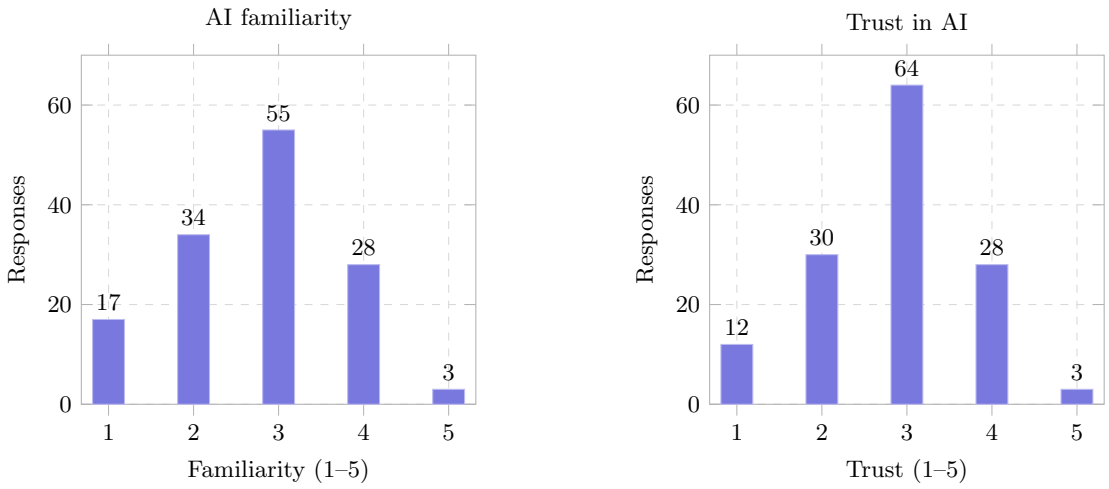


Figure 1.2: Public perception of AI: self-reported familiarity, and trust.

Acceptance of AI in dermatology. When asked about their willingness to rely on AI-assisted systems for lesion screening, 64% of participants indicated they would use such a tool if available, and more than 75% expressed interest in trying an AI model capable of recognizing potentially malignant lesions with 90% accuracy (Figure 1.3). While only a minority (5%) believed that AI could replace a physician, most respondents viewed AI as a valuable complementary resource for early screening and triage prior to dermatological examination.

Overall, the survey results confirm a high level of public interest in prevention and a willingness to adopt AI-based tools when transparency and clinical validation are ensured. The findings support the motivation of this thesis: developing interpretable, clinically relevant AI models for skin lesion analysis is not only a scientific objective but also a socially desirable goal. These insights underscore the importance of reliable systems that support dermatologists while empowering individuals to engage in proactive health monitoring.

Research Gaps and Questions

Despite remarkable advances in automated skin lesion analysis, several critical challenges remain unresolved. First, the *generalizability* of deep learning models remains limited: algorithms trained on one dataset often fail to maintain their performance when tested on data acquired under different imaging conditions, clinical contexts, or populations. Second, current benchmarks predominantly rely on isolated, lesion-centered images that overlook the broader spatial and clinical context considered by dermatologists during diagnosis. This lack of contextual representation restricts the ability of models to reason about lesion distribution, asymmetry, and evolution—factors that are essential in real-world decision-making. Third, while segmentation is frequently used as a preprocessing step to support classification, the *interaction and ordering between these tasks* remain poorly understood, particularly within Transformer-based architectures. Finally, although the integration of heterogeneous data modalities (e.g., images and clinical metadata) holds promise, the develop-

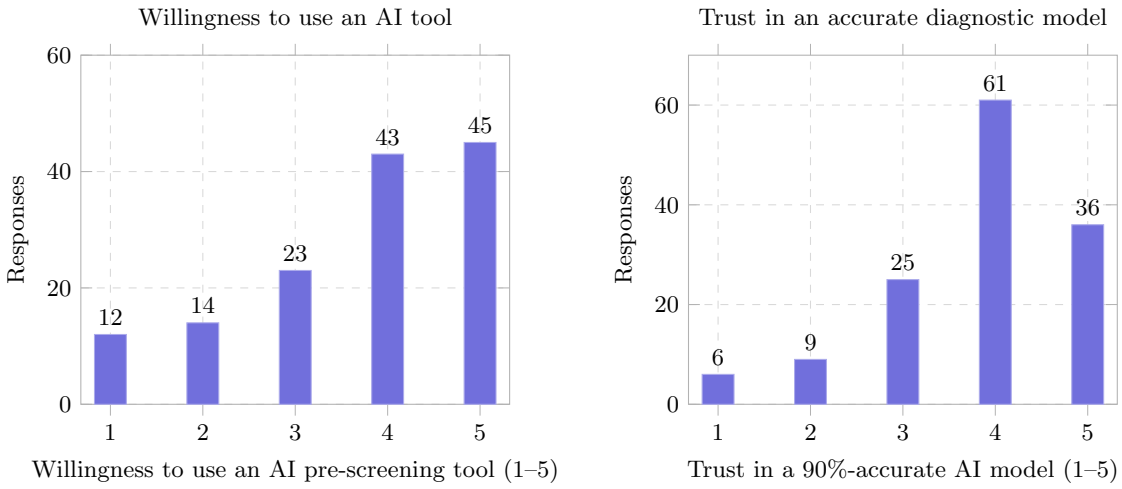


Figure 1.3: Acceptance of AI in dermatology: willingness to use an AI-based pre-screening system and trust in a highly accurate diagnostic model.

ment of robust multimodal frameworks remains an open research direction rather than a mature, consolidated field.

These gaps motivate the following guiding research questions:

- How can Transformer-based architectures be optimized to improve classification performance and generalization across heterogeneous dermatological datasets?
- What is the specific contribution of segmentation to classification, and how does the *order of learning* between these tasks affect model performance and interpretability?
- How can new dataset designs—incorporating contextual and clinically meaningful annotations—contribute to the development of more reliable and transferable AI systems?

Main Contributions of the Thesis

The central contributions of this doctoral work can be summarized as follows:

- Development of baseline Transformer-based models for skin lesion classification across multiple benchmark datasets.
- Critical assessment of the role of segmentation in classification and the proposal of a **Sequential Transfer Learning** strategy to integrate these tasks.
- Construction of **SKINPAN**, a high-resolution panoramic dermatology dataset collected in collaboration with dermatologists from *Ospedale di Circolo e Fondazione Macchi, University of Insubria (Varese)*, reflecting real clinical examination scenarios and supporting the development of context-aware models.
- Critical analysis of limitations in current datasets and evaluation protocols, providing guidelines for future research on multimodal and clinically grounded AI systems.

The overarching hypothesis of this doctoral project is that **Transformer-based architectures**, when properly structured and trained through sequential task learning, can enhance both the accuracy and robustness of medical image classification. The research focuses on single-modality image analysis and investigates how the sequential ordering of segmentation and classification tasks affects model performance and generalization. The study further explores the capability of such models to generalize across different datasets and imaging conditions, emphasizing the impact of architectural design and training strategy on interpretability.

This study helps bridge the gap between AI research and its practical application in dermatology. By examining the interaction between segmentation and classification, assessing model robustness across datasets collected in different clinical contexts, and introducing a dataset that more closely reflects real examination conditions, this thesis contributes to the development of more reliable, transparent, and generalizable computer-aided diagnostic systems.

This thesis is structured as follows. Chapter 2 presents a comprehensive review of the current state of the art, including available datasets, classification and segmentation approaches, and recent advances in deep learning for dermatology. Particular emphasis is placed on identifying the limitations and open challenges in the literature, thereby motivating the contributions of this research. Chapters 3 through 6 outline the core contributions of this thesis, each grounded in peer-reviewed or under review articles developed during the doctoral program. These chapters cover a range of topics, from baseline classification models to task sequencing strategies. Chapter 7 synthesizes the results and offers a critical discussion, comparing proposed methods with existing approaches and analyzing their strengths and weaknesses. Finally, Chapter 8 concludes the thesis with a summary of the main findings and outlines several promising directions for future research.

Chapter 2

Background and State of the art

In this chapter, we review the state of the art in automated skin lesion analysis with a unified perspective on both classification and segmentation. We begin by surveying publicly available datasets (Section 2.1) with explicit attention to their suitability for each task—i.e., whether they provide image-level diagnostic labels, pixel-level masks, or both. As will be shown, only a limited subset offers paired annotations, a constraint that has shaped methodological choices across the literature.

Building on this dataset overview, we then examine CNN-based (Section 2.2) and Transformer-based methods (Section 2.3) for skin lesion classification, followed by dedicated segmentation approaches (Section 2.4). We subsequently discuss multi-task frameworks that couple segmentation and classification within a single model and their trade-offs (Section 2.5.1), as well as multimodal pipelines that integrate images with metadata (Section 2.6). Finally, we synthesize cross-cutting limitations and open challenges that emerge across these lines of work (Section 2.7).

2.1 Publicly Available Datasets for Skin Lesion Analysis

A wide range of publicly available datasets has been proposed in the literature to support the development of automated skin lesion analysis methods. These datasets differ substantially in terms of the number of diagnostic categories, the size of the image collections, the acquisition modalities (dermatoscopic or macroscopic), and the presence of metadata or histopathological annotations. In particular, the choice of acquisition modality has a direct impact on the visual information available to learning algorithms and on the type of diagnostic cues that can be exploited. Dermatoscopic and macroscopic images differ substantially in both acquisition process and visual content. Dermatoscopic images are captured using a dermatoscope, which provides magnification and controlled illumination—often polarized—allowing visualization of subsurface skin structures such as pigment networks, globules, streaks, and vascular patterns. These images emphasize fine-grained morphological details that are critical for dermatological diagnosis but are not visible to the naked eye.

In contrast, macroscopic (or clinical) photographs are acquired using standard cameras or smartphones without magnification or optical enhancement. They capture the lesion in its broader anatomical context, including surrounding skin, body location, and scale, but lack the microstructural detail of dermoscopy. Macroscopic images are therefore more representative of real-world

clinical settings, especially in primary care or teledermatology scenarios, yet they pose additional challenges due to variations in lighting, focus, background clutter, and camera quality. These modality-dependent differences contribute to the heterogeneity observed across public datasets and partly explain the variability in model performance when transferring across acquisition settings. As a consequence of this heterogeneity, some datasets include only a small number of well-defined classes, while others feature a broader and more heterogeneous taxonomy encompassing various benign, malignant, and ambiguous skin conditions.

Among the numerous diagnostic categories encountered across these datasets, seven classes recur most frequently and are commonly adopted as benchmarks for classification tasks:

- **Melanoma (MEL)**: A malignant tumor derived from melanocytes, responsible for skin pigmentation and potentially lethal if not diagnosed early.
- **Melanocytic Nevi (NV)**: Benign proliferations of melanocytes, typically presenting as pigmented moles or lesions.
- **Basal Cell Carcinoma (BCC)**: The most prevalent form of skin cancer, originating in the basal layer of the epidermis.
- **Actinic Keratoses (AKIEC)**: Premalignant skin lesions caused by chronic exposure to ultraviolet radiation, often appearing as scaly or rough patches.
- **Benign Keratosis-like Lesions (BKL)**: A group of non-cancerous conditions, including seborrheic keratosis, clear cell acanthoma, and sebaceous hyperplasia.
- **Dermatofibroma (DF)**: A benign fibrous skin nodule, usually resulting from minor trauma or insect bites.
- **Vascular Lesions (VASC)**: Lesions related to blood or lymphatic vessel abnormalities, such as hemangiomas or angiokeratomas.

The distribution of these classes varies widely across datasets, both in terms of absolute counts and balance between categories. Some datasets are relatively balanced, while others are heavily skewed toward more prevalent conditions such as melanocytic nevi.

Moreover, only a limited number of datasets provide pixel-level segmentation masks, and when available, these annotations are typically binary, distinguishing only between the lesion area and the surrounding background. Such simplified masks are sufficient for delineating lesion boundaries but do not capture intra-lesional structures or other important information.

2.1.1 The HAM10000 Dataset

The HAM10000 dataset (HAM) (Human Against Machine with 10,000 training images), introduced by Tschandl et al. [27], is one of the most influential and widely adopted public resources for skin lesion analysis. It was released with the goal of providing a large, diverse, and well-annotated collection of dermatoscopic images to foster the development of machine learning algorithms for automated diagnosis. The dataset comprises a total of 10,015 dermatoscopic images, collected from two primary sources: the Department of Dermatology at the Medical University of Vienna and a skin cancer clinic in Queensland, Australia.

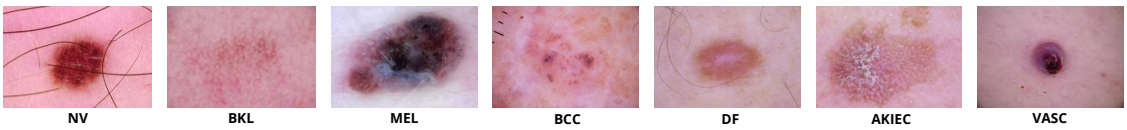


Figure 2.1: Representative skin lesion images from the HAM dataset, annotated with their respective diagnostic classes. In order, from left to right, we have: Nevus (NV), Benign Keratosis-like Lesions (BKL), Melanoma (MEL), Basal Cell Carcinoma (BCC), Dermatofibroma (DF), Actinic Keratosis / Intraepithelial Carcinoma (AKIEC), and Vascular Lesions (VASC).

Each image in the dataset belongs to one of seven diagnostic categories, which have become standard labels in most benchmark studies and challenges. Representative examples of each class are shown in Figure 2.1. As illustrated in Figure 2.2, the dataset exhibits a marked class imbalance: NV alone accounts for 6,705 images—over 66% of the total—while all other classes are significantly less represented. The number of images per class ranges from 1,113 in the MEL class to as few as 115 in the DF class and 142 in the VASC, with BCC, AKIEC, and BKL lesions also contributing modestly.

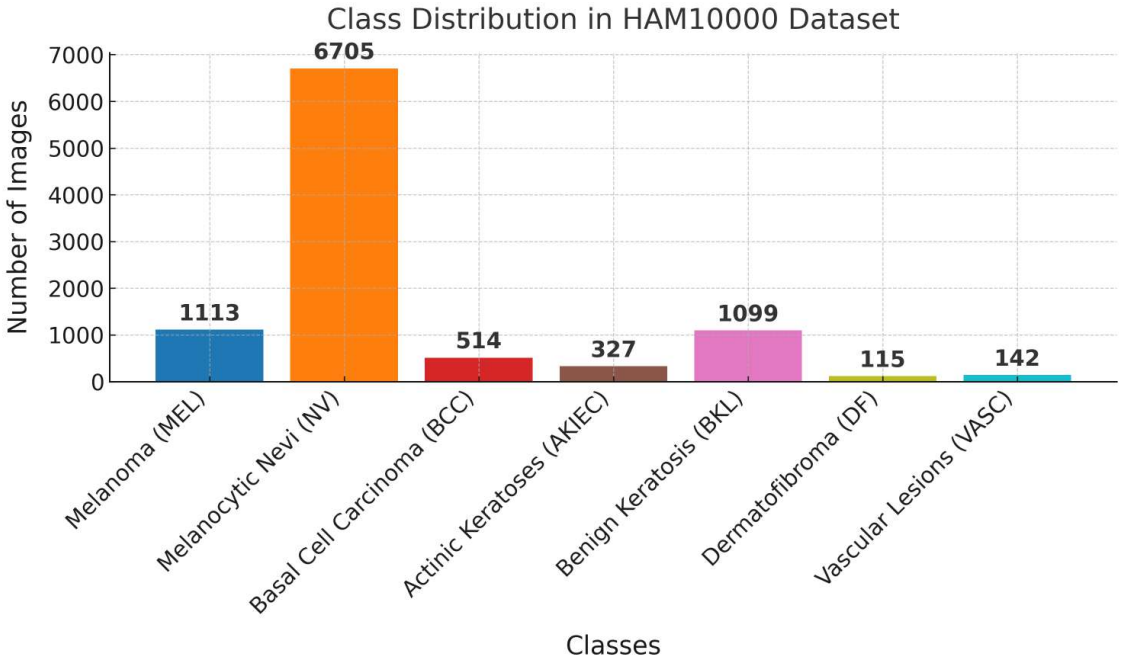


Figure 2.2: Class distribution in the HAM dataset.

All images are dermatoscopic and provided in JPEG format at a resolution of 600×450 pixels. The dataset includes both macroscopic and close-up views of lesions, manually selected and curated to reduce noise and artifacts. Diagnoses are verified through various mechanisms: approximately

53% of the cases are histopathologically confirmed, while others rely on follow-up imaging, expert consensus, or in vivo confocal microscopy.

HAM is accompanied by a structured metadata file in CSV format, which includes information such as patient age, sex, lesion anatomical site (e.g., torso, scalp, extremities), image type, and acquisition source.

A subset of HAM images has been annotated with pixel-level segmentation masks as part of the ISIC 2018 challenge [28]. These masks delineate the precise lesion boundaries and were manually traced by dermatologists. The availability of corresponding classification and segmentation annotations makes HAM particularly well-suited and used in literature for different research purposes, such as segmentation and classification tasks.

In addition to the original 10,015 training images, a complementary test set (HAMt) has been used in follow-up studies to enable cross-dataset validation. This test set comprises approximately 2,500 dermatoscopic images sourced from the same institutions and annotated using the same diagnostic taxonomy. While not officially part of the original release, it was released a few years after, and it has become an important standard for evaluating generalization and robustness beyond the training set.

HAM has served as the reference dataset in a wide array of publications employing classical CNNs, Transformer-based architectures, segmentation pipelines, and multimodal systems. It was also used in the official ISIC 2018 challenge, where participants tackled classification and segmentation as separate tasks. Its balanced structure of data and metadata, together with wide community adoption, make it a cornerstone in the field of AI for dermatology.

Despite its strengths, HAM is not without limitations. Beyond class imbalance, some images are duplicated across other datasets (more details in 2.1.3), which can artificially inflate performance if not carefully controlled. Additionally, the dataset is limited to dermatoscopic images—no clinical or macroscopic photographs are included—reducing its applicability in broader diagnostic pipelines where multiple imaging modalities are used.

2.1.2 Other Dermatology Datasets

In addition to HAM, several other publicly available datasets have been developed to support the training and evaluation of skin lesion classification and segmentation models. These datasets vary significantly in size, class taxonomy, image modality (dermatoscopic or macroscopic), metadata availability, and the presence of segmentation masks. Table 2.1 summarizes the presence of different modalities and metadata, along with the total number of available images.

Some datasets were created in clinical or research environments with detailed metadata and histopathological labels (e.g., MSK series [29], Hospital Italiano de Buenos Aires [30]), while others were designed as public benchmarks for challenges (e.g., ISIC 2018 [28], ISIC 2020 [31]).

Among the most prominent datasets:

- **BCN20000** [32] (BCN) contains 12,413 dermatoscopic images across 8 classes, with a high overlap in taxonomy with HAM. It includes both metadata and histological confirmation for part of the dataset. A known issue is that some images overlap with HAM and ISIC archives.
- **PH2** [36] is a small, high-quality dataset with 200 dermatoscopic images annotated into 3 classes. It includes manual segmentations and detailed metadata (e.g., lesion diameter, symmetry, color), making it suitable for validation rather than large-scale training.

Table 2.1: Summary of key characteristics of the main dermatology datasets: presence of dermatoscopic (D) or macroscopic (M) images, availability of metadata (Meta), presence of segmentation masks (Segm), and the total number of images (FC).

Dataset	**D**	**M**	**Meta**	**Segm**	**FC**
HAM10000 [27]	✓	✓	✓	✓	10015
BCN20000 [32]	✓				12413
ISIC 2020 [31]	✓	✓	✓		33126
MSK (1-5) [29]	✓	✓			9185
Hospital Italiano de Buenos Aires [30]		✓	✓		1616
7-point criteria [33]	✓				2013
Consecutive biopsies [34]	✓		✓		1295
UDA (1-2) [29]	✓		✓		617
SKINL2 [35]	✓				437
PH2 [36]	✓		✓	✓	200
DermNetNZ [37]		✓			>20,000
Fitzpatrick17k [38]		✓	✓		16577
ISIC2016 [39]	✓			✓	900
ISIC2017 [28]	✓			✓	2000

- The **MSK datasets** (MSK1 to MSK5) [29] include thousands of images collected at Memorial Sloan Kettering Cancer Center. These datasets offer a broader diagnostic taxonomy (up to 25 classes) and include both dermatoscopic and macroscopic images. However, metadata availability is variable, and some classes are sparsely populated.
- **Hospital Italiano de Buenos Aires dataset** [30] offers 1,616 images with 10 diagnostic classes and associated clinical metadata. It includes both dermatoscopic and macroscopic images and aims to provide a more realistic, non-curated distribution of cases for algorithm testing.
- **ISIC challenge datasets** [40] are widely used benchmarks for classification and segmentation. The ISIC 2020 [31] training set includes 33,126 images annotated into 8 classes. However, the dataset is partially noisy, with label reliability ranging from histology to user-submitted tags.
- The **7-point criteria evaluation dataset** [33] (7-pt) contains 2,013 dermatoscopic images annotated with structured diagnostic criteria. Although not always used for classification, it is relevant for explainability and rule-based approaches.
- **UDA datasets (UDA-1 and UDA-2)** [29] comprise 557 and 60 dermatoscopic images, respectively, collected with histopathological or clinical confirmation. They include structured metadata such as patient age, sex, anatomical site, and diagnostic label. However, they do not provide segmentation masks, and some rare classes are only sparsely represented.
- **SKINL2** [35] is a multi-class dataset containing 437 dermatoscopic images distributed across 51 diagnostic categories. Although relatively small, it features a highly granular taxonomy and can support fine-grained classification tasks. No metadata or segmentation annotations are provided, limiting its applicability.

- **DermNetNZ** [37], an online repository containing thousands of dermatological images collected in New Zealand across a wide spectrum of diseases and skin types. Although not curated for a specific machine learning task, it represents one of the largest openly accessible dermatology collections and is often used for pretraining or as a supplementary dataset.
- **Fitzpatrick17k dataset** [38] provides 16,577 clinical images annotated with skin tone categories based on the Fitzpatrick scale. This dataset is particularly relevant for studying algorithmic fairness, as it enables the evaluation of model performance across different skin phototypes and highlights disparities in AI-driven dermatology.
- **ISIC2016 challenge dataset** [39] represents the first large-scale public benchmark introduced by the ISIC for the task of automated melanoma detection. It contains 900 dermatoscopic training images and 379 test images, each accompanied by expert-provided segmentation masks. The dataset was released as part of the ISBI 2016 challenge and defined three tasks: segmentation, dermoscopic feature detection (i.e., the localization of predefined dermoscopic structures used in clinical assessment), and binary classification (benign vs. malignant).
- **ISIC2017 challenge dataset** [28] expanded the previous edition (ISIC2016) by increasing the dataset size and complexity. It includes 2,000 dermatoscopic training images, 150 validation images, and 600 test images, each annotated for both segmentation and classification. The classification task comprises three diagnostic categories—melanoma, seborrheic keratosis, and benign nevi—making it a widely adopted benchmark for evaluating both segmentation and classification models.

In terms of modality, most datasets are limited to dermatoscopic images. Notable exceptions include the MSK datasets and the Hospital Italiano de Buenos Aires, which also provide macroscopic images. Regarding metadata, BCN, MSK datasets, Hospital Italiano, and PH2 include some form of structured clinical information (e.g., patient age, lesion location), whereas others, such as ISIC 2018 test set and UDA variants, are metadata-poor. Lastly, only HAM and PH2 are also providing the segmentation masks.

2.1.3 Data Issues: Class Imbalance, Duplicates, Annotation Noise

Despite the growing number of publicly available datasets for skin lesion analysis, several intrinsic limitations hinder their effective use for training robust and generalizable deep learning models. These issues span from severe class imbalance and duplicate samples to inconsistencies in label annotations and variations in image modality. A detailed understanding of these challenges is crucial when designing models intended for real-world deployment or cross-dataset evaluation.

One of the most pervasive issues across dermatology datasets is the imbalance between diagnostic classes. As shown in Figure 2.3, datasets such as HAM [27], BCN [32], and ISIC 2020 [31] are heavily skewed toward common benign lesions, especially NV, which can account for over 60% of the total samples. Conversely, malignant or rare conditions such as DF, VASC, or AKIEC are severely underrepresented, sometimes comprising less than 1% of the dataset.

This imbalance introduces significant biases during model training, especially when using standard loss functions that do not explicitly correct for label distribution. Models tend to overfit the dominant classes and underperform on rare but clinically critical ones like MEL. Table 2.2 provides a numerical overview of the distribution of the seven benchmark classes across major datasets.

Table 2.2: Overview of major dermatology datasets with reported data issues. For each dataset, we include image size, number of diagnostic classes, missing lesions (ML), number of macroscopic images (MI), number of duplicate images (DI), and total image count (Full Cardinality, FC).

Dataset	Size of images	N. Classes	ML	MI	DI	FC
HAM [27]	600 x 450	7.0				10 015.0
Consecutive biopsies for melanoma (2020) [34]	3264 x 2448	21.0	22.0			1295.0
MSK1 [29]	variable	17.0	248.0			1678.0
MSK2 [29]	variable	22.0	50.0			4880.0
MSK3 [29]	variable	25.0				466.0
MSK4 [29]	variable	24.0	10.0			2050.0
MSK5 [29]	variable	17.0				111.0
Hospital Italiano de Buenos Aires Dataset [30]	variable	10.0				1616.0
SKINL2 [35]	1920 x 1080	51.0				437.0
BCN [32]	1024 x 1024	8.0				12 413.0
UDA-1 [29]	variable	7.0				557.0
UDA-2 [29]	variable	7.0				60.0
7-pt [33]	768 x 512	20.0	48.0	1002.0		2013.0
ISIC 2020 Challenge Training Set [31]	variable	8.0	26 706.0		425.0	33 126.0
ISIC Challenge 2018: Task 1-2 Test [28]	variable	3.0			461.0	1000.0
ISIC Challenge 2018: Task 1-2 Validation[28]	variable	3.0			53.0	100.0
PH2 [36]	765 x 575	3.0				200.0
DermNetNZ [37]	variable	> 50.0				$> 20\,000.0$
Fitzpatrick17k [38]	variable	114.0		16 577.0		16 577.0

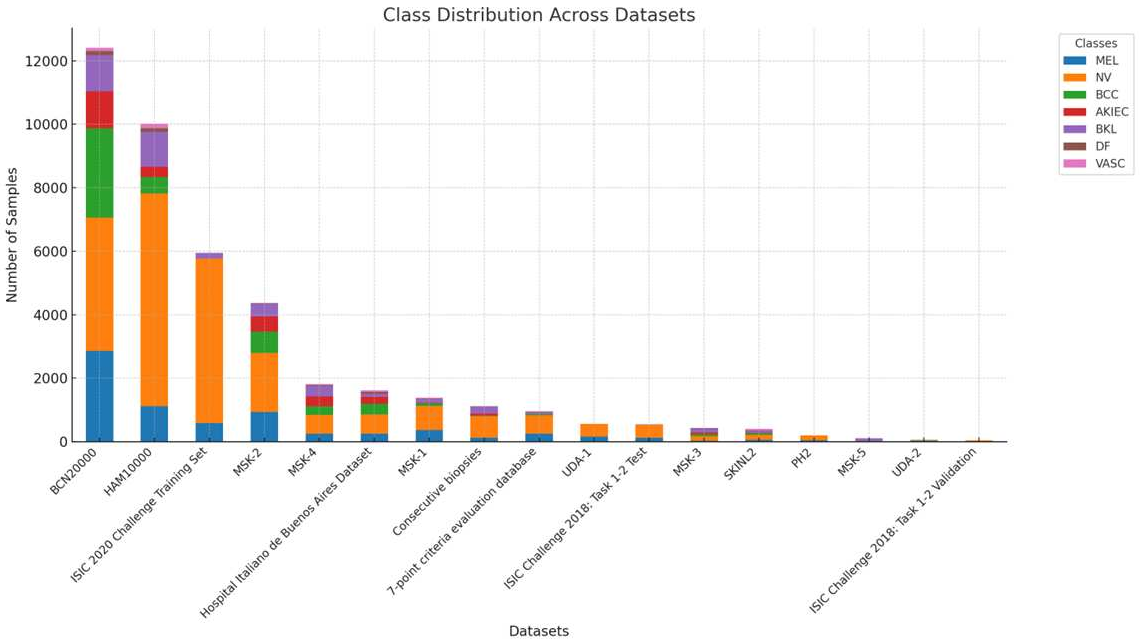


Figure 2.3: Distribution of the seven benchmark diagnostic classes across selected dermatology datasets.

Another significant concern is the presence of duplicate (multiple images of the same lesion acquired under different illumination conditions, orientations, or zoom levels) or overlapping images across datasets. Several studies [32, 27] have reported that many ISIC-related collections (e.g., HAM, BCN, ISIC challenges) partially overlap due to data sharing and re-uploading across platforms. Moreover, some datasets include multiple views of the same lesion (e.g., close-up and wider dermatoscopic captures), which can lead to data leakage when these images are split across training and test sets. These duplicates, if not properly filtered, may cause models to memorize lesion appearances rather than generalize to unseen cases. Furthermore, the lack of standardized patient identifiers across datasets prevents automated de-duplication.

The reliability of ground truth labels varies substantially among datasets. While some collections provide labels based on histopathological reports (e.g., PH2 [36], parts of MSK [29]), others include user-submitted or expert-reviewed tags without pathological confirmation. This results in annotation noise, label ambiguity, or the complete absence of a diagnosis for certain samples.

In datasets such as ISIC 2020, thousands of images are labeled without precise documentation of how the diagnosis was established. In the case of multi-center datasets, inter-observer variability can further exacerbate labeling inconsistencies. These factors introduce uncertainty during model training and may explain inconsistent performance when models are evaluated on new datasets.

The modality and resolution of images also vary widely across datasets. While most datasets focus on dermatoscopic images (e.g., HAM, PH2, 7-point criteria database [33]), others incorporate macroscopic photographs taken with standard cameras (e.g., Hospital Italiano [30], MSK series [29]). These differences in acquisition devices, lighting, and zoom levels introduce domain shifts that affect

model transferability.

Furthermore, as shown in Table 2.2, the spatial resolution of images ranges from 600×450 pixels to over 3000×2000 pixels. Models trained on high-resolution dermatoscopic images may not generalize well to lower-quality or macroscopic photographs, and vice versa. This heterogeneity in modality and resolution must be accounted for during preprocessing, data augmentation, or domain adaptation procedures.

2.2 CNN-based Approaches for Skin Lesion Classification

CNNs have played a foundational role in the development of automated skin lesion classification systems. This section provides an in-depth overview of influential works leveraging CNNs for this task, with a focus on their methodological contributions, architectural innovations, and reported performance. The HAM dataset is often adopted as a benchmark due to its scale, diversity, and clinical relevance.

Esteva et al. [18] pioneered the application of deep CNNs to skin cancer diagnosis. Their model was based on the Inception-v3 architecture, pre-trained on ImageNet and fine-tuned using a private dataset comprising over 129,000 clinical images of skin lesions. The network was evaluated on tasks distinguishing malignant from benign lesions and achieved performance comparable to that of board-certified dermatologists. This study was the first to demonstrate dermatologist-level diagnostic accuracy using deep learning and set a precedent for subsequent research in AI-assisted dermatology.

DeVries and Ramachandram [41] extended the Inception-v3 architecture into a multi-scale framework to enhance classification robustness. By processing both low- and high-resolution versions of each image, their network was able to learn complementary features at different spatial scales. Tested on the ISIC 2017 dataset, the model ranked among the top performers in the official challenge, highlighting the benefit of multi-resolution input for dermoscopic image classification.

Mahbod et al. [42] proposed a hybrid deep neural network in which features extracted from multiple CNN architectures (AlexNet, VGG16, and ResNet-18) were fused and classified using support vector machines. Their approach, evaluated on the ISIC 2017 dataset, demonstrated that combining heterogeneous CNN feature representations significantly improved robustness and classification performance compared to single-model baselines.

Carcagnì et al. [43] investigated the use of DenseNet-61 with a center loss function for the seven-class classification task of ISIC 2018. Their work emphasized the importance of compact feature embeddings to better separate visually similar lesion categories and reduce intra-class variability.

Rezvantalab et al. [44] performed a comparative evaluation of several deep CNN architectures, including ResNet-152, DenseNet-201, and Inception-ResNet-v2, on the HAM and PH2 datasets. Their study showed that deeper models like DenseNet-201 achieved superior classification performance, with an AUC of 94.4% for melanoma and 99.3% for basal cell carcinoma. The authors concluded that model depth and feature reuse were crucial in capturing subtle visual differences in skin lesion classes.

Mohamed and El-Behaidy [45] proposed an enhanced CNN for skin lesion classification, focusing on improved feature extraction and model depth to increase diagnostic accuracy.

Huang et al. [46] developed a lightweight CNN model tailored for cloud-based applications and remote skin cancer diagnosis, emphasizing computational efficiency while maintaining reliable classification performance.

Liu et al. [47] introduced a semi-supervised classification framework based on a relation-driven self-ensembling model, which reduces the dependency on large annotated datasets while improving medical image classification accuracy.

Gu et al. [48] proposed a progressive transfer learning strategy combined with adversarial domain adaptation for cross-domain skin disease classification, aiming to enhance generalization across heterogeneous datasets.

Khan et al. [49] presented a multi-model deep learning approach where features extracted from multiple CNN architectures are fused and optimally selected, leveraging complementary representations to improve skin lesion classification performance.

Calderón et al. [50] introduced BILSK, a bilinear CNN framework that applies bilinear pooling to model second-order interactions between features. This architecture is particularly suited for tasks involving fine-grained visual categorization, such as skin lesion classification. Their results on HAM demonstrated competitive accuracy and confirmed the effectiveness of bilinear representations in enhancing the discrimination of visually similar lesion types.

Nguyen et al. [51] addressed the issue of class imbalance and noise in dermoscopic datasets by incorporating soft attention mechanisms into their CNN architecture. Their model used attention maps to focus on informative regions within each image and included a data reweighting strategy to handle underrepresented classes. Evaluated on HAM, the model achieved an overall classification accuracy above 92%, showing increased robustness to both intra-class variation and dataset imbalance.

Shetty et al. [52] proposed a hybrid approach combining CNN-based feature extraction with traditional machine learning classifiers, such as SVMs and decision trees. After extracting features using a deep CNN, these were fed into ML classifiers to perform the final prediction. On HAM, the approach achieved a classification accuracy of 95.18%, outperforming pure ML pipelines and showcasing the value of integrating deep and shallow learning paradigms.

Ajabani et al. [53] developed a CNN model designed to mimic clinical evaluation and compared its diagnostic performance directly to that of human experts. The model was trained and evaluated on HAM, and results showed that it consistently outperformed the average human baseline across several lesion categories. The study emphasized the potential of AI to not only support but, in some scenarios, exceed human-level performance in dermatological tasks.

Rasel et al. [54] presented a novel classification pipeline combining CNN-based feature extraction with a support vector machine (SVM) classifier. Their method incorporated geometric lesion features and asymmetry descriptors, which are clinically relevant for melanoma diagnosis. The hybrid model achieved over 98% classification accuracy, with macro-F1 and weighted F1 scores of 95% and 97%, respectively, on combined ISIC and HAM datasets.

Nguyen et al. [55] proposed AC-MambaSeg, an architecture integrating adaptive convolutional layers and a Mamba-based attention mechanism within a segmentation framework. Although primarily designed for the segmentation of skin lesions, the model's high-quality masks were used to extract refined regions of interest that improved downstream classification performance.

He et al. [56] introduced ResNet, a deep residual network that enables training of substantially deeper CNNs through identity-based skip connections. ResNet has become a widely used backbone in medical image analysis, including dermatology, where its depth and feature abstraction capabilities support high-performance classification and segmentation tasks.

Ronneberger et al. [57] designed U-Net, an encoder-decoder CNN with symmetric skip connections that allow for high-resolution feature preservation during upsampling. Originally developed for biomedical image segmentation, U-Net has been extensively adapted for dermatological tasks

due to its spatial precision and architectural simplicity.

Younas et al. [58] proposed AIR-CNN, an attention-based Inception-Residual architecture that effectively handles intra-class variations with reduced parameters, achieving 91.63% accuracy on ISIC 2019.

Halder et al. [59] introduced a fuzzy rank-based deep ensemble methodology, leveraging transfer learning, soft attention, and feature fusion (e.g., DenseNet201, ResNet50, VGG19+HOG), reaching accuracy up to 98.85% (ResNet50) and 98.32% (DVFNet) on ISIC and HAM datasets.

Table 2.3: Summary of key CNN-based approaches for skin lesion classification.

Authors (Year)	**Architecture / Approach**	**Dataset(s)**	**Key Contributions and Metrics**
Ronneberger et al. (2015) [57]	U-Net (encoder-decoder)	Biomedical	Foundational segmentation architecture with skip connections
He et al. (2016) [56]	ResNet (deep residual CNN)	Generic medical	Backbone enabling deep CNNs; widely adopted in classification
Esteva et al. (2017) [18]	Inception-v3 fine-tuned	Private dataset	First CNN achieving dermatologist-level classification performance
DeVries et al. (2017) [41]	Multi-scale Inception-v3	ISIC 2017	Multi-resolution strategy; top performer in ISIC 2017 challenge
Mahbod et al. (2017) [42]	Hybrid CNN (AlexNet, VGG16, ResNet-18) + SVM	ISIC 2017	Feature fusion improved robustness; AUC ≈97% for SK class
Carcagnì et al. (2018) [43]	DenseNet-61 + Center Loss	ISIC 2018	Seven-class classification; compact feature embeddings
Rezvantalab et al. (2018) [44]	ResNet-152, DenseNet-201	HAM, PH2	AUC: melanoma 94.4%, BCC 99.3%; deep model comparison
Mohamed & El-Behaidy (2019) [45]	Enhanced CNN	ISIC	Improved depth and feature extraction for better accuracy
Gu et al. (2019) [48]	CNN + progressive TL + domain adaptation	Derm/Skin datasets	Cross-domain classification; adversarial learning improves generalization
Khan et al. (2019) [49]	Multi-model CNN + optimal feature selection	Derm/Skin datasets	Feature fusion from multiple CNNs; classification accuracy improvement
Liu et al. (2020) [47]	Semi-supervised CNN (self-ensembling)	ISIC 2018, others	Reduces reliance on labeled data; boosts classification accuracy
Huang et al. (2021) [46]	Lightweight CNN	Private dataset	Cloud-ready, efficient model for remote skin cancer diagnosis
Calderón et al. (2021) [50]	Bilinear CNN (BILSK)	HAM	Bilinear pooling enhances feature interactions; strong performance
Nguyen et al. (2022) [51]	CNN with soft attention	HAM	Attention maps + class reweighting; accuracy ¿92%
Shetty et al. (2022) [52]	CNN + ML hybrid	HAM	CNN features + SVM/DT classifiers; accuracy 95.18%
Rasel et al. (2024) [54]	CNN-SVM + geometric features	HAM, ISIC	Asymmetry descriptors; accuracy >98%, F1: 95%/97%
Nguyen et al. (2024) [55]	AC-MambaSeg (seg.+cls.)	ISIC, PH2	Adaptive conv. + attention; masks used for classification ROI
Ajabani et al. (2025) [53]	CNN vs human comparison	HAM	Outperformed average human baseline in multi-class task
Younas et al. (2025) [58]	Attention-based Inception-Residual CNN (AIR-CNN)	ISIC 2019	Lightweight attention model; accuracy 91.63%
Halder et al. (2025) [59]	Fuzzy rank-based CNN ensemble	ISIC, HAM	Ensemble with attention/feature fusion; accuracy up to 98.85%, 98.32% (DVFNet)

2.2.1 Limitations of CNN-Based Approaches

Although CNNs have played a foundational role in automated skin lesion classification, several inherent limitations restrict their clinical applicability. First, these models are strongly data-driven, and their performance depends heavily on the quality and diversity of training datasets. Publicly

available benchmarks such as ISIC and HAM, while valuable, often suffer from significant class imbalance and under-representation of rare lesion types. Moreover, demographic biases in the data can lead to fairness issues: Daneshjou et al. [60, 22] reported that CNN-based classifiers trained predominantly on lighter skin tones exhibited performance drops of up to 36% in ROC-AUC when evaluated on darker skin tones, underscoring disparities in dermatology AI.

Another critical limitation concerns the tendency of CNNs to overfit spurious correlations and dataset-specific artifacts. Bissoto et al. [23] demonstrated that networks may exploit non-diagnostic cues, such as surgical markers or image borders, to achieve inflated accuracy scores that do not reflect genuine clinical understanding. This phenomenon, often referred to as *shortcut learning*, undermines the robustness of CNNs when deployed in real-world settings where such artifacts are absent or vary substantially.

From an architectural perspective, CNNs are highly effective at local feature extraction but remain limited in their ability to capture long-range dependencies and global structural context. In dermoscopic analysis, clinically relevant patterns often span broader spatial regions than those captured by standard receptive fields, which may contribute to misclassification of visually similar lesions. This issue also relates to the broader problem of confidence calibration, with CNNs frequently producing overconfident yet incorrect predictions [61].

Interpretability further poses a major barrier to the adoption of CNN-based diagnostic systems. Although visualization techniques such as saliency maps and class activation mappings have been proposed, their reliability is contested. Adebayo et al. [62] showed that many popular saliency methods are insensitive to model parameters, raising concerns about their validity as trustworthy clinical explanations. In high-stakes scenarios such as melanoma detection, the inability to provide robust interpretability reduces clinical trust.

Finally, CNNs often exhibit poor generalization under domain shift, performing well on curated benchmarks but deteriorating when applied to external datasets or images acquired under different clinical conditions. Pacheco et al. [63] emphasized that variations in imaging devices, acquisition settings, and patient populations can significantly affect performance, creating a generalization gap that hinders safe deployment in diverse healthcare environments.

While CNNs have established the foundations of deep learning in dermatology, their reliance on balanced datasets, vulnerability to shortcut learning, limitations in modeling global context, interpretability issues, and sensitivity to domain shift collectively limit their clinical impact. These challenges motivate the transition toward more flexible paradigms, including transformer-based architectures, multimodal learning, and sequential transfer strategies, which are explored in subsequent chapters of this thesis.

2.3 Transformer-based Approaches for Skin Lesion Classification

The recent shift from CNN architectures to transformer-based models reflects the growing recognition of the power of self-attention mechanisms to capture both local and global context. Vision Transformers (ViTs) [26] enable models to attend to all parts of an image through learned attention weights, thus reducing the locality bias of CNNs. This transition has fostered a range of Transformer-based models for skin lesion classification, often outperforming or complementing CNNs in variants of the HAM and ISIC datasets.

Aladhadh et al. [64] proposed a Medical Vision Transformer (MVT) which splits dermoscopic

images into patches, embeds them, and processes them via transformer encoders. Trained on HAM, the MVT model achieved approximately 96.14% accuracy, 96.5% recall, and 97% F1-score, surpassing many existing CNN-based baselines.

Jain et al. [65] explored transfer learning using ViTs and CNNs (e.g. ResNet50), fine-tuned on HAM. While CNNs retained strong baseline performance, ViT models demonstrated improved generalization when pretrained on large-scale datasets and yielded more interpretable attention maps.

Lungu-Stan et al. (SkinDistilViT) [66] introduced a lightweight ViT model via knowledge distillation tailored for mobile deployment. Despite its reduced size, it retained 98.3% of the teacher's balanced multi-class accuracy and significantly lower inference time and memory footprint.

Flosdorf et al. [67] compared ViT-L16 and ViT-L32 models for melanoma vs benign classification, achieving accuracy between 91.6% and 92.8%, with melanoma recall around 56%, demonstrating robustness across configurations.

Krishna et al. (LesionAid) [68] developed a hybrid ViT+ViT-GAN framework that generates synthetic lesions to balance classes, applies explainable AI techniques, and achieves strong performance in both synthesis and classification on dermoscopic datasets.

Zhou and Luo [69] proposed a deep feature fusion method with a mutual attention Transformer for skin lesion diagnosis, combining CNN-based feature extraction with Transformer attention to capture both local and global image patterns.

Zhang et al. (DermViT) [70] designed a diagnosis-guided transformer that simulates clinician decision logic through attention maps, resulting in improved interpretability and classification performance.

Iqbal et al. [71] proposed TESL-Net, a hybrid model combining a CNN encoder-decoder with Bi-ConvLSTM and a Swin Transformer to capture local, temporal, and global features. Designed to segment skin lesions, TESL-Net addresses irregular shapes and artefacts by leveraging both spatial and contextual cues. Tested on ISIC 2016–2018, it achieved state-of-the-art Jaccard scores, outperforming U-Net and FCN-based models.

Yadav et al. [72] introduced a dual-scale lightweight cross-attention ViT (DSCATNet) that processes 8×8 and 16×16 patches in parallel via cross-attention fusion. On HAM, it achieved 97.8% accuracy.

Kumar et al. [73] proposed GS-TransUNet, integrating 2D Gaussian splatting with Transformer UNet for joint segmentation and classification on ISIC-2017 and PH2. The unified architecture yielded superior accuracy and segmentation metrics across both datasets.

Xin et al. [74] proposed a Transformer network enhanced with multi-scale and overlapping sliding window patch extraction, showing marked improvements in classification accuracy over traditional patch-based ViTs.

Ravi et al. [75] introduced DEEPSCAN, a ViT-based architecture benchmarked on a novel "DermVisD" dataset (15,000 high-resolution annotations). The model showed an 18% increase in accuracy compared to CNN baselines, achieving 97.8%.

Mahbod et al. [76] explored model fusion between foundation models (PanDerm) and transformer architectures (ViT, SwinV2). On HAM and MSKCC, the fusion outperformed individual models, demonstrating the value of leveraging domain-specific pretraining jointly with fine-tuned transformer features.

Flosdorf et al. [67] performed a direct comparison between ViT-L16 and ViT-L32 models on HAM, yielding accuracy in the 91.6–92.8% range but with melanoma recall remaining modest (around 56–58%), highlighting ongoing challenges in class imbalance and diagnostic sensitivity.

In SkinSwinViT, Tang et al. [77] delivered a lightweight Swin tailored for multiclass lesion classification, emphasizing improved generalization and computational efficiency, making it promising for clinical integration.

Shafiq et al. [78] introduced ViT-GradCAM, a transformer architecture augmented with GradCAM interpretability. This methodology supports transparent classification decisions, enhancing trust in real-time diagnostic scenarios.

Krishna et al. [68] proposed LesionAid, a ViT-GAN framework that augments training data by generating synthetic skin lesion images and employs explainable AI for interpretability. On HAM, the model achieved a validation accuracy of 97.4%, demonstrating both classification accuracy and quality of generated samples.

Collectively, these transformer-based studies demonstrate advancements in classification accuracy and generalization, solidifying the role of ViTs as a powerful alternative and complement to CNNs in dermoscopic image analysis.

Table 2.4: Summary of key Vision Transformer-based approaches in skin lesion classification.

Authors (Year)	**Architecture / Approach**	**Dataset(s)**	**Key Contributions and Metrics**
Jain et al. (2021) [65]	ViT + CNN transfer learning	HAM	Improved generalization and interpretable attention
Zhou et al. (2021) [69]	CNN + Transformer with mutual attention	ISIC	Deep feature fusion with mutual attention; captures local and global context for diagnosis
Aladhadh et al. (2022) [64]	Medical Vision Transformer (MVT)	HAM	Accuracy ~96.1%, Recall ~96.5%, F1 ~97%
Xin et al. (2022) [74]	Multi-scale sliding window ViT	HAM	Improved accuracy via overlapping patches
Krishna et al. (2023) [68]	LesionAid (ViT + GAN + XAI)	HAM	Synthetic oversampling, explainability; val acc 97.4%
Lungu-Stan et al. (2023) [66]	Distilled lightweight ViT	HAM	Retains 98.3% of teacher accuracy; faster inference
Flosdorf et al. (2024) [67]	ViT-L16 / L32 comparison	HAM	Accuracy 91.6–92.8%, melanoma recall ~56%
Yadav et al. (2024) [72]	DSCATNet (Dual-scale cross-attention ViT)	HAM	Accuracy 97.8%
Iqbal et al. (2024) [71]	TESL-Net (CNN + Bi-ConvLSTM + Swin Transformer)	ISIC 2016–2018	Hybrid local-global segmentation; state-of-the-art Jaccard index
Ravi et al. (2024) [75]	DEEPSCAN (ViT on DermVisD)	DermVisD (~15k)	+18% accuracy vs CNNs; achieved 97.8%
Flosdorf et al. (2024) [67]	ViT-L16 / ViT-L32 comparison	HAM	Accuracy 91.6–92.8%; melanoma recall ~56–58%
Tang et al. (2024) [77]	SkinSwinViT (lightweight Swin Transformer)	Multiple lesions	Enhanced generalization; efficient inference
Shafiq et al. (2024) [78]	ViT-GradCAM (ViT + GradCAM)	Skin lesions	Interpretability-focused classification method
Zhang et al. (2025) [70]	DermViT: diagnosis-guided ViT	HAM, ISIC	Interpretability via attention aligned to diagnostic workflow
Kumar et al. (2025) [73]	GS-TransUNet (TransUNet + Gaussian Splatting)	ISIC-2017, PH2	Joint segmentation–classification with superior metrics
Mahbod et al. (2025) [76]	PanDerm foundation + ViT/SwinV2 fusion	HAM, MSKCC	Fusion improves performance over standalone models

2.3.1 Limitations of Transformer-Based Methods

While transformer architectures have introduced powerful mechanisms for modeling long-range dependencies, several limitations constrain their widespread adoption in skin lesion analysis. A

primary challenge is their strong dependence on large-scale training data. The original ViT [26] demonstrated excellent performance on ImageNet, but only when trained on millions of images. Medical imaging datasets such as HAM1 or ISIC contain only tens of thousands of examples, and data scarcity becomes a critical bottleneck. Raghu et al. [79] and Khan et al. [80] emphasize that without extensive pretraining or transfer learning, ViTs often underperform compared to CNNs on smaller datasets due to overfitting.

Another major limitation is the high computational cost of transformers. Unlike CNNs, which leverage local convolutions efficiently, ViTs require quadratic complexity in relation to input sequence length, leading to heavy memory and runtime demands. This makes deployment on resource-constrained or real-time clinical systems difficult [80]. Although lightweight variants (e.g., SkinDistilViT [66]) attempt to address this, performance–efficiency trade-offs remain unresolved.

Transformers also face challenges of robustness and generalization. Bhojanapalli et al. [81] and Naseer et al. [82] reported that ViTs are more sensitive to adversarial perturbations and distribution shifts than CNNs, raising concerns for clinical safety where patient populations and imaging devices vary. In dermatology specifically, this sensitivity amplifies existing issues of dataset bias and fairness.

Finally, while attention maps are often presented as an interpretability advantage, their reliability as faithful explanations is debated. Valanarasu et al. [83] observed that attention weights may not always align with clinically relevant features, limiting trust in such models for high-stakes decision support. This mirrors broader concerns about explainability in AI for healthcare.

Although transformer-based methods represent a major step forward by enabling global context modeling and improved performance on benchmark datasets, their reliance on large-scale pretraining, computational overhead, vulnerability to domain shifts, and contested interpretability limit their immediate clinical integration. These constraints highlight the need for hybrid strategies and specialized adaptations to fully leverage transformers in medical imaging.

2.4 Segmentation Approaches

Segmentation of skin lesions from dermoscopic images is crucial for early melanoma detection, yet it remains challenging due to irregular lesion shapes, fuzzy borders, and artifacts such as hairs and ink. Traditional CNN-based models (e.g., U-Net, FCN) struggle with long-range context and produce suboptimal Jaccard scores when architectures become overly deep or redundant. Early CNN-based approaches, such as DeepLabV3+ [84], adopted by Codella et al. [28], achieved a mean Intersection over Union (mIoU) of 78.6%, highlighting the importance of accurate boundary detection. Later methods integrated attention mechanisms and multi-scale features to address challenges such as lesion heterogeneity and low contrast. Masood et al. [85] combined ASSP-based DeepLabv3+ with UNet variants and encoders like VGG-16 [86], VGG-19, and DenseNet [87], reaching a Dice score of 91% and Jaccard index of 84% across multiple ISIC datasets.

Transformer-based and hybrid models overcome many of these limitations by capturing both local and global dependencies through self-attention. Wang et al. [88] proposed the Boundary-Aware Transformer (BAT), introducing a Boundary-wise Attention Gate (BAG) to enhance edge detection. Wu et al. [89] developed FAT-Net, integrating Transformer branches into a lightweight CNN architecture. Hu et al. [90] adapted the Segment Anything Model (SAM) [91] for segmentations of the skin lesions, achieving a mean Dice of 88.79% and mIoU of 78.43% on HAM. CA-Net [92] proposed spatial, channel, and scale attention modules, reporting a Dice score of 92.08% on ISIC-2018.

Iqbal et al. [71] introduced TESL-Net, which combines a CNN encoder-decoder with Bi-Conv LSTM and a Swin Transformer. This hybrid architecture models local features, temporal uncertainty, and global contextual relationships. Evaluated on ISIC 2016–2018, TESL-Net achieved state-of-the-art Jaccard indices, surpassing U-Net and FCN baselines.

Khan et al. [93] proposed TAFM-Net, which integrates self-adaptive transformer attention within an encoder (EfficientNetV2B1) and uses focal modulation within densely connected decoder skip paths. With a dynamic boundary-aware loss function, it attained Jaccard scores of 93.64%, 86.88%, and 92.88% on ISIC 2016, 2017, and 2018, respectively.

Qamar et al. presented ScaleFusionNet [94], combining Swin blocks via a Cross-Attention Transformer Module (CATM) and an AdaptiveFusionBlock to refine encoder-decoder feature fusion. This approach achieved Dice scores of 92.94% (ISIC 2016) and 91.65% (ISIC 2018).

Xu et al. introduced SkinFormer [95], a Vision Transformer that extracts statistical texture representations via a Kurtosis-guided statistical operator and transformer blocks. On ISIC 2018, it achieved a Dice score of 93.2%.

Benvcevic et al. [96] explored the use of polar transformation centered on the lesion to simplify the segmentation task, reporting a Dice score of 92.53% and IoU of 87.53%.

Bozorgpour et al. [97] proposed Dermosegdiff, a diffusion-based segmentation model that integrates boundary-awareness to improve lesion delineation. Their approach demonstrated superior edge preservation compared to conventional CNN- and Transformer-based methods.

Perera et al. [98] introduced MobileUNETR, a lightweight hybrid architecture that combines convolutional and Vision Transformer components. The model is optimized for efficiency and end-to-end medical image segmentation, making it suitable for resource-constrained or real-time applications.

Narayanan et al. [99] developed IARS-SegNet, an interpretable segmentation model that incorporates attention and residual skip connections. The method enhances melanoma boundary detection while improving explainability for clinical use.

Ding et al. [100] proposed CTH-Net, a hybrid architecture combining CNN and Transformer encoders to enhance global-local feature integration for segmentation tasks.

Azad et al. [101] extended U-Net by incorporating attention gates and multi-scale processing pathways. These modifications improved the model's ability to handle high variability in lesion appearance and background noise. Their work highlights how modern adaptations of U-Net continue to push the boundaries of segmentation accuracy in complex medical images.

FAT-Net by Wu et al. [89] uses feature-adaptive transformers in an encoder–decoder setup to improve segmentation, while BAT—Boundary-aware Transformer, proposed by Wang et al. [88] —integrates boundary-aware attention gates (BAG) to better capture ambiguous lesion edges. The subsequent XBound-Former [102] extends this with cross-scale boundary modeling, delivering SOTA performance on ISIC-2016/PH2 and ISIC-2018 with a pure transformer architecture.

2.4.1 Limitations of Segmentation Approaches

Despite the remarkable progress of CNN- and Transformer-based segmentation models, several limitations continue to hinder their effectiveness in real-world dermatological applications. A first challenge is the high variability of lesion appearance in terms of color, size, and shape, as well as the presence of artefacts such as hairs, rulers, or ink markings. These factors reduce boundary clarity and lead to inconsistent annotations, limiting model generalization across datasets [27, 28].

Table 2.5: Summary of Transformer-based and hybrid CNN-Transformer models for segmentation of the skin lesions.

Authors (Year)	**Model / Approach**	**Dataset(s)**	**Metrics / Contributions**
Codella et al. (2018) [28]	DeepLabV3+	ISIC 2017	mIoU: 78.6%
Gu et al. (2020) [92]	CA-Net (Channel/Scale Attention)	ISIC 2018	Dice: 92.08%
Benvcevic et al. (2021) [96]	Polar transform-based segmentation	ISIC 2018	Dice: 92.53%, IoU: 87.53%
Wang et al. (2021) [88]	BAT (Boundary-Aware Transformer)	ISIC/PH2	Adds boundary attention gate for better edge modeling
Wu et al. (2022) [89]	FAT-Net (feature-adaptive Transformer)	ISIC	Improved segmentation/classification via adaptive attention
Wang et al. (2023) [102]	XBound-Former (cross-scale boundary ViT)	ISIC-16, PH2, ISIC-18	Pure ViT with boundary modeling; SOTA performance
Bozorgpour et al. (2023) [97]	Dermosegdiff (Diffusion + boundary-aware)	ISIC	Boundary-preserving diffusion model; improved delineation accuracy
Hu et al. (2023) [90]	SAM fine-tuning for dermatology	HAM	Dice: 88.79%, mIoU: 78.43%
Iqbal et al. (2024) [71]	TESL-Net (CNN + Bi-ConvLSTM + Swin Transformer)	ISIC 2016–2018	Hybrid local-global modeling; top Jaccard index
Khan et al. (2024) [93]	TAFM-Net (Transformer Attention + Focal Modulation)	ISIC 2016–2018	Jaccard: 93.64%, 86.88%, 92.88%
Masood et al. (2024) [85]	DeepLabV3+ + U-Net + VGG/DenseNet encoders	ISIC 2016–2018	Dice: 91%, Jaccard: 84%
Xu et al. (2024) [95]	SkinFormer (Statistical Texture ViT)	ISIC 2018	Dice: 93.2%
Ding et al. (2024) [100]	CTH-Net (CNN + Transformer hybrid)	ISIC	Enhanced global-local segmentation modelling
Perera et al. (2024) [98]	MobileUNETR (lightweight CNN+ViT hybrid)	Medical	Efficient segmentation for cloud/edge; lightweight design
Narayanan et al. (2024) [99]	IARS-SegNet (attention residual skip)	ISIC	Interpretable segmentation; improved melanoma boundary detection
Azad et al. (2024) [101]	U-Net + attention, multi-scale	Medical, dermatology	Improved U-Net with attention gating and multi-scale paths
Qamar et al. (2025) [94]	ScaleFusionNet (Cross-Attn + Adaptive Fusion)	ISIC 2016, 2018	Dice: 92.94%, 91.65%

Another critical limitation is the sensitivity of segmentation models to dataset size and imbalance. Transformer-based models in particular require large annotated datasets to learn robust attention patterns [26], but most skin lesion datasets contain only a few thousand pixel-level annotations. This scarcity leads to overfitting and hampers transferability across acquisition settings. Furthermore, manual annotation of lesion masks is labor-intensive and subject to inter-observer variability, introducing noise into the ground truth [32].

Model complexity and computational requirements also remain key obstacles. Hybrid CNN – Transformer models, while achieving state-of-the-art accuracy, often require high GPU memory and lengthy training times [73, 88]. This makes them impractical for clinical deployment in low-resource or real-time settings. In addition, segmentation performance frequently varies across lesion types: for example, melanomas with irregular borders or low-contrast lesions are systematically harder to delineate, resulting in lower Dice or Jaccard indices even with advanced models [103, 20].

Finally, although attention mechanisms improve boundary detection, several studies report that Transformer-based models can be unstable under domain shifts and adversarial perturbations [81, 82]. This fragility raises concerns about robustness and reliability in diverse clinical environments, where images are captured under heterogeneous conditions.

Overall, while segmentation constitutes a fundamental step for skin lesion analysis, its limitations—including dataset scarcity, annotation variability, computational cost, and sensitivity to artefacts—underscore the need for approaches that either reduce reliance on explicit segmentation or integrate it more effectively into broader diagnostic pipelines.

2.5 Multitask Learning for Skin Lesion Classification

As discussed in Section 2.1 and summarized in Table 2.1, only a limited number of dermatology datasets provide paired annotations—namely, image-level diagnostic labels together with pixel-level segmentation masks. In practice, the main sources of such paired supervision are the HAM dataset (with a subset of segmentation masks released through the ISIC 2018 challenge) and PH2, while most other datasets lack either masks or aligned multi-class labels. The scarcity of jointly annotated data has significantly influenced the development of multi-task learning (MTL) strategies, which are typically evaluated using these two datasets.

Building upon this context, MTL has emerged as a powerful paradigm in skin lesion analysis, enabling models to simultaneously perform segmentation of the lesions and disease classification within a unified architecture. By optimizing both tasks simultaneously, MTL facilitates the learning of shared representations, reduces model redundancy, and improves both accuracy and generalization across tasks.

One of the earliest contributions is by Yang et al. [104], who proposed a CNN-based MTL architecture capable of segmenting lesions and classifying them into two categories. Their model achieved a Jaccard index of approximately 0.724 and classification AUCs of 0.88 and 0.97 on the ISIC-2017 dataset.

Chen et al. [105] introduced **MT-TransUNet**, a Transformer-based framework that mediates between segmentation and classification tasks via task-specific token decoders and regional consistency loss. The model surpassed state-of-the-art performance on ISIC-2017 and PH2, requiring around 48 million parameters and 0.17 seconds per inference.

Khan et al. [73] developed **GS-TransUNet**, which incorporates Gaussian Splatting into a Transformer-based U-Net structure. This end-to-end model demonstrated strong segmentation and

classification accuracy on both ISIC-2017 and PH2 datasets.

Amin et al. [106] proposed an integrated model combining BASNet for boundary-aware segmentation with a Convolutional Compact Transformer Model (CCTM) for classification. This fusion effectively enhanced both tasks by aligning lesion borders and class semantics.

Himel et al. [107] proposed an end-to-end pipeline combining segmentation via SAM with ViT-based classification. On HAM, their ViT-Google patch-32 model achieved 96.15% accuracy, with IoU 96.0% and Dice 98.1% across segmented lesion masks.

Manzoor et al. [108] proposed a two-stage pipeline composed of a U-Net (with VGG16 encoder) for segmentation and an EfficientFormer or SwiftFormer for classification. The pipeline attained a segmentation accuracy of 97.59% and a Jaccard index of 0.891 on HAM and ISIC-2018.

Zhang et al. [109] proposed **SkinM2Former**, a multimodal, multi-label Transformer that integrates dermoscopic images, clinical photographs, and patient metadata. The model addresses class imbalance using task-specific attention and achieved a mean diagnostic accuracy of 77.27% on the Derm7pt dataset.

Xie et al. [110] introduced the Mutual Bootstrapping Deep CNN (MB-DCNN), where segmentation provides masks to guide classification and vice versa, achieving Jaccard indices of 0.804 (ISIC-2017) and 0.894 (PH2), and classification AUCs of 0.938 and 0.977.

Saha et al. [111] proposed YoTransViT, a hybrid CNN-Transformer architecture that integrates segmentation of the lesions with classification. The model leverages segmentation cues to refine feature extraction, thereby improving classification robustness. Evaluated on multiple dermoscopic datasets, YoTransViT demonstrated strong performance with efficient training and inference, highlighting its potential for real-time clinical use.

Al-Masni et al. [112] proposed a unified multi-task model using joint reverse optimization, resulting in Dice scores of 89.48% and 88.81% (ISIC-2016, PH2) and improvements in classification F1 from 78.26% to 82.07% (ISIC-2016) and from 82.38% to 85.50% (PH2).

Al Mahmud et al. [113] presented SkinNet-14, a Transformer-mediated MTL framework achieving 91.2% classification accuracy and robust segmentation. Ahmed et al.'s DuaSkinSeg [114] employs a dual-encoder architecture combining MobileNetV2 and ViT-CNN, demonstrating competitive segmentation and classification performance across ISIC datasets.

2.5.1 Limitations of Multi-Task Approaches

Although MTL offers clear advantages by jointly optimizing segmentation and classification, several limitations hinder its robustness and clinical applicability. A primary issue is the strong interdependence between tasks: errors in the segmentation stage can propagate to the classification branch, leading to misdiagnoses. This dependency has been reported in early CNN-based pipelines [104, 110], where inaccurate lesion boundaries directly degraded classification performance.

Another challenge lies in dataset annotation requirements. Training reliable MTL models necessitates both pixel-level masks and class labels, which are expensive and time-consuming to obtain in dermatology [27, 39]. Moreover, annotation variability across experts introduces label noise that can affect both tasks simultaneously [32]. This dual dependency makes MTL approaches more vulnerable to inconsistencies compared to single-task counterparts.

From a computational perspective, MTL frameworks are often heavier than single-task models, as they combine segmentation decoders and classification heads into a single architecture. Models such as MT-TransUNet [105] or GS-TransUNet [73] significantly increase memory and inference demands, limiting their suitability for deployment in real-time or resource-constrained clinical environ-

Table 2.6: Summary of Multi-Task segmentation–classification models for skin lesion analysis.

Authors (Year)	Model / Approach	Datasets	Metrics / Contributions
Yang et al. (2017) [104]	Early MTL CNN (seg+cls)	ISIC-2017, PH2	Jaccard: 0.724; AUC: 0.88 / 0.97
Xie et al. (2020) [110]	MB-DCNN (mutual bootstrapping CNN)	ISIC-2017, PH2	Jaccard ~0.804/0.894; AUC 0.938/0.977
Chen et al. (2021) [105]	MT-TransUNet (Transformer MTL)	ISIC-2017, PH2	48M parameters, 0.17 s/image; SOTA joint performance
Khan et al. (2024) [73]	GS-TransUNet (Gaussian + TransUNet)	ISIC-2017, PH2	Accurate unified Transformer-based architecture
Al Mahmud et al. (2024) [113]	SkinNet-14 (Transformer MTL)	Skin lesion datasets	Accuracy: 91.2%
Himel et al. (2024) [107]	ViT with SAM segmentation + classification	HAM	Accuracy ~96.15%; segmentation IoU ~96%, Dice ~98%
Saha et al. (2024) [111]	YoTransViT (CNN + Transformer)	Dermoscopic datasets	Joint segmentation–classification; improved robustness and efficient inference
Al-Masni et al. (2025) [112]	Unified reverse-learning MTL	ISIC-2016, PH2	Dice: 89.48%/88.81%; F1: 82.07%/85.50%
Ahmed et al. (2025) [114]	DuaSkinSeg (dual-encoder ViT–CNN)	ISIC-2016/17/18	Competitive segmentation and classification performance
Amin et al. (2025) [106]	BASNet + CCTM	ISIC variants	Boundary-aware segmentation with Transformer classifier
Manzoor et al. (2025) [108]	U-Net + EfficientFormer	HAM, ISIC-2018	Accuracy: 97.11%; Jaccard: 0.891
Zhang et al. (2025) [109]	SkinM2Former (tri-modal ViT)	Derm7pt	Accuracy: 77.27%; robust multi-label prediction

ments. Recent works have attempted to address this with lighter hybrids (e.g., YoTransViT [111]), but efficiency–performance trade-offs remain unresolved.

Finally, while MTL can foster shared representations, it does not always guarantee performance improvements for both tasks. As noted by Ruder [115] and Vandenhende et al. [116], negative transfer may occur when one task dominates the optimization process, hindering the learning of the other. In dermatology, this can manifest when segmentation quality improves but classification accuracy stagnates, or vice versa. Addressing task balancing remains an open research problem, requiring adaptive weighting or dynamic optimization strategies.

In summary, despite their promise, multi-task solutions face limitations related to error propagation, annotation scarcity, computational overhead, and potential negative transfer between tasks. These constraints underscore the necessity for carefully designed architectures and training strategies to fully leverage the potential of MTL in skin lesion analysis.

2.6 Multimodal Approaches: Integrating Images and Metadata

Early works such as Ge et al. [117] and Kawahara et al. [33] demonstrated that integrating dermoscopic and clinical information, or metadata aligned with the seven-point checklist, can improve diagnostic prediction and interpretability. Subsequent CNN-based methods explored progressively more sophisticated fusion strategies: Bi et al. [118] introduced hyper-connected networks for multi-label learning, Gessert et al. [119] combined metadata with multi-resolution image encoders, and Pacheco & Krohling [120] proposed an attention-based fusion block to better weight clinical attributes. Tang et al. [121], Wang et al. [122], and Ou et al. [123] extended these designs with cross-attention and adversarial mechanisms to integrate metadata or smartphone-acquired clini-

cal images. Recent transformer-based architectures further advanced multimodal learning: Xu et al. [124, 125], Cai et al. [126], Cheslerean-Boghiu et al. [127], Zhang et al. [128, 109], Adebiyi et al. [129], and Christopoulos et al. [130] introduced various cross-attention, distillation, and contrastive frameworks for joint image–metadata representation learning. Other notable contributions include multimodal ensembles [131], dual-image fusion [132, 122], and metadata-driven fairness assessments [133]. Collectively, these works highlight the growing interest in multimodal fusion for skin lesion analysis and the diverse strategies proposed to achieve robust integration of heterogeneous data sources.

Despite their promise, multimodal models face several limitations. Their performance critically depends on the quality, completeness, and consistency of metadata, which in public datasets are often missing, heterogeneous, or biased toward specific populations [27, 32]. Metadata can also introduce spurious correlations and shortcut learning effects [63, 134], particularly if patient-level separation is not enforced during dataset splitting. Additional challenges include the high annotation and curation effort required to collect reliable metadata [27, 32], demographic imbalance leading to fairness disparities [22], and the risk of negative transfer or over-reliance on a dominant modality [135]. Furthermore, attention-based interpretability in fusion networks may not always reflect the true reasoning process [62], and multimodal systems typically demand higher computational resources while being sensitive to missing modalities at inference [136].

Image–metadata fusion can improve diagnostic accuracy and interpretability compared to image-only models [120], yet its benefits remain contingent on rigorous dataset curation, leakage-safe evaluation, and fairness-aware model design. While not the main focus of this thesis, these multimodal frameworks provide valuable context for understanding how complementary non-visual data can further enhance future dermatological AI systems.

2.7 General Limitations and Open Challenges

The limitations discussed in the previous sections highlight several cross-cutting challenges that continue to hinder the reliable and equitable deployment of AI in dermatological practice.

A primary and persistent limitation is *dataset bias and limited diversity*. Widely adopted benchmarks such as ISIC and HAM over-represent lighter skin tones and common benign lesions, while under-representing darker phototypes and rarer malignancies [22, 60]. These imbalances lead to substantial drops in accuracy when models are evaluated on underrepresented groups. Class imbalance further compounds the issue, with malignant lesions systematically outnumbered by benign ones. Although data augmentation and GAN-based synthesis have been explored to counteract this [137], generated images can introduce artifacts that compromise clinical realism and generalization.

A second recurring challenge is *robustness under domain shift*. Models trained on a single dataset or institution often fail to generalize across different imaging devices, acquisition settings, or populations [63]. The lack of standardized evaluation protocols, consistent metrics, and truly external test sets exacerbates this issue, resulting in inflated or non-comparable performance reports [138, 139]. For instance, certain studies report validation accuracy as benchmark performance, blurring the boundary between model tuning and independent testing [140].

Another major limitation concerns *interpretability and clinical trust*. Although saliency maps and attention mechanisms are widely used for visual explanation, several studies have shown that these tools may not faithfully represent the model's decision process [62, 83]. Moreover, shortcut learning remains a risk: models may rely on spurious correlations such as patient age or lesion

location instead of pathognomonic features [134]. This problem is amplified by the presence of duplicated patients or overlapping metadata between training and test sets, potentially inflating results if not properly controlled [139].

From a methodological standpoint, *model complexity and computational cost* also limit the applicability of models in real-world workflows. Transformer-based architectures, while powerful, impose substantial memory and compute demands [80]. This restricts their use in low-resource clinical settings and complicates deployment on edge devices. Furthermore, reproducibility remains a systemic weakness: inconsistent splits, incomplete reporting, and unshared code hinder transparent comparison and replication of published results [138].

In summary, despite remarkable progress, deep learning for skin lesion analysis remains constrained by dataset bias, domain shift, interpretability challenges, limited reproducibility, and computational cost. Addressing these limitations requires both methodological and systemic efforts: curating larger and demographically balanced datasets, enforcing standardized evaluation and reporting practices, developing lightweight yet robust architectures, and designing models that enhance interpretability and clinical reliability.

Within this thesis, particular attention is devoted to mitigating the effects of dataset bias and domain shift through sequential training strategies, improving generalization across datasets and tasks, and promoting robustness and transparency in model evaluation. These directions form the foundation for the experimental framework and analyses presented in the following chapters.

Chapter 3

A Large Dataset to Enhance Skin Cancer Classification with Transformer-Based Deep Neural Networks[1]

3.1 Introduction and Motivation

As discussed in Chapter 2, publicly available datasets for skin lesion analysis often suffer from significant limitations, including class imbalance, duplicated images, inconsistent labeling, and a lack of diversity in image modalities. These issues substantially impact the generalization capability of ML models, particularly in clinical scenarios where robustness is crucial.

To address these challenges, this chapter presents an approach that leverages Transformer-based deep learning models, specifically the Swin Transformer architecture, combined with a curated and unified Large Dataset (LD) to improve classification performance for skin lesions. Transformer models, originally developed for natural language processing, have demonstrated exceptional capabilities in capturing long-range dependencies in image data through the use of self-attention mechanisms. However, their application in medical imaging remains relatively underexplored, particularly in dermatology.

Beyond evaluating the potential of Transformer architectures, a key assumption underpinning this study is that increasing the amount and diversity of training data can significantly improve a model's generalization capabilities. Given that Transformer models typically require large datasets to express their full potential, we aimed to investigate how classification performance evolves as we expand and balance the available data. At the same time, we remain mindful of the persistent class imbalance and its potential to bias learning outcomes.

This first study focuses on constructing the LD by aggregating and refining multiple public

[1]This chapter is based on the published article: **M. Gallazzi, S. Biavaschi, A. Bulgheroni, T. M. Gatti, S. Corchs, and I. Gallo**, "A Large Dataset to Enhance Skin Cancer Classification With Transformer-Based Deep Neural Networks," *IEEE Access*, vol. 12, pp. 109544–109559, 2024. DOI: 10.1109/ACCESS.2024.3439365.

archives, while also examining the impact of dataset imbalance. We proposed strategies, such as data augmentation and class downsampling, to counteract this imbalance and systematically tested these choices using the Swin Transformer model. The following sections detail the dataset creation, model architecture, and evaluation framework used to explore these questions. Parts of this chapter are based on work published in [141].

3.2 Dataset Construction: From Fragmented Archives to a Unified Large Dataset

The widely used HAM dataset, while highly influential, is limited by its size (10,015 images) and pronounced class imbalance, with more than 66% of images belonging to the benign NV class. Moreover, several datasets used in the literature—such as BCN20000, MSK1-5, and PH2—contain images that overlap with HAM or with each other, raising the risk of data leakage during training and testing.

To address these issues, we created a unified LD by integrating multiple publicly available datasets, selecting only dermatoscopic images annotated with one of the seven canonical diagnostic categories already described in Subsection 2.1. A rigorous filtering process was applied to remove duplicate images, non-dermatoscopic modalities, samples without confirmed diagnoses, and irrelevant classes. In Table 2.2, we have already presented the cardinality of the datasets and some of the problematic characteristics that are taken into consideration in this work, such as the number of duplicate images (DI) or the variability in the image sizes.

To complement this, Table 3.1 shows all the characteristics and cardinalities of the datasets used to create the LD. For each dataset, we report the image size and the number of images per class, with the final column indicating the total number of usable images. As shown, this integration process resulted in a dataset composed of 41,975 images from diverse sources, contributing to improved generalization potential. However, we also observe that the class distribution remains heavily imbalanced, with the NV class dominating the dataset. This pronounced imbalance motivated the implementation of dedicated strategies in the preprocessing and training phases to mitigate its impact. In the next subsection, after introducing the model architecture, we will explore the techniques adopted in more detail to address this issue.

Such variety across acquisition settings, equipment, and labeling standards is expected to make the LD more representative of real-world conditions. Nevertheless, despite the increase in scale and diversity, the class distribution remains heavily skewed toward benign lesions, particularly the NV class. This pronounced imbalance motivated the implementation of specific mitigation strategies in both the preprocessing and training stages. In the next subsection, after introducing the architecture of the employed Transformer model, we will explore the data augmentation and sampling techniques devised to address the imbalance challenge.

3.3 Methodology: Transformer-Based Learning for Skin Lesion Classification

Following the construction of the LD dataset, we designed a comprehensive experimental framework to evaluate the performance of Transformer-based models—specifically, the Swin Transformer

Table 3.1: Datasets used to create the LD proposal. For each dataset, we report the image size, the number of images present for each of the seven classes considered here, and the total number of images.

DATASETS	Size of images	MEL	NV	BCC	AKIEC	BKL	DF	VASC	TOT
HAM10000 [27]	600 x 450	1113.0	6705.0	514.0	327.0	1099.0	115.0	142.0	10 015.0
Consecutive biopsies [34] for melanoma (2020)	3264 x 2448	117.0	691.0	12.0	60.0	223.0	6.0	7.0	1116.0
MSK-1 [29]	variable	368.0	760.0	71.0	17.0	116.0	9.0	39.0	1380.0
MSK-2 [29]	variable	937.0	1861.0	672.0	480.0	392.0	9.0	23.0	4374.0
MSK-3 [29]	variable	27.0	148.0	61.0	56.0	124.0	11.0	7.0	434.0
MSK-4 [29]	variable	247.0	595.0	278.0	303.0	343.0	26.0	14.0	1806.0
MSK-5 [29]	variable	0.0	0.0	0.0	0.0	109.0	0.0	2.0	111.0
Hospital Italiano de Buenos Aires Dataset [30]	variable	253.0	602.0	340.0	221.0	88.0	61.0	51.0	1616.0
SKINL2 [35]	1920 x 1080	53.0	151.0	52.0	14.0	64.0	17.0	46.0	397.0
BCN20000 [32]	1024 x 1024	2857.0	4206.0	2809.0	1168.0	1138.0	124.0	111.0	12 413.0
UDA-1 [29]	variable	159.0	396.0	0.0	0.0	0.0	0.0	0.0	555.0
UDA-2 [29]	variable	34.0	12.0	3.0	0.0	7.0	2.0	2.0	60.0
7-pt [33]	768 x 512	252.0	575.0	42.0	0.0	45.0	20.0	29.0	963.0
ISIC 2020 Challenge Training Set [31]	variable	581.0	5191.0	0.0	0.0	177.0	0.0	0.0	5949.0
ISIC Challenge 2018: Task 1-2 Test [28]	variable	118.0	410.0	0.0	0.0	11.0	0.0	0.0	539.0
ISIC Challenge 2018: Task 1-2 Validation[28]	variable	11.0	35.0	0.0	0.0	1.0	0.0	0.0	47.0
PH2 [36]	765 x 575	40.0	160.0	0.0	0.0	0.0	0.0	0.0	200.0
TOT		7167.022	498.0	4854.0	2646.0	3937.0	400.0	473.0	41 975.0

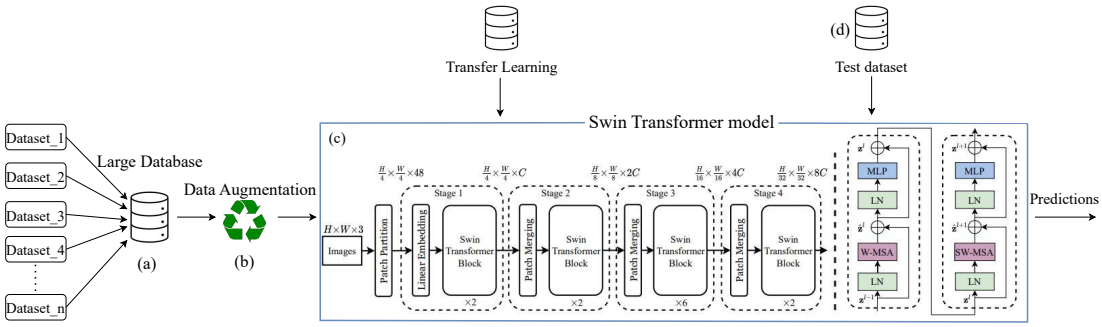


Figure 3.1: Overview of the proposed model and dataset for skin lesion classification with the key components: (a) creation of an LD assembled from multiple existing datasets; (b) application of various data augmentation techniques to improve the robustness of the model; (c) the architecture of the pre-trained Swin model used [1]; (d) use of a standardized test set.

(Swin) [1]—in the task of skin lesion classification. Our approach aimed to systematically investigate how increasing dataset size and diversity affect classification accuracy, while also addressing class imbalance through data augmentation and fine-tuning strategies.

The overall pipeline proposed in this work is illustrated in Figure 3.1, and involves: (i) dataset partitioning and preprocessing, (ii) model training using different dataset configurations, (iii) validation - based model selection, and (iv) evaluation on an external test set.

We designed three groups of experiments. The first group uses the HAM dataset alone for training, the second experiment used a combination of the two most populated datasets, HAM and BCN, while the third leverages the proposed LD dataset. In all cases, the same external test dataset (HAMt) is used for evaluation to ensure a fair comparison of model performance across seven classes.

3.3.1 Dataset Preprocessing and Augmentation

A key component of this work involves managing the class imbalance and variability inherent in real-world skin lesion datasets. For all experiments, the datasets were randomly divided into training (80%) and validation (20%) subsets, maintaining the original class distribution to ensure consistency across splits.

To address the limited number of samples—particularly in underrepresented classes—and to enhance model generalization, we employed a set of data augmentation (DA) techniques during training. These included:

- *Resize:* All images were adjusted to the input resolution required by the model architecture.
- *Crop:* Either centered or random cropping was applied, depending on the experimental phase.
- *Horizontal Flip:* Random flipping along the vertical axis.
- *Rotation:* Random rotations within a predefined range and probability.
- *Normalization:* Applied using ImageNet statistics (mean: 0.485, 0.456, 0.406; std: 0.229, 0.224, 0.225).

For validation and testing, only basic preprocessing (resize and crop) was applied to avoid information leakage and ensure fair performance evaluation. In Table 3.2, all the techniques used in our experiments are reported.

Beyond standard augmentation, we systematically curated multiple dataset configurations to investigate the impact of class imbalance, duplication, and dataset scale on classification performance. Specifically, we generated several experimental variants of HAM, BCN, and the proposed LD by removing repeated samples and using a downsampling strategy to reduce the over-representation of the NV class. In addition, merged datasets—such as HAM+BCN—were created to assess the effect of increased sample diversity and cardinality.

Table 3.3 summarizes the class-wise distributions of all dataset variants used in our experiments. Each configuration reflects a distinct strategy for mitigating dataset biases and evaluating model robustness. Unified datasets further enable analysis of scalability and generalization across heterogeneous sources.

These curated configurations, combined with the augmentation strategies described above, constitute the experimental framework used to evaluate the effectiveness and generalization capability of Transformer-based models for skin lesion classification.

Table 3.2: Augmentation techniques used in the different experiments. For each strategy, the resize, crop size, horizontal flip, rotation, and normalization parameters are indicated. The parameter "p" represents the probability with which that technique is applied.

Identifier	**Resize**	**Crop**	**Horizontal Flip**	**Rotation**	**Normalization**
DA_Train	224×280	RandomCrop $(224, 224)$	RandomHorizontalFlip $(p = 0.5)$	RandomRotation $(-180, 180),\ p = 0.99$	mean $(0.485, 0.456, 0.406)$; std $(0.229, 0.224, 0.225)$
DA_Train2	$600 \cdot 0.65$ $450 \cdot 0.65$	RandomCrop $(224, 224)$	RandomHorizontalFlip $(p = 0.5)$	RandomRotation $(-180, 180),\ p = 0.99$	mean $(0.485, 0.456, 0.406)$; std $(0.229, 0.224, 0.225)$
DA_Train3	$600 \cdot 0.56$ $450 \cdot 0.56$	RandomCrop $(224, 224)$	RandomHorizontalFlip $(p = 0.5)$	RandomRotation $(-180, 180),\ p = 0.99$	mean $(0.485, 0.456, 0.406)$; std $(0.229, 0.224, 0.225)$
DA_Train4	256×280	RandomCrop $(256, 256)$	RandomHorizontalFlip $(p = 0.5)$	RandomRotation $(-180, 180),\ p = 0.99$	mean $(0.485, 0.456, 0.406)$; std $(0.229, 0.224, 0.225)$
DA_Val	224×280	RandomCrop $(224, 224)$	–	–	mean $(0.485, 0.456, 0.406)$; std $(0.229, 0.224, 0.225)$
DA_Val2	256×280	RandomCrop $(256, 256)$	–	–	mean $(0.485, 0.456, 0.406)$; std $(0.229, 0.224, 0.225)$
DA_Test	224×280	RandomCrop $(224, 224)$	–	–	mean $(0.485, 0.456, 0.406)$; std $(0.229, 0.224, 0.225)$
DA_Test2	256×280	RandomCrop $(256, 256)$	–	–	mean $(0.485, 0.456, 0.406)$; std $(0.229, 0.224, 0.225)$

Table 3.3: Class-wise distribution of the datasets used in our experiments. Variants include deduplicated versions, NV-downsampled subsets to reduce class imbalance, and merged datasets (e.g., HAM+BCN or LD unified). The last column reports the total number of samples per configuration.

ACRONYM REFERENCE	MEL	NV	BCC	AKIEC	BKL	DF	VASC	TOT
HAM_noDuplicates	614.0	5403.0	327.0	228.0	727.0	73.0	98.0	7470.0
HAM_Duplicates	1113.0	6705.0	514.0	327.0	1099.0	115.0	142.0	10 015.0
HAM_NV_Downsampled	1113.0	5403.0	514.0	327.0	1099.0	115.0	142.0	8713.0
BCN_noDuplicates	524.0	1281.0	983.0	358.0	353.0	40.0	37.0	3576.0
BCN_Duplicates	2857.0	4206.0	2809.0	1168.0	1138.0	124.0	111.0	12 413.0
HAM_BCN_Duplicates	3970.0	10 911.0	3323.0	1495.0	2237.0	239.0	253.0	22 428.0
HAM_Duplicates_BCN_noDuplicates	1637.0	7986.0	1497.0	685.0	1452.0	155.0	179.0	13 591.0
HAM_BCN_noDuplicates	1138.0	6684.0	1310.0	586.0	1080.0	113.0	135.0	11 046.0
LARGE_DATASET_Derm_Duplicates	7167.0	22 498.0	4854.0	2646.0	3937.0	400.0	473.0	41 975.0
LARGE_DATASET_Derm_NV_Downsampled	7167.0	13 306.0	4854.0	2646.0	3937.0	400.0	473.0	32 783.0
LARGE_DATASET_Derm_NV_30Balanced	7167.0	9314.0	4854.0	2646.0	3937.0	400.0	473.0	28 791.0
LARGE_DATASET_Derm_NV_20Balanced	7167.0	10 644.0	4854.0	2646.0	3937.0	400.0	473.0	30 121.0
LARGE_DATASET_Unified_Duplicates	8413.0	24 929.0	6936.0	5034.0	4811.0	637.0	1373.0	52 133.0
LARGE_DATASET_Unified_NV_Downsampled	8413.0	13 521.0	6936.0	5034.0	4811.0	637.0	1374.0	40 726.0

3.3.2 Swin Transformer Architecture

CNNs have traditionally dominated image classification tasks by learning hierarchical spatial features. However, their inherent limitation in capturing long-range dependencies has led to growing interest in Transformer-based models. The Swin Transformer (Swin) [1], a hierarchical vision Transformer, was selected for this work due to its excellent trade-off between efficiency and representational power.

Unlike standard Vision Transformers, which compute global self-attention across the entire image—often at a high computational cost—Swin introduces shifted window-based self-attention. This approach limits attention computations to non-overlapping local windows while allowing cross-window interactions through window shifting. This mechanism drastically reduces computational complexity while retaining the ability to model long-range dependencies across layers.

The Swin architecture is composed of four stages, each reducing the spatial resolution of feature maps and increasing the channel dimension, mimicking the pyramidal structure of CNNs. Each stage contains multiple Swin Blocks, which consist of:

- Layer Normalization (LN) before each operation;
- Window-based Multi-head Self-Attention (W-MSA) and Shifted Window MSA (SW-MSA) modules;
- Multi-layer Perceptron (MLP) with GELU activation.

The core attention mechanism follows the standard Transformer formulation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{3.1}$$

Where

- Q, K, and V are the query, key, and value matrices, respectively;
- softmax represents the softmax function;
- $(\frac{QK^T}{\sqrt{d}})$ denotes the matrix multiplication of Q, K transposed, scaled by the square root of d, where d is the dimensionality of the query and key matrices;
- The entire expression inside the softmax function is then multiplied by V (value matrices).

To allow the model to capture both local and global representations, the window partitioning is shifted in alternate blocks (cyclic shift), enabling information exchange between adjacent windows. This hierarchical representation allows the network to first capture fine details and progressively build abstract concepts at coarser levels.

Between stages, patch merging layers are applied for downsampling, where feature maps are concatenated and linearly projected to reduce dimensionality. The number of stages, window size, and depth of the network vary depending on the model variant.

Table 3.4 summarizes the configurations used in our experiments, from the lightweight Swin-Tiny to the deeper Swin-Large and SwinV2-Large. Notably, SwinV2 introduces improvements in positional encoding, normalization, and attention scaling, resulting in enhanced performance, especially on larger datasets.

These characteristics make Swin particularly suitable for our task, which demands both high-resolution detail (e.g., lesion borders and textures) and global context (e.g., lesion symmetry and distribution).

Table 3.4: Swin Transformer parameters

Model	**Layers per Stage**	**#Layers**	**Param.**
Swin-Tiny	2, 2, 6, 2	12	28M
Swin-Small	2, 2, 18, 2	24	50M
Swin-Base	2, 2, 18, 2	24	88M
Swin-Large	2, 2, 18, 2	24	197M
SwinV2-Large	2,2,18,18	48	197M

3.3.3 Training Strategy and Evaluation

All models were trained using an early stopping criterion based on the validation loss. After each training epoch, the model was evaluated on the validation set, and weights were saved if an improvement in validation loss was observed. The training process continued until no further improvement was detected for 20 or 30 consecutive epochs, depending on the experimental configuration.

Once training was complete, the best model—identified as the one with the lowest validation loss—was reloaded and evaluated on the external test set. This evaluation allowed us to assess the model's performance on an unseen and independent dataset. We report four standard classification metrics used to evaluate the different performances: accuracy, precision, recall, and F1-score, which are defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.4}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.5}$$

Where "TP" indicates the "True Positives", "TN" the "True Negatives", "FP" the "False Positives", and " FN " the "False Negatives", respectively.

In selected experiments, we also applied a fine-tuning strategy to explore the model's transfer learning capabilities. Specifically, after training the model on a source dataset (e.g., HAM or LD), the learned weights were used to initialize training on a second dataset. This procedure allowed the model to retain previously acquired representations while adapting to new data distributions. The evaluation process for fine-tuned models remained unchanged, including early stopping and assessment on the external test set.

3.4 Experimental Results and Analysis

To systematically evaluate the impact of dataset characteristics, augmentation strategies, and model configurations on skin lesion classification, we organized the subsequent subsections to describe our experiments into three main parts:

- The first part includes experiments conducted exclusively on the HAM dataset.
- The second group of experiments investigates fine-tuning strategies between the HAM and BCN datasets.
- The third group evaluates classification performance on the proposed Large Dataset (LD), built by unifying multiple publicly available datasets.

Each experiment explores a specific combination of model architecture (Swin variants), data augmentation pipeline, and dataset configuration (e.g., with or without duplicate removal, with balanced or imbalanced class distributions). A summary of all experiments, including the identifiers (MEL1–MEL20), is provided in Table 3.5. The Swin Large (SwL), SwinV2 Base (SwV2B), and SwinV2 Large (SwV2L) models were tested across different settings.

Performance evaluation was consistently carried out on an external test set, with results reported in terms of weighted average accuracy, precision, recall, and F1-score. Unless otherwise specified, all training phases used the Adam optimizer with a learning rate of 1×10^{-4}, CrossEntropy loss, and early stopping based on validation loss. The batch size was generally set to 128, except for the SwV2L (MEL7), which required a batch size of 64 due to memory constraints.

3.4.1 Results on the HAM Dataset

The first group of experiments (MEL1–MEL7) investigates the classification performance of Swin models trained exclusively on the HAM dataset.

In experiment **MEL1**, we trained the SwL model using the unfiltered HAM dataset, which contains duplicated samples. Duplicates typically refer to multiple images of the same lesion captured under different lighting conditions, orientations, or zoom levels. Despite their prevalence, such duplicates can introduce bias by causing the model to overfit to repeated visual patterns. Using standard data augmentation (DA_Train), MEL1 achieved a test accuracy of **83.06%**, representing the initial baseline for comparison.

In subsequent experiments, we explicitly removed duplicate samples. In **MEL2**, the same SwL model was trained on the HAM dataset without duplicates, leading to a modest improvement in accuracy (**83.85%**). Similarly, in **MEL3**, we evaluated the SwV2B model under the same training conditions, yielding a slightly lower accuracy (**82.13%**), possibly due to the reduced model capacity compared to SwL.

Experiments **MEL4** and **MEL5** investigated the impact of varying resizing scales in the augmentation pipeline. In MEL4, a resize factor of 65% (DA_Train2) was applied, while in MEL5 the factor was reduced to 56% (DA_Train3). Both experiments used the SwL model. MEL4 achieved the highest accuracy in this subgroup (**84.51%**), followed by MEL5 (**83.52%**). These findings suggest that moderate resizing enhances classification by better adapting the lesion scale to the network input, although excessive downscaling may result in the loss of discriminative detail.

Table 3.5: This table collects all the experiments in this paper. The lines divide the three groups of experiments in detail: MEL1-7 are the experiments where HAM was used, MEL8-12 are the experiments where HAM and BCN were used for fine-tuning, and MEL13-20 are the experiments involving the use of the proposed LD. Values in bold are the best values obtained for each type of experiment. The acronyms TA, TP, TR, and TF1 represent Test Accuracy, Test Precision, Test Recall, and Test F1 Score, respectively, and are calculated as a weighted average, taking into account the attendance of each class.

ID	MODEL	DATASET	DA	TA	TP	TR	TF1
MEL1	Swin Large	HAM_Duplicates	DA_Train	83.1	83.3	83.1	83.0
MEL2	Swin Large	HAM_noDuplicates	DA_Train	83.9	83.5	83.9	83.5
MEL3	SwinV2 Base	HAM_noDuplicates	DA_Train	82.1	82.2	82.1	81.7
MEL4	Swin Large	HAM_noDuplicates	DA_Train2	84.5	84.4	84.5	84.0
MEL5	Swin Large	HAM_noDuplicates	DA_Train3	83.5	83.4	83.5	83.4
MEL6	Swin Large	HAM_NV_Downsampled	DA_Train	84.3	84.7	84.3	84.4
MEL7	SwinV2 Large	HAM_NV_Downsampled	DA_Train4	**84.64**	84.8	84.6	84.6
MEL8	Swin Large	HAM_BCN_noDuplicates	DA_Train3	**85.70**	85.5	85.7	85.3
MEL9	Swin Large	HAM_BCN_noDuplicates	DA_Train	79.8	80.8	79.8	78.8
MEL10	Swin Large	HAM_BCN_Duplicates	DA_Train	83.5	83.5	83.5	82.8
MEL11	Swin Large	HAM_Duplicates_BCN_noDuplicates	DA_Train	80.7	80.3	80.7	80.2
MEL12	Swin Large	HAM_BCN_noDuplicates	DA_Train	83.7	83.6	83.7	83.5
MEL13	Swin Large	LARGE_DATASET_Derm_Duplicates	DA_Train	85.8	85.7	85.8	85.7
MEL14	Swin Large	LARGE_DATASET_Derm_NV_Downsampled	DA_Train	**86.37**	86.8	86.4	86.4
MEL15	Swin Large	LARGE_DATASET_Derm_NV_30Balanced	DA_Train	84.7	86.1	84.7	85.0
MEL16	Swin Large	LARGE_DATASET_Derm_NV_20Balanced	DA_Train	84.0	84.9	84.0	84.2
MEL17	SwinV2 Large	LARGE_DATASET_Derm_NV_Downsampled	DA_Train4	83.7	83.6	83.7	83.5
MEL18	Swin Large	LARGE_DATASET_Unified_Duplicates	DA_Train	74.4	74.2	74.4	73.9
MEL19	Swin Large	LARGE_DATASET_Unified_NV_Downsampled	DA_Train	84.6	86.1	84.6	85.0
MEL20	Swin Large	LARGE_DATASET_Unified_NV_Downsampled	DA_Train	70.7	73.7	70.7	71.5

Given the significant imbalance in HAM, where the NV (benign nevi) class dominates the dataset, experiments **MEL6** and **MEL7** addressed this issue by applying class-specific downsampling. In MEL6, duplicate removal was applied only to the NV class to reduce its disproportionate weight, while all other classes retained their original size. This approach led to a test accuracy of **84.32%**, demonstrating the positive impact of targeted rebalancing.

In **MEL7**, we employed the SwV2L model with a larger input resolution (256×256) and downsampled the NV class. With the modified DA_Train4 pipeline and adjusted input sizes, MEL7 achieved the highest accuracy in this group: **84.64%**. The associated confusion matrix (see Figure 3.2, top-center) reveals a substantial reduction in misclassifications toward the dominant NV class, particularly for MEL and BKL, confirming the benefits of rebalancing.

3.4.2 Results on HAM and BCN Fine-Tuning

To evaluate the transferability of learned representations and the impact of dataset diversity, we conducted a second series of experiments (MEL8–MEL12) involving both the HAM and BCN datasets.

In experiment **MEL8**, we merged the HAM and BCN datasets, removing all duplicates prior to training. The resulting dataset comprised 10,015 images from HAM and 12,413 from BCN, maintaining seven common diagnostic classes. The Swin Large model was trained using the DA_Train3 augmentation strategy, previously shown to yield strong results. This experiment achieved the best performance in this group, with a test accuracy of **85.70%**. The confusion matrix (Figure 3.2, top-right) indicates a noticeable improvement in classification consistency across all classes. This

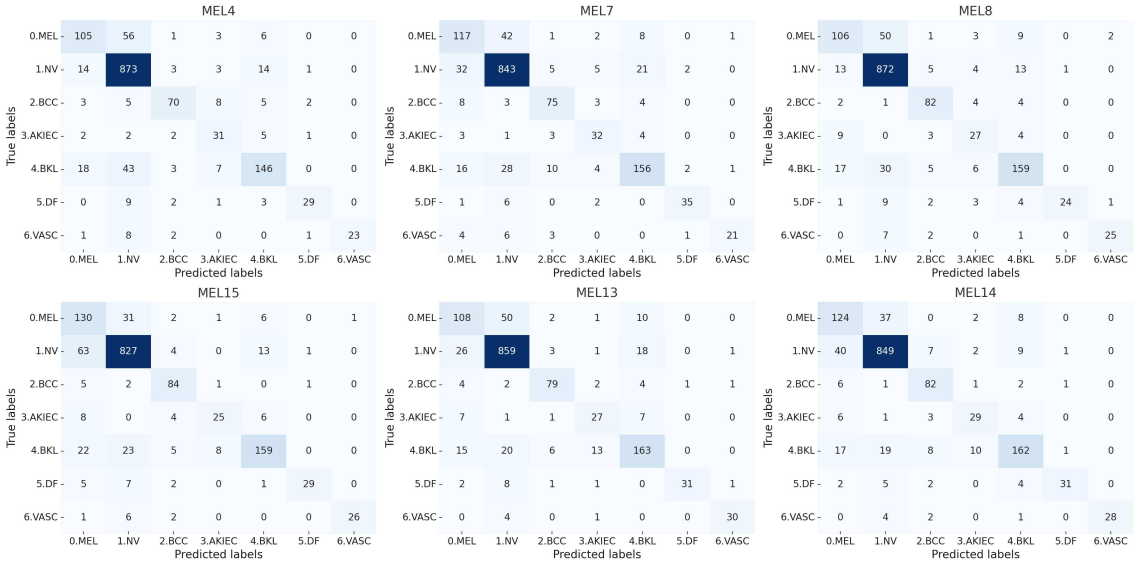


Figure 3.2: Confusion matrices of some of the experiments described in Table 3.5. From left to right, first row: MEL4, MEL7, and MEL8, and second row: MEL15, MEL13, and MEL14.

result demonstrates that increasing the size and variability of the training data can have a positive impact on model generalization for unseen test data.

Experiments **MEL9** through **MEL12** explored sequential fine-tuning between HAM and BCN datasets. Two configurations were tested: training from scratch on one dataset and fine-tuning on the other, with or without duplicate filtering.

In **MEL9**, the SwL model was trained on HAM and fine-tuned on BCN (no duplicates), resulting in an accuracy of **79.81%**. Conversely, **MEL10** reversed the order (trained on BCN, fine-tuned on HAM), achieving **83.45%**. In **MEL11**, the model was trained on HAM (with duplicates) and fine-tuned on BCN (no duplicates), obtaining **80.68%**. **MEL12** involved training on HAM (no duplicates) and fine-tuning on BCN (no duplicates), leading to a performance of **83.65%**.

These results suggest that both the pretraining order and the presence of duplicated samples have a significant influence on the outcome. In general, training on larger and more diverse data (e.g., BCN) before fine-tuning on cleaner subsets (e.g., deduplicated HAM) tends to yield higher performance. However, none of the sequential fine-tuning strategies outperformed the direct combined training used in MEL8.

3.4.3 Results on the Proposed LD

To further investigate the impact of dataset scale, diversity, and class balancing on model performance, we introduced and evaluated a curated and unified LD. This experimental group (MEL13 – MEL20) aimed to assess whether increasing data volume and reducing imbalance could lead to improved generalization, especially when combined with Transformer-based models such as Swin.

In **MEL13**, the SwL model was trained on the LD containing duplicates, totaling 41,975 images. Data augmentation strategy DA_Train was applied. The test accuracy reached **85.84%**, demon-

strating the benefits of larger data volume. This result is comparable to the best performances achieved with combined HAM+BCN datasets, indicating that a single, unified dataset—despite containing some redundant samples—can effectively support high-performance classification.

Recognizing the strong dominance of the NV class within the dataset (with over 22,000 samples), we conducted a series of controlled experiments to mitigate class imbalance. In **MEL14**, all duplicate images from the NV class were removed, reducing its cardinality to 13,306. This significantly improved test accuracy to **86.37%**—the highest among all experiments—suggesting that reducing NV prevalence helps the model better learn minority classes. In **MEL15** and **MEL16**, 30% and 20% of NV images were randomly removed, leading to test accuracies of **84.71%** and **83.98%**, respectively. These results indicate a clear trade-off: while some reduction of imbalance is beneficial, excessive downsampling may decrease the available training signal and lead to lower performance. The confusion matrices for these experiments (Figure 3.2, bottom row) confirm that balancing the NV class improves classification across other categories such as MEL and BKL, which otherwise tend to be misclassified as NV due to dataset skew.

Experiment **MEL17** tested the SwV2L on the best-performing LD configuration (NV - downsampled), using DA_Train4 to accommodate the required 256×256 input resolution. The result was a test accuracy of **83.65%**, lower than the one obtained with SwL in MEL14. This outcome suggests that simply increasing model complexity (in terms of parameters and input size) does not guarantee better generalization, particularly when training data is already optimized for a specific architecture.

Lastly, experiments **MEL18**–**MEL20** evaluated the classification performance on an extended version of the LD, including both dermatoscopic and clinical images ("Unified"). Despite the increased variety in imaging modalities, the inclusion of additional non-dermatoscopic images had a significant impact on classification performance. **MEL18**, using the Unified dataset with duplicates, achieved only **74.39%** accuracy. **MEL19**, applying NV downsampling, improved performance to **84.58%**. **MEL20**, with further filtering, obtained the lowest result in this group: **70.68%**.

These experiments reveal the challenges of integrating heterogeneous imaging modalities without tailored preprocessing or domain adaptation techniques. While diversity may enhance generalization in some cases, it can also introduce noise and domain shift that hinder performance.

3.4.4 Classification Optimization and Robustness Evaluation

To assess and further improve the classification robustness of the trained models, we implemented two complementary evaluation strategies. These aimed to reduce variability due to training stochasticity and to simulate realistic conditions where lesion images may be captured under different orientations or affected by input noise.

For each of the three best-performing experiments—**MEL7** (HAM), **MEL8** (HAM+BCN), and **MEL14** (LD)—we repeated the training process five times using different random seeds, while keeping all other hyperparameters constant. This allowed us to compute the mean accuracy and standard deviation over multiple training cycles, providing insights into the consistency and reliability of each configuration.

Table 3.7 shows the statistical results. The MEL8 experiments yielded the highest average accuracy of **86.00%** with a low standard deviation of **0.39%**, indicating excellent stability across training sessions. MEL14 reached an average of **85.43%**, but with a slightly higher variability (**std = 0.66%**), while MEL7 showed the lowest average (**83.80%**) and higher inconsistency (**std = 0.57%**).

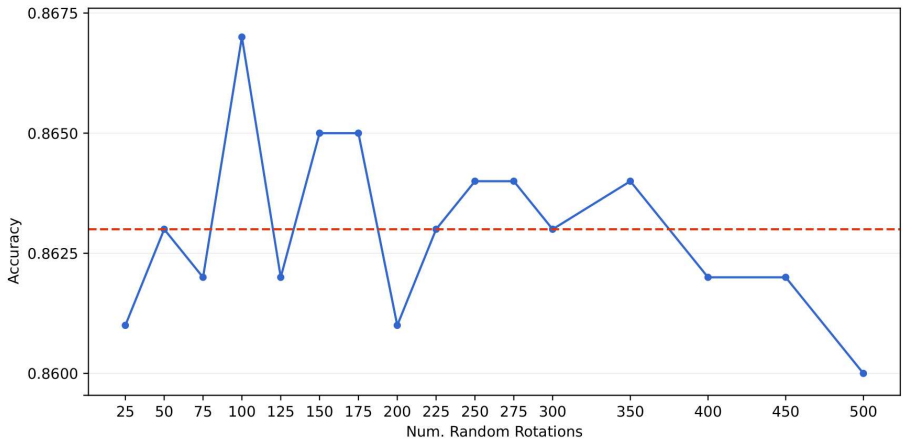


Figure 3.3: Classification accuracy as a function of the number of image rotations. The red line represents the test accuracy, while the blue line represents the test accuracy obtained after applying the rotations.

These results suggest that both dataset size and heterogeneity contribute positively to robustness, provided that the data distribution is properly balanced. In contrast, smaller or less diverse datasets are more sensitive to initialization and training randomness.

To further enhance model reliability, especially in real-world scenarios where lesion orientation may vary, we introduced a test-time augmentation strategy based on random image rotations. The goal was to simulate orientation variability and utilize an ensemble-like aggregation of predictions to enhance classification confidence. This analysis is not meant to "learn" rotational invariance (already encouraged during training via data augmentation), but to quantify prediction stability under controlled perturbations at inference time. The aggregation provides a robustness-oriented consensus estimate and reveals a saturation point beyond which excessive distortions can degrade accuracy.

The procedure consisted of rotating each test image n=100 times using random angles, obtaining a prediction for each rotation. The final class label was assigned by computing the mode (i.e., most frequent prediction) across all 100 runs. Figure 3.3 shows that applying more than 100 rotations often led to a plateau or even degradation in accuracy, likely due to excessive input distortion. Thus, we selected 100 as the optimal number of rotations.

The results, shown in Table 3.6, highlight a consistent improvement across most runs. For MEL7, performance improved in 4 out of 5 runs (maximum improvement: +1.13%). For MEL8, the improvements were modest but consistent (+0.06% to +0.13%). For MEL14, accuracy increased in 4 out of 5 runs, with a maximum gain of +0.53%.

Aggregated results (Table 3.7) show that rotation-augmented predictions lead to higher mean accuracy in all cases. MEL7 improved from **83.80%** to **84.17%**, MEL8 from **85.94%** to **86.00%**, MEL14 from **85.43%** to **85.85%**, with the lowest standard deviation (**0.25%**) across all tests. These results confirm that test-time augmentation via random rotations enhances classification reliability and can be particularly effective when combined with large and well-balanced datasets.

Finally, we investigated the influence of duplicate images in the test set. Following the removal of duplicate lesions from the external test set (identified via shared lesion IDs), the number of

Table 3.6: Additional experiments where the best models are towed five times, tested on the normal test dataset, and using rotations. EXP represents the Experiment's name, TA stands for Test Accuracy, TAwR is Test Accuracy with Rotations, and RV stands for Result Variation between TA and TAwR.

EXP	TA	TAwR	RV
MEL7_1	84.64	84.78	+0.14
MEL7_2	84.25	84.65	+0.4
MEL7_3	83.59	83.12	-0.47
MEL7_4	83.45	84.58	+1.13
MEL7_5	83.06	83.72	+0.66
MEL8_1	85.70	85.70	+0.0
MEL8_2	86.69	86.76	+0.07
MEL8_3	85.84	85.84	+0.0
MEL8_4	85.84	85.90	+0.06
MEL8_5	85.64	85.77	+0.13
MEL14_1	86.37	86.11	-0.26
MEL14_2	85.77	86.03	+0.26
MEL14_3	84.38	85.51	+0.13
MEL14_4	85.51	86.04	+0.53
MEL14_5	85.11	85.57	+0.46

unique test images dropped from 1511 to 1222. When re-evaluating the best model (MEL14) on this cleaned test set, accuracy increased from **86.37%** to **87.88%**, confirming the observation by Cassidy et al. [142] that duplicates can artificially inflate performance.

We also measured the inference times for training and validation across the top three experiments. The **MEL7** training inference time was around 160–165s, validation = 13–14s. In **MEL8**, the training inference was between 150–155s, while validation was around 40–45s. Lastly, in **MEL14**, the inference in training was 830–860s, and for validation was 185–215s. These results highlight the computational demands of larger datasets, but also support the notion that increased data quantity and diversity are correlated with higher model performance and robustness.

3.5 Discussion

The results presented in this study demonstrate the effectiveness of Transformer-based architectures—specifically, the Swin—for skin lesion classification. Despite the challenges posed by real-world dermatological datasets, such as class imbalance, duplicated samples, and visual artifacts, the Swin model consistently achieved high accuracy and robustness. Its hierarchical attention mechanism and window-based processing not only ensured good performance on relatively small datasets, such as HAM, but also scaled effectively to larger and more diverse datasets, like the proposed LD.

Furthermore, the model exhibited competitive inference efficiency. As reported in Section 3.4.4, training and validation times remained acceptable across different dataset scales, showing that Swin offers a viable compromise between complexity and speed. However, beyond computational considerations, our primary goal was to push classification performance to its limits. For this reason, we explored several optimization strategies—including rotation-based evaluation and repeated training

Table 3.7: Additional experiments. EXP represents the name of the group of Experiments, and MEAN and STD are the statistical values calculated from the previous Table 3.6.

EXP	MEAN	STD
MEL7(1-5)	83.80	0.57
MEL7(1-5)Rotations	84.17	0.64
MEL8(1-5)	85.94	0.38
MEL8(1-5)Rotations	86.00	0.39
MEL14(1-5)	85.43	0.66
MEL14(1-5)Rotations	85.85	0.25

sessions—to capture potential variabilities in performance and to ensure robustness in our findings.

To gain deeper insights into the internal representation learned by the model, we employed the t-distributed Stochastic Neighbor Embedding (t-SNE) technique [143]. t-SNE is a non-linear dimensionality reduction method widely used for visualizing high-dimensional data. In the context of deep learning, it enables us to project the learned feature embeddings into a two-dimensional space, allowing for a visual assessment of how well the model separates different classes.

Figure 3.4 shows the t-SNE projection of the feature embeddings extracted from the best-performing model (MEL14). Each point corresponds to an image in the test set, colored according to its ground-truth class. The plot reveals an important insight: while some classes—such as VASC and DF—form distinct and compact clusters, others, including MEL, NV, and BKL, display a significant degree of overlap. This confirms what was already partially visible from the confusion matrices: despite data augmentation, balanced training strategies, and the use of large datasets, certain lesion types remain visually similar and difficult to disentangle based on appearance alone.

This overlap suggests that, even though classification accuracy improves with more data and better augmentation strategies, the underlying feature space still suffers from ambiguity between clinically similar classes. In particular, misclassifications between MEL and NV remain critical, given the serious implications of failing to correctly identify malignant lesions.

These observations raise an important question: *Can we further improve the model's discriminative capability by enhancing the quality of the input features?* The hypothesis we propose is that incorporating segmentation as a preprocessing step could help mitigate this issue. By explicitly localizing and isolating the lesion regions, segmentation might reduce background noise and direct the model's attention toward the most informative areas, thereby reducing class confusion and improving representation separability.

In the following chapter, we will explore this idea in detail. We will present a new learning pipeline where segmentation precedes classification and evaluate how this modification impacts the overall performance and feature organization of the model. This transition builds directly on the insights gained from the current study, opening the door to more refined and semantically guided approaches for skin lesion analysis.

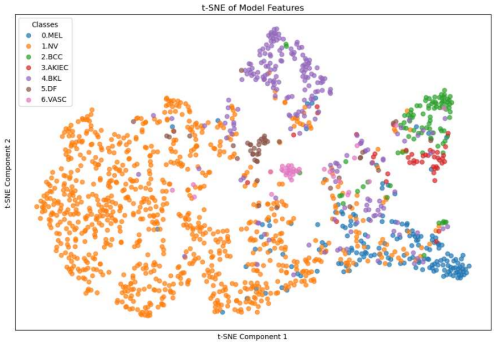


Figure 3.4: t-SNE plot of the learned feature embeddings from experiment MEL14. Some classes, such as VASC and DF, form distinct clusters, while others, including MEL, NV, and BKL, show substantial overlap, highlighting the need for better class disambiguation.

Chapter 4

Improving Classification in Skin Lesion Analysis through Segmentation[1]

4.1 Introduction and Motivation

Building on the findings presented in the previous chapter (2), we now turn our attention to the role of segmentation in improving skin lesion classification. As previously noted, the analysis of t-SNE projections (Figure 3.4 revealed that, despite the use of powerful Transformer-based models and large-scale, augmented datasets, certain classes—particularly MEL, NV, and BKL—continue to exhibit a high degree of overlap in the feature space. This latent ambiguity suggests that even the most optimized classification pipelines might still struggle to fully disentangle visually similar lesion types.

To address this limitation, we propose incorporating segmentation as an explicit preprocessing step prior to classification. This approach is motivated by both empirical evidence and prior literature discussed in 2.4: several works in the field have demonstrated that isolating the lesion area from the surrounding skin and artifacts can help reduce noise, emphasize the most discriminative regions, and provide spatial context that is otherwise diluted in full-image analysis. Unlike methods that rely purely on global appearance, segmentation introduces a form of structured attention by highlighting the object of interest—namely, the lesion—before it is passed to the classifier.

Importantly, segmentation is treated here as an independent module, operating upstream of the classification task. This modular strategy aligns with many recent pipelines in medical image analysis, where segmentation serves as a means to extract cleaner, semantically relevant image patches that can then be used to train or fine-tune a classification model. In this work, we retain the Swin as our core classification architecture to maintain continuity with prior experiments, while exploring different segmentation models—Swin, YOLOv8 [144], and DeepLabV3 [56]—for their

[1]This chapter is based on the published article: **M. Gallazzi, A. U. Rehman, S. Corchs, and I. Gallo**, "Improving Classification in Skin Lesion Analysis through Segmentation," *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, SciTePress, pp. 696–703, 2025. ISBN: 978-989-758-730-6, ISSN: 2184-4313. DOI: 10.5220/0013247900003905.

ability to localize skin lesions on the HAM dataset.

Our investigation thus aims to answer the following research question: *Does segmenting lesions prior to classification improve the model's ability to distinguish between visually similar skin lesion types?* By comparing classification performance on raw versus segmented input data, we assess whether this preprocessing step contributes to greater feature separability and overall accuracy.

The remainder of this chapter is organized as follows. We first describe the segmentation models and training procedures employed. We then present quantitative and qualitative results on both segmentation and classification tasks, including an analysis of class-level performance and confusion matrices. Finally, we offer a critical discussion of the observed outcomes and their implications for future research. Parts of this chapter are based on work published in [2].

4.2 Methodology

This study investigates whether incorporating the segmentation of the different lesions as a preprocessing step can improve the separability of skin lesion classes in a classification pipeline. To this end, we adopted a two-stage framework: (i) training deep segmentation models on dermatoscopic images to isolate lesion regions, and (ii) using the segmented outputs as inputs to a Swin classifier.

4.2.1 Dataset Selection

For the segmentation stage, we focused exclusively on the HAM dataset. Among the publicly available skin lesion datasets, HAM is the only one that provides both a substantial number of images and associated expert-annotated segmentation masks for a wide range of lesion types. This dual availability makes it uniquely suited for training and evaluating both segmentation and classification models in a consistent and controlled setting. Additionally, HAM is widely adopted in the literature for segmentation tasks, enabling comparative benchmarking with previous approaches. More details were discussed in 2.1.1.

4.2.2 Segmentation Architectures and Training Setup

We investigated three state-of-the-art architectures representing distinct deep learning paradigms:

- **DeepLabV3+ :** A convolutional model employing Atrous Spatial Pyramid Pooling (ASPP) and residual blocks. The encoder extracts multi-scale semantic features while the decoder restores resolution through bilinear upsampling. Each residual block is computed as:

$$F(X) = \sigma(WX + b) + X \tag{4.1}$$

  where σ is a non-linear activation function, W and b are trainable parameters, and X is the input feature map. This structure supports stable training and precise boundary localization.

- **Swin Transformer:** Based on the original Swin architecture, this model uses hierarchical transformer blocks in an encoder-decoder format for segmentation. Attention is computed within local windows to efficiently capture spatial relationships. The attention mechanism is the same, defined in the Formula 3.1. Swin's hierarchical structure uses patch merging and shifting windows to enable multi-scale representation with controlled complexity:

$$O((H \times W) \times d^2) \tag{4.2}$$

where H and W denote the height and width of the input feature map, and d the window size. The decoder reconstructs full-resolution segmentation masks through patch expansion layers and skip connections, effectively integrating coarse-to-fine contextual features.

- **YOLOv8-Seg:** A real-time model capable of simultaneous detection and segmentation. Its architecture comprises three stages. The first is a CSP-based backbone for feature extraction, followed by a PANet-based neck to enhance multiscale feature fusion, and ended with an anchor-free detection and segmentation head, predicting bounding boxes, classes, and binary masks. The segmentation mask is generated alongside the bounding box, and training is optimized via Generalized Intersection over Union (GIoU) loss:

$$L_{\text{GIoU}} = 1 - \text{IoU} + \frac{|C - (B_p \cup B_g)|}{|C|} \tag{4.3}$$

where B_p and B_g are the predicted and ground truth bounding boxes, and C is the smallest enclosing box.

Each model was trained using binary lesion masks and standard augmentations (flip, rotation, scaling). Segmentation performance was evaluated using:

Intersection over Union (IoU):

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{4.4}$$

4.3 Experiments and Results

This section presents the core experimental findings of our investigation into the role of segmentation in improving skin lesion classification using the HAM dataset. We evaluated the three deep learning models described above—YOLOv8, DeepLabV3, and Swin—on both segmentation and classification tasks. Our goal was to explore whether using segmented lesion images as inputs could enhance classification accuracy, particularly for visually similar skin lesion types.

4.3.1 Segmentation Results

Each model was trained on the HAM segmentation training set and evaluated on the HAM segmentation challenge test set using the IoU metric (Formula 4.4. Our Swin model achieved the best performance with an IoU of 82.75%, surpassing the 80.2% benchmark set in the original challenge. DeepLabV3 followed with 81.66%, while YOLOv8 reached 77.0%. These results, summarized in Table 4.1 confirm the strong segmentation capabilities of both transformer-based and residual architectures, especially in delineating complex lesion boundaries.

4.3.2 Classification with Segmented Inputs

Using the masks generated by each segmentation model, we produced three new datasets by cropping the lesion regions in HAM. These cropped images were used to train only Swin for classification,

Table 4.1: Segmentation Results on HAM Test Set.

Experiment Type	Model	IoU (%)
Segmentation(Challenge)	MaskRcnn2	80.2
Segmentation	DeepLabV3	81.66
Segmentation	Swin	82.75
Segmentation	YOLO	77.0

with performance compared against Swin trained on the original HAM dataset. The results were evaluated using metrics including Accuracy, Precision, Recall, and F1-score (Formulas from 3.2 to 3.5).

The highest classification performance, as shown in Table 4.2, was achieved using the original, non-segmented HAM dataset: 84.64% Accuracy and 84.57% F1-score. Surprisingly, classification performance slightly declined when using segmented images, regardless of the model that produced them. For example, YOLO-based cropped images yielded an Accuracy of 84.01% and an F1-score of 83.53%, while DeepLabV3 and Swin crops produced similarly minor declines.

These findings suggest that segmentation as a preprocessing step—while enhancing spatial focus—may inadvertently discard relevant contextual or textural information critical for classification. This is particularly impactful for lesions that are visually ambiguous or have subtle differences.

Table 4.2: Classification results on the HAM test set. TA, TP, TR, and TF1 indicate Test Accuracy, Test Precision, Test Recall, and Test F1 Score (%). The "Dataset" column shows the HAM dataset segmented by different models.

Experiment	Dataset	Model	TA (%)	TP (%)	TR (%)	TF1 (%)
Classification	HAM	Swin	84.64	84.77	84.64	84.57
Classification	HAM_segmented_YOLO	Swin	84.01	84.09	84.12	83.53
Classification	HAM_segmented_DeepLabV3	Swin	84.12	84.38	84.12	83.75
Classification	HAM_segmented_Swin	Swin	84.13	84.41	84.12	83.58

4.4 Discussion

The experimental results obtained in this study provide valuable insights into the role of segmentation in skin lesion classification. While segmentation models—particularly Swin and DeepLabV3—demonstrated good performance on the HAM dataset, with Swin achieving the highest IoU (82.75%), integrating segmentation outputs as a preprocessing step for classification did not lead to improved accuracy. On the contrary, classification performance slightly declined when training the Swin on segmented versions of the dataset produced by YOLO, DeepLabV3, or Swin itself.

This outcome highlights a critical limitation: treating segmentation as an isolated preprocessing task may inadvertently remove contextual or background information that is implicitly important for accurate classification. Cropping around the lesion might lead to a loss of surrounding visual cues or structural context, which can be especially detrimental in cases where lesion types exhibit subtle visual similarities.

Nonetheless, these findings are far from discouraging. Rather, they generate two important observations that guide the next stage of our research. First, the Swin model exhibited excellent per-

formance in the segmentation task despite being originally designed for classification, confirming its robustness and adaptability across different medical imaging challenges—even in data-constrained settings. This reinforces the Swin as a strong backbone for unified, multi-task medical image analysis pipelines.

Second, and more crucially, the experiment supports a paradigm shift in the role of segmentation. Instead of being treated as a static preprocessing step, segmentation can serve a more integrated function: guiding the model's internal feature representations. By training the model first on segmentation and subsequently fine-tuning it for classification, the learning process can be enriched with spatially aware features that improve class separability—especially among visually overlapping lesions, as observed in the t-SNE embedding analysis discussed in the previous chapter.

Although the segmentation-based classification setup did not surpass the current benchmarks in the literature, it provided crucial insight into the interdependence between visual localization and semantic classification. This understanding lays the groundwork for the next chapter, in which we investigate a sequential training strategy that integrates segmentation and classification into a coherent and progressive learning pipeline aimed at improving both performance and generalization.

Several recent studies have also highlighted the benefits of such a sequential approach. Works by Paulsen and Casey [145], Chan et al. [146], and Wang et al. [147] have shown that sequential training can enhance class separability by leveraging task-specific pretraining. The underlying hypothesis is that fine-grained spatial features learned during segmentation can effectively guide attention mechanisms during downstream classification, leading to more robust and discriminative representations. These findings provide further theoretical and empirical support for the framework that we discuss in the following chapter.

Chapter 5

A Sequential Segmentation and Classification Learning Approach for Skin Lesion Images[1]

5.1 Introduction and Motivation

In the previous chapter 4, we investigated whether using segmentation as a standalone preprocessing step—by cropping and isolating lesion areas—could enhance classification performance. Although segmentation itself proved effective, the results showed that this strategy did not lead to a significant improvement in classification accuracy. In some cases, it even resulted in the loss of valuable contextual information, suggesting that a purely segmented representation may not always capture the global cues necessary for accurate diagnosis. These findings naturally raised a new question: *can segmentation still play a meaningful role in improving classification, if it is integrated differently within the learning process?*

This chapter addresses that question by rethinking the relationship between segmentation and classification, not as two isolated stages, but as interdependent tasks within a broader learning paradigm. Instead of using segmentation merely to crop or mask the lesion, we explore whether learning segmentation before or after classification can influence the quality and transferability of the learned representations. This perspective extends beyond preprocessing and considers segmentation as a task that can benefit from or be guided by the inductive biases of another.

The motivation for this investigation comes from the growing interest in transfer learning between related medical imaging tasks [148]. While several studies have shown that joint or multitask training can exploit complementary information [149, 150, 151], the effect of task ordering—which task should be learned first—has received limited attention, especially in Transformer-based architectures. Understanding this aspect is crucial, since the representations learned during segmentation emphasize spatial and boundary features, whereas classification encourages abstraction and semantic discrimination. Determining how these forms of knowledge interact may reveal whether one task

[1]This chapter is based on the published article: **M. Gallazzi, I. Gallo, and S. Corchs**, A Sequential Segmentation and Classification Learning Approach for Skin Lesion Images. Appl. Sci. 2025, 15, 12614. https://doi.org/10.3390/app152312614

can effectively bootstrap the other.

To systematically study this, we introduce the **Sequential Swin Transformer (SST)** framework, a modular architecture that supports both segmentation and classification via task-specific heads and a shared Swin backbone [141]. The SST allows us to train the same backbone sequentially in two opposite directions: from segmentation to classification (SST_SC) and from classification to segmentation (SST_CS). By comparing these two learning sequences, we can quantify how the order of tasks affects the transfer of useful representations and the model's generalization ability across different domains.

In addition to quantitative evaluation, we also analyze the interpretability of the learned representations using Gradient-weighted Class Activation Mapping (Grad-CAM) [152]. This qualitative analysis helps visualize whether the model trained in different orders tends to focus more accurately on lesion-relevant regions rather than background information.

Finally, to ensure a consistent and fair evaluation, all experiments are conducted on datasets that include both segmentation masks and classification labels, such as HAM [27, 28] for dermatological images and Kvasir [153] for gastrointestinal data. These datasets allow us to assess both within-domain and cross-domain generalization, revealing how the sequential learning strategy behaves across different medical contexts.

The main contributions of this chapter are as follows:

- We propose the SST framework, which enables modular and task-sequential learning over a shared Transformer backbone.
- We conduct a systematic comparison between two sequential configurations, SST_SC and SST_CS, to evaluate how task ordering influences performance and feature transfer.
- We validate the generalizability of the approach through experiments on multiple datasets and qualitative analyses of model interpretability.

5.2 Sequential Learning Framework

This section introduces the Sequential Swin Transformer (SST) framework, which is designed to explore the impact of task ordering—specifically, whether segmentation is followed by classification or vice versa—in skin lesion analysis. The architecture builds upon the Swin backbone and adopts a modular approach to support both segmentation and classification through task-specific heads operating on shared representations.

5.2.1 Model Architecture

The SST is designed to isolate the effect of task ordering in medical image analysis. It reuses a single Swin Transformer backbone for both segmentation and classification, and couples it with lightweight, interchangeable heads. Swapping heads does not modify the backbone parameters; therefore, any performance difference between configurations can be attributed to the training order and the manner in which representations are transferred, rather than to architectural changes.

The shared backbone is based on the `swin_large_patch4_window7_224` model [1]. Given an input RGB image $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$, with $C = 3$, the network first partitions the image into non-overlapping 4×4 patches, embeds them linearly, and processes them through four hierarchical

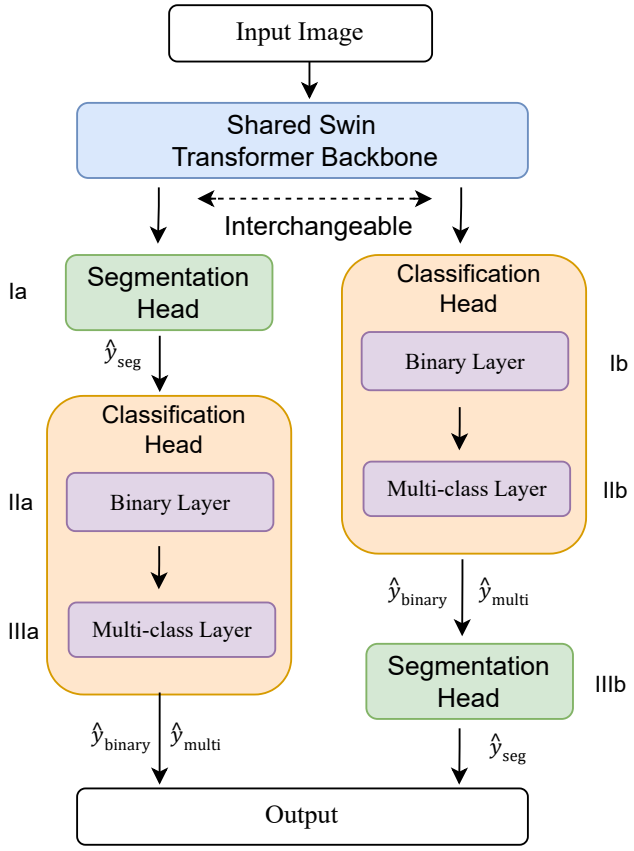


Figure 5.1: Visual overview of the two sequential training strategies investigated in this study. **Left (SST_SC)**: the model is first trained on segmentation (Ia), then transferred to binary classification (IIa), and finally fine-tuned for multi-class classification (IIIa). **Right (SST_CS)**: training starts with binary classification (Ib), followed by multi-class classification (IIb), and ends with segmentation (IIIb). In both settings, the SWIN backbone is shared across tasks, and task-specific heads are modularly swapped to enable sequential transfer learning.

stages with shifted-window self-attention. Each stage halves the spatial resolution and increases the channel dimension, so that for inputs of size 224×224, the final feature map has a spatial size of $H' = W' = 7$ and a channel depth of $D = 1536$. By removing the default classification head, the backbone acts as a generic feature extractor:

$$\mathbf{F} = \mathcal{B}(\mathbf{x}), \qquad \mathbf{F} \in \mathbb{R}^{D \times H' \times W'}, \tag{5.1}$$

where $\mathcal{B}(\cdot)$ denotes the Swin backbone, including patch embedding, shifted-window attention, and MLP blocks, and $\mathbf{F}$ provides a shared representation for all downstream tasks.

Segmentation Head. The segmentation head reconstructs spatial detail by upsampling $\mathbf{F}$ back to the input resolution through a sequence of transposed convolutions. Channels are progressively reduced according to the following pattern:

$$1536 \to 512 \to 256 \to 128 \to 64 \to 32 \to 1.$$

Each transposed convolution is followed by Batch Normalization and ReLU activation. A final 1×1 convolution produces a single-channel logit map

$$\mathbf{M} \in \mathbb{R}^{1 \times H'' \times W''}, \tag{5.2}$$

which is then resized via bilinear interpolation to match the original image resolution (H, W). For each transposed convolution, the output size O along one spatial dimension is given by

$$O = (I - 1) \cdot S - 2P + K + \mathrm{OP}, \tag{5.3}$$

where I is the input size, S the stride (set to 2), P the padding, K the kernel size, and OP the output padding. The final sigmoid activation is applied at inference time to obtain a probabilistic lesion map, which highlights the lesion region at pixel level.

Classification Head. The classification head operates on the same shared representation $\mathbf{F}$. After global average pooling over the spatial dimensions, we obtain a descriptor vector

$$\mathbf{f} = \frac{1}{H'W'} \sum_{i=1}^{H'} \sum_{j=1}^{W'} \mathbf{F}_{:,i,j} \in \mathbb{R}^D, \tag{5.4}$$

which compactly summarizes the multi-scale features produced by the backbone. The vector $\mathbf{f}$ is first passed through a dropout layer and then fed to two fully connected layers arranged in cascade: a multiclass layer and a binary layer.

The multiclass head produces logits and probabilities over C_{multi} diagnostic categories (e.g., MEL, NV, BCC, etc.):

$$\begin{aligned} \mathbf{z}_{\text{multi}} &= \mathbf{W}_{\text{multi}}\mathbf{f} + \mathbf{b}_{\text{multi}}, \qquad \mathbf{W}_{\text{multi}} \in \mathbb{R}^{C_{\text{multi}} \times D}, \\ \mathbf{y}_{\text{multi}} &= \mathrm{Softmax}(\mathbf{z}_{\text{multi}}) \in [0,1]^{C_{\text{multi}}}. \end{aligned} \tag{5.5}$$

The binary head is cascaded on top of the multiclass logits:

$$\begin{aligned} \mathbf{z}_{\text{bin}} &= \mathbf{W}_{\text{bin}}\mathbf{z}_{\text{multi}} + \mathbf{b}_{\text{bin}}, \qquad \mathbf{W}_{\text{bin}} \in \mathbb{R}^{C_{\text{bin}} \times C_{\text{multi}}}, \\ \mathbf{y}_{\text{bin}} &= \mathrm{Softmax}(\mathbf{z}_{\text{bin}}) \in [0,1]^{C_{\text{bin}}}, \qquad C_{\text{bin}} = 2, \end{aligned} \tag{5.6}$$

where $C_{\text{bin}} = 2$ corresponds to benign vs. malignant classification. This cascaded design mirrors the actual implementation and supports a hierarchical reasoning scheme in which the multiclass logits guide the binary prediction.

Role and Motivation of Binary Classification. The inclusion of a binary classification head has both clinical and methodological motivations. Clinically, the benign–malignant decision is often the first and most critical triage step, directly influencing whether a lesion requires immediate intervention. From a learning perspective, exposing the backbone to a coarse-grained binary task before (or together with) the fine-grained multiclass decision helps the model capture broad visual differences related to malignancy (e.g., asymmetry, border irregularity, color variation). These coarse representations are then refined during multiclass training. By placing the binary head on top of the multiclass logits, the SST enforces a dependency between the two decision levels and encourages consistency between fine-grained and coarse predictions.

Task Modularity. The SST architecture is explicitly designed for sequential multi-task learning through modular head replacement. The Swin backbone remains shared across all stages, while the task-specific heads — segmentation, binary classification, and multiclass classification — can be attached or detached without modifying the core representation. This modularity enables flexible execution of the learning pipeline in both directions:

- **Segmentation → Binary → Multiclass (SST_SC):** the model first learns to localize the lesion at pixel level, then exploits these spatial priors for binary diagnosis, and finally refines the representation for multiclass prediction.
- **Binary → Multiclass → Segmentation (SST_CS):** the model starts from coarse diagnostic cues, refines them for fine-grained classification, and eventually learns to delineate lesion contours on top of the resulting representation.

This design isolates the role of task ordering in feature transfer, as summarized in Figure 5.1.

5.2.2 Training Pipeline

The training procedure of the SST framework is structured into sequential phases, dictated by the chosen task order. Each phase involves training a specific task head (segmentation, binary classification, or multiclass classification) while reusing and progressively refining the shared backbone weights. This modular training design enables effective knowledge transfer between tasks.

Two distinct training configurations are considered:

SST_SC (Segmentation → Binary → Multiclass):

1. **Segmentation Phase:** The model is trained to generate binary masks of the lesion regions using the Binary Cross-Entropy (BCE) loss function. The goal is to learn spatially-aware features and precise boundary delineation. The best checkpoint is selected based on the highest validation IoU.
2. **Binary Classification Phase:** The segmentation head is replaced by the classification head. The model is fine-tuned to distinguish between benign and malignant lesions using BCE loss. This step leverages previously learned spatial features for high-level diagnosis. The best model is selected by validation accuracy.
3. **Multiclass Classification Phase:** The head is extended to perform multi-class lesion classification (e.g., MEL, NV, BCC, etc.) using categorical Cross-Entropy loss. All model parameters, including the backbone, are fine-tuned during this phase to capture fine-grained diagnostic categories.

SST_CS (Binary → Multiclass → Segmentation):

1. **Binary Classification Phase:** The model is first trained on the binary task (benign vs. malignant) using BCE loss. The objective is to learn coarse diagnostic patterns. The best checkpoint is selected by validation accuracy.
2. **Multiclass Classification Phase:** The classification head is extended to output seven diagnostic categories. Training continues using categorical Cross-Entropy loss. The model refines its representation to capture more detailed class-specific features.
3. **Segmentation Phase:** The classification head is replaced by the segmentation head. The model is fine-tuned to generate pixel-wise lesion masks using the same BCE loss and training settings adopted in the SST_SC configuration.

During the multi-class phase, both binary and multiclass outputs are computed jointly. The total classification loss is expressed as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{multi}} \cdot \mathcal{L}_{\text{multi}} + \lambda_{\text{bin}} \cdot \mathcal{L}_{\text{bin}} \tag{5.7}$$

where $\mathcal{L}_{\text{multi}}$ is the multi-class Cross-Entropy loss, $\mathcal{L}_{\text{bin}}$ is the binary Cross-Entropy loss, and $\lambda_{\text{multi}}, \lambda_{\text{bin}} \in \mathbb{R}^{+}$are the scalar weights balancing the two tasks.

The weights are defined as hyperparameters and remain fixed during training. Only the loss corresponding to the currently active task is used for backpropagation in each phase, while the other is computed but not optimized.

Model Selection Criteria. At the end of each training phase, the best model is selected based on task-specific validation metrics:

- **Accuracy** for classification tasks as defined in 3.2;
- **IoU** for segmentation already described in 4.4;
- **Dice Score** (Dice defined as:

$$\text{Dice} = \frac{2 \cdot \text{Intersection}}{\text{Prediction Size} + \text{Ground Truth Size}} \tag{5.8}$$

These checkpoints are then reused to initialize the following task in the sequence, ensuring optimal transfer of learned representations and maximizing the benefit of task ordering.

5.2.3 Interpretability Analysis with Grad-CAM

To further investigate how task ordering affects the internal representations learned by SST, we analyze the model using Gradient-weighted Class Activation Mapping (Grad-CAM) [152]. Grad-CAM produces post-hoc spatial heatmaps that highlight the image regions contributing most to a specific class decision, thus providing a qualitative assessment of whether the model focuses on lesion-relevant areas rather than on background structures.

In this study, Grad-CAM is applied to both sequential configurations (SST_SC and SST_CS), with a focus on the **multiclass classification head**, which is the most informative for assessing fine-grained diagnostic behaviour. Given an input image and its target class t, we extract the feature

maps $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ from the last backbone block and compute the gradients of the corresponding class logit y_t with respect to these features. Channel-wise importance weights are obtained by global average pooling of the gradients:

$$\alpha_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\frac{\partial y_t}{\partial \mathbf{F}_{c,i,j}}, \tag{5.9}$$

where α_c denotes the contribution of channel c. The Grad-CAM heatmap is then given by

$$\mathbf{H}_t = \mathrm{ReLU}\left(\sum_{c=1}^{C}\alpha_c \mathbf{F}_c\right), \tag{5.10}$$

where $\mathbf{F}_c$ is the c-th feature channel and $\mathrm{ReLU}(\cdot)$ discards negative contributions, preserving only features that positively support class t. The map $\mathbf{H}_t$ is normalized to $[0,1]$, upsampled to the input resolution via bilinear interpolation, and superimposed on the original image to visualize the attention intensity.

To improve stability, Grad-CAM is computed using a memory-efficient implementation adapted to Transformer backbones, with multi-crop smoothing (multiple random crops per image) to reduce variability in the resulting heatmaps. Qualitative analyses are primarily performed on the HAM dataset, comparing correctly and incorrectly classified samples under both SST_SC and SST_CS.

These visualizations complement the quantitative results reported in Section **??**. In particular, they allow us to assess whether the segmentation-first configuration encourages more lesion-centric attention and sharper focus along lesion boundaries, while the classification-first configuration tends to rely more on broader contextual cues. Representative examples are reported in the experimental section, illustrating how activation regions align with annotated lesion areas and how this alignment changes with the training order.

5.2.4 Implementation Details

All experiments were implemented in PyTorch 2.1.0 (CUDA 11.8), using the same backbone, optimizer, and learning schedule across all sequential configurations to ensure a fair comparison. Input images were resized to a fixed resolution of 224×224 and normalized using the ImageNet mean and standard deviation, in line with the Swin pretraining protocol.

Data Partitioning. For HAM, data were split into training, validation, and test sets with subject-level independence, ensuring that all images from the same patient (including multi-view acquisitions of the same lesion) appear in a single subset only. Segmentation masks were matched to their corresponding image IDs across all splits. Patient-level metadata were used to enforce this grouping. No explicit near-duplicate removal was performed, since multi-view samples reflect realistic clinical conditions and were intentionally retained to enrich intra-class variability. The external HAMt test set is fully independent and was never used for training or validation.

Data Augmentation. To promote generalization and reduce overfitting, a dynamic data augmentation pipeline was applied on-the-fly during training. For each training image, the following transformations were sampled stochastically:

- **Random resize and crop:** the shortest side was resized in the range $[224, 280]$ and a random or centered crop of size 224×224 was extracted;

- **Horizontal flip:** applied with probability 0.5 to simulate left–right invariance;
- **Affine transformations:** random translations up to 10% of the image size, applied with probability 0.5;
- **Random rotation:** angles uniformly sampled from $[-180°, 180°]$, applied with probability 0.99;
- **Normalization:** pixel intensities normalized using ImageNet statistics.

This dynamic augmentation ensures that the model rarely sees the same image twice in the exact same form, increasing data diversity, partially mitigating class imbalance, and improving robustness to variations in scale, orientation, and illumination.

Validation images were resized to 224×280 and center-cropped to 224×224, while test images were directly resized to 224×224 without additional augmentations.

Optimization and Scheduling. Unless otherwise specified, training used the Adam optimizer with an initial learning rate of $1 \cdot 10^{-4}$ for all tasks. A `LambdaLR` scheduler was employed to decay the learning rate exponentially with the epoch index e. The total number of epochs and batch sizes were kept consistent across SST_SC and SST_CS; in particular, batch size was set to 144 for segmentation-first runs and 128 for classification-first runs, reflecting GPU memory constraints. At the end of each phase, the best-performing checkpoint (according to the criteria described above) was saved and used to initialize the next task in the sequence.

5.3 Experimental Setup

We conduct our evaluation using the already described HAM dataset [27]. For segmentation, we utilize the 2,594 samples provided with pixel-wise masks in the ISIC 2018 Challenge [154], using a fixed set of 1,000 images as test data. To assess generalization, we also include the HAMt dataset [155], a 2023 test set containing 1,512 previously unreleased images.

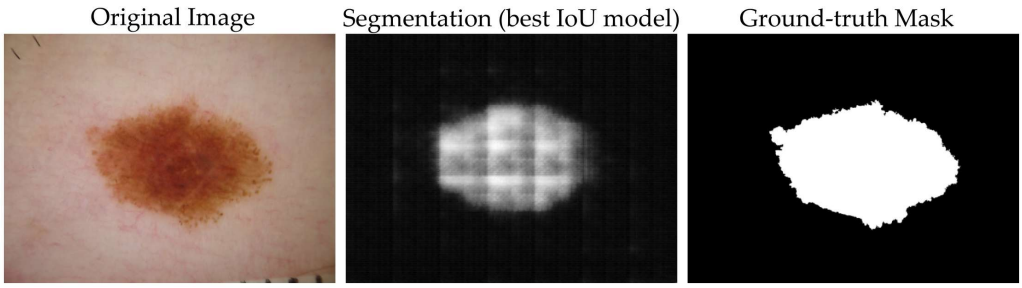
We have already described the two sequential configurations, **SST_SC** and **SST_CS**, in Section 5.2.2, each of which was tested on both internal (HAM) and external (HAMt) settings. Experiments are repeated over five independent runs, and we report the mean and standard deviation for all metrics.

Classification is evaluated using accuracy, precision, recall, F1-score (Formulas from 3.2 to 3.5, and AUROC/ROC (macro and weighted). Segmentation is evaluated using IoU (4.4) and the Dice (5.8).

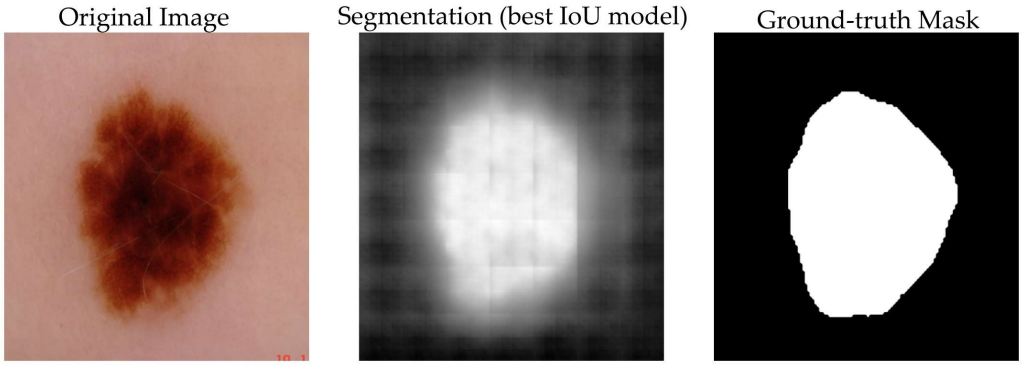
5.3.1 Results and Analysis

We report four experimental setups: **Our_A/Our_B** are trained and evaluated on HAM (internal); **Our_C/Our_D** are trained on HAM and evaluated on HAMt (external). Within each pair, the two SST orders are compared head-to-head.

Table 5.1 summarizes the performance across four experimental setups. Overall, segmentation accuracy (IoU/Dice) remains stable across task orders, with minor differences between SST_SC and SST_CS. Figure 5.2 shows visual comparisons of predicted masks under both pipelines, confirming consistent lesion localization.



(a) SST_SC (Our_D)



(b) SST_CS (Our_C)

Figure 5.2: Visual comparison of segmentation results for a representative lesion from the test set, evaluated under two sequential learning configurations. Each row presents, from left to right, the original input image, the predicted segmentation map from the best-performing model, and the ground-truth mask. (a) SST_SC (Our_D): segmentation followed by classification. (b) SST_CS (Our_C): classification followed by segmentation.

Table 5.1: Benchmark results of the SST model on the HAM dataset. TA denotes test accuracy, with the Dataset column distinguishing between HAM (full dataset) and HAMt (external test set). Jaccard and Dice scores evaluate segmentation performance, while TAb and TAm report binary and multiclass classification accuracy, respectively. TPm, TRm, and TF1m correspond to multiclass precision, recall, and F1-score. All values are averaged over five runs.

Author	Dataset	Jaccard (%)	Dice (%)	TAb (%)	TAm (%)	TPm (%)	TRm (%)	TF1m (%)
Our_A	HAM	86.13 ± 0.27	89.61 ± 0.15	93.23 ± 0.59	92.16 ± 0.69	89.75 ± 0.98	84.90 ± 0.89	86.44 ± 1.13
Our_B	HAM	86.03 ± 0.18	89.48 ± 0.22	93.76 ± 0.29	91.94 ± 0.49	89.54 ± 1.40	86.45 ± 1.14	87.68 ± 1.24
Our_C	HAM+Ht	86.58 ± 0.19	89.50 ± 0.25	90.80 ± 0.55	86.63 ± 0.39	82.37 ± 0.51	79.42 ± 0.30	79.74 ± 1.80
Our_D	HAM+Ht	86.26 ± 0.16	89.48 ± 0.23	91.07 ± 0.52	86.87 ± 0.45	84.94 ± 0.69	78.47 ± 1.48	80.57 ± 0.10

Classification, however, is more sensitive to task ordering. On HAM, both models perform similarly, but when evaluated on HAMt, SST_SC (Our_D) achieves higher multiclass accuracy (86.87%) and macro F1-score (80.57) compared to SST_CS (Our_C, 86.63% / 79.74). A Wilcoxon signed-rank test confirms the statistical significance of this improvement (p=0.03125). Confusion matrices and ROC curves (Figures 5.3–5.4) further illustrate the superior discrimination achieved by segmentation-first training, especially for challenging classes like MEL and AKIEC.

5.3.1.1 Qualitative Analysis with Grad-CAM

To qualitatively assess how task ordering influences model attention, we employed Gradient-weighted Class Activation Mapping (Grad-CAM) [152]. For both sequential configurations—**SST_SC** and **SST_CS**—Grad-CAM was applied to the multiclass classification head by backpropagating the gradients of the predicted class score to the final attention/convolutional blocks, producing heatmaps that highlight the most influential spatial regions for each decision.

Figure 5.5 illustrates examples from two representative test cases (BCC and MEL). Each row shows, from left to right, the input image with ground-truth mask overlay, the predicted segmentation map, and the Grad-CAM maps for the two sequential configurations. The **SST_SC** model (segmentation-first) consistently centers its attention within the lesion boundaries, whereas **SST_CS** exhibits a broader and less discriminative activation spread extending into peri-lesional skin. This effect is especially pronounced for melanoma, where the segmentation-first configuration tightly focuses on the lesion body, while the classification-first model attends to surrounding non-diagnostic areas.

Beyond static analysis, we further examined how Grad-CAM activations evolve across training epochs. Figure 5.6 shows the temporal evolution of attention for two lesions. In **SST_SC**, activations progressively contract toward the diagnostically relevant lesion core, demonstrating how spatial priors learned during segmentation guide the classifier throughout training. By the best-performing epoch, saliency becomes compact, boundary-aligned, and morphology-aware. This progression provides direct visual evidence that segmentation-first training stabilizes and constrains the classifier's attention.

Two key effects emerge: (i) focus tightening, where activations converge onto the lesion area, and (ii) background spillover reduction, with fewer non-lesion activations compared to **SST_CS**. These observations align with the improved external performance of **SST_SC** and provide a mechanistic explanation for its greater robustness under domain shift.

Overall, these qualitative findings complement the quantitative results in Table 5.1 and the

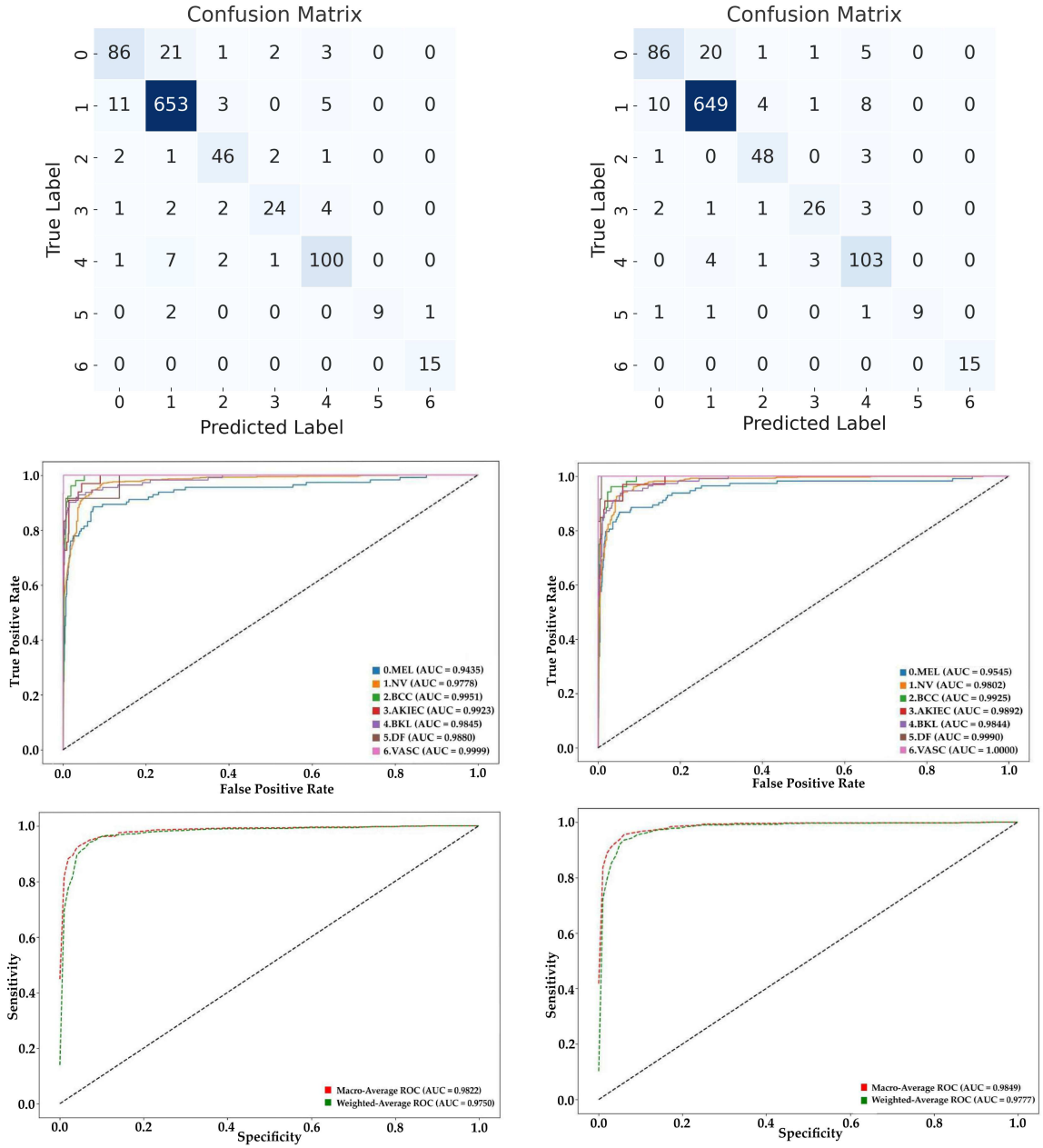


Figure 5.3: Confusion matrices, per-class ROC curves (with legend), and macro/weighted ROC curves (from top to bottom) for the models Our_A (left column) and Our_B (right column), as reported in Table 5.1. Each row corresponds to the respective experimental setup, based on the best-performing checkpoint on the validation set. In the confusion matrices (top row), true labels are shown on the vertical axis and predicted labels on the horizontal axis; classes are indexed from 0 to 6, corresponding respectively to MEL, NV, BCC, AKIEC, BKL, DF, and VASC. The ROC curves in the middle row display the True Positive Rate (TPR, sensitivity) on the y-axis versus the False Positive Rate (FPR) on the x-axis for each class. In the bottom row, macro and weighted average ROC curves are reported with the y-axis representing sensitivity and the x-axis representing specificity.

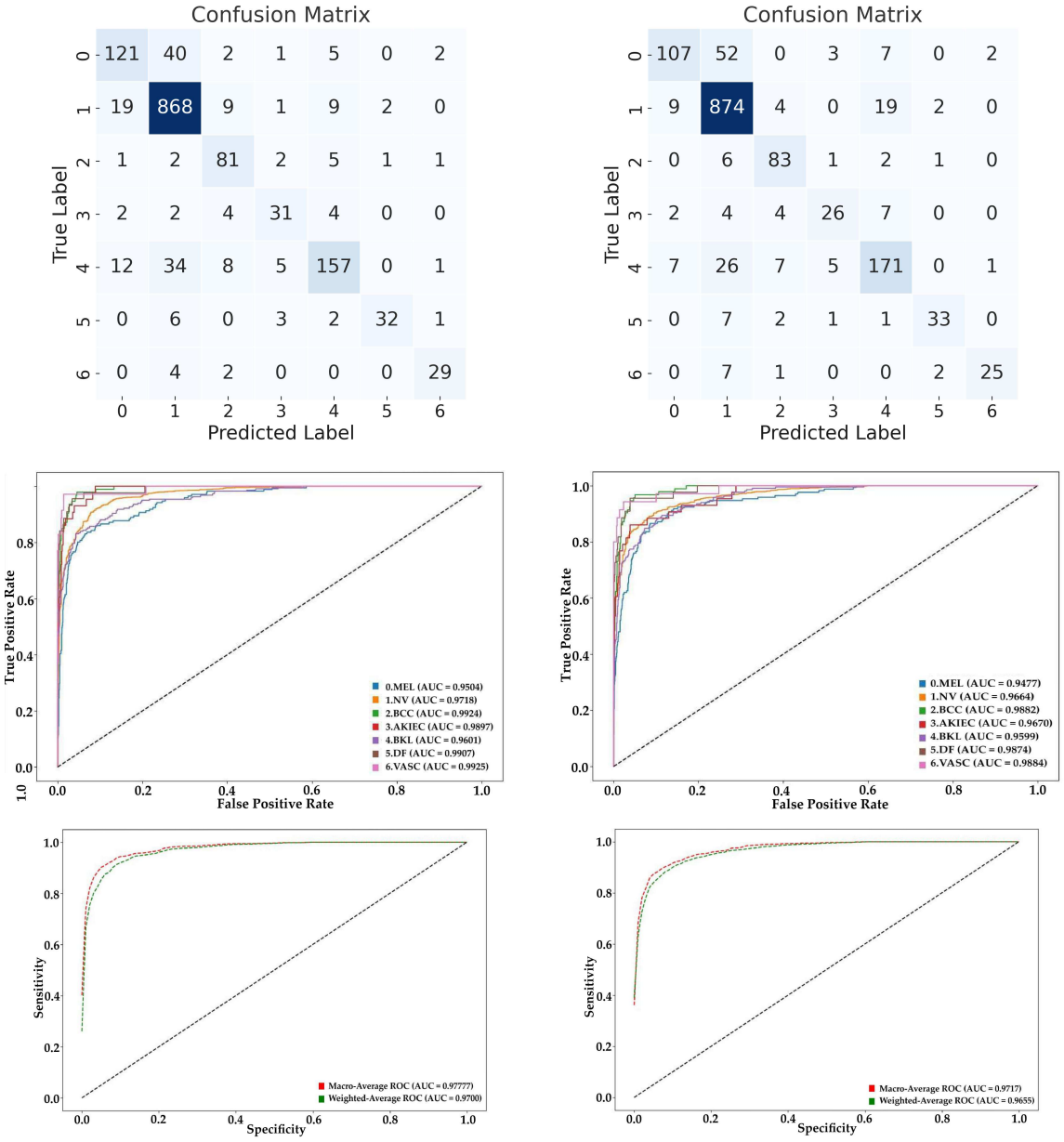


Figure 5.4: Confusion matrices, per-class ROC curves (with legend), and macro/weighted ROC curves (from top to bottom) for the models Our_C (left column) and Our_D (right column), as reported in Table 5.1. Each row corresponds to the respective experimental setup, based on the best-performing checkpoint on the validation set. In the confusion matrices (top row), true labels are shown on the vertical axis and predicted labels on the horizontal axis; classes are indexed from 0 to 6, corresponding to MEL, NV, BCC, AKIEC, BKL, DF, and VASC, respectively. The ROC curves in the middle row display the True Positive Rate (TPR, sensitivity) on the y-axis versus the False Positive Rate (FPR) on the x-axis for each class. In the bottom row, macro and weighted average ROC curves are reported with the y-axis representing sensitivity and the x-axis representing specificity.

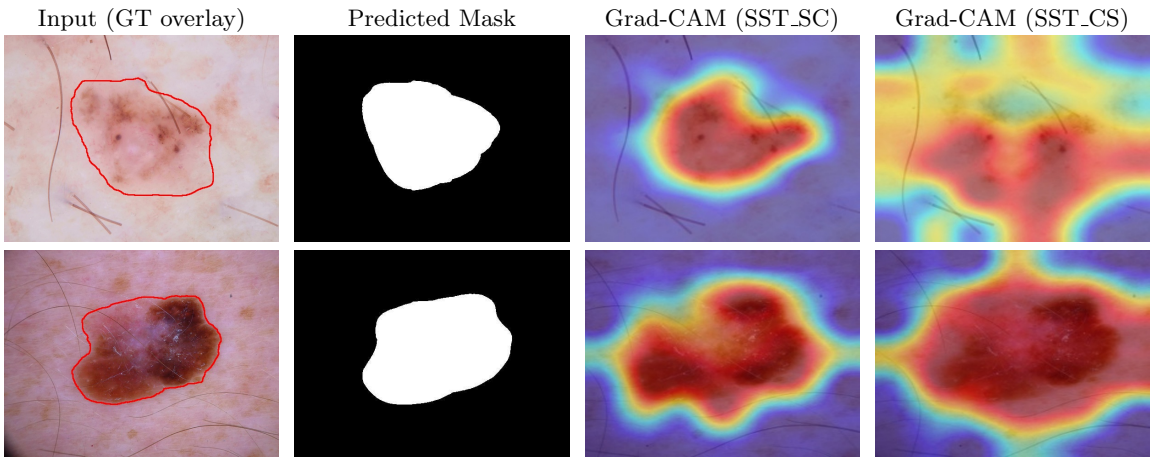


Figure 5.5: Grad-CAM visualizations on two representative HAM cases comparing **SST_SC** and **SST_CS**.

latent-space analysis in Figs. 5.7–5.8. The stronger alignment between Grad-CAM activations and lesion boundaries observed in **SST_SC** supports the conclusion that segmentation-first training promotes more lesion-centric attention and enhances the discriminative structure of the learned representations.

5.3.2 Latent Space Analysis via t-SNE

To further investigate how task ordering shapes feature representations, we analyze the latent space using t-distributed Stochastic Neighbor Embedding (t-SNE). We extract embeddings from the penultimate layer of the multiclass head for all samples in the external HAMt test set and reduce them to 2D using t-SNE (perplexity 30, 1000 iterations).

Figure 5.7 compares a Swin Transformer baseline trained only for classification with the sequential SST_CS configuration. The baseline embedding space shows substantial overlap among visually similar classes such as MEL, BKL, and NV. In contrast, the sequential model exhibits tighter intra-class clusters and clearer inter-class margins. Although t-SNE is non-metric, this qualitative separation consistently reflects the multiclass accuracy trends reported earlier.

To analyze how the embedding structure evolves during training, we visualize t-SNE projections at multiple checkpoints (Fig. 5.8). Clusters progressively tighten and drift apart as training progresses, especially after the segmentation phase. This confirms that segmentation-first training guides the backbone toward a more structured, morphology-aware representation before classification starts.

To complement the qualitative t-SNE analysis, we additionally quantify the geometry of the latent space using class-wise centroid distances. For each checkpoint, feature vectors from the penultimate multiclass layer are grouped by class, and the centroid of each class is computed as the mean feature vector. Pairwise Euclidean distances between class centroids, then provide:

(i) the mean inter-class distance, capturing overall separation, and (ii) the minimum centroid distance, identifying the closest pair of classes. Table 5.2 reports these values at representative

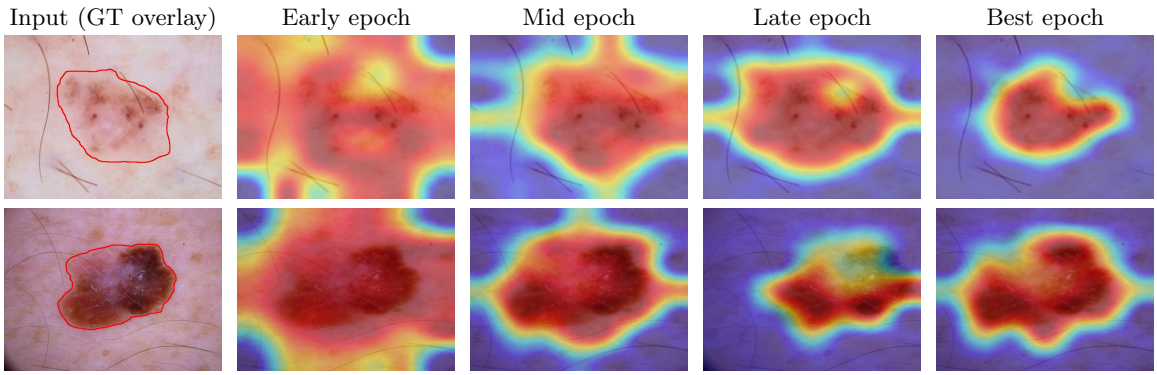


Figure 5.6: Temporal evolution of Grad-CAM activations for two HAM lesions across training epochs in the SST_SC configuration.

Table 5.2: Centroid-based quantification for the segmentation-followed-by-classification configuration. Distances are computed on the validation set at representative training epochs.

Epoch	**Mean inter-class distance**	**Closest class pair**	**Minimum distance**
1	28.64	AKIEC–DF	8.05
6	44.98	AKIEC–BCC	13.97
18	47.01	AKIEC–BCC	18.50
32	53.42	BKL–DF	22.57

epochs.

The mean inter-class distance increases from **28.64** (epoch 1) to **53.42** (epoch 32), indicating that lesion categories become progressively more separated as training advances. Similarly, the minimum centroid distance increases from **8.05** to **22.57**, showing that even the most similar classes grow more distinguishable over time. Together, these quantitative findings directly support the qualitative trends observed in the t-SNE plots: early segmentation shapes the latent space toward more structured, discriminative representations, thereby facilitating more robust downstream classification.

Overall, the progressively increasing separation and compactness of clusters provide a complementary explanation for the improved generalization of **SST_SC**, corroborating the Grad-CAM analysis. Taken together, these observations indicate that the benefit of the segmentation-first configuration is not limited to a single "better" checkpoint, but emerges as a training trajectory effect. As the model alternates between segmentation and classification, the latent space evolves from a noisy and entangled configuration toward a progressively more structured and stable geometry, where class centroids drift apart and intra-class variability shrinks. This representation convergence across epochs, consistently observed in both t-SNE projections and centroid-based distances, provides a mechanistic explanation for the improved robustness of **SST_SC** under domain shift, and links the quantitative behaviour of the backbone to the lesion-centred attention patterns highlighted by Grad-CAM.

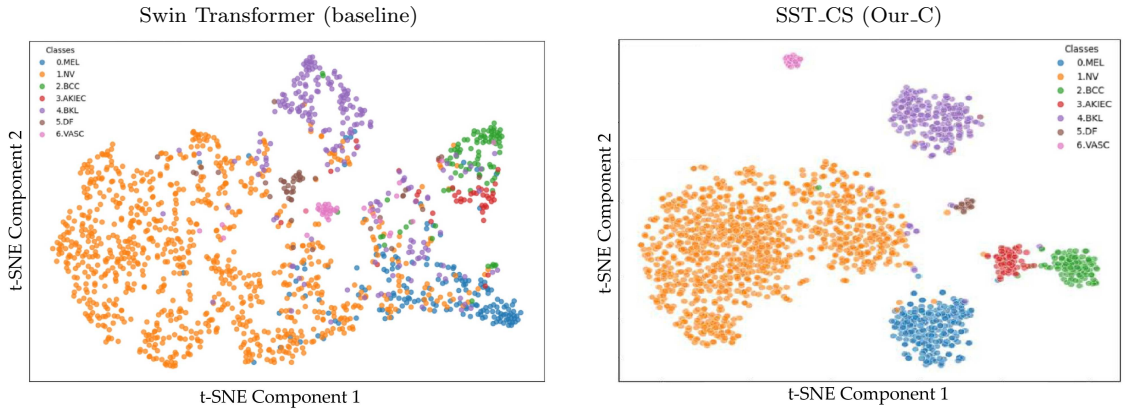


Figure 5.7: t-SNE projections of multiclass embeddings on the external HAMt test set for the Swin baseline (left) and SST_CS (right).

5.3.3 Comparison with State-of-the-Art Methods

Table 5.3: Segmentation benchmark on the HAM dataset. All values are percentages. Asterisks (*) denote values as originally reported. The Jaccard index refers to the Intersection over Union (IoU) metric.

Method	Jaccard (%)	Dice (%)
GAN-UNET [85]	77.00	85.20
SkinSAM [90]	78.43*	88.79*
DenseNet121-UNET [85]	83.50	89.70
DeepLabV3+ [85]	82.80	89.80
CA-Net [92]	–	92.08
BAT [88]	84.30	92.10
Polar Image Transformation [96]	87.43	92.53
SST_SC (Our_A)	86.13 ± 0.27	89.61 ± 0.15
SST_CS (Our_B)	86.03 ± 0.18	89.48 ± 0.22
SST_CS (HAM+Ht, Our_C)	86.58 ± 0.19	89.50 ± 0.25
SST_SC (HAM+Ht, Our_D)	86.26 ± 0.16	89.48 ± 0.23

Regarding segmentation performance, Table 5.3 reports a comparison of our SST framework with several recent segmentation models on the HAM dataset. Both SST_SC and SST_CS achieve high Jaccard scores (86.26% and 86.58%, respectively), outperforming classical architectures such as GAN-UNET (77.00%) and DeepLabV3+ (82.80%). While a few specialized models—such as CA-Net [92] and BAT [88]—report higher Dice scores (up to 92.10%), they are typically tailored to segmentation and often lack full cross-task compatibility.

It is important to note that SST was not explicitly optimized for segmentation; nonetheless, it achieves highly competitive performance in terms of Jaccard index, which we consider a more robust metric for evaluating overlap in challenging medical image segmentation scenarios. Furthermore, the SST model offers the advantage of reusing a single backbone across tasks, reducing the need for multiple dedicated models.

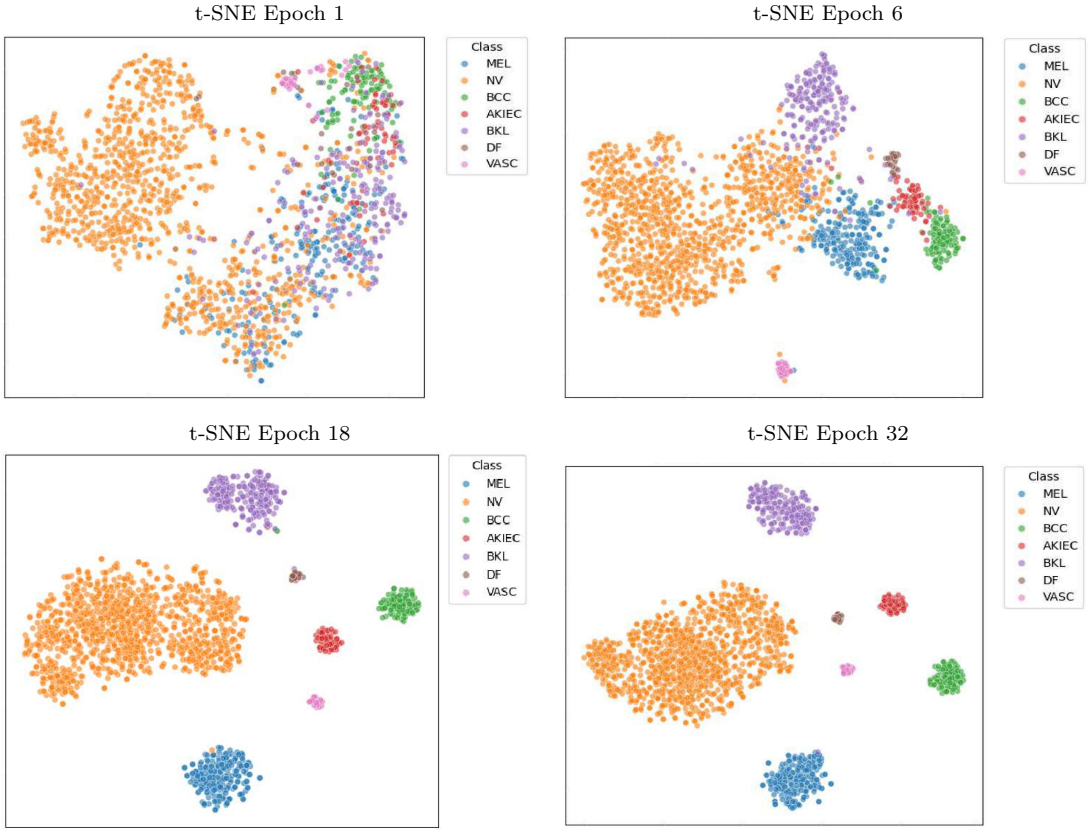


Figure 5.8: Temporal evolution of t-SNE embeddings for the SST model on HAMt.

When training on the combined HAM+HAMt dataset (Our_C and Our_D), segmentation results improve slightly, suggesting that increased data variability contributes to more robust boundary delineation and generalization across domains.

On the classification task, Table 5.4 benchmarks our classification results against other recent methods on the HAM dataset. Internally, SST_SC (Our_A) achieves a multiclass accuracy of 92.16%, which is on par with or slightly below some high-performing deep models such as InceptionResNetV2 (93.20%) and FixCaps (96.49%). However, several of these models report results under unclear or inconsistent train/test splits, often lacking an external validation component—raising concerns about overfitting and inflated performance metrics.

To address this, we evaluate generalization using the HAMt external test set. In this setting, SST_SC (Our_D) achieves 86.87% multiclass accuracy, significantly outperforming the FixCaps baseline re-evaluated under the same conditions (Exp1–Exp3, averaging ~75%). This sharp performance drop for FixCaps, when tested on unseen data, highlights a key strength of our approach: the use of a standardized external test protocol that more accurately reflects real-world performance.

Moreover, the SST architecture integrates both binary and multiclass classification heads. Binary classification, though often overlooked, plays a vital role in early triage scenarios. Our binary

Table 5.4: Multiclass classification benchmark results on the HAM dataset. The "Split" column indicates the dataset partitioning into training, validation, and test sets. For entries with parentheses (e.g., "80 (90/10) – 20"), the portion inside represents the training/validation split, while the value outside refers to the test set. An asterisk (*) indicates that no test set was used or reported. TA_multiclass (%) reports the accuracy for multiclass classification.

Author	**Method**	**Dataset**	**Split**	**TA_multiclass (%)**
Mushtaq et al. [156]	Ensemble VGG16	HAM	80–20 + 15% not duplicate in Test	89
Jain et al. [65]	TL on Xception	HAM	80 (90–10) – 20	89.66
Chaturvedi et al. [157]	InceptionV3	HAM	88–12 *	91.56
Chaturvedi et al. [157]	InceptionResNetV2	HAM	88–12 *	93.20
Shetty et al. [52]	CNN models	HAM	80–20	95.18
Aladhadh et al. [64]	Medical Vision Transformer	HAM	70–20–10	96.14
Lan et al. [140]	FixCaps	HAM	85–15	96.49
Exp1	FixCaps	HAM+HAMt	80 + HAMt	76.19
Exp2	FixCaps	HAM+HAMt	80 + HAMt	75.28
Exp3	FixCaps	HAM+HAMt	80 + HAMt	75.60
Our_A	SST_SC	HAM	75–15–10	92.16 ± 0.69
Our_B	SST_CS	HAM	75–15–10	91.94 ± 0.49
Gallazzi et al. [141]	Swin Transformer	Large Dataset	80–20 + HAMt	86.37
Our_C	SST_CS	HAM+HAMt	80–20 + HAMt	86.63 ± 0.39
Our_D	SST_SC	HAM+HAMt	80–20 + HAMt	86.87 ± 0.45

accuracy exceeds 93% on HAM and approaches 91% on HAMt, making SST suitable for both coarse- and fine-grained diagnostic applications.

Lastly, Table 5.5 presents a direct comparison between our previous pipeline [2]—which used segmented masks as preprocessing input to separate classifiers—and the SST model, which jointly learns segmentation and classification. SST_SC (Our_D) improves multiclass accuracy by more than 2.7 percentage points (from 84.13% to 86.87%) and Jaccard by over 3.5 points (from 82.75% to 86.26%). In addition, SST achieves more balanced precision, recall, and F1-scores, indicating improved stability and class-wise discrimination.

These improvements demonstrate the benefit of sequential task learning over disjoint preprocessing pipelines. By training the segmentation and classification heads within a shared backbone, SST enables cross-task feature reuse and promotes more structured latent representations, as further validated in the t-SNE analyses.

5.3.3.1 Ablation and Discussion

Beyond raw performance metrics, the ablation analysis aims to clarify how task ordering reshapes the shared backbone rather than merely how much accuracy is gained. The combined evidence from Grad-CAM, t-SNE, and centroid-based distances shows that the segmentation-first configuration does not simply yield slightly different feature vectors at convergence, but drives a distinct and more stable representational pathway: spatial priors learned during segmentation progressively constrain attention to lesion regions, tighten class clusters in the latent space, and enlarge inter-class margins over training epochs. In contrast, starting from classification tends to produce broader, less lesion-centric activations and a less clearly structured embedding, which ultimately translates into lower robustness on external data.

Table 5.5: Comparative results between our previous preprocessing-based pipeline [2]—which used segmentation outputs as inputs to classification—and the proposed sequential SST models trained on HAM and tested on HAMt. The first three rows report the baseline strategy, while SST_CS and SST_SC represent the joint sequential learning configurations from Table 5.1. Metrics include segmentation performance (Jaccard %) and multiclass classification results: accuracy (TAm %), precision (TPm), recall (TRm), and F1-score (TF1m).

Author	**Method**	**Jaccard (%)**	**TAm (%)**	**TPm (%)**	**TRm (%)**	**TF1m (%)**
Gallazzi et al. [2]	HAM+YOLO	77.00	84.01	84.09	84.12	83.53
Gallazzi et al. [2]	HAM+DeepLabV3	81.66	84.12	84.38	84.12	83.75
Gallazzi et al. [2]	HAM+ST	82.75	84.13	84.41	84.12	83.58
Our_C	SST_CS	86.58 ± 0.19	86.63 ± 0.39	82.37 ± 0.51	79.42 ± 0.30	79.74 ± 1.80
Our_D	SST_SC	86.26 ± 0.16	86.87 ± 0.45	84.94 ± 0.69	78.47 ± 1.48	80.57 ± 0.10

Task ordering. Across all experimental configurations, segmentation metrics (IoU/Dice) remain stable between the two sequential orders, indicating that segmentation quality is not substantially affected by whether it precedes or follows classification. Classification, however, is consistently more sensitive to task ordering. On the external HAMt set, the **SST_SC** configuration (Our_D) achieves higher multiclass accuracy (86.87%) and macro F1-score (80.57) than **SST_CS** (Our_C, 86.63% / 79.74). A Wilcoxon signed-rank test confirms the statistical significance of this improvement (W = 15.0, p = 0.03125). These findings support the hypothesis that segmentation-first learning provides a spatial inductive bias that benefits downstream classification, particularly under distribution shift.

Qualitative evidence. Grad-CAM visualizations (Fig. 5.5, Fig. 5.6) complement the quantitative results by showing distinct attention behaviours between the two orders. The segmentation-first model (**SST_SC**) tends to focus activation on lesion interiors and boundaries, closely aligning with annotated regions. In contrast, the classification-first configuration (**SST_CS**) often exhibits a broader, less lesion-centric focus that extends toward surrounding skin areas. This difference becomes more pronounced for difficult categories such as melanoma and actinic keratosis, where precise localization is crucial for correct classification. The improved lesion focus in **SST_SC** explains its higher robustness and generalization observed in the external HAMt evaluation.

Latent-space structure. The t-SNE analyses (Figs. 5.7–5.8) provide complementary insight into how task ordering influences feature organization. While the baseline Swin Transformer shows overlapping and poorly defined class clusters, the sequential models—particularly **SST_SC**—produce more compact intra-class groups and clearer inter-class boundaries. This structural separation indicates that the segmentation-first order encourages the model to encode more discriminative and semantically meaningful representations, aligning with its higher classification accuracy and lesion-centric Grad-CAM activations.

Interpretation and implications. Together, these findings support the hypothesis that segmentation - first training acts as an effective inductive bias for subsequent classification, by embedding spatial and morphological cues into the shared backbone. Rather than improving segmentation per se, the benefit emerges from how segmentation shapes the internal representation space, guiding the network to attend to diagnostically relevant features. This mechanism explains the stronger

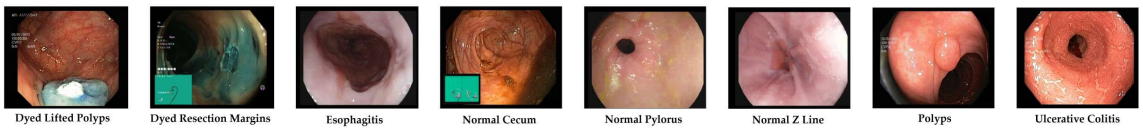


Figure 5.9: Representative gastrointestinal images from the Kvasir dataset, each annotated with its corresponding class label.

generalization performance of **SST_SC** across datasets, as well as the improved interpretability observed through Grad-CAM. In practice, this suggests that task ordering should be explicitly considered when designing sequential or multitask frameworks for medical imaging, particularly when both spatial precision and diagnostic discrimination are required.

Although both orders perform competitively, the segmentation-first approach consistently yields more lesion-focused attention, better feature disentanglement, and improved generalization under distribution shift. These advantages make SST_SC a more reliable configuration for clinical scenarios where interpretability and robustness are equally important.

5.4 Generalization to a Different Medical Domain: The Kvasir Dataset

To assess the generalization capability of the SST framework beyond dermatology, we extended its application to a different clinical imaging domain—gastrointestinal endoscopy—by evaluating its performance on the Kvasir dataset [153]. This experiment aims to determine whether the sequential learning paradigm can adapt to alternative anatomical structures, imaging modalities, and clinical targets.

5.4.1 Kvasir Dataset

The **Kvasir** dataset [153] is a publicly available collection of medical images captured through gastrointestinal (GI) endoscopy. It consists of several subsets tailored to different computer vision tasks, including classification and segmentation.

The *Kvasir-CLASSIFICATION* (KvasirC) subset comprises **8,000 RGB images**, evenly distributed among **8 classes**, including both anatomical landmarks (e.g., Z-line, pylorus, cecum) and pathological findings (e.g., esophagitis, polyps, dyed-lifted polyps). Each class contains exactly 1,000 images, all labeled by medical experts and validated through a consensus process. Representative examples of each class are shown in Figure 5.9. The images vary in resolution, typically ranging from 720×576 to 1920×1072 pixels, and are stored in JPEG format. No patient-identifiable metadata is included, but each image is associated with its class label, ensuring balanced and well-curated data.

In addition to the classification dataset, the *Kvasir-SEGMENTATION* (KvasirS) subset [158] provides **1,000 polyp images** paired with pixel-wise **segmentation masks**. These images are derived from endoscopic recordings and are annotated with binary masks delineating the polyp boundaries. The dataset is frequently utilized in polyp segmentation tasks and serves as a benchmark for models aiming to achieve object-level delineation in gastrointestinal images.

Although developed for gastrointestinal disease analysis, the Kvasir dataset introduces a significantly different visual distribution compared to dermatoscopic datasets as it includes internal tissue views, high vascular content, mucus artifacts, and camera motion blur. This modality divergence makes Kvasir an ideal testbed to assess the generalization capacity of models trained on skin lesion images.

The dataset is employed in our framework to examine the extent to which representations learned on dermatological images retain semantic relevance when applied to a different yet medical image domain.

5.4.2 Experimental Protocol on Kvasir

The two subsets are disjoint (no image overlap) and are used separately to mirror the sequential scheme adopted on HAM. We split each subset into **75%/15%/10%** for training/validation/testing. For classification, we apply stratified sampling to preserve class balance; for segmentation, we split at the image level while ensuring no leakage across folds. This setup satisfies SST requirements—i.e., the availability of both a classification dataset and a segmentation dataset with the relevant labels and masks—allowing a like-for-like evaluation of sequential training beyond dermatology.

5.4.3 Related Work on the Kvasir Dataset

Before describing the dataset in detail, it is worth briefly reviewing how Kvasir has been employed in the medical image analysis literature. Several studies have investigated both segmentation and classification tasks on this dataset, primarily relying on convolutional neural network (CNN) architectures.

For segmentation, [159] introduced the Deep Understanding Convolutional Kernel (DUCK) model, a CNN-based approach tailored for colonoscopy images, which achieved a Jaccard index above 90%. Comparable performance was later reported by [160], who employed an EfficientNet backbone pre-trained on ImageNet. These works highlight the feasibility of reaching high segmentation accuracy on gastrointestinal imagery using CNN-based pipelines.

In the context of multiclass classification, several CNN-based methods have also been proposed [161, 162, 163, 164, 165]. Reported accuracies range across different settings, with one of the most competitive results achieving over 93% in the eight-class classification task [165]. These findings indicate that Kvasir has become a well-established benchmark for endoscopic image analysis and provide a solid baseline for evaluating the generalization of novel frameworks such as SST.

5.4.4 Experimental Setup

To ensure comparability with previous results, we retained the same SST architecture and training pipeline described in Section 5.2.2, with minor adjustments to accommodate the Kvasir task definitions:

- **Segmentation Task**: Predict binary polyp masks on KvasirS.
- **Classification Task**: Perform 8-class multiclass prediction on KvasirC.

Each configuration was tested using the same STL strategy and task ordering paradigms:

- **Our_E (SST_CS)**: Classification → Segmentation

- **Our_F (SST_SC)**: Segmentation → Classification

Input images were resized to 224×224. All models were trained for 100 epochs per task, using Adam optimizer with learning rate $1 \cdot 10^{-4}$ and batch size 144. Data augmentation followed the same pipeline described before in Section 5.2.4. Performance metrics included accuracy, precision, recall, F1-score, Dice, Jaccard, and AUC, averaged over five runs for statistical robustness.

5.4.5 Results and Evaluation

Table 5.6: Overall benchmark results of the SST model on the Kvasir datasets. TA denotes test accuracy. Jaccard and Dice scores evaluate segmentation performance, while TAb and TAm indicate binary and multiclass classification accuracy. TPm, TRm, and TF1m represent multiclass precision, recall, and F1-score.

Model	Jaccard (%)	Dice (%)	TAb (%)	TAm (%)	TPm (%)	TRm (%)	TF1m (%)
Our_E	90.95 ± 0.21	93.12 ± 0.39	99.40 ± 0.14	94.41 ± 0.37	94.54 ± 0.38	94.40 ± 0.38	94.38 ± 0.38
Our_F	91.12 ± 0.45	93.60 ± 0.17	98.90 ± 0.14	94.59 ± 0.41	94.16 ± 0.59	94.05 ± 0.60	94.03 ± 0.60

The experimental evaluation on the Kvasir dataset offers a comprehensive view of SST's ability to generalize across clinical imaging domains. Tables 5.7 and 5.8 report the quantitative performance for segmentation and classification tasks, respectively. Visual support is provided in Figure 5.10, which displays confusion matrices and ROC curves per configuration.

Table 5.7: Segmentation benchmark on the KvasirS dataset. Evaluation based on Jaccard and Dice metrics.

Method	Jaccard (%)	Dice (%)
DUCK-net [159]	90.51	95.02
EffiSegNet-B4 [160]	90.56	94.83
EffiSegNet-B6 [160]	90.60	94.77
EffiSegNet-B5 [160]	90.65	94.88
SST_CS (Our_E)	90.95 ± 0.21	93.12 ± 0.39
SST_SC (Our_F)	91.12 ± 0.45	93.60 ± 0.17

Segmentation. On the segmentation task (KvasirS), the SST_SC configuration (**Our_F**) yielded the highest performance, achieving a Jaccard Index of **91.12%** and a Dice score of **93.60%**. These results slightly surpass those of the SST_CS configuration (**Our_E**), which reached 90.95% and 93.12%, respectively. The observed difference, although modest in absolute terms, is consistent and statistically significant across runs (standard deviation < 0.5), suggesting that pretraining the shared backbone with segmentation allows the model to learn more detailed spatial priors, such as lesion contours and localized structures.

Importantly, SST_SC outperformed state-of-the-art architectures tailored specifically for polyp segmentation, such as DUCK-net and multiple EffiSegNet variants. This confirms the versatility of the Swin backbone and the effectiveness of the sequential training strategy, even when transferred to a different anatomical context.

Table 5.8: Multiclass classification accuracy on the KvasirC dataset. TA_multiclass (%) indicates classification accuracy.

Method	TA_multiclass (%)
Multi-model classification [161]	90.20
Single Shot MultiBox Detector [162]	90.40
Transfer Learning framework [163]	93.00
Deep CNN-based SAM [164]	93.19
Spatial-attention ConvMixer [165]	93.37
SST_CS (Our_E)	94.41 ± 0.37
SST_SC (Our_F)	94.59 ± 0.41

Classification. In the multiclass classification task (KvasirC), both SST configurations achieved high performance, with SST_SC again slightly outperforming SST_CS. Specifically, SST_SC reached a mean accuracy of **94.59%**, while SST_CS followed closely at **94.41%**. Additionally, binary classification accuracy exceeded **98.9%** for both configurations, confirming the backbone's capacity to distinguish lesion vs. non-lesion classes with high confidence.

Beyond accuracy, ROC curve analyses (Figure 5.10) further emphasize the consistency of the results. All models achieved AUC values above 0.97 across classes. SST_SC displayed slightly improved per-class ROC profiles and macro-average AUCs, indicating better discrimination performance across the heterogeneous class set.

Interpretation. These results validate several key assumptions. First, the ordering of tasks plays a measurable role in downstream performance. Initiating training with segmentation (SST_SC) appears to enhance the model's capacity to encode spatially-aware features that are then transferred and reused in classification, leading to more precise and robust decisions. Second, the Swin backbone proves capable of learning high-quality representations even when shifted to new domains and tasks. Despite the domain-specific nature of gastrointestinal imagery, the shared backbone generalized well, further demonstrating SST's potential as a modular and adaptable learning framework.

Finally, the competitive results obtained on Kvasir — both in segmentation and classification — underscore the generalizability of sequential learning and support its application beyond dermatology. This motivates further exploration into robustness under domain shift, which will be the focus of the following chapter.

5.4.6 Discussion

The results demonstrate that the SST maintains strong performance across different modalities and clinical domains. In particular, initiating training with segmentation (SST_SC) consistently provides marginally higher scores across both segmentation and classification tasks. This finding aligns with the trends observed in the HAM dataset, confirming that early exposure to spatial and morphological priors facilitates the development of robust, transferable representations. Such behaviour is further supported by the ablation analysis, where the segmentation-first configuration exhibited more lesion-centric Grad-CAM activations and a better-structured latent space, evidencing its capacity to encode discriminative and semantically meaningful features.

Beyond numerical metrics, the qualitative and latent-space analyses reinforce this interpretation. Grad-CAM visualizations revealed that **SST_SC** directs attention more precisely toward lesion boundaries and diagnostically relevant areas, while **SST_CS** often includes surrounding con-

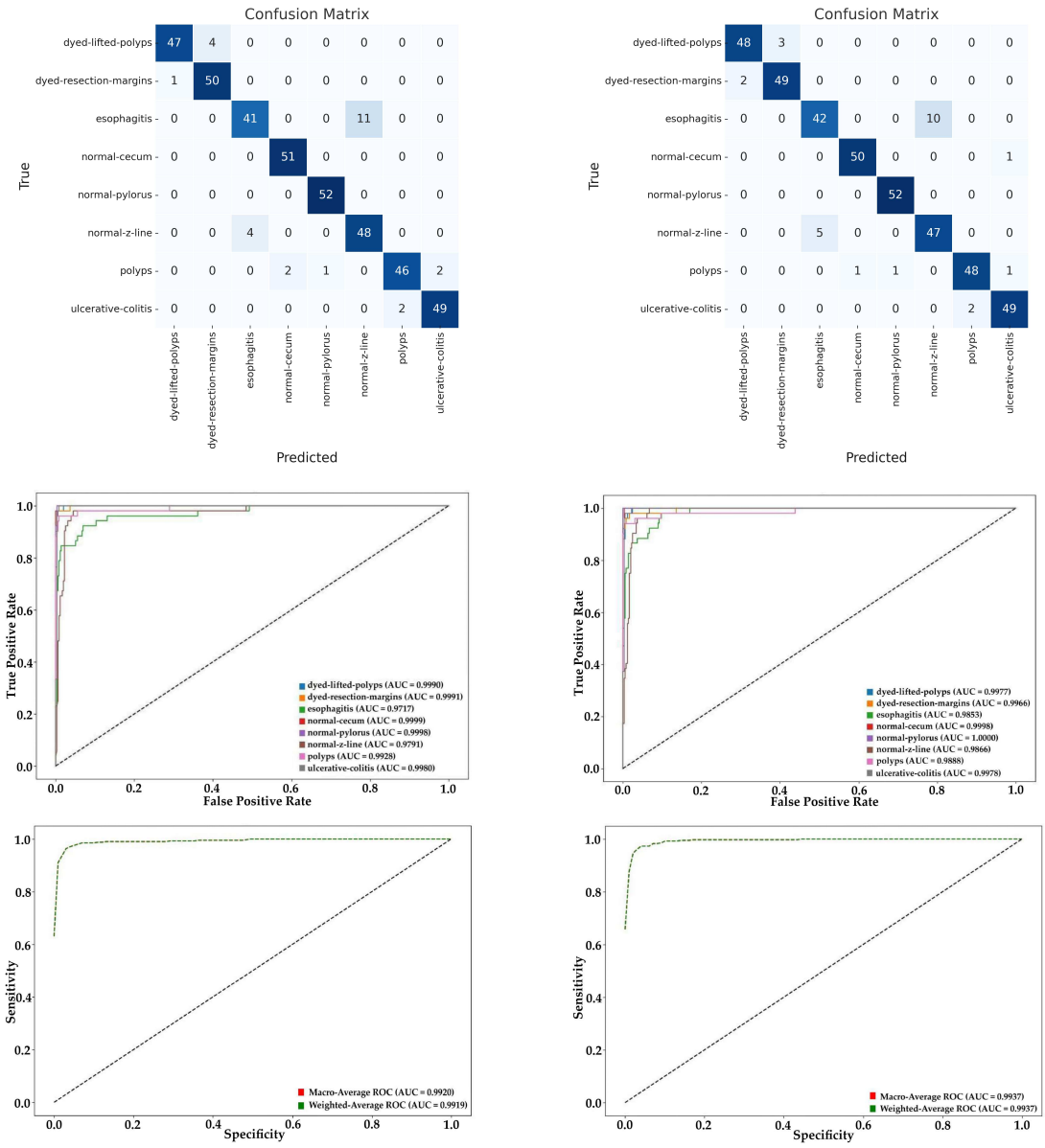


Figure 5.10: Confusion matrices, per-class ROC curves (with legend), and macro/weighted ROC curves (from top to bottom) for the models Our_E (left column) and Our_F (right column), as reported in Table 5.6. Each row corresponds to the respective experimental setup, based on the best-performing checkpoint on the validation set. In the confusion matrices (top row), true labels are shown on the vertical axis and predicted labels on the horizontal axis; classes are indexed from 0 to 6, corresponding respectively to MEL, NV, BCC, AKIEC, BKL, DF, and VASC. The ROC curves in the middle row display the True Positive Rate (TPR, sensitivity) on the y-axis versus the False Positive Rate (FPR) on the x-axis for each class. In the bottom row, macro and weighted average ROC curves are reported with the y-axis representing sensitivity and the x-axis representing specificity.

textual regions. Similarly, the t-SNE projections showed clearer inter-class separation and tighter intra-class clusters under segmentation-first training, confirming that task ordering influences how representations are organized within the shared backbone. Together, these findings suggest that the performance advantage of **SST_SC** stems not only from higher accuracy, but also from a more effective internal representation that facilitates generalization across datasets.

Comparisons with recent literature highlight the competitiveness of our approach across both tasks. In segmentation, **SST_SC** outperforms all baselines, including DUCK-net and several EffiSegNet variants, while maintaining a single unified architecture. In classification, it achieves superior multiclass accuracy despite the substantial domain shift from dermatological to gastrointestinal images, validating the cross-domain adaptability of the sequential framework.

However, a recurring issue emerged during literature comparison: many published results lack sufficient methodological transparency or reproducibility. Several works report internal validation only, with limited details on dataset splits or access to pretrained models, hindering fair benchmarking and replication. Among the few open-sourced approaches—such as FixCaps—a notable performance degradation is observed when evaluating on truly unseen external datasets, underscoring the necessity of standardized and domain-shifted validation protocols.

Overall, this study highlights how a sequentially organized learning strategy not only achieves strong performance but also enhances interpretability, robustness, and transferability across clinical contexts. Having established SST's adaptability beyond skin lesion analysis, the next chapter extends this evaluation to external validation scenarios, further investigating robustness under cross-dataset and cross-modal conditions.

Chapter 6

A Panoramic Dermatology Image Dataset for Clinically Suspicious Lesions[1]

6.1 Introduction and Motivation

Building on the findings of previous chapters, where the SST framework demonstrated robust performance and strong generalization in the dermatological and gastrointestinal domains, this chapter shifts the focus toward data representation and clinical context. Although previous experiments were conducted on well-established dermatoscopic datasets such as HAM [27], these datasets capture isolated lesions acquired under controlled imaging conditions. Such a setting, although useful for algorithmic benchmarking, only partially reflects the complexity of real clinical assessment, where dermatologists reason about the relative characteristics of multiple lesions distributed across the skin.

Recent efforts in dermatological imaging—such as SLICE-3D [166] and iToBoS [167]—have begun to address this gap by providing wide-field or body-region acquisitions. However, these datasets remain limited in scope: SLICE-3D focuses primarily on cropped lesion classification, while iToBoS predicts bounding boxes on body tiles that may not consistently contain the lesion of interest. Consequently, the broader visual and diagnostic context, which is central to clinical decision-making, remains underrepresented.

To overcome these limitations, we introduce **SKINPAN** (*Skin Panoramic*), a novel dataset of **10,050 high-resolution panoramic dermatology images**. Each image was captured in a hospital setting and contains one or more lesions explicitly identified by expert dermatologists as *clinically suspicious* or *worthy of monitoring*. These annotations reflect clinical decision-making and provide a unique opportunity for context-aware AI models.

[1]This chapter is based on the manuscript: **M. Gatti, M. Gallazzi, I. Gallo, S. Corchs, A. Carugno, and N. Zerbinati**, "SKINPAN: A Panoramic Dermatology Image Dataset for Identification of Clinically Suspicious Skin Lesions," Submitted for review in *Scientific Reports*, Nature Portfolio, 2025.

6.2 Dataset Creation

6.2.1 Image Acquisition

From September 2014 through June 2025, panoramic dermatology photographs were collected at the *Circolo Hospital and Macchi Foundation, ASST Sette Laghi, University of Insubria, Varese (Italy)* during standard consultations and follow-up visits. Ten board-certified dermatologists operated the FotoFinder Universe imaging system (FotoFinder Systems GmbH [168]) equipped with a *Medicam 800HD* dermatoscope, which provides continuous optical zoom up to 140× and a field of view ranging from 2.5 to 25 mm. This system allows both panoramic acquisition through a motor-arm setup, which captures broad anatomical regions, and dermatoscopic close-ups with polarized and non-polarized illumination.

When considered clinically relevant, dermatologists acquired standardized panoramic RGB photographs, ensuring consistent lighting, patient positioning, and framing. These wide-field overviews functioned as reference maps, guiding dermatoscopic inspections and documenting lesion context over time.

6.2.2 Annotation Protocol

All images were anonymized before annotation by obscuring personally identifiable features such as eyes, tattoos, and jewelry. Annotation was performed through a semi-automated pipeline built on the open-source platform *X-AnyLabeling* [169], integrated with the Segment Anything Model (SAM) [91, 170]. During acquisition, dermatologists marked lesions of concern using arrows. The arrow-tip coordinates were extracted and used as spatial prompts for SAM to generate segmentation proposals.

Annotators then evaluated the proposals: if accurate, they were accepted; otherwise, they were refined or manually redrawn by dermatologists. Each annotated lesion was labeled as *selected for observation*, reflecting the dermatologist's judgment at the time of the consultation. This protocol ensured that annotations were both precise and clinically meaningful.

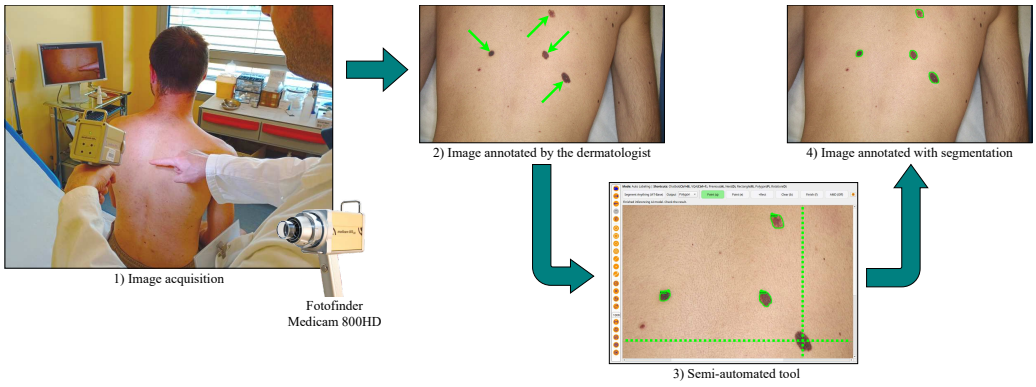


Figure 6.1: Annotation workflow in SKINPAN. Dermatologists acquire panoramic images and mark lesions with arrows. The coordinates are used as prompts for SAM, which generates segmentation proposals. Annotators validate or refine the masks, resulting in expert-approved annotations.

6.2.3 Synthetic Inpainting

Since panoramic images were generally acquired when suspicious lesions were present, the dataset lacked lesion-free examples. To overcome this, synthetic inpainting was applied to a subset of annotated images. Lesion masks defined the inpainting regions, and a diffusion-based model generated realistic, lesion-free skin textures consistent with surrounding areas. All inpainted images were reviewed by dermatologists, and only clinically plausible outputs were retained.

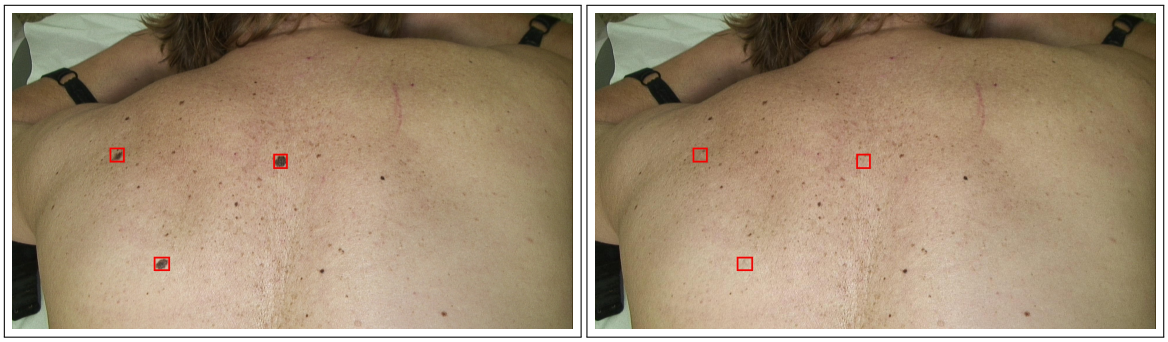


Figure 6.2: Example of lesion inpainting in SKINPAN. Left: original panoramic image with multiple annotated lesions. Right: inpainted version where lesions are replaced by realistic skin texture.

6.2.4 Ethics

The dataset was collected under the approval of the Institutional Ethics Committee of the University of Insubria (protocol **0026287**, 21 February 2025). All data were anonymized prior to release.

6.3 Dataset Composition

SKINPAN comprises **10,050 panoramic RGB images** with expert-validated segmentation masks and bounding boxes. Each image may contain one or multiple annotated lesions, with some including more than five. Metadata provides patient age, sex, anatomical region (generic and specific), and lesion count.

The dataset covers diverse anatomical regions, including head, trunk (anterior and posterior), upper limbs, and lower limbs. The trunk accounts for the majority of images, reflecting both its clinical importance and difficulty in self-monitoring. Age distribution shows a predominance of adult patients, with lighter Fitzpatrick skin types (I–III) more represented due to regional demographics.

6.4 Preliminary Results

To validate the dataset's utility, two state-of-the-art instance segmentation models were trained: YOLO11 [171, 172] and Mask DINO [173]. Performance was assessed using COCO metrics [174].

Both models achieve strong performance, with Mask DINO reaching an AP of 0.66 for segmentation. While accuracy decreases for very small lesions (APs), results confirm that SKINPAN is technically valid and provides a non-trivial benchmark for modern algorithms.

(a) By anatomical region

Localization	Images
Head	856
Trunk anterior	2,449
Trunk posterior	4,115
Upper limbs	902
Lower limbs	1,928

(b) By polygon count

Polygons	Images
1	8,371
2	1,007
3	386
4	114
5	47
6	17
7	4
8	4
≥ 9	6

(c) Age distribution

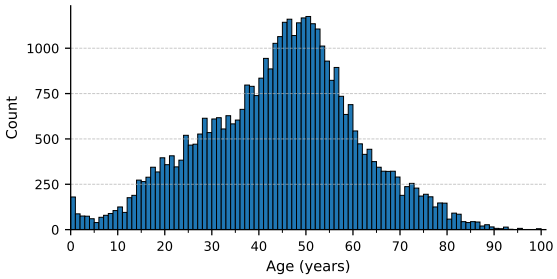


Figure 6.3: Dataset composition by body region and annotation count, and the distribution of subject ages.

Table 6.1: Preliminary results on SKINPAN using COCO evaluation metrics.

Model	Task	AP	AP50	AP75	APs	APm	APl
YOLO11	bbox	0.619	0.796	0.704	0.509	0.722	0.723
YOLO11	segm	0.461	0.789	0.495	0.324	0.565	0.676
Mask DINO	bbox	0.647	0.851	0.755	0.554	0.722	0.727
Mask DINO	segm	0.658	0.853	0.762	0.568	0.734	0.739

6.5 Contribution to Literature and Impact

SKINPAN advances the field in several ways:

- **Context-aware analysis:** enabling algorithms to evaluate lesion distribution, symmetry, and co-occurrence.
- **Clinical realism:** annotations reflect dermatologists' real-time clinical concerns rather than retrospective biopsy outcomes.
- **Benchmarking:** providing the first large-scale, public dataset of panoramic dermatology images with clinically grounded lesion annotations.

Unlike SLICE-3D [166], which focuses on isolated lesion crops, and iToBoS [167], which includes lesion-free tiles, SKINPAN uniquely emphasizes lesions explicitly selected by clinicians for monitoring.

6.6 Role in the Thesis

This dataset complements the thesis work on sequential classification and segmentation (3–5) by addressing the data dimension. It expands the research scope to context-rich imagery, supporting

future multimodal pipelines where panoramic overviews guide dermatoscopic inspection. SKIN-PAN thus represents a bridge between algorithmic advances and clinically realistic applications, reinforcing the thesis vision of generalizable, context-aware AI in dermatology.

Chapter 7

General Discussion

This chapter synthesizes and interprets the empirical evidence presented across Chapters 3–6. Rather than restating the research questions, it consolidates insights from the sequential learning experiments, cross-domain evaluations, and dataset analyses, highlighting both methodological and clinical implications. The discussion is organized around six central themes: evaluation rigor, the role of segmentation, task ordering and representation dynamics, cross-domain robustness, interpretability, and future research directions.

7.1 Evaluation Rigor and External Generalization

One of the most consequential outcomes of this work is the clear divergence between internal validation and performance on truly unseen data. Under the standardized HAM/HAMt protocol [27], several published baselines exhibit substantial accuracy drops on HAMt, confirming that many prior claims of generalization are overestimated. This observation reinforces that *protocolization*—the use of transparent splits, external test sets, and controlled experimental budgets—is a prerequisite for credible benchmarking in medical imaging.

Across the consolidated evaluations (Table 5.1, Table 5.3, Table 5.4), the proposed SST demonstrates strong external consistency. The segmentation-first configuration (**SST_SC**) maintains performance under domain shift and preserves balanced gains across all classes, as illustrated in the ROC and confusion plots (Figures 5.3–5.4). This contrasts with earlier CNN-based baselines, where improvements often concentrated in majority classes, a pattern also highlighted in prior analyses of dermatological datasets [22, 23].

Technically, the unified protocol mitigates three frequent confounders: *(i)* non-standard or implicit data splits, *(ii)* inconsistent metric computation or thresholding, and *(iii)* selective reporting of best internal checkpoints. Where applicable, paired experiments were conducted with identical augmentation pipelines, optimizers, and epoch budgets; non-parametric significance tests (e.g., Wilcoxon signed-rank) confirmed the reliability of the observed ordering effects. These methodological precautions elevate external generalization from an ancillary check to a central performance criterion, in line with recent calls for reproducibility and transparent evaluation in medical AI research.

7.2 Dataset Composition and Population-Aware Generalization

An additional and often underexplored dimension of generalization in dermatological AI concerns the diversity of skin appearance represented in training and evaluation datasets. Beyond acquisition protocols and lesion categories, dataset composition implicitly encodes population-level characteristics, including variations in skin tone, texture, pigmentation, and contrast, which directly influence visual cues available to learning algorithms.

As discussed in Section 2.1, most publicly available datasets for skin lesion analysis exhibit non-uniform population coverage, reflecting the demographics and clinical practices of their source institutions. Differences in skin pigmentation and background appearance may affect lesion visibility, boundary contrast, and color-based features, potentially leading to uneven model behavior when confronted with underrepresented skin types. This challenge is intrinsic to dermatological imaging and cannot be fully captured by standard cross-dataset validation alone.

From this perspective, generalization should be interpreted not only as transfer across datasets or imaging domains, but also as robustness across heterogeneous skin presentations. Future evaluation frameworks may therefore benefit from population-aware analyses that consider stratified performance, calibration behavior, and uncertainty estimates across visually distinct data subgroups. Such considerations are particularly relevant in preventive and screening-oriented settings, where reliable performance across diverse patient populations is a prerequisite for safe clinical deployment.

7.3 What Segmentation Contributes (and What It Does Not)

Early experiments (Chapter 4) demonstrated that using segmentation merely as a preprocessing step—for instance, cropping lesions from input images—does not consistently improve classification. This finding aligns with previous reports that excessive reliance on cropped or masked inputs can remove contextual cues crucial for diagnosis. By contrast, when segmentation is *learned first* within a shared backbone, as in the **SST_SC** configuration (Figure 5.1), it systematically enhances external classification accuracy and macro F1-score (Table 5.1).

The mechanism underlying this improvement is representational rather than geometric: early spatial supervision constrains the feature space around lesion morphology, boundary precision, and internal texture patterns, producing embeddings that are better structured for downstream classification. This insight clarifies that segmentation adds value not through explicit mask quality alone, but through how spatial cues shape the encoder's internal representations. As discussed in the Ablation and Grad-CAM analyses (Section 5.3.3.1, Figure 5.5), this spatial pretraining yields more lesion-centered attention maps and tighter latent-space clustering, confirming that segmentation-first learning provides an effective inductive bias for discriminative reasoning.

7.4 Task Order, Representation Dynamics, and Negative Transfer

Comparative experiments across both task orders confirm that **SST_SC** outperforms **SST_CS** under identical computational budgets. Quantitatively, segmentation-first training achieves higher multiclass accuracy and macro F1 scores on external datasets (Table 5.1). Qualitatively, it yields

clearer inter-class separations in the latent space visualizations (Figures 5.7–5.8). These convergent signals support a *representational scaffolding* hypothesis: by first learning spatial priors and boundary structures, the model anchors subsequent high-level feature extraction in geometrically stable regions of the input space.

Conversely, classification-first training (**SST_CS**) often produces noisier embeddings and broader Grad-CAM activations that extend beyond lesion boundaries, suggesting that abstract semantic features learned early may not easily re-adapt to spatially constrained tasks. This behavior reflects an instance of *negative transfer*, where conflicting optimization signals across tasks hinder specialization. The evidence thus supports the premise that the order of learning tasks is a key determinant of representation quality in sequential architectures, especially when both global and local reasoning are required.

7.5 Cross-Domain Robustness: From Skin to Gastrointestinal Imaging

The evaluation on the Kvasir dataset [153] extends the generalization analysis beyond the dermatological domain. Despite large differences in acquisition modality, scale, and texture statistics, the sequential strategy remains effective: the **SST_SC** configuration achieves competitive segmentation (Table 5.7) and strong multiclass classification (Table 5.8). The consolidated view (Table 5.6) and per-class ROC curves (Figure 5.10) confirm consistent discrimination across lesion categories.

This cross-domain transfer reinforces the interpretation that early spatial supervision captures modality-agnostic structure — such as boundaries, symmetry, and shape regularities — which are beneficial even in anatomically and visually distinct contexts. The finding that the same sequential recipe generalizes from skin to gastrointestinal images underscores the potential of segmentation-first learning as a general strategy for medical image analysis, not confined to a specific clinical domain.

7.6 Context-Aware Dataset Design and Clinical Realism

The introduction of the **SKINPAN** dataset (Chapter 6) addresses an often-overlooked limitation in existing resources: the lack of contextual, panoramic information in dermatological imaging. Unlike lesion-centered datasets such as HAM [27] or BCN [32], SKINPAN captures full-body or body-region views where multiple lesions coexist, allowing clinical reasoning to extend beyond a single crop. By focusing on lesions identified by experts as *clinically suspicious* or *worthy of monitoring*, SKINPAN encodes the decision-making context underlying real diagnostic workflows.

From a methodological standpoint, SKINPAN provides a foundation for studying large-scale lesion localization, selection, and monitoring under real-world variability. It also facilitates the development of models capable of contextual reasoning—integrating global and local cues—thus complementing the findings of the sequential learning framework, where the ability to attend to relevant spatial regions was a key factor in performance. In this sense, SKINPAN bridges data-centric and model-centric innovation, aligning dataset design with clinical interpretability.

7.7 Explainability and Model Understanding

Interpretability emerged as a central diagnostic tool throughout the thesis. Grad-CAM analyses (Figure 5.5) revealed that the segmentation-first configuration (**SST_SC**) consistently concentrates activation within lesion contours, closely matching the corresponding masks. This alignment between predicted and ground-truth regions supports the hypothesis that segmentation-first learning improves not only quantitative performance but also spatial reasoning. t-SNE projections (Figures 5.7–5.8) further demonstrated that sequentially trained models organize the latent space into compact, semantically meaningful clusters, providing a complementary perspective on feature disentanglement.

Despite these advances, explainability remains a practical limitation. The current qualitative tools, while useful for visual inspection, do not provide fully faithful or quantitative attribution. Future work should integrate gradient-based (Grad-CAM), path-based (Integrated Gradients), and propagation-based (LRP) methods, along with quantitative validation (sanity checks, deletion–insertion curves, and pointing-game IoU with segmentation masks). Such integration would elevate model interpretability from illustrative to verifiable evidence, closing the loop between prediction, explanation, and clinical reliability.

7.8 Error Patterns and Clinical Implications

Residual misclassifications across HAM/HAMt primarily involve the {MEL, BKL, NV} classes, mirroring known diagnostic overlap in dermatology. Relative to **SST_CS**, the **SST_SC** configuration reduces these confusions, particularly for melanoma, consistent with its lesion-focused Grad-CAM activations and refined latent clusters. Maintaining both a binary and a multiclass head within the same model, as adopted in the SST framework, enables flexible calibration for different clinical use cases: binary outputs for screening or triage, and multiclass outputs for detailed differential diagnosis.

7.9 Clinical Metadata as Contextual Priors

The analysis of residual errors suggests that visually similar lesion categories can remain challenging to disambiguate from images alone, particularly when classes overlap in morphology and color patterns. In routine dermatological assessment, however, decisions are rarely made in an image-only setting: contextual factors such as patient age, sex, anatomical site, and lesion distribution contribute to risk stratification and to the choice of follow-up actions.

In this thesis, clinical metadata should therefore be regarded as contextual priors rather than as independent predictors. Within the sequential learning framework, metadata can be integrated in ways that support image-based representations — e.g., by gently modulating decision boundaries or uncertainty estimates — while preserving the central role of lesion-specific visual evidence. Conceptually, this aligns with clinical reasoning, in which context can increase or decrease suspicion but does not replace visual inspection and, when needed, dermoscopic examination.

At the same time, metadata integration introduces non-trivial risks, including shortcut learning and population-dependent biases when demographic attributes correlate spuriously with labels. For this reason, any metadata-aware extension should be coupled with robust evaluation practices, including patient-level separation, stress tests under missing or noisy metadata, and calibration

checks across subgroups. When treated as a supportive context, rather than as a primary decision driver, clinical metadata represent a principled pathway toward more clinically grounded and reliable AI systems.

7.10 Actionable Future Work

(1) Quantitative explainability and attribution. Extend current interpretability analyses beyond Grad-CAM by incorporating complementary attribution paradigms such as Integrated Gradients and Layer-wise Relevance Propagation (LRP). Couple these qualitative techniques with quantitative validation metrics (e.g., sanity checks, deletion–insertion curves, and pointing-game IoU against segmentation masks) to ensure that saliency patterns faithfully correspond to diagnostically relevant regions. The release of standardized explanation benchmarks and reference heatmaps would enable reproducibility and external auditing.

(2) Higher-capacity and specialized heads. Investigate the effect of increased head complexity on both performance and generalization. Future work should evaluate deeper classification and segmentation heads (e.g., multi-layer MLPs with attention pooling, class-specific prototypes, or lightweight mixture-of-experts) to test the limits of feature specialization within the shared backbone. Measuring the trade-off between accuracy, calibration, and interference across sequential tasks would clarify how representational capacity influences task transfer.

(3) Learning paradigms and loss functions. Explore alternative optimization strategies that could further improve representation learning and task balance. Potential directions include contrastive pretraining, self-distillation, or hybrid losses that combine pixel-wise (e.g., Dice, Tversky) and feature-level objectives. Task-aware or dynamically weighted losses could also mitigate negative transfer, enabling smoother transitions between segmentation and classification phases. Such investigations would contribute to a deeper understanding of how inductive biases and optimization dynamics impact the stability of sequential learning.

(4) Task-interaction analysis. Quantify and mitigate gradient interference between segmentation and classification heads using conflict-aware optimization (e.g., PCGrad, GradNorm) or adapter-based partial sharing. This analysis would help delineate when joint learning becomes detrimental and when staged fine-tuning offers superior stability. A systematic study of task coupling, combined with representational similarity measures, would clarify how features evolve across phases.

(5) Beyond dermoscopy: macroscopic and panoramic imaging. Evaluate the proposed sequential framework on non-dermatoscopic data, including clinical close-ups and full-body panoramic images from datasets such as SKINPAN (Chapter 6). This would allow assessing robustness to illumination, background variability, and scale changes, as well as the model's ability to reason over contextual spatial relationships. Experiments should compare zero-shot transfer, few-shot adaptation, and full fine-tuning, coupled with prevalence-aware metrics suitable for real clinical deployment.

(6) Multimodal integration. Building on the discussion in Section 7.9, extending sequential learning to multimodal scenarios remains a promising avenue for future research. Combining der-

moscopic or panoramic images with structured clinical metadata (e.g., patient age, sex, lesion site) could enhance diagnostic reasoning, provided that data alignment, fairness, and robustness against missing modalities are rigorously addressed. The relevance of this direction is further supported by the strong classification results achieved by multimodal frameworks reviewed in Chapter 2.6. Future multimodal architectures may leverage transformer-based fusion or cross-attention mechanisms to integrate complementary modalities without compromising interpretability or transparency.

(7) Context-aware and longitudinal learning. Leveraging the SKINPAN dataset, future studies should explore learning strategies that incorporate temporal and contextual cues—such as tracking lesion evolution across follow-up visits or reasoning over lesion distributions across the body. Integrating spatial attention with temporal modeling (e.g., Transformer encoders with recurrent or memory components) could enable predictive monitoring of disease progression.

(8) Reproducibility and standardization. Ensure long-term reproducibility by releasing dataset splits, configuration files, trained weights, and calibration scripts for all evaluated datasets (HAM, HAMt, Kvasir, SKINPAN). Implement uncertainty estimation and attribution sanity-check pipelines to reduce reporting variance and foster independent replication. This open-science approach will promote transparency, facilitate benchmarking, and support clinical auditing.

(9) Preventive and Risk-Oriented Use of SKINPAN. In addition to its use as a benchmarking resource, SKINPAN is intended to support research on preventive and risk-oriented dermatological AI based on macroscopic imaging. By modeling global skin context and lesion distribution, the dataset enables the study of predictive signals that may indicate elevated risk or the need for closer clinical monitoring, even in the absence of dermoscopic detail. This perspective positions SKINPAN as a complementary tool to dermoscopic datasets, focusing on early screening and prioritization rather than fine-grained diagnosis.

Collectively, these directions highlight how sequential, segmentation-first learning can serve as a foundation for broader advancements in medical AI—from architectural innovation and optimization paradigms to multimodal integration and clinically realistic datasets. The evidence accumulated throughout this thesis supports the notion that combining structured task ordering with rigorous evaluation protocols yields models that are not only accurate but also interpretable, reproducible, and generalizable across domains.

Chapter 8

Conclusions

This final chapter provides concise answers to the research questions, summarizes the key scientific contributions, translates methodological insights into design rules for practitioners, and outlines a forward-looking research agenda. It concludes with reflections on the scientific and translational significance of this work.

Answers to the Research Questions

RQ1 — How can Transformer-based architectures be optimized to improve classification performance and generalization across heterogeneous dermatological datasets?
Transformer-based architectures, when trained under *standardized and leakage-safe protocols*, can achieve strong external generalization. The proposed Swin-based framework demonstrated stable performance across internal (HAM) and external (HAMt) test sets, bridging the gap between controlled validation and clinical realism. Key to this success were consistent preprocessing, fixed budgets for all comparisons, and transparent reporting of results. These findings support the view that reproducibility and methodological rigor are essential preconditions for generalizable AI in dermatology (see Section 7.1).

RQ2 — What is the specific contribution of segmentation to classification, and how does the *order of learning* between these tasks affect model performance and interpretability?
Segmentation contributes meaningfully to classification only when employed as a learning phase rather than a preprocessing step. The sequential order **SST_SC** (Segmentation → Classification) enhances both quantitative accuracy and qualitative interpretability by embedding spatial priors early in the backbone. This structured learning enhances feature alignment with lesion morphology and texture, resulting in sharper class boundaries in the latent space and more localized Grad-CAM activations. Conversely, reversing the task order (**SST_CS**) can introduce mild negative transfer and less focused attention maps (Sections 7.3–7.7). Overall, the results demonstrate that task ordering is a decisive factor in shaping transferable and clinically meaningful representations.

RQ3 — How can new dataset designs—incorporating contextual and clinically meaningful annotations—contribute to the development of more reliable and transferable AI systems?

The introduction of the **SKINPAN** dataset (Chapter 6) highlights the importance of contextual and clinically grounded data design. Unlike traditional lesion-centered datasets, SKINPAN captures panoramic clinical views and expert annotations identifying lesions *worthy of monitoring*, reflecting real-world diagnostic reasoning. This context-aware dataset enables research on lesion selection, distribution, and follow-up, providing a crucial bridge between algorithmic performance and clinical applicability. It demonstrates the reliability and transferability of medical AI depend not only on model architecture but also on the representational depth and realism of the data used for training and evaluation.

Design Rules for Practice

Protocolization. Always employ transparent dataset splits and include external test sets to prevent information leakage. Publish configuration files, code, and preprocessing pipelines to ensure full reproducibility (see Section 7.1).
Sequential Learning. Prefer the segmentation-first order (**SST_SC**) when aiming for robust generalization; use classification-first only when data constraints or deployment simplicity dictate it (Sections 7.4, 7.7).
Segmentation as learning, not preprocessing. Treat segmentation as a task for representation shaping, not as a spatial filter. Its greatest impact arises when lesion boundaries and internal textures are subtle (Section 7.3).
Dataset realism. Employ datasets with clinically meaningful and context-aware annotations, as they encourage more transferable representations and better alignment with diagnostic reasoning (Section 7.6).
Interpretability. Use visual explanations (Grad-CAM, t-SNE) as tools to audit model focus and verify lesion-centered learning rather than as end goals of explainability (Section 7.7).

Research Agenda: Immediate Next Steps

We highlight two high-priority research directions that naturally extend this work.

(1) Higher-capacity heads. Future studies should investigate increased head complexity to probe the limits of specialization within shared Transformer backbones. This includes deeper MLP heads with attention pooling, class-specific prototypes, or lightweight mixture-of-experts. Evaluations should quantify accuracy–calibration trade-offs, robustness under domain shift (HAMt, Kvasir), and task interference in sequential and joint settings. Representation probes (CKA, linear probes) and freeze–unfreeze ablations can reveal how knowledge evolves across training stages while maintaining computational efficiency.

(2) Multimodal integration. Building on the observations discussed in Chapter 2.6, future work should explore how sequential learning can extend to multimodal scenarios. Integrating dermoscopic or panoramic images with structured clinical metadata (e.g., age, sex, lesion site) may enhance diagnostic reasoning, provided that alignment, fairness, and missing-modality robustness are explicitly addressed. Transformer-based fusion strategies and cross-attention mechanisms could offer effective solutions while preserving interpretability and maintaining protocolized evaluation standards.

Closing Remarks

The collective evidence presented in this thesis demonstrates that the ordering of learning tasks is not a secondary implementation choice, but a central determinant of model robustness and interpretability. By positioning segmentation as a foundation for classification, this research redefines the interplay between structural and semantic learning in medical imaging. The Sequential Swin Transformer framework, validated across multiple datasets and clinical domains, represents a step toward reliable and reproducible AI systems that mirror human diagnostic reasoning.

Beyond technical advances, this work underscores a broader principle: generalization emerges from rigor, not complexity. Carefully designed protocols, transparent evaluation, and clinically meaningful datasets are the true enablers of trustworthy AI. While challenges remain—in multimodal integration, explainability, and longitudinal modeling—the trajectory outlined here provides a concrete and sustainable path forward. In closing, this thesis contributes both methodological insights and practical guidance toward the development of next-generation, domain-aware, and clinically grounded medical AI systems.

Bibliography

[1] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10 012–10 022.

[2] M. Gallazzi, A. U. Rehman, S. Corchs, and I. Gallo, "Improving classification in skin lesion analysis through segmentation," in Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods, 2025, pp. 696–703.

[3] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: a cancer journal for clinicians, vol. 71, no. 3, pp. 209–249, 2021.

[4] F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, and A. Jemal, "Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA: a cancer journal for clinicians, vol. 74, no. 3, pp. 229–263, 2024.

[5] R. L. Siegel, A. N. Giaquinto, and A. Jemal, "Cancer statistics, 2024." CA: a cancer journal for clinicians, vol. 74, no. 1, 2024.

[6] N. C. Institute, "Seer cancer statistics review 1975–2020," https://seer.cancer.gov/csr/1975_2020/, 2023.

[7] J. E. Gershenwald, R. A. Scolyer, K. R. Hess, V. K. Sondak, G. V. Long, M. I. Ross, A. J. Lazar, M. B. Faries, J. M. Kirkwood, G. A. McArthur et al., "Melanoma staging: evidence-based changes in the american joint committee on cancer eighth edition cancer staging manual," CA: a cancer journal for clinicians, vol. 67, no. 6, pp. 472–492, 2017.

[8] W. H. Organization et al., Artificial tanning devices: public health interventions to manage sunbeds. World Health Organization, 2017.

[9] C. Conforti and I. Zalaudek, "Epidemiology and risk factors of melanoma: a review," Dermatology practical & conceptual, vol. 11, no. Suppl 1, p. e2021161S, 2021.

[10] M. A. Tucker and A. M. Goldstein, "Melanoma etiology: where are we?" Oncogene, vol. 22, no. 20, pp. 3042–3052, 2003.

[11] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, "Diagnostic accuracy of dermoscopy," The lancet oncology, vol. 3, no. 3, pp. 159–165, 2002.

[12] M. Vestergaard, P. Macaskill, P. Holt, and S. Menzies, "Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting," British Journal of Dermatology, vol. 159, no. 3, pp. 669–676, 2008.

[13] C. Robert, J. Schachter, G. V. Long, A. Arance, J. J. Grob, L. Mortier, A. Daud, M. S. Carlino, C. McNeil, M. Lotem et al., "Pembrolizumab versus ipilimumab in advanced melanoma," New England Journal of Medicine, vol. 372, no. 26, pp. 2521–2532, 2015.

[14] G. V. Long, D. Stroyakovskiy, H. Gogas, E. Levchenko, F. de Braud, J. Larkin, C. Garbe, T. Jouary, A. Hauschild, J. J. Grob et al., "Combined braf and mek inhibition versus braf inhibition alone in melanoma," New England Journal of Medicine, vol. 371, no. 20, pp. 1877–1888, 2014.

[15] J. Larkin, V. Chiarion-Sileni, R. Gonzalez, J.-J. Grob, P. Rutkowski, C. D. Lao, C. L. Cowey, D. Schadendorf, J. Wagstaff, R. Dummer et al., "Five-year survival with combined nivolumab and ipilimumab in advanced melanoma," New England journal of medicine, vol. 381, no. 16, pp. 1535–1546, 2019.

[16] G. Salerni, C. Carrera, L. Lovatto, J. A. Puig-Butille, C. Badenas, E. Plana, S. Puig, and J. Malvehy, "Benefits of total body photography and digital dermatoscopy ("two-step method of digital follow-up") in the early diagnosis of melanoma in patients at high risk for melanoma," Journal of the American Academy of Dermatology, vol. 67, no. 1, pp. e17–e27, 2012.

[17] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Medical image analysis, vol. 42, pp. 60–88, 2017.

[18] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," nature, vol. 542, no. 7639, pp. 115–118, 2017.

[19] T. J. Brinker, A. Hekler, A. H. Enk, J. Klode, A. Hauschild, C. Berking, B. Schilling, S. Haferkamp, D. Schadendorf, S. Fröhling et al., "A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task," European Journal of Cancer, vol. 111, pp. 148–154, 2019.

[20] H. Haenssle, C. Fink, R. Schneiderbauer et al., "Man against machine: diagnostic performance of a deep learning cnn for dermoscopic melanoma recognition in comparison to 58 dermatologists," Annals of Oncology, vol. 29, no. 8, pp. 1836–1842, 2018.

[21] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy et al., "Human–computer collaboration for skin cancer recognition," Nature medicine, vol. 26, no. 8, pp. 1229–1234, 2020.

[22] R. Daneshjou, K. Vodrahalli, W. Liang, R. A. Novoa, M. Jenkins, V. Rotemberg, J. Ko, S. M. Swetter, E. E. Bailey, O. Gevaert et al., "Disparities in dermatology ai: Assessments using diverse clinical images," arXiv preprint arXiv:2111.08006, 2021.

[23] A. Bissoto, M. Fornaciali, E. Valle, and S. Avila, "(de) constructing bias on skin lesion datasets," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, no. 7553, pp. 436–444, 2015.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, 2012.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[27] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific data, vol. 5, no. 1, pp. 1–9, 2018.

[28] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler 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.

[29] "International skin imaging collaborations. isic archive," last accessed on 2024-07-04. [Online]. Available: https://api.isic-archive.com/images/?query=&collections=289

[30] M. A. Ricci Lara, M. V. Rodríguez Kowalczuk, M. Lisa Eliceche, M. G. Ferraresso, D. R. Luna, S. E. Benitez, and L. D. Mazzuoccolo, "A dataset of skin lesion images collected in argentina for the evaluation of ai tools in this population," Scientific Data, vol. 10, no. 1, p. 712, 2023.

[31] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, L. Caffery, E. Chousakos, N. Codella, M. Combalia, S. Dusza, P. Guitera, D. Gutman et al., "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," Scientific data, vol. 8, no. 1, p. 34, 2021.

[32] M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig et al., "Bcn20000: Dermoscopic lesions in the wild," arXiv preprint arXiv:1908.02288, 2019.

[33] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 2, pp. 538–546, 2019.

[34] "Consecutive biopsies for melanoma across year 2020." [Online]. Available: https://doi.org/10.34970/151324

[35] S. M. de Faria, J. N. Filipe, P. M. Pereira, L. M. Tavora, P. A. Assuncao, M. O. Santos, R. Fonseca-Pinto, F. Santiago, V. Dominguez, and M. Henrique, "Light field image dataset of skin lesions," in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2019, pp. 3905–3908.

[36] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "Ph 2-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.

[37] A. Oakley, M. Duffill, and M. Rademaker, "Dermnet new zealand trust," 1996, [Online. Accessed November 12, 2024]. [Online]. Available: https://dermnetnz.org/

[38] M. Groh, C. Harris, L. Soenksen, F. Lau, R. Han, A. Kim, A. Koochek, and O. Badri, "Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset," 2021, pp. 1820–1828.

[39] D. Gutman, N. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection," in International Symposium on Biomedical Imaging (ISBI),(International Skin Imaging Collaboration (ISIC), 2016), 2016.

[40] "Isic challenge webpage," https://challenge.isic-archive.com, Last accessed on 2024-07-04.

[41] T. DeVries and D. Ramachandram, "Skin lesion classification using deep multi-scale convolutional neural networks," arXiv preprint arXiv:1703.01402, 2017.

[42] A. Mahbod, G. Schaefer, C. Wang, R. Ecker, and I. Ellinge, "Skin lesion classification using hybrid deep neural networks," in ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2019, pp. 1229–1233.

[43] P. Carcagnì, A. Cuna, and C. Distante, "A dense cnn approach for skin lesion classification," arXiv preprint arXiv:1807.06416, 2018.

[44] A. Rezvantalab, H. Safigholi, and S. Karimijeshni, "Dermatologist level dermoscopy skin cancer classification using different deep learning convolutional neural networks algorithms," arXiv preprint arXiv:1810.10348, 2018.

[45] E. H. Mohamed and W. H. El-Behaidy, "Enhanced skin lesions classification using deep convolutional networks," in 2019 Ninth international conference on intelligent computing and information systems (ICICIS). IEEE, 2019, pp. 180–188.

[46] H.-W. Huang, B. W.-Y. Hsu, C.-H. Lee, and V. S. Tseng, "Development of a light-weight deep learning model for cloud applications and remote diagnosis of skin cancers," The Journal of dermatology, vol. 48, no. 3, pp. 310–316, 2021.

[47] Q. Liu, L. Yu, L. Luo, Q. Dou, and P. A. Heng, "Semi-supervised medical image classification with relation-driven self-ensembling model," IEEE transactions on medical imaging, vol. 39, no. 11, pp. 3429–3440, 2020.

[48] Y. Gu, Z. Ge, C. P. Bonnington, and J. Zhou, "Progressive transfer learning and adversarial domain adaptation for cross-domain skin disease classification," IEEE journal of biomedical and health informatics, vol. 24, no. 5, pp. 1379–1393, 2019.

[49] M. A. Khan, M. Y. Javed, M. Sharif, T. Saba, and A. Rehman, "Multi-model deep neural network based features extraction and optimal selection approach for skin lesion classification," in 2019 international conference on computer and information sciences (ICCIS). IEEE, 2019, pp. 1–7.

[50] C. Calderón, K. Sanchez, S. Castillo, and H. Arguello, "Bilsk: A bilinear convolutional neural network approach for skin lesion classification," Computer Methods and Programs in Biomedicine Update, vol. 1, p. 100036, 2021.

[51] V. D. Nguyen, N. D. Bui, and H. K. Do, "Skin lesion classification on imbalanced data using deep learning with soft attention," Sensors, vol. 22, no. 19, p. 7530, 2022.

[52] B. Shetty, R. Fernandes, A. P. Rodrigues, R. Chengoden, S. Bhattacharya, and K. Lakshmanna, "Skin lesion classification of dermoscopic images using machine learning and convolutional neural network," Scientific Reports, vol. 12, no. 1, p. 18134, 2022.

[53] D. Ajabani, Z. A. Shaikh, A. Yousef, K. Ali, and M. A. Albahar, "Enhancing skin lesion classification: a cnn approach with human baseline comparison," PeerJ Computer Science, vol. 11, p. e2795, 2025.

[54] M. Rasel, S. A. Kareem, Z. Kwan, N. A. A. Faheem, W. H. Han, R. K. J. Choong, S. S. Yong, and U. Obaidellah, "Asymmetric lesion detection with geometric patterns and cnn-svm classification," Computers in Biology and Medicine, vol. 179, p. 108851, 2024.

[55] V.-T. Nguyen, V.-T. Pham, and T.-T. Tran, "Ac-mambaseg: An adaptive convolution and mamba-based architecture for enhanced skin lesion segmentation," in International Conference on Green Technology and Sustainable Development. Springer, 2024, pp. 13–26.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[57] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Germany, 2015, proceedings, part III 18. Springer, 2015, pp. 234–241.

[58] S. Younas, A. B. Sargano, L. You, and Z. Habib, "Attention-based inception-residual cnn: Skin cancer diagnosis with attention-based inception-residual cnn model," Information, vol. 16, no. 2, p. 120, 2025.

[59] A. Halder, A. Dalal, S. Gharami, M. Wozniak, M. F. Ijaz, and P. K. Singh, "A fuzzy rank-based deep ensemble methodology for multi-class skin cancer classification," Scientific Reports, vol. 15, no. 1, p. 6268, 2025.

[60] R. Daneshjou, K. Vodrahalli, R. A. Novoa, M. Jenkins, W. Liang, V. Rotemberg, J. Ko, S. M. Swetter, E. E. Bailey, O. Gevaert et al., "Disparities in dermatology ai performance on a diverse, curated clinical image set," Science advances, vol. 8, no. 31, p. eabq6147, 2022.

[61] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in International conference on machine learning. PMLR, 2017, pp. 1321–1330.

[62] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," Advances in neural information processing systems, vol. 31, 2018.

[63] A. G. Pacheco and R. A. Krohling, "The impact of patient clinical information on automated skin cancer detection," Computers in biology and medicine, vol. 116, p. 103545, 2020.

[64] S. Aladhadh, M. Alsanea, M. Aloraini, T. Khan, S. Habib, and M. Islam, "An effective skin cancer classification mechanism via medical vision transformer," Sensors, vol. 22, no. 11, p. 4008, 2022.

[65] S. Jain, U. Singhania, B. Tripathy, E. A. Nasr, M. K. Aboudaif, and A. K. Kamrani, "Deep learning-based transfer learning for classification of skin cancer," Sensors, vol. 21, no. 23, p. 8142, 2021.

[66] V.-C. Lungu-Stan, D.-C. Cercel, and F. Pop, "Skindistilvit: Lightweight vision transformer for skin lesion classification," in International Conference on Artificial Neural Networks. Springer, 2023, pp. 268–280.

[67] C. Flosdorf, J. Engelker, I. Keller, and N. Mohr, "Skin cancer detection utilizing deep learning: Classification of skin lesion images using a vision transformer," arXiv preprint arXiv:2407.18554, 2024.

[68] G. S. Krishna, K. Supriya, M. Sorgile et al., "Lesionaid: Vision transformers-based skin lesion generation and classification," arXiv preprint arXiv:2302.01104, 2023.

[69] L. Zhou and Y. Luo, "Deep features fusion with mutual attention transformer for skin lesion diagnosis," in 2021 IEEE international conference on image processing (ICIP). IEEE, 2021, pp. 3797–3801.

[70] X. Zhang, Y. Liu, G. Ouyang, W. Chen, A. Xu, T. Hara, X. Zhou, and D. Wu, "Dermvit: Diagnosis-guided vision transformer for robust and efficient skin lesion classification," Bioengineering, vol. 12, no. 4, p. 421, 2025.

[71] S. Iqbal, M. Zeeshan, M. Mehmood, T. Khan, and I. Razzak, "Tesl-net: a transformer-enhanced cnn for accurate skin lesion segmentation," in 2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2024, pp. 313–320.

[72] D. P. Yadav, B. Sharma, S. Chauhan, J. L. Webber, and A. Mehbodniya, "Dual scale light weight cross attention transformer for skin lesion classification," PloS one, vol. 19, no. 12, p. e0312598, 2024.

[73] A. Kumar, K. R. Kanthen, and J. John, "Gs-transunet: integrated 2d gaussian splatting and transformer unet for accurate skin lesion analysis," in Medical Imaging 2025: Computer-Aided Diagnosis, vol. 13407. SPIE, 2025, pp. 790–800.

[74] C. Xin, Z. Liu, K. Zhao, L. Miao, Y. Ma, X. Zhu, Q. Zhou, S. Wang, L. Li, F. Yang et al., "An improved transformer network for skin cancer classification," Computers in Biology and Medicine, vol. 149, p. 105939, 2022.

[75] V. Ravi, T. J. Alahmadi, T. Stephan, P. Singh, M. Diwakar et al., "Deepscan: Integrating vision transformers for advanced skin lesion diagnostics," The Open Dermatology Journal, vol. 18, no. 1, 2024.

[76] A. Mahbod, R. Ecker, and R. Woitek, "Fusion of foundation and vision transformer model features for dermatoscopic image classification," arXiv preprint arXiv:2505.16338, 2025.

[77] K. Tang, J. Su, R. Chen, R. Huang, M. Dai, and Y. Li, "Skinswinvit: A lightweight transformer-based method for multiclass skin lesion classification with enhanced generalization capabilities," Applied Sciences, vol. 14, no. 10, p. 4005, 2024.

[78] M. Shafiq, K. Aggarwal, J. Jayachandran, G. Srinivasan, R. Boddu, and A. Alemayehu, "Retracted: A novel skin lesion prediction and classification technique: Vit-gradcam," Skin Research and Technology, vol. 30, no. 9, p. e70040, 2024.

[79] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" Advances in neural information processing systems, vol. 34, pp. 12 116–12 128, 2021.

[80] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," ACM computing surveys (CSUR), vol. 54, no. 10s, pp. 1–41, 2022.

[81] S. Bhojanapalli, A. Chakrabarti, D. Glasner, D. Li, T. Unterthiner, and A. Veit, "Understanding robustness of transformers for image classification," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10 231–10 241.

[82] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. Shahbaz Khan, and M.-H. Yang, "Intriguing properties of vision transformers," Advances in Neural Information Processing Systems, vol. 34, pp. 23 296–23 308, 2021.

[83] J. M. J. Valanarasu, P. Oza, I. Hacihaliloglu, and V. M. Patel, "Medical transformer: Gated axial-attention for medical image segmentation," in International conference on medical image computing and computer-assisted intervention. Springer, 2021, pp. 36–46.

[84] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.

[85] H. Masood, A. Naseer, and M. Saeed, "Optimized skin lesion segmentation: Analysing deeplabv3+ and assp against generative ai-based deep learning approach," Foundations of Science, pp. 1–25, 2024.

[86] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.

[87] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

[88] J. Wang, L. Wei, L. Wang, Q. Zhou, L. Zhu, and J. Qin, "Boundary-aware transformers for skin lesion segmentation," in Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24. Springer, 2021, pp. 206–216.

[89] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, and Z. Wen, "Fat-net: Feature adaptive transformers for automated skin lesion segmentation," Medical image analysis, vol. 76, p. 102327, 2022.

[90] M. Hu, Y. Li, and X. Yang, "Skinsam: Empowering skin cancer segmentation with segment anything model," arXiv preprint arXiv:2304.13973, 2023.

[91] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo et al., "Segment anything," in Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 4015–4026.

[92] R. Gu, G. Wang, T. Song, R. Huang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, "Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation," IEEE transactions on medical imaging, vol. 40, no. 2, pp. 699–711, 2020.

[93] T. M. Khan, D. Lin, S. Iqbal, and E. Meijering, "Tafm-net: A novel approach to skin lesion segmentation using transformer attention and focal modulation," arXiv preprint arXiv:2411.17556, 2024.

[94] S. Qamar, S. F. Qadri, R. Alroobaea, G. M. M. Alshmrani, and R. Jiang, "Scalefusionnet: Transformer-guided multi-scale feature fusion for skin lesion segmentation," arXiv preprint arXiv:2503.03327, 2025.

[95] R. Xu, C. Wang, J. Zhang, S. Xu, W. Meng, and X. Zhang, "Skinformer: Learning statistical texture representation with transformer for skin lesion segmentation," IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 10, pp. 6008–6018, 2024.

[96] M. Benčević, I. Galić, M. Habijan, and D. Babin, "Training on polar image transformations improves biomedical image segmentation," IEEE access, vol. 9, pp. 133 365–133 375, 2021.

[97] A. Bozorgpour, Y. Sadegheih, A. Kazerouni, R. Azad, and D. Merhof, "Dermosegdiff: A boundary-aware segmentation diffusion model for skin lesion delineation," in International workshop on predictive intelligence in medicine. Springer, 2023, pp. 146–158.

[98] S. Perera, Y. Erzurumlu, D. Gulati, and A. Yilmaz, "Mobileunetr: A lightweight end-to-end hybrid vision transformer for efficient medical image segmentation," arXiv preprint arXiv:2409.03062, 2024.

[99] V. S. Narayanan, O. Sikha, and R. Benitez, "Iars segnet: interpretable attention residual skip connection segnet for melanoma segmentation," IEEE access, 2024.

[100] Y. Ding, Z. Yi, J. Xiao, M. Hu, Y. Guo, Z. Liao, and Y. Wang, "Cth-net: A cnn and transformer hybrid network for skin lesion segmentation," Iscience, vol. 27, no. 4, 2024.

[101] R. Azad, E. K. Aghdam, A. Rauland, and A. Bozorgpour, "Medical image segmentation review: The success of u-net," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[102] J. Wang, F. Chen, Y. Ma, L. Wang, Z. Fei, J. Shuai, X. Tang, Q. Zhou, and J. Qin, "Xbound-former: Toward cross-scale boundary modeling in transformers," IEEE Transactions on Medical Imaging, vol. 42, no. 6, pp. 1735–1745, 2023.

[103] M. E. Celebi, N. Codella, and A. Halpern, "Dermoscopy image analysis: overview and future directions," IEEE journal of biomedical and health informatics, vol. 23, no. 2, pp. 474–478, 2019.

[104] X. Yang, Z. Zeng, S. Y. Yeo, C. Tan, H. L. Tey, and Y. Su, "A novel multi-task deep learning model for skin lesion segmentation and classification," arXiv preprint arXiv:1703.01025, 2017.

[105] J. Chen, J. Chen, Z. Zhou, B. Li, A. Yuille, and Y. Lu, "Mt-transunet: Mediating multi-task tokens in transformers for skin lesion segmentation and classification," arXiv preprint arXiv:2112.01767, 2021.

[106] J. Amin, M. Azhar, H. Arshad, A. Zafar, and S.-H. Kim, "Skin-lesion segmentation using boundary-aware segmentation network and classification based on a mixture of convolutional and transformer neural networks," Frontiers in Medicine, vol. 12, p. 1524146, 2025.

[107] G. M. S. Himel, M. M. Islam, K. A. Al-Aff, S. I. Karim, and M. K. U. Sikder, "Skin cancer segmentation and classification using vision transformer for automatic analysis in dermatoscopy-based noninvasive digital system," International Journal of Biomedical Imaging, vol. 2024, no. 1, p. 3022192, 2024.

[108] K. Manzoor, N. U. Gilal, M. Agus, and J. Schneider, "Dual-stage segmentation and classification framework for skin lesion analysis using deep neural network," Digital Health, vol. 11, p. 20552076251351858, 2025.

[109] Y. Zhang, Y. Xie, H. Wang, J. C. Avery, M. L. Hull, and G. Carneiro, "A novel perspective for multi-modal multi-label skin lesion classification," in 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025, pp. 3549–3558.

[110] Y. Xie, J. Zhang, Y. Xia, and C. Shen, "A mutual bootstrapping model for automated skin lesion segmentation and classification," IEEE transactions on medical imaging, vol. 39, no. 7, pp. 2482–2493, 2020.

[111] D. K. Saha, A. M. Joy, and A. Majumder, "Yotransvit: A transformer and cnn method for predicting and classifying skin diseases using segmentation techniques," Informatics in Medicine Unlocked, vol. 47, p. 101495, 2024.

[112] M. A. Al-Masni, A. K. Al-Shamiri, D. Hussain, and Y. H. Gu, "A unified multi-task learning model with joint reverse optimization for simultaneous skin lesion segmentation and diagnosis," Bioengineering, vol. 11, no. 11, p. 1173, 2024.

[113] A. Al Mahmud, S. Azam, I. U. Khan, S. Montaha, A. Karim, A. Haque, M. Zahid Hasan, M. Brady, R. Biswas, and M. Jonkman, "Skinnet-14: a deep learning framework for accurate skin cancer classification using low-resolution dermoscopy images with optimized training time," Neural Computing and Applications, vol. 36, no. 30, pp. 18 935–18 959, 2024.

[114] A. Ahmed, G. Sun, A. Bilal, Y. Li, and S. A. Ebad, "Precision and efficiency in skin cancer segmentation through a dual encoder deep learning model," Scientific Reports, vol. 15, no. 1, p. 4815, 2025.

[115] S. Ruder, "An overview of multi-task learning in deep neural networks," arXiv preprint arXiv:1706.05098, 2017.

[116] S. Vandenhende, S. Georgoulis, W. Van Gansbeke, M. Proesmans, D. Dai, and L. Van Gool, "Multi-task learning for dense prediction tasks: A survey," IEEE transactions on pattern analysis and machine intelligence, vol. 44, no. 7, pp. 3614–3633, 2021.

[117] Z. Ge, S. Demyanov, R. Chakravorty, A. Bowling, and R. Garnavi, "Skin disease recognition using deep saliency features and multimodal learning of dermoscopy and clinical images," in Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20. Springer, 2017, pp. 250–258.

[118] L. Bi, D. D. Feng, M. Fulham, and J. Kim, "Multi-label classification of multi-modality skin lesion via hyper-connected convolutional neural network," vol. 107, p. 107502, 2020.

[119] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution efficientnets with meta data," MethodsX, vol. 7, p. 100864, 2020.

[120] A. G. Pacheco and R. A. Krohling, "An attention-based mechanism to combine images and metadata in deep learning models applied to skin cancer classification," IEEE journal of biomedical and health informatics, vol. 25, no. 9, pp. 3554–3563, 2021.

[121] P. Tang, X. Yan, Y. Nan, S. Xiang, S. Krammer, and T. Lasser, "Fusionm4net: A multi-stage multi-modal learning algorithm for multi-label skin lesion classification," Medical Image Analysis, vol. 76, p. 102307, 2022.

[122] Y. Wang, Y. Feng, L. Zhang, J. T. Zhou, Y. Liu, R. S. M. Goh, and L. Zhen, "Adversarial multimodal fusion with attention mechanism for skin lesion classification using clinical and dermoscopic images," Medical Image Analysis, vol. 81, p. 102535, 2022.

[123] C. Ou, S. Zhou, R. Yang, W. Jiang, H. He, W. Gan, W. Chen, X. Qin, W. Luo, X. Pi et al., "A deep learning based multimodal fusion model for skin lesion diagnosis using smartphone collected clinical images and metadata," Frontiers in Surgery, vol. 9, p. 1029991, 2022.

[124] J. Xu, Y. Gao, W. Liu, K. Huang, S. Zhao, L. Lu, X. Wang, X.-S. Hua, Y. Wang, and X. Chen, "Remixformer: a transformer model for precision skin tumor differential diagnosis via multi-modal imaging and non-imaging data," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2022, pp. 624–633.

[125] J. Xu, K. Huang, L. Zhong, Y. Gao, K. Sun, W. Liu, Y. Zhou, W. Guo, Y. Guo, Y. Zou et al., "Remixformer++: A multi-modal transformer model for precision skin tumor differential diagnosis with memory-efficient attention," IEEE Transactions on Medical Imaging, 2024.

[126] G. Cai, Y. Zhu, Y. Wu, X. Jiang, J. Ye, and D. Yang, "A multimodal transformer to fuse images and metadata for skin disease classification," The Visual Computer, vol. 39, no. 7, pp. 2781–2793, 2023.

[127] T. Cheslerean-Boghiu, M.-E. Fleischmann, T. Willem, and T. Lasser, "Transformer-based interpretable multi-modal data fusion for skin lesion classification," arXiv preprint arXiv:2304.14505, 2023.

[128] Y. Zhang, F. Xie, and J. Chen, "Tformer: A throughout fusion transformer for multi-modal skin lesion diagnosis," Computers in biology and medicine, vol. 157, p. 106712, 2023.

[129] A. Adebiyi, N. Abdalnabi, E. H. Smith, J. Hirner, E. J. Simoes, M. Becevic, and P. Rao, "Accurate skin lesion classification using multimodal learning on the ham10000 dataset," MedRxiv, pp. 2024–05, 2024.

[130] D. Christopoulos, S. Spanos, E. Baltzi, V. Ntouskos, and K. Karantzalos, "Skin lesion phenotyping via nested multi-modal contrastive learning," arXiv preprint arXiv:2505.23709, 2025.

[131] N. L. T. Pham, D. D. Pham, T. D. Le, and K. T. Huynh, "A multimodal deep ensemble framework for skin lesion classification," in International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making. Springer, 2025, pp. 100–111.

[132] B. Banothu, J. Tulasiram, N. S, and G. Patil, "Multimodal deep learning framework for skin lesion classification," in International Conference on Computer Vision and Image Processing. Springer, 2024, pp. 206–216.

[133] S. Ahammed, X. Cui, W. Lu, and M. H. Yap, "Skin lesion classification using dermoscopic images and clinical metadata: Insights from multimodal models," in Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 222–230.

[134] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," Nature Machine Intelligence, vol. 2, no. 11, pp. 665–673, 2020.

[135] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 2, pp. 423–443, 2018.

[136] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "Hemis: Hetero-modal image segmentation," in International conference on medical image computing and computer-assisted intervention. Springer, 2016, pp. 469–477.

[137] A. Bissoto, E. Valle, and S. Avila, "Gan-based data augmentation and anonymization for skin-lesion analysis: A critical review," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 1847–1856.

[138] A. Simkó, A. Garpebring, J. Jonsson, T. Nyholm, and T. Löfstedt, "Reproducibility of the methods in medical imaging with deep learning." in Medical Imaging with Deep Learning. PMLR, 2024, pp. 95–106.

[139] J. Mongan, L. Moy, and C. E. Kahn Jr, "Checklist for artificial intelligence in medical imaging (claim): a guide for authors and reviewers," p. e200029, 2020.

[140] Z. Lan, S. Cai, X. He, and X. Wen, "Fixcaps: An improved capsules network for diagnosis of skin cancer," IEEE Access, vol. 10, pp. 76 261–76 267, 2022.

[141] M. Gallazzi, S. Biavaschi, A. Bulgheroni, T. Gatti, S. Corchs, and I. Gallo, "A large dataset to enhance skin cancer classification with transformer-based deep neural networks," IEEE Access, 2024.

[142] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the isic image datasets: Usage, benchmarks and recommendations," Medical image analysis, vol. 75, p. 102305, 2022.

[143] T. T. Cai and R. Ma, "Theoretical foundations of t-sne for visualizing high-dimensional clustered data," Journal of Machine Learning Research, vol. 23, no. 301, pp. 1–54, 2022.

[144] Ultralytics, "Yolov8 release notes," https://github.com/ultralytics/yolov8, 2024, available: https://github.com/ultralytics/yolov8.

[145] S. Paulsen and M. Casey, "Sequential transfer learning to decode heard and imagined timbre from fmri data," arXiv preprint arXiv:2305.13226, 2023.

[146] J. Y.-L. Chan, K. T. Bea, S. M. H. Leow, S. W. Phoong, and W. K. Cheng, "State of the art: a review of sentiment analysis based on sequential transfer learning," Artificial Intelligence Review, vol. 56, no. 1, pp. 749–780, 2023.

[147] Y. Wang, J. Su, Q. Xu, and Y. Zhong, "A collaborative learning model for skin lesion segmentation and classification," Diagnostics, vol. 13, no. 5, p. 912, 2023.

[148] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" IEEE transactions on medical imaging, vol. 35, no. 5, pp. 1299–1312, 2016.

[149] X. He, Y. Wang, S. Zhao, and X. Chen, "Joint segmentation and classification of skin lesions via a multi-task learning convolutional neural network," Expert Systems with Applications, vol. 230, p. 120174, 2023.

[150] S. Mustafa, A. Jaffar, M. Rashid, S. Akram, and S. M. Bhatti, "Deep learning-based skin lesion analysis using hybrid resunet++ and modified alexnet-random forest for enhanced segmentation and classification," PloS one, vol. 20, no. 1, p. e0315120, 2025.

[151] M. A. Khan, M. Sharif, T. Akram, R. Damaševičius, and R. Maskeliūnas, "Skin lesion segmentation and multiclass classification using deep learning features and improved moth flame optimization," Diagnostics, vol. 11, no. 5, p. 811, 2021.

[152] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in 2018 IEEE winter conference on applications of computer vision (WACV). IEEE, 2018, pp. 839–847.

[153] K. Pogorelov, K. R. Randel, C. Griwodz, S. L. Eskeland, T. de Lange, D. Johansen, C. Spampinato, D.-T. Dang-Nguyen, M. Lux, P. T. Schmidt et al., "Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection," in Proceedings of the 8th ACM on Multimedia Systems Conference, 2017, pp. 164–169.

[154] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti et al., "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," arXiv preprint arXiv:1902.03368, 2019.

[155] I. C. 2018, "Ham10000 test set release," https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T, Last accessed on 2024-05-2020.

[156] S. Mushtaq and O. Singh, "A deep learning based architecture for multi-class skin cancer classification," Multimedia Tools and Applications, vol. 83, no. 39, pp. 87 105–87 127, 2024.

[157] S. S. Chaturvedi, J. V. Tembhurne, and T. Diwan, "A multi-class skin cancer classification using deep convolutional neural networks," Multimedia Tools and Applications, vol. 79, no. 39, pp. 28 477–28 498, 2020.

[158] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. De Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in MultiMedia modeling: 26th international conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, proceedings, part II 26. Springer, 2020, pp. 451–462.

[159] R.-G. Dumitru, D. Peteleaza, and C. Craciun, "Using duck-net for polyp image segmentation," Scientific reports, vol. 13, no. 1, p. 9803, 2023.

[160] I. A. Vezakis, K. Georgas, D. Fotiadis, and G. K. Matsopoulos, "Effisegnet: Gastrointestinal polyp segmentation through a pre-trained efficientnet-based network with a simplified decoder," arXiv preprint arXiv:2407.16298, 2024.

[161] R. Fonolla, F. van der Sommen, R. M. Schreuder, E. J. Schoon, and P. H. de With, "Multimodal classification of polyp malignancy using cnn features with balanced class augmentation," in 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, 2019, pp. 74–78.

[162] X. Zhang, F. Chen, T. Yu, J. An, Z. Huang, J. Liu, W. Hu, L. Wang, H. Duan, and J. Si, "Real-time gastric polyp detection using convolutional neural networks," PloS one, vol. 14, no. 3, p. e0214133, 2019.

[163] X. Liu, C. Wang, J. Bai, and G. Liao, "Fine-tuning pre-trained convolutional neural networks for gastric precancerous disease classification on magnification narrow-band imaging images," Neurocomputing, vol. 392, pp. 253–267, 2020.

[164] Z. M. Lonseko, P. E. Adjei, W. Du, C. Luo, D. Hu, L. Zhu, T. Gan, and N. Rao, "Gastrointestinal disease classification in endoscopic images using attention-guided convolutional neural networks," Applied Sciences, vol. 11, no. 23, p. 11136, 2021.

[165] A. A. Demirbaş, H. Üzen, and H. Fırat, "Spatial-attention convmixer architecture for classification and detection of gastrointestinal diseases using the kvasir dataset," Health Information Science and Systems, vol. 12, no. 1, p. 32, 2024.

[166] N. R. Kurtansky, B. M. D'Alessandro, M. C. Gillis, B. Betz-Stablein, S. E. Cerminara, R. Garcia, M. A. Girundi, E. V. Goessinger, P. Gottfrois, P. Guitera et al., "The slice-3d dataset: 400,000 skin lesion image crops extracted from 3d tbp for skin cancer detection," Scientific Data, vol. 11, no. 1, p. 884, 2024.

[167] A. Saha, J. Adeola, N. Ferrera, A. Mothershaw, G. Rezze, S. Gaborit, B. D'Alessandro, R. Voskanyan, G. Szabó, B. Pataki et al., "Skin region images extracted from 3d total body photographs for lesion detection," Scientific Data, vol. 12, no. 1, p. 1442, 2025.

[168] F. S. GmbH, "Fotofinder systems – cutting-edge skin imaging technology," https://www.fotofinder-systems.com/, Bad Birnbach, Germany, n.d., accessed July 2025.

[169] W. Wang, "Advanced auto labeling solution with added features," https://github.com/CVHub520/X-AnyLabeling, CVHub, 2023.

[170] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," 2024. [Online]. Available: https://arxiv.org/abs/2408.00714

[171] G. Jocher and J. Qiu, "Ultralytics yolo11," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics

[172] N. Jegham, C. Y. Koh, M. Abdelatti, and A. Hendawi, "Evaluating the evolution of yolo (you only look once) models: A comprehensive benchmark study of yolo11 and its predecessors," arXiv preprint arXiv:2411.00201, 2024.

[173] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask DINO: Towards a unified transformer-based framework for object detection and segmentation," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 3041–3050.

[174] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in Computer Vision – ECCV 2014, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

Acknowledgements

These acknowledgments are intentionally concise but deeply sincere, aiming to express my genuine appreciation to all those who have contributed, directly or indirectly, to this work and to my personal growth during the Ph.D.

I would first like to express my deepest gratitude to my supervisor, Prof. Silvia Corchs, and my co-supervisor, Prof. Ignazio Gallo, for their guidance and support throughout these three years of doctoral research. Their feedback, advice, and encouragement have been invaluable, and every discussion and critique has helped me grow both academically and personally.

I would also like to sincerely thank Dr. Nicola Zerbinati and Dr. Andrea Carugno from the Department of Medicine and Technological Innovation at the University of Insubria, Varese, Italy, for their openness and willingness to collaborate throughout this Ph.D. journey. Their clinical perspective and enthusiasm made it possible to build something meaningful and, hopefully, useful.

I would also like to acknowledge Prof. Francesca Gasparini and Aurora Saibene for their availability during the first year of my Ph.D. and for the opportunity to collaborate on research topics beyond those addressed in this dissertation.

I would also like to express my sincere gratitude to the members of the examination committee, Prof. Jordi Solè Casals and Prof. Domenico Giorgio Sorrenti, for the time and attention they devoted to reviewing my dissertation and for their insightful comments, which have greatly contributed to refining this work.

My heartfelt thanks go to my parents for their constant love and support, for listening to my endless worries, and for being by my side through every challenge. A very special thank you goes to my mother, who, for better or worse, has borne the full extent of my frustrations with infinite patience.

I am also deeply grateful to all my friends — the Fridayay group, Trasher, Ninja Norvegesi, and everyone else I may have forgotten to mention — for being part of my life, sharing adventures, and bringing laughter and lightness along the way.

A special thank you goes to Elena, Eleonora, Giorgia, Roberto, and Simone — the people who have endured most of my struggles, each in their own way, some from the beginning, while others joined the journey along the way. Thank you from the heart.

I would like to thank all my colleagues and fellow Ph.D. students for making the department such a welcoming and inspiring place, even during the most challenging times. A special mention goes to Mattia and his always captivating stories, which brightened even the longest days.

I would also like to thank all the people at the Tempio dei Gladiatori, who, over the past year

and a half, have welcomed me into their CrossFit community — an endless source of fatigue, yet essential for releasing all the accumulated tension.

I would also like to thank the Nuotatori del Carroccio for welcoming me into the team and for rekindling my love for swimming, allowing me to rediscover the joy of practicing sport in the water within a motivating and supportive group.

Finally, I would like to thank myself for holding on, for not giving up in the hardest moments, and for bringing this chapter of my life to completion.