OmniFusion:
A Hybrid Deep Learning Foundation Model for Skin Cancer Diagnostics

by

Joshua D. Arnow

Submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Computer Science

at

Seidenberg School of Computer Science
and Information Systems
Pace University

May 2026

Abstract

OmniFusion:
A Hybrid Deep Learning Foundation
Model for Skin Cancer Diagnostics

by

Joshua D. Arnow

Submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Computer Science

May 2026

Skin cancer is the most prevalent form of cancer globally, with millions of new cases diagnosed annually. Early detection is critical for reduced mortality, improved patient outcomes, and minimization of associated healthcare costs. However, access to dermatological expertise is often limited, particularly in underserved regions. While deep learning models have achieved expert-level diagnostic accuracy, their deployment in real-world clinical settings has been hindered by challenges related to model robustness, architectural bottlenecks, and concerns regarding cost-effectiveness and safety of implementation. This study introduces OmniFusion, a highly robust, hybrid deep learning foundation model designed to address these challenges by synthesizing the most effective architectural components of leading computer vision models while advancing the state of the art in computer-aided medical diagnostics. Two top-performing architectures, a pure vision transformer approach used by the PanDerm model and an adaptive attention-based feature fusion mechanism used by the SkinEHDLF model, are integrated into the OmniFusion framework to leverage and compare their unique strengths. This study facilitates a rigorous four-phase comparative evaluation between the two architectural approaches across unimodal and multimodal training data. To optimize performance for high-variance clinical environments, the OmniFusion framework integrates a linear probing then fine-tuning (LP-FT) pipeline alongside advanced spatial augmentation techniques, namely Mixup and CutMix.

Empirical results definitively confirm the architectural superiority of the adaptive attention-based feature fusion mechanism, which achieved a peak AUROC of 95.01% on the SLICE-3D (ISIC 2024) dataset, significantly outperforming the pure vision transformer approach on the binary classification task. Furthermore, this study investigates the complex dynamics of dataset multimodality, revealing that while multimodal training holds immense potential, the inclusion of severely domain-shifted modalities can induce modality collapse and degrade cross-domain generalization. Finally, this study examines the pre-clinical viability of the OmniFusion model by comparing its performance against peer-reviewed quantitative thresholds for sensitivity, specificity, and cost-effectiveness, ultimately demonstrating that it is suitable for further validation and deployment. This research establishes OmniFusion as a scalable, decentralized diagnostic foundation model, bridging the gap between theoretical performance and deployable, real-world healthcare applications.

Contents

Figures

Tables

Acknowledgments

I would like to express my deepest appreciation to those who have supported me throughout my journey of completing this thesis.

I am especially grateful to God for blessing me with the perseverance and patience to undertake this pursuit, as well as for the countless blessings and opportunities that He has provided outside of academia.

I am very thankful for my family, namely my mom, dad, Grandma Fran, Grandpa Joel, and Aunt Janice, for their unwavering love, encouragement, and support in this and all other endeavors in my life.

I greatly appreciate the immense guidance that my thesis advisor, Dr. Juan Shan, has consistently provided throughout this process, as well as the aid and time of the remaining members of my thesis committee, Dr. Lixin Tao and Dr. Paul Benjamin.

This thesis would not have been possible without the pioneering work of the authors of the PanDerm and SkinEHDLF models, Yan et al. (2025) and Lilhore et al. (2025), whose groundbreaking research propelled the state of the art in computer-aided medical diagnostics and laid the foundation for this study and the OmniFusion model.

I also appreciate my friends who have provided encouragement throughout this process and remained understanding when I forwent socializing in order to focus on this project.

Thank you all for being an integral part of this journey!

Abbreviations

AI	Artificial Intelligence
AUROC	Area Under the Receiver Operating Characteristic (Curve)
BAcc	Balanced Accuracy
CLAHE	Contrast Limited Adaptive Histogram Equalization
CLIP	Contrastive Language-Image Pretraining
CNN	Convolutional Neural Network
DTC	Direct-to-Consumer
FN	False Negative
FP	False Positive
HAM10000	Human Against Machine with 10,000 training images
HPC	High-Performance Computing
ICER	Incremental Cost-Effectiveness Ratio
ISIC	International Skin Imaging Collaboration
LP-FT	Linear Probing then Fine-tuning
NTK	Neural Tangent Kernel
TBP	Total Body Photography
SkinEHDLF	Skin Enhanced Deep Learning Framework
SLICE-3D	Skin Lesion Image Crops Extracted from 3D TBP

TN	True Negative
TNR	True Negative Rate (Specificity)
TP	True Positive
TPR	True Positive Rate (Sensitivity)
ViT	Vision Transformer

Chapter 1
Introduction

1.1 Background and Problem Statement

Skin cancer is the most common form of cancer in the United States and the world (*Skin Cancer Facts* 2026). For melanoma, metastasis to the lymph nodes is associated with a 5-year survival rate of 76% whereas distant metastasis substantially cuts the survival rate to 35% (*Skin Cancer Facts* 2026). However, when detected early, the 5-year survival rate for melanoma is over 99% (*Skin Cancer Facts* 2026). Early melanoma detection is not only important to those affected by the disease, but to national healthcare systems as well due to the knock-on effects stemming from high-cost, potentially long-term treatment regimens for metastatic cancer. Every cancer detected and treated early corresponds with a patient who has better odds for survival at a lower cost for treatment, as well as at least one clinician who has more capacity to treat other patients. Achieving this widespread early detection requires scalable diagnostic tools. In recent years, state-of-the-art deep learning models designed to detect skin cancer have improved to the extent that their performance has been found to exceed that of the most skilled dermatologists (Esteva et al. 2017; Yan et al. 2025). The proliferation of smartphone technology and availability of high-resolution mobile cameras present an unprecedented opportunity to shift initial diagnostic screening into pre-clinical environments via direct-to-consumer (DTC) model deployment. Such a maneuver could enable more accessible healthcare not only for the affected individual but for every potential patient through reduced insurance costs and more efficient triaging, which reduces clinician caseload.

Despite the incredible promise of decentralized diagnostics, deployment of automated skin lesion classification systems in uncontrolled environments introduces several challenges. An image captured by a layman can suffer from domain shifts, poor lighting, occlusion, and inferior resolution compared to standard clinical dermatoscopy. Therefore, a model placed in the hands of laymen must possess a high degree of robustness and generalizability to prevent false positives from

overburdening the healthcare system and, more importantly, to prevent false negatives that can endanger patient lives. Additionally, laymen must be educated on the implications of the results reported by the model and maintain regular checkups with a trained clinician to mitigate risks associated with false negative results. If those conditions are met, the mass adoption of automated skin lesion classification systems could lead to immeasurable benefits for public health at large. The primary aim of this thesis is to determine the most effective aspects of state-of-the-art models and to synthesize them into a single, cohesive predictive diagnostic tool. Furthermore, this thesis aims to isolate and evaluate the impact of multimodal training data on model performance to provide deeper insight into its role in model development. Once implemented, the performance of the model will be analyzed to determine if it is suitable for real-world, DTC deployment based on an established set of criteria.

1.2 Research Gap

Although state-of-the-art diagnostic models have managed to exceed the performance of skilled dermatologists, they have done so through substantially different approaches (Esteva et al. 2017; Lilhore et al. 2025; Yan et al. 2025). Model training on massive, diverse datasets that have only recently been made available is likely a significant factor in the performance exhibited by the models, yet the degree to which dataset multimodality impacts performance has not been closely studied. A fair comparison of state-of-the-art models that controls for training, validation, and test datasets has not been made either; therefore, it remains unclear as to which model holds the architectural advantage and thus which has the potential for superior performance. In addition, the readiness of state-of-the-art models for pre-clinical deployment has only been superficially analyzed across fragmented qualitative and quantitative studies. Ultimately, the literature lacks a cohesive framework that synthesizes the advantages of one model with the advantages of the other, which is a prerequisite in designing the most effective pre-clinical diagnostic tool optimized for high-variance data inputs.

1.3 Proposed Solution

This study attempts to address several important questions. First, it seeks to discover the significance of multimodal training data in deep learning models designed to classify skin lesions in photographs as benign or malignant. By training each of the two most performant state-of-the-art backbone architectures, PanDerm and SkinEHDLF, on unimodal and multimodal datasets, intra-architectural performance differences can be measured and used to assess the impact of multimodal data. The study also seeks to discover which architecture leads to superior performance when comparing PanDerm and SkinEHDLF; by controlling for the datasets used by the models, a comparison can be made that reveals architectural advantages. Third, the study seeks to determine if the models are ready for pre-clinical deployment using their performance metrics and generally agreed-upon performance criteria for DTC diagnostic tools.

This study also proposes OmniFusion: a hybrid deep learning model that combines the multimodal data pipeline supported by PanDerm with the adaptive attention-based fusion mechanism of SkinEHDLF, all while incorporating additional techniques for improved robustness and generalizability that were not used by the original PanDerm and SkinEHDLF authors. By integrating “linear probing then fine-tuning” (LP-FT) rather than simply fine-tuning, time to convergence can be reduced, which is critical given limited resources. The introduction of the Mixup and CutMix spatial augmentations during training also aids in maximizing confidence calibration and out-of-distribution generalizability, which are crucial when considering mass DTC deployment.

The name “OmniFusion" was chosen to reflect a synthesis of PanDerm and SkinEHDLF. Like the Greek-derived “Pan" prefix used for the PanDerm model, “Omni" is a Latin-derived prefix that means “all" or “every," emphasizing the model’s ability to integrate various data modalities through a well-defined data pipeline. The “Fusion" component of the name reflects the model’s integration of a Feature Fusion Layer adopted from the SkinEHDLF model, which facilitates the adaptive attention-based fusion of features from three independent convolutional neural networks (CNNs) and vision transformers.

It must be noted that when PanDerm and SkinEHDLF are discussed in the context of experimentation within this study, it refers to their custom implementation as part of the OmniFusion model. That is, the LP-FT and Mixup/CutMix techniques that make the OmniFusion model distinct are applied for each experiment of this study but were not used in either of the original publications of the models. The use of "PanDerm" and "SkinEHDLF" in this context refers to the classification architectures used by the original authors of those models: the Vision Transformer (ViT) Large encoder and the adaptive feature fusion approach, respectively.

1.4 Primary Contributions

The specific contributions of this research are as follows:

- **Impact of Dataset Multimodality**: Training the PanDerm and SkinEHDLF models on the large, unimodal SLICE-3D dataset as well as a combined multimodal dataset consisting of the SLICE-3D dataset and other sources will allow for intra-architectural comparisons that elucidate the significance of dataset multimodality on the binary classification task of detecting skin cancer.
- **Performance Comparison of PanDerm and SkinEHDLF Architectures**: The relative performance of the PanDerm and SkinEHDLF architectures will be determined by comparing inter-architectural performance on the same datasets.
- **Pre-Clinical Translation Assessment**: The viability of DTC pre-clinical deployment will be evaluated by comparing the performance of PanDerm and SkinEHDLF against diagnostic thresholds required for safe patient triage.
- **Architectural Synthesis**: The development of OmniFusion, a hybrid deep learning model that bridges the gap between massive multimodal foundation pretraining and highly efficient, adaptive feature fusion networks.
- **Pre-Clinical Optimization**: Advanced data augmentation techniques and training methodologies will be applied to OmniFusion specifically to enhance model robustness

against the noise and domain shifts inherent to photographs taken by laymen in diverse environments.

1.5 Document Outline

This thesis is broken into several chapters. In Chapter 2, a comprehensive literature review is performed to establish the current state of machine learning as applied to skin cancer diagnostics, the viability of such tools when deployed to pre-clinical environments, the foundational datasets used in the field, and modern deep learning model optimization strategies. In Chapter 3, the methodology of the study is elaborated upon and baseline performance of the PanDerm and SkinEHDLF models on the unimodal SLICE-3D dataset is established for further comparison in subsequent chapters. Chapter 4 reports on the performance of the various models implemented in this study. Chapter 5 provides insight into the significance of the results in a discussion format. Therein, the performances of the models are compared against one another and against established benchmarks to determine the impact of dataset multimodality, model architecture, and the feasibility of DTC deployment. In addition, broader implications of the research and potential avenues for future work are discussed. Finally, the key findings of the study are summarized in Chapter 6.

Chapter 2
Literature Review

Recent academic research has revealed that deep learning models can outperform the diagnostic performance of dermatologists when distinguishing between benign and malignant skin lesions (Esteva et al. 2017; Yan et al. 2025). Models have proven to be highly effective both in augmenting the diagnostic capabilities of physicians and as independent tools. While design approaches differ, creators often highlight specific architectural advantages to rationalize their implementations. However, a common thread that runs between them is their reliance on features that have only recently been made possible: very large datasets, modern algorithms, and hardware capable of performing model training in a reasonable amount of time. A primary aim of this study is to synthesize the best aspects of those approaches into a cohesive, state-of-the-art model suitable for deployment in clinical and pre-clinical settings.

2.1 Foundational Datasets and Model Architectures

2.1.1 HAM10000 Dataset

In 2018, a major milestone in the development of clinically viable automated diagnostic systems for skin cancer occurred with the publication of a new dataset by Tschandl, Rosendahl, and Kittler (2018). The Human Against Machine with 10,000 training images (HAM10000) dataset, which consists of 10,015 dermatoscopic images, was published with the intent of "boost[ing] the research on automated diagnosis of dermatoscopic images" and to "serve as [a] benchmark set for the comparisons of humans and machines" (Tschandl, Rosendahl, and Kittler 2018). HAM10000 differed from prior datasets through its combination of large size and its diversity of data, which are both crucial features for training models designed for deployment in clinical or pre-clinical settings. Much like future ISIC datasets such as the SLICE-3D dataset, the HAM10000 dataset would go on to be used in research contexts beyond the scope of the challenge. Because ISIC datasets contain

carefully curated images intended to encourage the development of models trained on them, the ISIC datasets used in this study were selected for closer examination from among all the datasets used.

When Binder et al. (1994) successfully trained an artificial neural network to differentiate melanomas from benign nevi using just 200 images, graphics card capabilities and machine learning techniques at the time were nowhere near as advanced as they would become. By 2018, advances enabled the processing of exponentially larger datasets which in turn would result in substantially better model performance, but high-quality, sufficiently diverse datasets at such a large size did not exist. Although contemporaneous datasets with a larger image count existed, most notably the International Skin Imaging Collaboration (ISIC) archive dataset of 13,786 dermatoscopic images (Tschandl, Rosendahl, and Kittler 2018), they were severely biased towards melanocytic lesions. In the case of the ISIC archive dataset, melanocytic lesions represented roughly 94% of the entire dataset (Tschandl, Rosendahl, and Kittler 2018).

Tschandl, Rosendahl, and Kittler (2018) mitigated the imbalanced nature of past datasets by including images that would fall into one of seven diagnostic categories, of which "more than 95% of all lesion[s] encountered during clinical practice . . . fall into." Their attempt was not perfect: benign nevi still dominated the total image count at about 66.9%, with melanoma images sitting at around 11.1% and the remaining image categories at smaller percentages (Tschandl, Rosendahl, and Kittler 2018) as shown in Table 2.1. However, roughly 78% of images falling into the category of melanocytic lesions was still a substantial improvement over the 94% in the ISIC archive dataset. The significance of diverse categories cannot be overstated: research has shown that "the mismatch between the small diversity of available training data and the variety of real life data result[s] in a moderate performance of automated diagnostic systems in the clinical setting despite excellent performance in experimental settings" (Codella et al. 2018; Dreiseitl et al. 2009; Kharazmi et al. 2017; Mendonca et al. 2013; Tschandl, Rosendahl, and Kittler 2018). The improvements in image diversity that the HAM10000 dataset brought to the field provided a means for machine learning researchers to train models that achieved results indicating they were closer to clinical viability than ever before possible.

Table 2.1: Comparison of Public Dermatology Datasets in 2018

Dataset	Total	Ver. (%) [a]	akiec[b]	bcc[c]	bkl[d]	df[e]	mel[f]	nv[g]	vasc[h]
PH2	200	20.5%	-	-	-	-	40	160	-
Atlas	1,024	unknown	5	42	70	20	275	582	30
ISIC 2017[i]	13,786	26.3%	2	33	575	7	1,019	11,861	15
Rosendahl	2,259	100%	295	296	490	30	342	803	3
ViDIR Legacy	439	100%	0	5	10	4	67	350	3
ViDIR Current	3,363	77.1%	32	211	475	51	680	1,832	82
ViDIR MoleMax	3,954	1.2%	0	2	124	30	24	3,720	54
HAM10000	10,015	53.3%	327	514	1,099	115	1,113	6,705	142

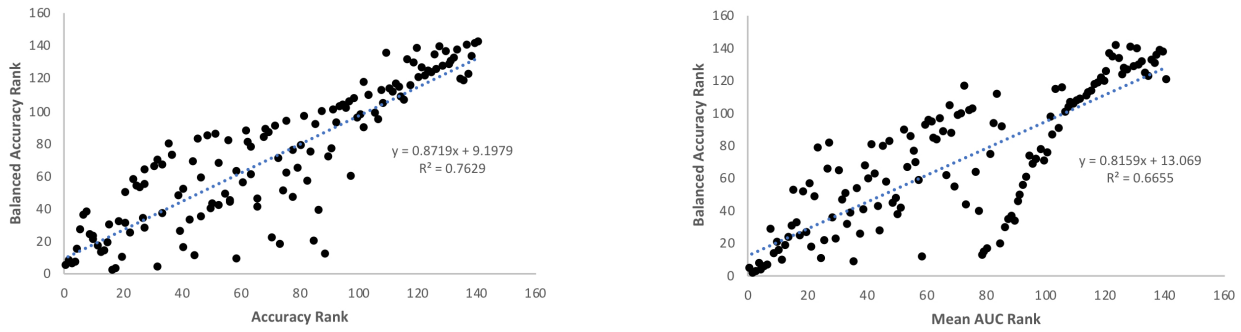
[a]Pathological verification; [b]Actinic Keratoses (Solar Keratoses) and Intraepithelial Carcinoma (Bowen's disease); [c]Basal cell carcinoma; [d]Benign keratosis; [e]Dermatofibroma; [f]Melanoma; [g]Melanocytic nevi; [h]Vascular skin lesions; [i]Eight different datasets with CC-0 licensing combined as available on February 12th 2018.
Source: Adapted from Tschandl, Rosendahl, and Kittler (2018).

2.1.2 Establishment of Algorithmic Baselines

Following its publication, the HAM10000 dataset became the foundation for the 2018 ISIC challenge titled "Skin Lesion Analysis Toward Melanoma Detection" (Codella et al. 2019). It would be the third challenge hosted by ISIC with the aim of "increasing the accuracy and scale of diagnostic methods" to further the early detection of melanoma through machine learning tools (Codella et al. 2019), with several notable changes compared to previous years. Aside from the HAM10000 dataset enabling the use of "considerabl[y]" more training data as well as more diagnostic labels, the challenge organizers implemented balanced accuracy (BAcc) for classification decisions and included external test data from institutions not present in the training dataset (Codella et al. 2019).

The shift from precision and AUC to BAcc reflected a desire to ensure the evaluation of models' performances reflected their robustness across clinical settings (Codella et al. 2019). Since the HAM10000 dataset is imbalanced in its composition of benign and malignant images, weighing each diagnostic category equally regardless of its frequency in the dataset prevents models from appearing performant when in fact they are overfitting to the majority class (Codella et al. 2019). Likewise, the inclusion of external test data allowed researchers to "better assess how algorithms generalize beyond the environments for which they were trained" (Codella et al. 2019).

141 participants engaged in the lesion disease classification task with the highest BAcc recorded at 88.5% (Codella et al. 2019). Notably, results demonstrated that the choice of evaluation metric makes a substantial impact in demonstrating the true performance of a model. As shown in Figure 2.1, the coefficients of determination comparing BAcc with accuracy ($R^2 = 0.76$) and mean AUC ($R^2 = 0.67$) indicate a high degree of divergence between the metrics. The imperfect correlation validates the organizers' concern of relying on standard metrics: by not using accuracy or mean AUC the competitors were ranked in a fundamentally different way, one which arguably is more capable of robustly detecting minority diagnoses in clinically imbalanced settings. The ISIC 2018 challenge proved to be important in establishing new performance benchmarks from a public dataset of unprecedented scope, especially as the HAM10000 dataset continued to be used in research contexts beyond the scope of the challenge.



(a) **(b)**

Figure 2.1: Impact of New Balanced Accuracy (BAcc) Metric on Participant Ranking in the ISIC 2018 Challenge

Source: Adapted from Codella et al. (2019).
Note: **(a)** Comparison of participant rankings using Balanced Accuracy vs. Standard Accuracy. **(b)** Comparison of participant rankings using Balanced Accuracy vs. Mean AUC.

2.1.3 SLICE-3D Dataset

Seven medical institutions from around the world alongside Kurtansky et al. recently facilitated the release of a massive, novel dataset containing 3D total body photographs (3D TBP) with the aim of "facilitat[ing] the development of open-source AI algorithms capable of rendering diagnostic decisions from reduced quality, clinical photos resembling the resolution of smartphone

images" (Kurtansky et al. 2024). Each medical institution is a member of ISIC, a partnership intending to "reduce skin cancer morbidity and mortality through the development and use of digital skin imaging applications" (Kurtansky et al. 2024). The dataset of 401,059 images is referred to as Skin Lesion Image Crops Extracted from 3D TBP (SLICE-3D) in its original publication but is also referred to as ISIC 2024 elsewhere, particularly in research that has made use of it. The creation of clinically suitable, widely deployable AI models for skin cancer detection had previously been hindered by the quality of available datasets; their selection bias, lack of standardization, and reliance on dermatoscopic images typically captured by skilled clinicians negatively impacted their ability to train a robust model (Kurtansky et al. 2024). The SLICE-3D dataset was released with the intention of overcoming those limitations.

Kurtansky et al. (2024) strongly emphasized that the creation of AI algorithms using their dataset has "great potential impact" in improving patient outcomes and reducing the burden on health systems. More specifically, the authors indicated that algorithms trained on it can "improve clinical workflows and detect skin cancers earlier if deployed in primary care or non-clinical settings, where photos are captured by non-expert physicians or patients" (Kurtansky et al. 2024). The authors took steps to standardize the images to "mitigate biases that arise from differences in capture condition specific to the image source" (Kurtansky et al. 2024), furthering their objective of preventing overfitting and producing robust algorithms capable of deployment in non-clinical settings. The unprecedented size and quality of the dataset, along with the explicit encouragement of its authors, was the primary inspiration for this study. The use of the dataset to train two state-of-the-art AI models in the year following its publication further demonstrated its usefulness and a need for further research incorporating it. For a more detailed breakdown of the dataset, refer to Subsection 3.1.1.

2.1.4 PanDerm Model

By 2024, substantial strides had been made both in the quality of public dermatologic databases and in the performance of machine learning models relying on them. The advantages of utilizing multimodal inputs were self-evident when applied to a clinical practice like dermatology,

where various imaging types are used in diagnosing conditions; such advantages had been demonstrated in contemporary research as well (Luo et al. 2023; Yap, Yolland, and Tschandl 2018). Despite the clinical advantages, unimodality was the norm when it came to machine learning models designed to diagnose skin cancer (Yan et al. 2025). In response to a dearth of multimodal dermatological diagnostic models and the plethora of public imaging data that had become available in recent years, Yan et al. developed a multimodal vision foundation model dubbed PanDerm, which was trained on over 2 million images. Consequently, the model managed to achieve state-of-the-art performance across a range of diagnostic tasks (Yan et al. 2025). Although PanDerm was designed to engage in multiple dermatological tasks (Yan et al. 2025), its ability to perform binary melanoma diagnosis is the focus of this study.

PanDerm was designed with four imaging modalities in mind: 3D TBP images, dermatopathology images, clinical images, and dermatoscopic images (Yan et al. 2025). The data used to train the model were derived from eleven sources, many of which were in-house and thus not publicly available. The largest single source of publicly available data was the SLICE-3D dataset; 352,034 images were used during pretraining, representing roughly 16.4% of the complete pretraining dataset (Yan et al. 2025). Yan et al. broke training into two distinct phases: pretraining, where masked latent modeling and contrastive language-image pretraining (CLIP) were used for self-supervised learning, and training, where linear probing or fine-tuning was used to specialize the model to perform specific tasks after loading weights from the pretraining stage (Yan et al. 2025). Notably, the authors further leveraged transfer learning by initializing the model on ImageNet-1K weights prior to pretraining (Yan et al. 2025). The authors also relied on data augmentation techniques such as random resized cropping and horizontal flipping to boost performance during pretraining (Yan et al. 2025). A complete breakdown of the datasets used in Yan's study can be found in Table 2.2.

Table 2.2: Complete Breakdown of PanDerm Datasets

Dataset Name	**Modality**	**Availability**	**Image Count**
Stage 1: Pretraining			
MYM and MYM cohort	TBP	Private	405,856
ISIC 2024 (SLICE-3D)	TBP	Public	352,034
WSI (derived from TCGA-SKCM)	Dermatopathology	Public	377,764
Edu1	Clinical	Private	81,947
Edu2	Clinical	Private	67,430
MMT dataset	Clinical	Private	310,951
ACEMID pathology pilot study	Dermatopathology	Private	80,312
UAH89k	Dermatopathology	Public	88,971
NSSI	Dermoscopic	Private	29,832
MYM and HOP cohort	Dermoscopic	Public	38,110
MMT dataset	Dermoscopic	Private	316,399
Subtotal			*2,149,606*
Stage 2: Training / Evaluation			
ISIC 2024 (SLICE-3D)	TBP	Public	49,025
PATCH16	Dermatopathology	Public	129,364
WSI (derived from TCGA-SKCM)	Dermatopathology	Public	302
DDI	Clinical	Public	647
Derm7pt	Clinical	Public	839
DermNet	Clinical	Public	19,559
Fitzpatrick17K	Clinical	Public	16,577
Med-Node	Clinical	Public	170
MMT-09	Clinical	Private	38,476
MMT-74	Clinical	Private	38,476
PAD-UFES-20	Clinical	Public	2,298
PH2	Clinical	Public	200
SD-128	Clinical	Public	5,619
BCN20000	Dermoscopic	Public	12,413
HAM10000	Dermoscopic	Public	10,015
MSKCC	Dermoscopic	Public	8,984
Derm7pt	Dermoscopic	Public	839
HIBA	Dermoscopic	Public	1,635
SDDI-Alfred	Dermoscopic	Private	730
Subtotal			*336,168*
Grand Total			**2,485,774**

Source: Reconstructed from Yan et al. (2025).

The architecture of the PanDerm model in the pretraining phase is distinct from the training phase. In the pretraining phase, self-supervised learning is performed through a "student-teacher" model to accomplish masked latent alignment and visible latent alignment loss (Yan et al. 2025). First, the input image is masked with a ratio proportional to the complexity of the student encoder. A ViT-Large model acts as the student encoder, which "processes visible patches to produce latent representations [features]" using 224 x 224-pixel masked images as inputs (Yan et al. 2025). A regressor takes the output from the student encoder along with "learnable mask tokens" as queries to "infer the content of masked regions based on the context provided by visible areas," similar to how a cross-attention mechanism functions (Yan et al. 2025). Simultaneously, an unmasked 196 x 196-pixel version of the image is fed through a CLIP-Large teacher which "generate[s] supervision according to visible and masked patch locations" (Yan et al. 2025). The "target supervision" provided by the teacher model acts as a fixed ground truth that the student model attempts to mimic; that is, "Optimization primarily focuses on aligning visible and masked patch predictions with their corresponding CLIP latent supervisions" (Yan et al. 2025). Figure 2.2 provides a simple visualization of the pretraining mechanism.

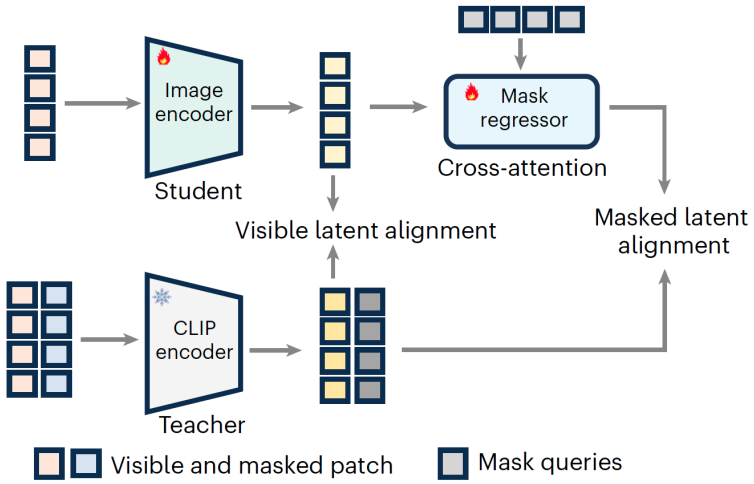


Figure 2.2: Pretraining Architecture of PanDerm

Source: Adapted from Yan et al. (2025).

The training phase simply uses the ViT-Large encoder model as a backbone combined with a linear classifier head to perform either linear probing or fine-tuning (Yan et al. 2025). Weights from the pretraining phase are loaded into the encoder at the start of the training phase which allows it to better understand skin features through transfer learning. During the forward pass, the encoder extracts features from image inputs before passing them to the linear classifier head to assign logits for each diagnostic class. By assigning logits, the head essentially acts to diagnose conditions based on the features passed to it. A softmax function is used to convert the logits into probabilities, and the probabilities are fed into a cross-entropy loss function which produces a loss calculation (Yan et al. 2025); the loss calculation is a representation of how far off the prediction from the forward pass was from the truth. During the backward pass, gradients are calculated via backpropagation for each weight that contributed to the loss, and the gradients are fed into an AdamW optimizer to adjust the model's trainable weights accordingly (Yan et al. 2025). If linear probing is being performed, the encoder weights are frozen and only the weights in the linear head are adjusted; alternatively, both the encoder and classifier head weights are adjusted if fine-tuning is being performed.

Yan et al. compared the relative performance of the PanDerm model against three "representative AI models" (Yan et al. 2025): SL-Imagenet, DINOv2, and SwAVDerm. Exact numerical tabulations of the results were not provided but figures were used to show the superior performance of PanDerm over other models, with results broken down by datasets used in the training phase following pretraining. Estimates for these results can be found in Table 2.3.

Table 2.3: Comparative Performance of PanDerm by Evaluation Metric

Dataset	**SL-Imagenet**	**DINOv2**	**SwAVDerm**	**PanDerm**
Metric: Weighted F1-Score				
HAM10000	0.89	0.89	0.87	0.92
BCN20000	0.70	0.72	0.70	0.78
MSKCC	0.70	0.70	0.70	0.72
HIBA	0.90	0.90	0.88	0.92
PAD	0.68	0.70	0.65	0.77
DDI	0.78	0.76	0.74	0.79
DermC	0.76	0.76	0.76	0.79

Continued on next page

Table 2.3, continued from previous page

Dataset	SL-Imagenet	DINOv2	SwAVDerm	PanDerm
ISIC 2024 (SLICE-3D)	0.88	0.85	0.87	0.92
PATCH16	0.87	0.86	0.81	0.90
WSI	0.94	0.94	0.93	0.94
Average Performance	*0.810*	*0.808*	*0.791*	*0.845*
Metric: AUROC				
HAM10000	0.98	0.98	0.98	0.99
BCN20000	0.91	0.92	0.91	0.95
MSKCC	0.72	0.70	0.70	0.74
HIBA	0.88	0.88	0.86	0.94
PAD	0.89	0.89	0.86	0.93
DDI	0.73	0.73	0.73	0.84
DermC	0.80	0.80	0.76	0.88
ISIC 2024 (SLICE-3D)	0.85	0.83	0.85	0.89
PATCH16	0.94	0.94	0.93	0.94
WSI	0.94	0.94	0.94	0.94
Average Performance	*0.864*	*0.861*	*0.852*	*0.904*
Metric: AUPR				
HAM10000	0.91	0.91	0.91	0.95
BCN20000	0.75	0.79	0.74	0.84
MSKCC	0.48	0.49	0.50	0.58
HIBA	0.61	0.60	0.50	0.75
PAD	0.75	0.75	0.70	0.85
DDI	0.40	0.45	0.41	0.55
DermC	0.69	0.71	0.65	0.80
ISIC 2024 (SLICE-3D)	0.86	0.85	0.84	0.92
PATCH16	0.94	0.95	0.89	0.95
WSI	0.95	0.94	0.93	0.98
Average Performance	*0.734*	*0.744*	*0.707*	*0.817*
Metric: Balanced Accuracy (BAcc)				
HAM10000	0.65	0.70	0.59	0.80
BCN20000	0.59	0.58	0.50	0.65
MSKCC	0.64	0.67	0.63	0.66
HIBA	0.69	0.75	0.63	0.80
PAD	0.61	0.60	0.53	0.69
DDI	0.66	0.61	0.56	0.71
DermC	0.70	0.71	0.70	0.74
ISIC 2024 (SLICE-3D)	0.73	0.68	0.69	0.80
PATCH16	0.83	0.82	0.73	0.88
WSI	0.93	0.93	0.90	0.96

Continued on next page

Table 2.3, continued from previous page

Dataset	SL-Imagenet	DINOv2	SwAVDerm	PanDerm
Average Performance	*0.703*	*0.705*	*0.646*	*0.769*

Source: Reconstructed from Yan et al. (2025).

The results produced by the PanDerm model reinforce the importance of dataset size and multimodality in maximizing performance and robustness. Dataset size proved to have a substantial impact on performance, with AUROC strongly scaling with pretraining data as it increased from 0.8 to 1.8 million skin images; this improvement is illustrated in Figure 2.3. PanDerm also managed to significantly outperform unimodal models across modalities with average gains of 5.1%, 8.0%, 4.2%, and 0.9% on dermatoscopic, clinical, TBP, and pathology datasets, respectively (Yan et al. 2025). The comparative results suggest that cross-modality performance gains could have been borne from the multimodal nature of PanDerm; in fact, the authors attribute its advantages over other models to "pretraining on varied dermatological image modalities and conditions, leading to consistent and significant performance improvements across tasks and modalities" (Yan et al. 2025). Furthermore, the results indicate that multimodality positively correlates with the generalizability of the model given its comparative performance improvements over unimodal models coupled with its ability to take different imaging modalities as inputs.

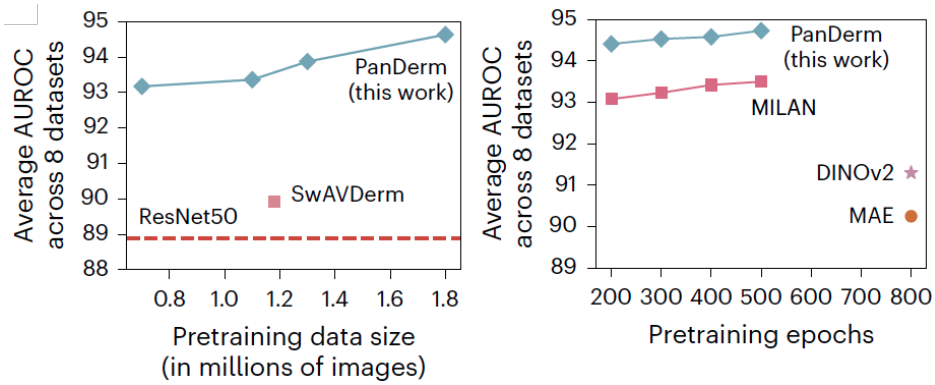


Figure 2.3: PanDerm Performance Versus Pretraining Data Size and Epochs (Average AUROC on 8 Benchmarks) Compared with Alternative Strategies

Source: Adapted from Yan et al. (2025).

2.1.5 SkinEHDLF Model

PanDerm was not the only skin cancer diagnosis model to claim state-of-the-art binary classification performance in 2024; Lilhore et al. (2025) produced an arguably more performant model using a vastly different approach. Lilhore et al. utilized an adaptive attention-based fusion mechanism to synthesize the acquired features of three models within their novel Skin Enhanced Deep Learning Framework (SkinEHDLF) model. The authors rationalize their approach by highlighting the benefits and flaws of CNN- and ViT-based architectures deployed independently. While CNNs have been found at times to outperform human experts in accuracy (Jeyageetha et al. 2025; Natha et al. 2025; Pacal et al. 2025), feature extraction and generalization across lesions can be a challenge for them (Lilhore et al. 2025). The ability of ViTs to use self-attention mechanisms to identify local and global dependencies has made them a popular alternative to CNNs, but they have substantial processing overhead and are less performant in extracting low-level spatial features (Lilhore et al. 2025). Therefore, the authors set out to design a model that could harness the power of both CNN and ViT models to "combin[e] the best aspects of both approaches" and "provide a more reliable classification system for skin cancer" (Lilhore et al. 2025).

Unlike PanDerm, SkinEHDLF was trained on a unimodal dataset. More specifically, training and evaluation were performed using the full SLICE-3D dataset (Lilhore et al. 2025) which was also used for PanDerm pretraining and training. Lilhore et al. noted that future work to improve upon SkinEHDLF should include dataset expansion and multimodal data integration to improve accuracy and robustness (Lilhore et al. 2025). Several data augmentation techniques were used "to enhance the diversity of the training data and mitigate the risk of overfitting" (Lilhore et al. 2025). Those techniques included image rotation, scaling and cropping, flipping, color jittering, and gaussian noise, which "helped to artificially expand the dataset and provide more variety" (Lilhore et al. 2025). Each augmentation was applied to the original images and produced an equal number of augmented images, thus 2,005,295 augmented images were produced from the 401,059 original images (Lilhore et al. 2025). Of the 2,406,354 total images, 70% constituted the training set, 20% the validation set, and 10% the test set (Lilhore et al. 2025). To address the imbalance of the

SLICE-3D dataset, the authors used a class-weighted loss function during training; this allowed the model to assign higher weights to underrepresented malignant images without altering the dataset (Lilhore et al. 2025). The authors indicate that they use "transfer learning strategies" to fine-tune the model (Lilhore et al. 2025), which can be assumed to mean that ImageNet weights were used to initialize the model prior to training considering that authors stated that more advanced transfer learning techniques were not used (Lilhore et al. 2025).

The architecture of the SkinEHDLF model involves a single training phase consisting of a series of different layers. Raw images are first passed into an Input Layer which then passes them to a Data Preprocessing Layer, where images are converted "into a format suitable for feature extraction and classification" (Lilhore et al. 2025). The authors indicate that aside from image resizing, normalization, and the aforementioned augmentations, artifact removal, contrast enhancement, and noise reduction were applied to the images (Lilhore et al. 2025). Images are then passed to three feature extraction layers utilizing the ConvNeXt, EfficientNetV2, and Swin Transformer models (Lilhore et al. 2025). The ConvNeXt Layer extracts spatial and hierarchical features from images resulting in a feature map that "highlights the most significant visual cues in the input image" (Lilhore et al. 2025). ConvNeXt is a CNN that was chosen due to its "capacity to extract resilient, distinctive features from preprocessed images" for use in the classification task (Lilhore et al. 2025). Its extraction performance is a consequence of the depth-wise separable convolutions it employs to "minimize computational expenses while proficiently capturing spatial hierarchies in image data" (Lilhore et al. 2025). The EfficientNetV2 model is an efficient CNN that applies depth-wise separable convolutions while also leveraging MBConv blocks and attention mechanisms to "enhance[e] diagnostic precision and model efficacy in practical applications" while minimizing computational overhead (Lilhore et al. 2025). EfficientNetV2 provides alternative spatial and hierarchical feature extraction from the other two feature extraction layers and is used primarily for its scalability (Lilhore et al. 2025). Finally, the Swin Transformer ViT model uses a shifted window approach to enhance global context comprehensions and "captures long-range dependencies via its self-attention mechanism" (Lilhore et al. 2025). The model is known "for its

robust capacity to capture both local and global features" (Lilhore et al. 2025), making it a worthy addition to the selection of feature extraction layers.

Following feature extraction, the Feature Fusion Layer weights, combines, or concatenates the features from each of the feature extraction layers "while maintaining their unique significance" (Lilhore et al. 2025). By synthesizing the features from each layer, Lilhore et al. aim to "enhance the model's ability to differentiate between malignant and benign images by capturing many significant features in skin images" leveraging "a more comprehensive and resilient representation of the data" (Lilhore et al. 2025). Every neuron from the Feature Fusion Layer connects with neurons in the Fully Connected Layer, where extracted features are used to effectively generate predictions (Lilhore et al. 2025). Non-linearity that results from each neuron computing a weighted sum on input features, incorporating a bias, and processing the result through an activation function allows the model to "recognize intricate patterns that might not be immediately obvious in the data that has not been processed beforehand" (Lilhore et al. 2025). Ultimately, the layer produces a "collection of values representing the probability of each class," benign or malignant (Lilhore et al. 2025). The Final Output Layer, as its name suggests, produces the final binary classification prediction by passing the output from the Feature Fusion Layer into a sigmoid function (Lilhore et al. 2025). Further information about the architecture of SkinEHDLF can be found in Table 2.4 and Figure 2.4.

Table 2.4: SkinEHDLF Architecture Configuration for Binary Classification

Layer	Type	No. of Units	Activation Function	Kernel Size	Stride
Input Layer	Image Input	–	–	–	–
Feature Extraction (ConvNeXt)	Convolutional + Attention	Varies	ReLU	3×3	1
Feature Extraction (EfficientNetV2)	Convolutional + Attention	Varies	Swish	3×3	2
Feature Extraction (Swin Transformer)	Self-Attention + Convolutional	Varies	GELU	4×4	2
Feature Fusion Layer	Fusion of Extracted Features	–	–	–	–
Fully Connected (Dense) Layer	Dense Layer	1024	ReLU	–	–
Final Output Layer	Sigmoid	2	Sigmoid	1×1	1

Source: Reproduced from Lilhore et al. (2025).

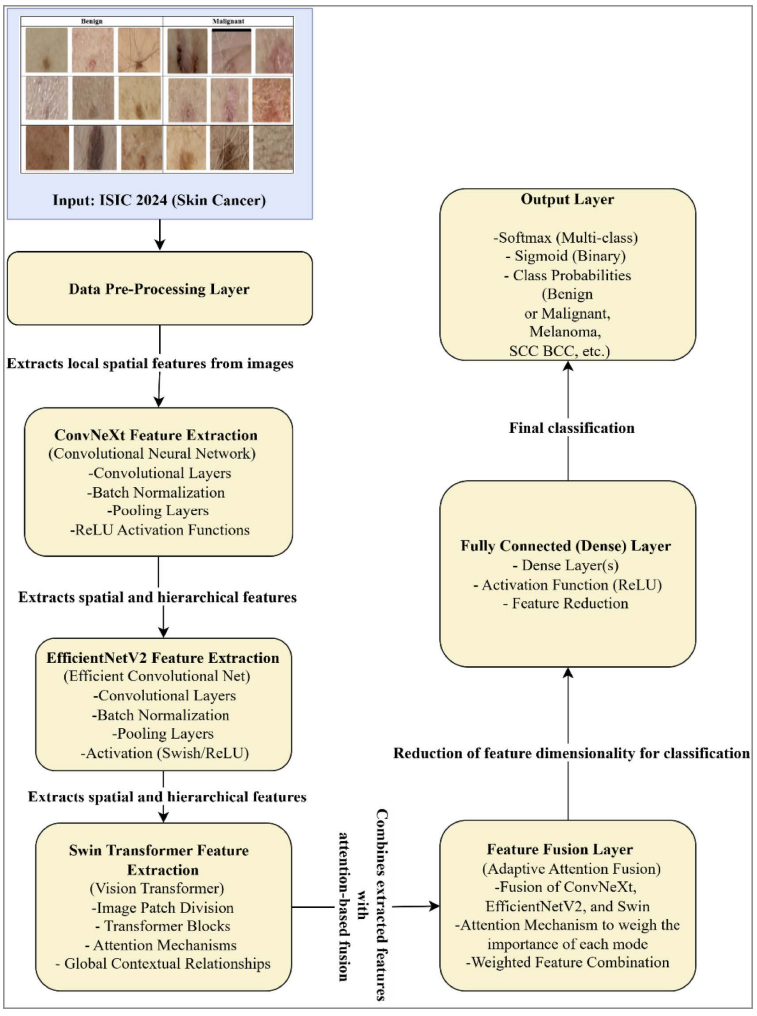


Figure 2.4: Architecture of SkinEHDLF

Source: Adapted from Lilhore et al. (2025).

Lilhore et al. provided a comparative analysis of SkinEHDLF against several other models, but PanDerm was not among them likely due to the time at which the results of the models were published. The following models were included for comparative analysis in the binary classification task for detecting skin cancer: ResNet-50, EfficientNet-B3, ViT-B16, ResNet-50 + EfficientNet, and ViT + CNN (Lilhore et al. 2025). The authors found that their model had significantly greater performance than all of those compared to. The results of the comparative analysis can be found in Table 2.5 and Table 2.6.

Table 2.5: Comparative Performance of SkinEHDLF for Binary Classification

Model	**Acc[a] (%)**	**Prec[b] (%)**	**Rec[c] (%)**	**F1[d] (%)**	**AUROC (%)**	**AUC (%)**	**Sens[e] (%)**	**Spec[f] (%)**
ResNet-50	90.2	89.4	88.1	88.7	92.6	91.3	88.1	91.0
EfficientNet-B3	93.5	92.1	91.3	91.7	94.5	94.2	91.3	93.8
ViT-B16	91.8	90.7	89.4	90.0	94.2	93.5	89.4	92.7
ResNet-50 + EfficientNet	92.3	91.5	90.2	90.8	93.8	93.2	90.2	92.5
ViT + CNN	92.1	91.2	90.0	90.5	94.1	93.7	90.0	92.3
SkinEHDLF	**98.76**	**99.2**	**98.3**	**98.7**	**99.8**	**99.5**	**98.3**	**99.0**

[a]Accuracy; [b]Precision; [c]Recall; [d]F1-Score; [e]Sensitivity; [f]Specificity.
Source: Adapted from Lilhore et al. (2025).

Table 2.6: Statistical Test Results Between SkinEHDLF and Comparative Models

Model Comparison	**Paired t-test (t)**	**p-Value (t-test)**	**Chi-Square (χ^2)**	**p-Value (χ^2)**
SkinEHDLF vs. ResNet-50	3.50	0.0015	8.25	0.004
SkinEHDLF vs. EfficientNet-B3	2.90	0.0060	7.00	0.010
SkinEHDLF vs. ViT-B16	4.10	0.0008	9.15	0.002
SkinEHDLF vs. ResNet-50 + EfficientNet	3.20	0.0022	7.85	0.003
SkinEHDLF vs. ViT + CNN	2.75	0.0075	6.70	0.005

Source: Reproduced from Lilhore et al. (2025).

The SkinEHDLF model produced results that indicate state-of-the-art performance can be achieved on the binary skin cancer classification task without relying on multimodality or a dataset consisting of millions of images. Given its superior performance compared to other models, the fusion layer approach appears to have been highly effective. Lilhore et al. (2025) also found that

data preprocessing “significantly enhance[d] the performance” of SkinEHDLF, improving accuracy from 95.8% to 98.9%, precision from 92.6% to 97.6%, and recall from 90.2% to 96.4%. Because the authors trained the SkinEHDLF model exclusively on a unimodal dataset and performed limited testing on domain-shifted datasets, without specifying whether target-domain fine-tuning was applied, its comparative robustness and generalizability relative to PanDerm remain unverified.

2.2 Model Optimization Strategies

2.2.1 Integration of Linear Probing and Fine-Tuning

Research has shown that combining linear probing with fine-tuning during training can result in performance improvements over training independently with either. Linear probing is a training technique that involves attaching a single-layer linear classifier, or “probe," to the end of a pretrained model and only updating the weights of the probe during training while keeping the weights of the model backbone frozen. A pretrained model that has learned robust foundational features should be able to map those features to the appropriate output classes through linear probing, which is a much simpler and faster task than fine-tuning the entire model. Kumar et al. (2022) introduced the feature distortion theory to explain the effectiveness of linear probing followed by fine-tuning, suggesting that fine-tuning using an optimized linear head from the linear probing stage preserves pretrained features and thus yields higher performance. While the scope of the feature distortion theory was initially limited, with its basis being a theoretical analysis of a two-layer model (Sato and Tomihari 2024), other researchers have been able to quantitatively substantiate the theory when applied to more complex architectures such as transformers.

Sato and Tomihari (2024) found that high accuracy and increased norms of the linear head during the linear probing stage reduced feature changes during a subsequent fine-tuning stage when applying neural tangent kernel (NTK) theory, which is consistent with feature distortion theory. Performance differences as measured by the Matthews correlation coefficient ranged from negligible on the PubMed 20k RCT dataset to about 3.2% improvement on the Commitment-Bank dataset (Sato and Tomihari 2024) when following linear probing with fine-tuning. Of note, only one epoch

of fine-tuning was analyzed (Sato and Tomihari 2024) using only natural language datasets. Despite these limitations, the results indicate that combining linear probing with fine-tuning might be an effective means of improving model performance in realms other than natural language processing.

2.2.2 Robustness and Generalizability Enhancement with Mixup and CutMix

Data augmentation techniques such as Mixup and CutMix have been shown to improve the robustness and generalizability of models in the medical imaging domain. To be clear, robustness refers to "a model's ability to maintain performance despite the variability in medical imaging environments" while generalizability refers to "a model's ability to perform effectively on entirely new, unseen datasets" (Tran, Zeevi, and Payabvash 2025). Both Mixup and CutMix force neural networks to learn generalized structural features instead of local artifacts or noise but do so through different methods.

Zhang et al. (2018) presented Mixup in 2018 as a means to "[extend] the training distribution by incorporating the prior knowledge that linear interpolations of feature vectors should lead to linear interpolations of the associated targets" and found that it improves the generalization of state-of-the-art architectures, reduces the memorization of corrupt labels, increases the robustness of adversarial examples, and stabilizes the training of generative adversarial networks. Mixup can be thought of more simply as a technique that blends two images together during training. CutMix was introduced by Yun et al. (2019) in 2019 as another regularization strategy in which "patches are cut and pasted among training images where the ground truth labels are also mixed proportionally to the area of the patches." The authors found that CutMix outperformed other state-of-the-art augmentation strategies, such as Mixup, on CIFAR and ImageNet classification tasks in addition to the ImageNet weakly-supervised localization task (Yun et al. 2019). Similar to Mixup, CutMix was found to "improve the model robustness against image corruptions and its out-of-distribution detection performances," indicating that the strategy was also an effective way to increase the generalizability of a model (Yun et al. 2019).

The importance of robustness when applying deep learning models to medicine is self-evident considering the varied imaging environments that a model could be deployed in.

Likewise, generalizability is important as no two patients will present symptoms exactly the same way. A model can be responsible for directing serious, potentially life-changing medical decisions based on its evaluation of the circumstances, so both robustness and generalizability are crucial factors that must be considered for any medical model. Apart from the general findings of the Mixup and CutMix designers, researchers have found that data augmentation techniques such as Mixup and CutMix enhance robustness and generalizability when specifically applied in a medical context. In neuroimaging, data augmentation techniques have been found to improve robustness, improve generalization, and address class imbalance issues, which is a common concern when handling medical datasets that can have negative downstream effects during training (Tran, Zeevi, and Payabvash 2025). These advantages also extend to other data augmentation techniques such as geometric transformations, but Mixup and CutMix are considered advanced methods that provide benefits beyond more traditional techniques (Tran, Zeevi, and Payabvash 2025). In addition to the aforementioned benefits when applying Mixup and CutMix in medical applications, the techniques have been found to improve the ability of models to provide reliable confidence estimations to clinicians through superior confidence calibration (Rao, Lee, and Aalami 2023). This is particularly important in clinical contexts since CNNs are "often poorly calibrated and tend to be overconfident or overly certain in predictions," which can subsequently result in "potentially false interpretations when deployed in the clinical setting" (Rao, Lee, and Aalami 2023).

2.3 Comparison of Diagnostic Accuracy: Algorithms vs. Clinical Experts

2.3.1 Expert-Level Parity in Diagnostic Classification

Seven years prior to the release of the SLICE-3D dataset, Esteva et al. (2017) created a CNN model that outperformed or matched the performance of 21 board-certified dermatologists in the binary classification task of distinguishing benign from malignant skin lesions. The model performance is particularly notable considering that the dataset used in training the model contained 129,450 proprietary clinical images, a fraction of the size of the SLICE-3D dataset. At the time, the research was quite novel with the dataset used being two orders of magnitude larger than previously

used datasets (Masood and Ali Al-Jumaily 2013). Anticipating later predictions made by Kurtansky et al. (2024), Esteva et al. (2017) noted that the remarkable performance of their model combined with widespread access to smartphones has the potential to "provide low-cost universal access to vital diagnostic care." The team had created a model distinct from previous models, not only in performance but in generalizability and readiness for mass deployment as well: unlike previous approaches, their model did not require extensive preprocessing, lesion segmentation, and extraction of domain-specific visual features prior to classification (Esteva et al. 2017).

Esteva et al. (2017) noted the advantages of very large datasets throughout their research. When their research was published, it was typical for researchers to use datasets containing under a thousand skin lesion images (Binder et al. 1998; Gutman et al. 2016; Kittler et al. 2002) for model training, consequently resulting in poor generalizability from those models. Esteva et al. were aware of the power of very large datasets when training models applied to other visual tasks (Deng et al. 2009), which served as a motivation for their model being trained on a dataset of hitherto unprecedented scale. The authors leveraged transfer learning (Pan and Yang 2010) through the pretraining of their model with an ImageNet library of approximately 1.28 million images, which coincided with their use of a very large dataset in training to produce a state-of-the-art model. Esteva et al. conclude that their "fast, scalable method" for skin lesion classification is deployable on mobile devices and therefore has tremendous potential for clinical impact through "broadening the scope of primary care practice and augmenting clinical decision-making for dermatology specialists"; ultimately, they acknowledge that their method is "primarily constrained by data" and has even greater potential should more training examples be provided (Esteva et al. 2017).

2.3.2 Superior Performance in Diagnostic Classification Over Experts

Yan et al. (2025) found their PanDerm model to achieve superior performance to expert clinicians in several diagnostic tasks, both when independently deployed and when deployed alongside clinicians. The authors performed three reader studies relating to early melanoma detection, human-PanDerm collaboration for skin cancer diagnosis, and human-PanDerm collaboration for differentiating between 128 skin conditions. Taken together, the results suggest

that state-of-the-art diagnostic models are not only ready for deployment in clinical settings but have surpassed the ability of humans in diagnostic performance, indicating that such models may also be ready for integration into the patient journey at the pre-clinical level.

In the early melanoma detection study, the performance of the model was compared against 7 experienced dermatologists and 5 dermatologist trainees with overall diagnostic accuracy and early melanoma detection capability being evaluated (Yan et al. 2025). PanDerm outperformed the average human reviewer by 10.2% and the best performing human by 3.2% in terms of overall accuracy (Yan et al. 2025). Early melanoma detection showed even greater differences with the model identifying 77.5% of melanoma lesions at the first imaging time point compared to 32.6% for human reviewers (Yan et al. 2025).

For the collaborative skin cancer diagnosis task, the impact of PanDerm on the accuracy of 41 clinicians with varying levels of competency was evaluated across seven pigmented dematoscopic images (Yan et al. 2025). Adding PanDerm as an assistant resulted in statistically significant improvement for diagnostic accuracy from 69% to 80%, with 17% improvement occurring for low competency clinicians, 12% improvement for medium competency clinicians, and 6% improvement for high competency clinicians (Yan et al. 2025). Notably, PanDerm actually achieved a slightly higher accuracy of 81% when used alone compared to the 80% when assisting a clinician (Yan et al. 2025). For the melanoma diagnosis task specifically, clinician accuracy improved from 69% to 83% when PanDerm assisted (Yan et al. 2025).

For the collaborative diagnostic task on the 128 skin conditions, 37 readers from 5 countries were included and broken into dermatology and generalist groups (Yan et al. 2025). Each reader "assessed up to 50 cases from a 200-case pool, providing their top 3 diagnoses both with and without PanDerm's assistance" (Yan et al. 2025). Statistically significant improvement in top-1 and top-3 diagnostic accuracy was found, with improvements in diagnostic scores for all readers from 2.83 to 3.08 and improvements in diagnostic accuracy from 54% to 63.4%, respectively (Yan et al. 2025). The generalist group also showed greater improvements over the specialist group, indicating the power of the model in the hands of people not regularly exposed to such diagnostic

tasks (Yan et al. 2025). Similar to the skin cancer diagnosis task, PanDerm achieved superior performance with a 3.6 top-1 score when used independently compared to unassisted readers, who achieved a score of 2.83, and when used in collaboration, which resulted in a score of 3.08 (Yan et al. 2025). The strength of the model when used independently is an important characteristic when considering pre-clinical viability since it is an environment where a clinician could not yet guide the diagnostic decision presented to a patient.

2.4 Efficacy of Machine Learning Tools in Pre-Clinical Environments

2.4.1 Measured Impact of Pre-Clinical Diagnostic Software Applications

AI has clearly been demonstrated to improve diagnostic outcomes when used in clinical settings both alongside clinicians and independently. Its power as an independent tool suggests its potential usefulness in pre-clinical settings, where laymen can utilize a classification model to inform their decision to seek further medical investigation by a human clinician. Academic research substantiates the notion as well while illuminating the potential pitfalls of such deployment.

In 2019, 2.2 million Dutch adults were given free access to an app utilizing a CNN computer vision model for skin cancer detection (Smak Gregoor et al. 2023). Smak Gregoor et al. (2023) performed a short-term cost-effectiveness analysis to determine the cost per additional detected (pre)malignancy and found that the app had a positive impact on detecting more cutaneous pre(malignancies), although with a higher cost of detection over the current standard of care. Researchers compared 18,960 app users against a control group of 56,880 people who did not use the app (Smak Gregoor et al. 2023). Approximately 6.0% of app users with (pre)malignant skin lesions made insurance claims for their conditions compared to 4.6% of those in the control group, indicating a statistically significant increase in cancer detection (Smak Gregoor et al. 2023). In other words, there was a 32% increase of (pre)malignancy detection for the app users compared to the control group (Smak Gregoor et al. 2023). However, the number of insurance claims for benign skin tumors and nevi were even higher at 5.9% compared to 1.7% for the experiment and control groups, respectively (Smak Gregoor et al. 2023). Ultimately the disproportionate number of claims for

non-malignancies resulted in a higher short-term cost of €2,567 per additional positive detection compared to the current standard of care (Smak Gregoor et al. 2023). Clearly, the specificity of the CNN was a limiting factor in its utility; its strong sensitivity of 87-95% was offset by a suboptimal specificity of 70-78% (Smak Gregoor et al. 2023). It should be noted that the additional cost includes the treatment of additionally detected (pre)malignancies; in fact, an app with perfect detection accuracy would result in a cost of €1,119 per additional positive detection (Smak Gregoor et al. 2023).

One must not only consider short-term costs as measured to understand the overall impact of the app. The authors acknowledge that the study has several limitations, most of which "[underestimate] [the] impact of the [app]" (Smak Gregoor et al. 2023). For one, the number of skin cancers detected were based on claims data which might not reflect all the people who sought treatment if billing codes were improperly input or if insurance was not used (Smak Gregoor et al. 2023). In addition, "no data were available on the number, type, and stage of skin cancers related to the claims," implying that multiple (pre)malignancies of varying severity might have been found on a physician visit preceded by detection on the app (Smak Gregoor et al. 2023). Perhaps most importantly, long-term costs were not analyzed due to the lack of longitudinal data, making it "impossible to estimate the health and cost related impact of early diagnosis of skin cancer" (Smak Gregoor et al. 2023). Considering that cancer is a notoriously expensive condition to treat, particularly in its later stages and if chemotherapy is required, the likelihood that short-term costs would be offset by long-term savings is high, particularly if a model with higher specificity were to be deployed. The authors also note that "the advantage of AI-based technology over other interventions is its scalability and that accuracy will improve as the number of users increases," which can lead to a reduction in cost over time (Smak Gregoor et al. 2023).

2.4.2 Physician Perspectives on Pre-Clinical Diagnostic Software Applications

While it is extremely important to analyze the quantitative impact of applications designed to detect skin cancer in a pre-clinical environment prior to mass deployment, qualitative factors

should also be considered. Given their expertise in medical diagnosis, input from physicians on the risks, benefits, and preconditions for endorsement of such applications is significant. In 2025, Sangers et al. (2025) published an in-depth qualitative online focus group study consisting of six focus groups to address the dearth of information regarding physician perspectives on the matter. Of the six focus groups, three consisted of Dutch dermatologists while three consisted of Dutch general practitioners (Sangers et al. 2025).

In terms of perceived risks, "incorrect diagnoses" was considered the most significant (Sangers et al. 2025). Study participants suggested that false negatives could lead to false security about users' health and increased delay in visiting a physician while false positives could result in increased patient anxiety and increased workload for physicians (Sangers et al. 2025). Participants also suggested that laymen are "incapable of accurately deciding which lesions are suspicious for skin cancer" which could thus increase the risk of missed diagnoses (Sangers et al. 2025). The second most significant risk was found to be "excluding specific subpopulations," specifically people with low digital literacy and with skin colors that are not analyzed well by AI models (Sangers et al. 2025). The third most significant perceived risk was "loss of [general practitioner] autonomy in clinical decision making" if the patient "refuse[s] to be reassured after a physical examination by the [general practitioner] or other primary healthcare providers and demand referral to a specialist" (Sangers et al. 2025). The fourth most significant risk was found to be "loss of [general practitioner] diagnostic experience regarding skin lesions"; in other words, the concern is that the applications will lead to "significantly fewer patients visiting [general practitioners]" and thus their ability to diagnose skin lesions will deteriorate over time (Sangers et al. 2025).

The most significant benefit was determined as "an increased skin cancer awareness" by educating users about skin cancer and its risks, specifically by "providing information about the multiform appearance of different types of skin cancer" (Sangers et al. 2025). This finding is particularly interesting because it contradicts the most significant perceived risk that laymen are "incapable of accurately deciding which lesions are suspicious for skin cancer," considering that "incapable" means they cannot be taught or have prior knowledge through means other than formal

medical education (Sangers et al. 2025). The second most significant perceived benefit was the "facilitation of early detection of skin cancer," with increased access to care and improvements in skin cancer detection accuracy coinciding (Sangers et al. 2025). The third most significant perceived benefit was "a streamlined patient journey" which coincides with a lower patient volume given adequate diagnostic accuracy and the referral of highly suspicious skin lesions directly to a dermatologist, ultimately leading to an "optimized flow of patients with high-risk lesions" (Sangers et al. 2025).

"Evidence-based verification of accuracy" was found to be the most significant precondition for the endorsement of skin cancer diagnostic apps deployed in a pre-clinical environment, but no specific accuracy was given (Sangers et al. 2025). Participants noted that the accuracy should be tested by an independent organization with a large representative sample for the sake of reliability of the evidence (Sangers et al. 2025). The second most significant precondition for endorsement, "successful integration within clinical practice," came with four sub-preconditions: "appropriate communication of medical information" in the form of a risk indication instead of explicit diagnosis given to the patient, report sharing functionality from the app to a clinician, adequate protection of patient data, and involvement of practitioners during the implementation process (Sangers et al. 2025). "Clarity about potential liability in case of adverse events" was identified as the third most significant precondition for endorsement (Sangers et al. 2025). Finally, accessible and inclusive app design was found to be the fourth most significant precondition for endorsement (Sangers et al. 2025).

Chapter 3
Methodology and Experimental Design

This chapter details the empirical framework, architectural implementations, and experimental design used throughout this study to address the research gaps discussed in Chapter 1 and Chapter 2. Ultimately, the potential of DTC diagnostic tools for skin cancer is evaluated through an analysis of the experimental results. The primary objective of this methodology is to provide a reproducible and highly structured pipeline for developing and evaluating the proposed OmniFusion model with support from either a PanDerm or SkinEHDLF backbone. The models produced in this study are defined in this chapter and assigned names for conciseness. Data acquisition and preprocessing steps are also defined as they relate to the unimodal SLICE-3D dataset and the supplemental multimodal datasets. In addition, cross-validation is employed to ensure reproducible separation of data and to avoid data leakage.

This chapter also defines the technical aspects of the machine learning backbone architectures under investigation. The implementations of the ViT-Large encoder backbone utilized by the PanDerm model and the adaptive attention-based feature fusion mechanism from the SkinEHDLF model are described in the context of the OmniFusion model. The unique optimizations applied to the OmniFusion model in the form of training pipeline and regularization strategies are also defined. Specifically, the integration of LP-FT as a mechanism to accelerate convergence and preserve pretrained features is detailed, along with the application of Mixup and CutMix spatial augmentations as a means to maximize confidence calibration and out-of-distribution generalizability.

Finally, this chapter outlines a four-phase experimental design, specifies the software and hardware used, and establishes the mathematical methods used for performance evaluation, namely BAcc, sensitivity, specificity, and Area Under the Receiver Operating Characteristic (AUROC) curve. The phased approach to experimental design isolates the multiple hypotheses being assessed, which further clarifies the implications of the results.

3.1 Dataset Acquisition

3.1.1 SLICE-3D Dataset (Unimodal)

The SLICE-3D dataset (Kurtansky et al. 2024) containing 401,059 3D TBP images was downloaded from https://challenge2024.isic-archive.com. Of note, the dataset is unimodal as it only contains TBP images. The "Training Ground Truth" CSV file provided by the dataset authors served to map malignancy status to each image in the dataset based on its filename. While the SLICE-3D dataset is in many ways superior to previously published skin lesion image datasets, particularly in its size, it is also severely imbalanced. Its imbalanced nature does pose a significant challenge to model training, but not an insurmountable one, as this study demonstrates. Further details regarding the SLICE-3D images can be found in Table 3.1 and Table 3.2. In addition, more information regarding the background of the SLICE-3D dataset and the motivation for its release can be found in Subsection 2.1.3.

Table 3.1: SLICE-3D Dataset Overview of Classes, Tags, and Image Sources

Category	**Overall Count**	**Athens**	**Barcelona**	**Basel**	**Brisbane**	**New York**	**Sydney**	**Vienna**
Total Images	401,059	7,976	105,724	65,218	51,768	129,068	28,665	12,640
Unique Lesions	401,059	7,976	105,724	65,218	51,768	129,068	28,665	12,640
Manual Tags	22,058	167	1,720	3,004	9,056	6,212	1,792	107
Diagnosis - Malignant	393	6	72	13	81	174	33	14
Diagnosis - Indeterminate	114	1	12	2	60	39	-	-
Diagnosis - Benign	400,552	7,969	105,640	65,203	51,627	128,855	28,632	12,626
Anatomic Site - Head/Neck	12,046	353	3,416	2,320	1,533	3,229	809	386
Anatomic Site - Anterior Torso	87,770	1,774	23,546	14,075	7,806	31,525	5,722	3,322
Anatomic Site - Posterior Torso	121,902	2,842	32,725	19,822	13,569	40,495	8,366	4,083
Anatomic Site - Upper Extremity	70,557	1,475	18,536	11,312	9,676	22,225	5,310	2,023
Anatomic Site - Lower Extremity	103,028	1,532	27,332	15,600	16,841	30,441	8,458	2,824
Anatomic Site - Unknown	5,756	-	169	2,089	2,343	1,153	-	2
Lighting - White Light	115,156	-	1	484	126	114,545	-	-
Lighting - XP Light	285,903	7,976	105,723	64,734	51,642	14,523	28,665	12,640

Source: Reconstructed from Lilhore et al. (2025).

Table 3.2: SLICE-3D Dataset Overview of Patient Ages and Image Sources

Category	Overall Count	Athens	Barcelona	Basel	Brisbane	New York	Sydney	Vienna
Total Patients	1,042	16	163	230	176	398	44	15
Male Patients	551	9	75	115	87	229	26	10
Female Patients	458	7	87	88	84	169	18	5
Unknown Sex	33	-	1	27	5	-	-	-
Age Group (<20)	8	-	-	6	1	1	-	-
Age Group [20-25)	16	1	2	5	1	7	-	-
Age Group [25-30)	28	-	4	5	2	17	-	-
Age Group [30-35)	56	2	4	16	10	23	1	-
Age Group [35-40)	59	-	3	21	11	22	1	1
Age Group [40-45)	107	2	10	24	20	43	7	1
Age Group [45-50)	85	-	17	20	16	27	4	1
Age Group [50-55)	119	5	26	23	17	40	5	3
Age Group [55-60)	133	2	16	35	26	48	4	2
Age Group [60-65)	122	2	21	21	21	46	9	2
Age Group [65-70)	129	-	18	16	32	56	5	2
Age Group [70-75)	76	1	12	18	12	29	3	1
Age Group [75-80)	46	-	12	9	1	21	3	-
Age Group [80-85)	33	1	10	5	-	14	2	1
Age Group [85+)	12	-	4	2	1	4	-	1
Unknown Age	13	-	4	4	5	-	-	-

Source: Reconstructed from Lilhore et al. (2025).

3.1.2 Supplementary Dataset (Multimodal)

The additional mix of multimodal data used by Yan et al. (2025) were chosen for use in this study based on their availability. The authors were contacted with a request for access to the private datasets used in their study, but the request was denied. The public datasets were acquired through links in the README file on the GitHub repository used by Yan et al. (2025) for their PanDerm model, which is located at https://github.com/SiyuanYan1/PanDerm. The "processed data" links were used for consistency with the PanDerm authors, with accompanying ground truth files used to map malignancy status to each image. Details regarding each dataset used can be found in Table 3.3. The aggregated datasets described in this section will hereafter be referred to as the supplementary dataset.

Table 3.3: Breakdown of Supplementary Dataset for OmniFusion

Source Dataset Name	**Modality**	**Image Count**
HAM10000[a]	Dermatoscopic	11,720
BCN20000[b]	Dermatoscopic	12,414
HIBA[c]	Dermatoscopic	1,635
MSKCC[d]	Dermatoscopic	10,847
DDI[e]	Clinical	656
Derm7pt[f]	Clinical	2,013
Dermnet[g]	Clinical	18,856
PAD-UFES-20[h]	Clinical	2,298
PATCH16[i]	Dermatopathology	40,533
Total		**100,972**

[a]Codella et al. (2019) and Tschandl, Rosendahl, and Kittler (2018)
[b]Perez et al. (2023)
[c]Hospital Italiano de Buenos Aires (HIBA Skin Lesions)
[d]Codella et al. (2018)
[e]Daneshjou et al. (2022)
[f]Kawahara et al. (2019)
[g]Goel (2024)
[h]Pacheco et al. (2020)
[i]Kriegsmann et al. (Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections [data])

3.2 Model Definitions

Five models were produced over the course of this study. Since Model 2 was involved in two different experiments, it is presented in two different configurations with varying test sets. Names of the models and their experimental configurations are provided in Table 3.4.

Table 3.4: OmniFusion Model Definitions and Experimental Configurations

Configuration	Architecture	Training Data	Validation Data	Test Data
Model 1	PanDerm	SLICE-3D	SLICE-3D	SLICE-3D
Model 2	SkinEHDLF	SLICE-3D	SLICE-3D	SLICE-3D
Model 3	PanDerm	SLICE-3D + Supp.	SLICE-3D	SLICE-3D
Model 4	SkinEHDLF	SLICE-3D + Supp.	SLICE-3D	SLICE-3D
Model 5	SkinEHDLF	Supplementary	SLICE-3D	SLICE-3D
Model 2-DS	SkinEHDLF	SLICE-3D	SLICE-3D	Supplementary

Model 2-DS represents an experimental configuration for testing domain shift. It utilizes the exact same model weights trained during the Model 2 configuration, but is evaluated strictly on out-of-distribution supplementary data (e.g., HAM10000) of different imaging modalities to measure cross-domain generalizability.

3.3 Data Pre-Processing and Augmentation Pipeline

Several steps were taken to preprocess the data in a similar fashion to Yan et al. (2025) and Lilhore et al. (2025). In experiments utilizing the PanDerm backbone architecture, images in the dataset were normalized and resized to 256x256px to improve model performance and standardize image resolution. In experiments utilizing the SkinEHDLF backbone architecture, images were normalized and resized to 256x256px as well. However, additional preprocessing was also performed in accordance with Lilhore et al. (2025); while the authors did explain some of the techniques used, they did not explain what specific algorithms or parameters were used in their study. Therefore, algorithms and parameters were chosen based on best practices. For instance, Contrast Limited Adaptive Histogram Equalization (CLAHE) was used to improve the contrast of images due

to its favored usage in medical imaging. Other techniques used to preprocess the images include the DullRazor algorithm to remove hair artifacts and bilateral filtering to reduce noise while preserving lesion borders. The full preprocessing procedure including specific parameters used can be viewed in the `phase_3_preprocessing_offline.py` script within the codebase for this study.

Data augmentation was performed to improve model robustness. Center cropping was applied to all validation and test set images in accordance with the PanDerm augmentation pipeline. For experiments utilizing the PanDerm backbone architecture, random cropping, random horizontal and vertical flipping, random rotation, and color jittering for hue were applied to training images. Experiments utilizing the SkinEHDLF backbone architecture had the same augmentations applied while adding color jittering for brightness, contrast, and saturation as well as Gaussian noise.

It should be noted that the way data augmentation was performed in this study fundamentally differs from the way Lilhore et al. (2025) performed their data augmentations. Although the techniques selected for this study resemble those chosen by Lilhore et al. (2025), Lilhore et al. (2025) opted to create a dataset using both unmodified and modified images. That is, each technique was applied to the unmodified images and the images that were output were added alongside the unmodified images to be processed as a larger dataset. Consequently, Lilhore et al. (2025) derived 2,005,295 additional images for their dataset on top of the 401,059 unmodified SLICE-3D images. Their procedure contrasts with the one used in this study, which follows the augmentation pipeline used by Yan et al. (2025) instead. In this study, all augmentations were applied simultaneously as an in-place operation. Since the augmentations were applied randomly, most images were modified in some way during training but many were likely unaltered, thereby increasing the diversity of the existing data without expanding the overall footprint of the dataset. Although Yan et al. (2025) do not explain why they chose to apply augmentations in place, the approach might have been chosen due to its substantial time and energy savings as well as the greater number of permutations associated with applying multiple augmentation techniques at random, which can potentially improve model generalizability. Those practical advantages, along

with the success that Yan et al. (2025) had applying the approach to the SLICE-3D dataset, served as the rationale for adopting the approach in this study.

3.4 K-Fold Cross-Validation and Separation of Data

K-fold cross-validation was used in all experiments to ensure that the performance of each model is generalizable. Data were split into ten folds for each experiment. In Model 3 and Model 4, only SLICE-3D data were rotated between training, validation, and test sets to allow for comparison against Model 1 and Model 2 by controlling for the contents of the validation and test sets. As a result, the effect of training set multimodality on model performance could be isolated and assessed.

In Model 5, SLICE-3D data were rotated between the validation and test sets while every fold used the entire supplementary dataset as the training set; consequently, the number of images per fold is lower than Model 3 and Model 4. This approach was designed to enhance understanding of the effect of training set multimodality and size on model performance by removing SLICE-3D data from the training set. Model 2-DS is identical to Model 2, except that each fold was tested on the entire supplementary dataset. As the only model using supplementary data solely in the test set, Model 2-DS provides insight into the generalizability of Model 2 in terms of evaluating on different imaging modalities. The separation of data for each model is described in Table 3.5.

Table 3.5: Separation of Data for Experimental OmniFusion Models

Configuration	Training Set Count	Validation Set Count	Test Set Count	Total Images
Model 1	306,810 (76.50%)	54,143 (13.50%)	40,106 (10.00%)	401,059
Model 2	306,810 (76.50%)	54,143 (13.50%)	40,106 (10.00%)	401,059
Model 3	407,782 (81.23%)	54,143 (10.78%)	40,106 (7.99%)	502,031
Model 4	407,782 (81.23%)	54,143 (10.78%)	40,106 (7.99%)	502,031
Model 5	100,972 (51.72%)	54,143 (27.73%)	40,106 (20.54%)	195,221
Model 2-DS	306,810 (66.42%)	54,143 (11.72%)	100,972 (21.86%)	461,925

3.5 Architectural Implementations

3.5.1 PanDerm (ViT-Large) Backbone

The PanDerm backbone architecture for this study was adopted from Yan et al. (2025); it is identical to the ViT-Large backbone used by the authors during the training stage for PanDerm. The vision transformer extracts features by dividing images into fixed-size patches, embedding them, and passing them through self-attention layers that capture local and global spatial dependencies. In this study, the optimizer and activation function remain as AdamW and softmax, respectively. A single, fully connected linear layer serves as the classification head.

3.5.2 SkinEHDLF (Adaptive Fusion) Backbone

The SkinEHDLF backbone architecture was adopted from Lilhore et al. (2025); it is implemented similarly to the adaptive attention-based fusion approach used by the authors, but is likely not identical due to a lack of publicly available code for the model. Reimplementation was accomplished by analyzing descriptions and algorithms provided by Lilhore et al. (2025). Unlike the PanDerm backbone, the SkinEHDLF backbone operates by passing images through three parallel feature extraction layers: ConvNeXt, EfficientNetV2, and a Swin Transformer. ConvNeXt and EfficientNetV2 are highly efficient CNNs that extract robust spatial and hierarchical features while the Swin Transformer uses a shifted-window approach similar to the PanDerm backbone. Latent representations from the three parallel networks are then aggregated in a Feature Fusion Layer that computes a weighted sum of the extracted features.

In this study, the activation function remains as sigmoid. However, the optimizer was modified from Adam to AdamW in an effort to improve performance and generalization by decoupling weight decay from the gradient update step. AdamW applies weight decay directly during the parameter update step rather than to the loss function, which prevents the weight decay from affecting adaptive learning rates. A six-layer dense block consisting of five hidden layers and one output layer acts as the classification head.

3.5.3 OmniFusion Model

This study introduces the OmniFusion model to facilitate a controlled comparison of the contrasting methodologies of Yan et al. (2025) and Lilhore et al. (2025) while introducing several changes intended to improve performance. Using the public codebase of Yan et al. (2025) as a framework, the OmniFusion model allows for simple swapping of feature extraction backbones between the aforementioned PanDerm and SkinEHDLF architectures depending on the task. The model is capable of training via linear probing or fine-tuning and has many adjustable hyperparameters. Images are first passed through the preprocessing and data augmentation pipelines described in this chapter before latent feature representations are extracted by the selected backbone. The latent feature representations are then passed to the appropriate classification head based on the selected backbone, where final diagnostic probabilities are computed and output. Model performance is maximized through decision threshold tuning during the test set evaluation stage, which is a process widely used in machine learning tools designed for medical diagnostics.

Critically, the OmniFusion model isolates the underlying backbone architecture and its accompanying classification head as the sole structural variables during comparative experiments. The OmniFusion model also enhances the base architectures adapted from Yan et al. (2025) and Lilhore et al. (2025) by enforcing advanced regularization and optimization techniques not utilized in the original implementations. The transfer learning pipelines, illustrated in Figure 3.1, also differ from the original implementations. Model weights for the PanDerm backbone are initially loaded from ImageNet-21K rather than ImageNet-1K in an effort to reduce the number of epochs required for training convergence.

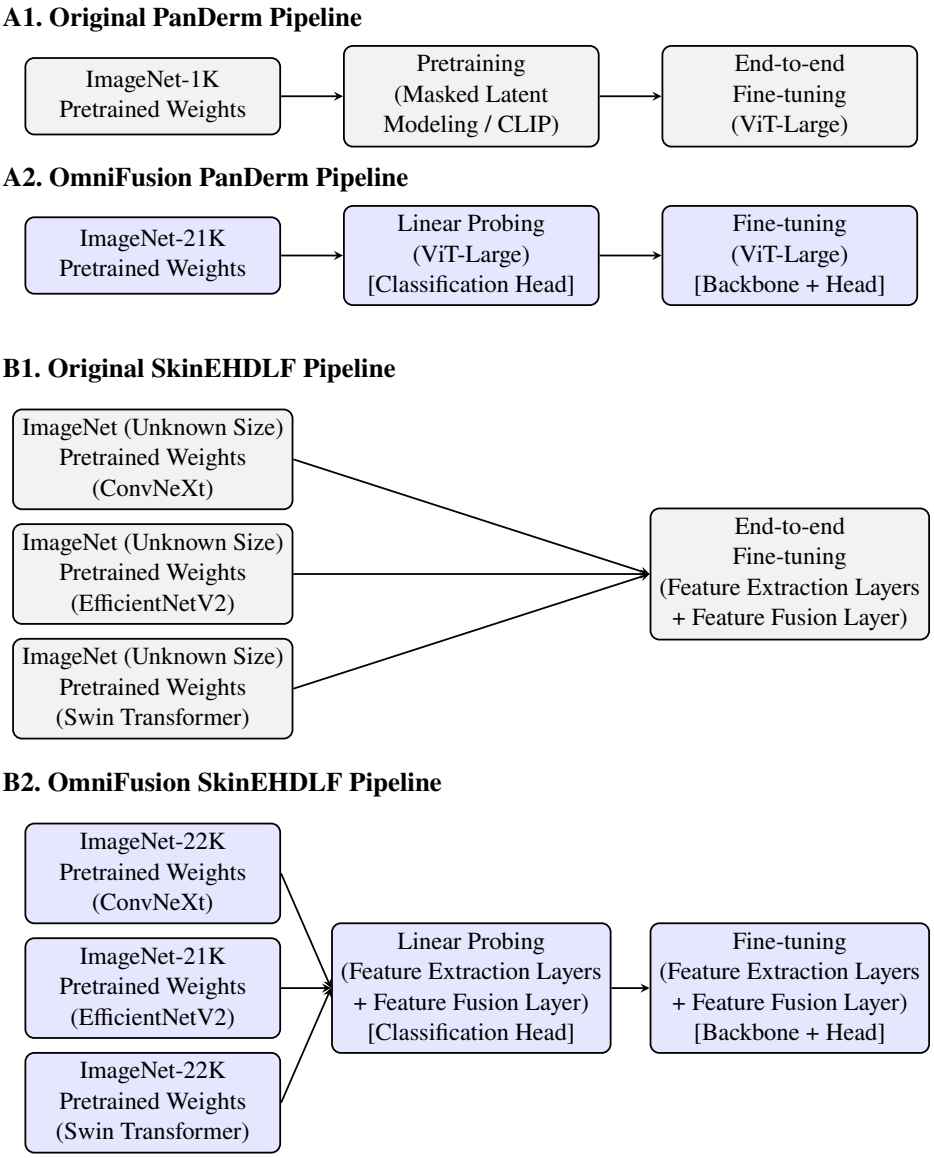


Figure 3.1: Comparison of Transfer Learning Pipelines Between OmniFusion and Original Implementations

3.6 Training and Optimization Strategies

3.6.1 Linear Probing then Fine-Tuning (LP-FT)

An LP-FT approach was used during training to improve performance through transfer learning. The training pipeline for each model involved loading pretrained weights, performing a linear probe using the training and validation sets, and using model weights from the linear probing stage to start a fine-tuning stage. The model weights loaded into the fine-tuning stage were selected from the best-performing linear probe epoch in terms of AUROC. The use of LP-FT was justified by recent research that is further explained in Subsection 2.2.1.

3.6.2 Weighted Random Sampling

A weighted random sampler was used during training to increase the number of malignant lesion samples seen by the model in each training epoch. The sampler was used to address the severe class imbalance in the datasets, particularly the SLICE-3D dataset, which contains over a thousand times more benign lesion samples compared to malignant lesion samples. This approach helped to prevent the model from becoming biased towards predicting the majority class and improved its ability to learn meaningful patterns associated with the minority class.

3.6.3 Spatial Augmentations

The Mixup and CutMix spatial augmentations were applied to each model during the training stage to improve their robustness and generalizability. Recent research provided the rationale for using Mixup and CutMix and is further explained in Subsection 2.2.2. The use of Mixup and CutMix also helped to mitigate the tendency towards overfitting associated with the use of a weighted random sampler, which causes the model to see the same minority class samples multiple times in each batch thus leading to memorization.

3.6.4 Hyperparameters and Loss Functions

Final hyperparameter settings and loss functions were chosen for each model after extensive experimentation to discover the combination that produced the highest AUROC for each backbone

and dataset. Details regarding hyperparameter settings and loss functions used for each model are outlined in Table 3.6.

Table 3.6: Hyperparameter Configurations and Loss Functions for OmniFusion Models

Model	Stage	Batch[a]	Epochs	Loss Function	LR[b]	WD[c]	Drop[d]	DP[e]	Mix[f]	Cut[g]	Weight Ratio[h]
Model 1	LP	1024	20	Focal Loss	1.0e-02	0.10	0.5	0.3	0.0	0.0	1:1
	FT	80	20	Focal Loss	5.0e-05	0.10	0.5	0.3	0.0	0.0	1:1
Model 2	LP	768	10	CE[i]Loss	1.0e-03	0.10	0.3	0.3	0.0	0.0	1:1
	FT	64	40	Weighted CE Loss	1.0e-05	0.05	0.3	0.2	0.8	1.0	2:1
Model 3	LP	1024	10	CE Loss	1.0e-02	0.10	0.5	0.3	0.0	0.0	1:1
	FT	80	20	Soft Target CE Loss	5.0e-05	0.10	0.5	0.3	0.8	1.0	1:1
Model 4	LP	768	10	CE Loss	1.0e-03	0.10	0.3	0.3	0.0	0.0	1:1
	FT	64	40	Weighted CE Loss	5.0e-05	0.05	0.3	0.2	0.8	1.0	2:1
Model 5	LP	768	10	CE Loss	1.0e-03	0.10	0.3	0.3	0.0	0.0	1:1
	FT	64	40	Weighted CE Loss	1.0e-05	0.05	0.3	0.2	0.8	1.0	2:1

[a]Batch Size; [b]Learning Rate; [c]Weight Decay; [d]Dropout; [e]Drop Path; [f]Mixup; [g]CutMix; [h]Ratio of weights applied to classes (minority:majority); [i]Cross-Entropy.

3.6.5 Test-Time Augmentation

Test-time augmentation (TTA) was applied to the test data for all models to enhance accuracy and robustness. Random transformations such as flipping and rotation are applied to the images to improve the confidence of predictions. Both the original and transformed images are evaluated by a trained model and the results of the predictions for each image are averaged to produce a more reliable final prediction.

3.7 Experimental Design and Phased Evaluation

To systematically address the research gaps identified in Chapter 1, the experimental methodology was divided into four distinct phases in this study. The phased approach was used to provide clarity and ensure that individual variables of interest are isolated and assessed independently. Each phase will be evaluated in detail in Chapter 4.

3.7.1 Phase 1: Baseline Establishment

The objective of the first phase was to establish rigorous control groups. Model 1 and Model 2, which respectively use the PanDerm and SkinEHDLF backbone architectures, were both trained exclusively on the unimodal SLICE-3D dataset. Their performance was compared against one another and against models in future phases to determine the effect of dataset and backbone choice. In addition, the use of the SLICE-3D dataset by both Yan et al. (2025) and Lilhore et al. (2025) to train their models allowed for direct comparison of their performance to the performance of the OmniFusion models. The comparison was necessary to establish the impact of the training and optimization strategies applied to the OmniFusion models.

3.7.2 Phase 2: Multimodality Impact Assessment

The second phase isolated the impact of image type multimodality in training. Model 3 and Model 4 were trained on both the SLICE-3D and supplemental datasets, integrating dematoscopic, clinical, TBP, and dermatopathology image modalities. Model 3 used the PanDerm backbone while Model 4 relied on the SkinEHDLF backbone. Intra-architectural comparison against the unimodal baselines established by Model 1 and Model 2 was performed to isolate and quantify the performance delta generated by dataset multimodality. In addition, the performance of Model 5 was evaluated to provide additional insight into the impact of training set multimodality and size by removing the SLICE-3D dataset from the training set.

3.7.3 Phase 3: Inter-Architectural Comparative Analysis

While Phase 2 compared how data affects a single model, Phase 3 compared the models against one another. Architectural superiority was established by cross-analyzing the performance of the PanDerm backbone against the SkinEHDLF backbone by controlling for the training, validation, and test sets used in Phase 1 and Phase 2. Inter-architectural comparison between Model 1 and Model 2 as well as between Model 3 and Model 4 was performed to determine which backbone is more performant when trained on unimodal data, on multimodal data, and in general. In

addition, an inter-architectural comparison between the performance deltas produced in Phase 2 was performed to discover the impact of backbone choice on the handling of multimodal data.

3.7.4 Phase 4: Pre-Clinical Feasibility Assessment

In the final phase, focus was shifted from performance comparison between models to comparison against clinically viable thresholds. The sensitivity and specificity of the best performing model were evaluated against that of the clinically-deployed and peer-reviewed model presented by Smak Gregoor et al. (2023). In addition, cost-effectiveness was determined in terms of the parameters defined by Smak Gregoor et al. (2023). The best performing model was also compared to a list of preconditions for physician endorsement as established by Sangers et al. (2025). Finally, the ability of the best performing model to generalize to domain-shifted data was evaluated.

3.8 Evaluation Metrics

A standardized set of evaluation metrics was employed to assess the performance of models across all experimental phases. Given the severe class imbalance towards benign lesions, the use of standard accuracy as a metric would be misleading since high accuracy could be achieved by predicting only the majority class. Therefore, performance was evaluated using BAcc, sensitivity, specificity, and the AUROC curve. The metrics rely on the foundational calculations of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), where a "positive" indicates a malignant lesion and a "negative" represents a benign lesion.

3.8.1 Sensitivity (Recall or True Positive Rate (TPR))

Sensitivity measures the proportion of actual malignant lesions correctly predicted by the model. It is calculated as:

$$TPR = \frac{TP}{TP + FN} \tag{3.1}$$

Sensitivity is arguably the most critical metric in the context of pre-clinical, DTC diagnostic tools. High sensitivity would indicate a low rate of missed cancer diagnoses and vice versa, in either

case strongly impacting patient outcomes by either expediting or potentially delaying medical intervention. Therefore, maximizing sensitivity is prioritized to ensure optimal patient outcomes.

3.8.2 Specificity (True Negative Rate (TNR))

Specificity measures the proportion of actual benign lesions correctly predicted by the model. It is calculated as:

$$TNR = \frac{TN}{TN + FP} \tag{3.2}$$

It is also important that medical diagnostic tools, particularly ones intended for DTC use, achieve high specificity. A high specificity would indicate a low rate of false cancer diagnoses and vice versa; in a scenario with low specificity, a high number of false positives could lead to unnecessary anxiety for laymen users and burden dermatologists with healthy patients. Therefore, a high specificity is necessary to ensure improved patient triaging when using a diagnostic tool in pre-clinical settings.

3.8.3 Balanced Accuracy (BAcc)

BAcc provides a single, holistic measure of performance that penalizes models for overfitting to the benign majority class. Its power lies in its definition as the arithmetic mean of sensitivity and specificity:

$$BAcc = \frac{TPR + TNR}{2} \tag{3.3}$$

By weighting the accuracy of the minority and majority classes equally, BAcc ensures the predictive power of the model is discriminative and robust across lesion types. Although AUROC is often used as the primary performance metric when evaluating diagnostic models like the one in this study, Codella et al. (2019) argued that BAcc could be a superior metric to ensure robustness across clinical settings when training on an imbalanced dataset.

3.8.4 Area Under the Receiver Operating Characteristic Curve (AUROC)

AUROC allows for the evaluation of aggregate classification performance of the models across all possible decision thresholds. It is defined by the following integration, where TPR and FPR represent the True Positive Rate and the False Positive Rate, respectively:

$$AUROC = \int_0^1 TPR(FPR)\, d(FPR) \tag{3.4}$$

AUROC represents the area under the receiver operating curve, hence its name and the use of integration in its calculation. It provides a threshold-invariant measure of the classification ability of a model, allowing for a performance comparison to other models without considering threshold tuning.

3.8.5 Weighted F1-Score

The standard F1-score serves as a metric to balance the trade-off between false positives and false negatives. It is represented as the harmonic mean of precision and sensitivity:

$$F1 = 2 \times \frac{Precision \times TPR}{Precision + TPR} \tag{3.5}$$

Precision is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{3.6}$$

Weighted F1-score was recorded in this study to adapt standard F1-score for the imbalanced datasets used. The F1-score for each class is independently calculated and aggregated, with the score of each class weighted by the number of true instances of the class within the dataset:

$$Weighted\ F1 = \sum_{i=1}^{C} W_i \times F1_i \tag{3.7}$$

In the calculation, C represents the number of classes and W_i is the proportion of true instances belonging to class i. Although Weighted F1-score provides a useful supplementary view of overall model stability, it is still influenced by the majority benign class due to its reliance on class proportions. Consequently, BAcc and AUROC were prioritized as the primary determinants of pre-clinical viability as well as algorithmic fairness, while Weighted F1-scores were computed for comprehensive benchmarking.

3.9 Hardware and Software Environment

3.9.1 Hardware Infrastructure

The training for all models was performed on the Pace University High-Performance Computing (HPC) cluster. Each compute node utilized in training was equipped with eight Intel Xeon CPU cores and a single Tesla V100 GPU with 32GB of VRAM. Due to the massive size of the training datasets, the ability to train multiple folds in parallel across nodes along with a relatively high amount of VRAM per node made completion of this study possible in a timely manner.

3.9.2 Software

The codebase for the OmniFusion model was initially forked from the PanDerm repository provided by Yan et al. (2025) at https://github.com/SiyuanYan1/PanDerm. Python 3.9.23 was used to run the code within a dedicated virtual environment on the HPC cluster. Job scheduling was managed using Slurm Workload Manager. Deep learning architectures and training mechanisms were implemented using PyTorch 2.8.0 for CUDA 12.8. Various other libraries such as scikit-learn, NumPy, and Pandas were used for preprocessing, data manipulation, and K-fold cross-validation splitting.

Dependencies were strictly version-controlled for most libraries within the virtual environment to ensure reproducibility. A repository for the OmniFusion codebase which includes version information for all dependencies is available at https://github.com/josharnow/OmniFusion.

Chapter 4

Results

This chapter reports the performance of each model across four experimental phases as outlined in Section 3.7. As described in Section 3.8, sensitivity, specificity, BAcc, AUROC, and weighted F1-score were selected for presentation. Various tables and figures are used to highlight differences in performance and facilitate discussion regarding their implications in Chapter 5.

Performance metrics are reported as the mean across cross-validation folds, with the ± symbol preceding the standard deviation to denote model variance across folds and illustrate model stability. Due to time constraints, decision threshold tuning was selectively performed on models that exhibited the highest AUROC; however, all models were evaluated at a default threshold of 0.5 for consistency and comparison. Additional tables and figures are included in Appendix A to provide additional insight into the performance of each model across all experimental phases, including aggregated metrics, fold-wise performance, confusion matrices, and training/validation curves.

4.1 Phase 1: Baseline Establishment

The key performance metrics for Phase 1 are presented in Table 4.1. In this phase, Model 1 and Model 2 were trained on the unimodal SLICE-3D training set using PanDerm and SkinEHDLF backbones, respectively. Model 2 demonstrated superior performance across all metrics compared to Model 1, regardless of the decision threshold used. Most notably, Model 2 achieved an AUROC of 95.01%, which is a significant improvement over Model 1's AUROC of 88.41%. Furthermore, Model 2's BAcc of 87.86% at the default threshold of 0.5 represents a substantial increase from Model 1's BAcc of 80.65%. Model 2 also exhibited lower variance across folds for all metrics, as demonstrated by its lower standard deviations compared to Model 1.

Decision threshold tuning was applied to Model 2 in order to optimize it for pre-clinical deployment, where maximizing sensitivity while balancing specificity is often a priority. Shifting the decision threshold from 0.5 to 0.4431 resulted in a 4.08% absolute increase in sensitivity from 84.21% to 88.29%, albeit with a 2.73% absolute decrease in specificity from 91.52% to 88.79%. A

marginal improvement in BAcc was also achieved, increasing from 87.86% to 88.54%. Given that AUROC is threshold-invariant, it remained unchanged at 95.01%, while the weighted F1-score decreased from 95.47% to 93.96%. These results demonstrate the inherent trade-off between sensitivity and specificity when tuning decision thresholds for pre-clinical deployment.

Table 4.1: OmniFusion Model Baseline Performance Using SLICE-3D Training Set

Model	Backbone	Threshold	BAcc (%)	Sens[a] (%)	Spec[b] (%)	AUROC (%)	F1[c] (%)
Model 1	PanDerm	0.5000	80.65 ± 5.72	74.54 ± 10.73	86.75 ± 3.95	88.41 ± 3.77	92.76 ± 2.33
Model 2	SkinEHDLF	0.5000	87.86 ± 3.98	84.21 ± 8.17	**91.52 ± 1.08**	**95.01 ± 1.73**	**95.47 ± 0.59**
Model 2	SkinEHDLF	0.4431	**88.54 ± 3.15**	**88.29 ± 5.89**	88.79 ± 1.29	**95.01 ± 1.73**	93.96 ± 0.72

[a]Sensitivity; [b]Specificity; [c]Weighted F1-Score.
Note: Bold values indicate superior performance relative to other models for the same metric.

4.2 Phase 2: Multimodality Impact Assessment

The key performance metrics for Phase 2 are presented in Table 4.2. Model 3 and Model 4 were trained on SLICE-3D data as well as the supplementary data using PanDerm and SkinEHDLF backbones, respectively. Consequently, both models were trained on a multimodal dataset as opposed to the unimodal dataset used in Phase 1. By doing so, the impact of multimodal training data on model performance can be assessed by comparing the results of Model 3 and Model 4 against their corresponding unimodal baselines from Phase 1. For the PanDerm backbone, the use of multimodal data in Model 3 yielded a substantial 6.93% increase in sensitivity alongside a marginal 0.41% improvement of AUROC. BAcc only increased by 1.27% due to an accompanying 4.39% decrease in specificity. Conversely, the SkinEHDLF backbone experienced a performance regression when using multimodal data; Model 4 demonstrated a marginal 0.65% decrease in AUROC and a substantial 5.59% decrease in sensitivity when using a 0.5 threshold. Although specificity increased by 0.61%, the overall BAcc decreased by 2.49%.

Decision threshold tuning partially mitigated the performance regression of Model 4. When optimized for sensitivity, the tuned Model 4 exhibited a smaller 0.76% deficit in sensitivity and a

1.58% deficit in BAcc compared to its tuned unimodal counterpart. Despite this multimodal regression, the SkinEHDLF backbone largely maintained its architectural advantage over the PanDerm backbone; the non-tuned Model 4 still displayed superior performance compared to Model 3 across all metrics except for sensitivity.

Table 4.2: OmniFusion Model Multimodal Performance and Baseline Variance

Model	**Backbone**	**Threshold**	**BAcc (%)**	**Sens[a] (%)**	**Spec[b] (%)**	**AUROC (%)**	**F1[c] (%)**
Model 3	PanDerm	0.5000	81.92 ± 3.79 (↑ 1.27)	81.47 ± 10.28 (↑ 6.93)	82.37 ± 6.24 (↓ 4.39)	88.83 ± 3.49 (↑ 0.41)	90.12 ± 3.83 (↓ 2.64)
Model 4	SkinEHDLF	0.5000	85.37 ± 6.38 (↓ 2.49)	78.62 ± 14.55 (↓ 5.59)	**92.13 ± 2.32** (↑ 0.61)	**94.36 ± 1.70** (↓ 0.65)	**95.79 ± 1.25** (↑ 0.32)
Model 4	SkinEHDLF	0.3884	**86.96 ± 2.26** (↓ 1.58)	**87.53 ± 6.19** (↓ 0.76)	86.39 ± 3.54 (↓ 2.39)	**94.36 ± 1.70** (↓ 0.65)	92.57 ± 2.04 (↓ 1.39)

[a]Sensitivity; [b]Specificity; [c]Weighted F1-Score.
Note: Values in parentheses denote the intra-architectural absolute change (Δ) in percentage points compared to the corresponding unimodal baseline model from Phase 1. Bold values indicate superior performance relative to other models for the same metric.

Model 5 was evaluated to investigate the cause of Model 4's regression under multimodal training conditions. The impact of training with the supplementary data was isolated by removing the SLICE-3D data from the training set and training Model 5 exclusively on the supplementary data using the same SkinEHDLF backbone. Model 5 achieved an AUROC of 53.06% and a BAcc of 49.31%, representing 41.95% and 38.55% drops in performance compared to Model 2, respectively. Despite the significant performance drop of Model 5 compared to Model 2, Model 4 only experienced a 0.65% decrease in AUROC and a 2.49% decrease in BAcc compared to Model 2. The performance of Model 5 suggests that the dip in Model 4's performance compared to Model 2 is more attributable to a severe domain mismatch between the supplementary training data and the SLICE-3D test data rather than an inability of the SkinEHDLF architecture to handle multimodal inputs.

4.3 Phase 3: Inter-Architectural Comparative Analysis

Inter-architectural performance metrics and the results of statistical significance testing are detailed in Table 4.3. To isolate the impact of architecture on classification performance, the PanDerm and SkinEHDLF backbones were evaluated against each other while controlling for dataset modality and decision threshold. Specifically, Model 1 was compared against Model 2 and Model 3 was compared against Model 4 at a decision threshold of 0.5. Statistical significance was evaluated using a dual-test approach: paired t-tests for continuous fold-wise metrics (AUROC, BAcc, F1) and McNemar's Chi-Square (χ^2) tests for discrete image-wise classifications (sensitivity, specificity).

Under unimodal training conditions, the SkinEHDLF architecture demonstrated highly significant superiority ($p < 0.01$) across all metrics. Most notably, SkinEHDLF established a 6.60% absolute advantage in AUROC (t-statistic = 6.013, $p = 0.0002$) and fixed a substantial proportion of PanDerm's classification errors across both malignant and benign sub-populations ($\chi^2 = 15.520$ and 7216.868, respectively; both $p < 0.0001$). The very large χ^2 value for specificity is a consequence of using a very highly imbalanced dataset that contains more benign examples than malignant ones by several orders of magnitude.

However, the architectural performance gap narrowed considerably under multimodal training conditions. While SkinEHDLF maintained highly significant advantages in overall AUROC ($p = 0.0008$), specificity ($p < 0.0001$), and weighted F1-score ($p = 0.0017$), the differences in BAcc and sensitivity were found to be statistically insignificant ($p = 0.1618$ and $p = 0.2724$, respectively). Therefore, the difference in BAcc and true-positive malignant detection between the two architectures should be considered negligible when both are trained using multiple imaging modalities. Nevertheless, the threshold-invariant superiority of the SkinEHDLF architecture's global classification capacity is demonstrated by its superior AUROC. The overlaid receiver operating characteristic curves for all evaluated models are displayed in Figure 4.1.

Table 4.3: Inter-Architectural Comparative Analysis and Statistical Testing

Modality	Metric	PanDerm	SkinEHDLF	Abs. Δ	Test Type[a]	Statistic[b]	p-value
Unimodal	BAcc (%)	80.65 ± 5.72	**87.86 ± 3.98**	↑ 7.21	t-test	5.616	0.0003
	Sens (%)	74.54 ± 10.73	**84.21 ± 8.17**	↑ 9.67	McNemar (χ^2)	15.520	< 0.0001
	Spec (%)	86.75 ± 3.95	**91.52 ± 1.08**	↑ 4.77	McNemar (χ^2)	7216.868	< 0.0001
	AUROC (%)	88.41 ± 3.77	**95.01 ± 1.73**	↑ 6.60	t-test	6.013	0.0002
	F1 (%)	92.76 ± 2.33	**95.47 ± 0.59**	↑ 2.71	t-test	3.410	0.0078
Multimodal	BAcc (%)	81.92 ± 3.79	85.37 ± 6.38	↑ 3.45	t-test	1.524	0.1618
	Sens (%)	81.47 ± 10.28	78.62 ± 14.55	↓ 2.85	McNemar (χ^2)	1.205	0.2724
	Spec (%)	82.37 ± 6.24	**92.13 ± 2.32**	↑ 9.76	McNemar (χ^2)	25214.251	< 0.0001
	AUROC (%)	88.83 ± 3.49	**94.36 ± 1.70**	↑ 5.53	t-test	4.932	0.0008
	F1 (%)	90.12 ± 3.83	**95.79 ± 1.25**	↑ 5.67	t-test	4.405	0.0017

[a]Continuous metrics were evaluated via paired t-tests evaluating cross-validation fold means. Categorical metrics were evaluated globally via McNemar's test (χ^2) on concatenated image-wise predictions (N = 401,059); [b]Test statistic corresponding to the specified test type (either t-statistic or χ^2).
Note: All models evaluated at the default 0.5 decision threshold. The Absolute Δ column denotes the inter-architectural percentage point difference achieved by transitioning from the PanDerm backbone to the SkinEHDLF backbone. Bold values indicate statistically significant superior performance ($p < 0.05$) and underlined bold values indicate highly significant performance ($p < 0.01$) for that specific metric and modality pairing.

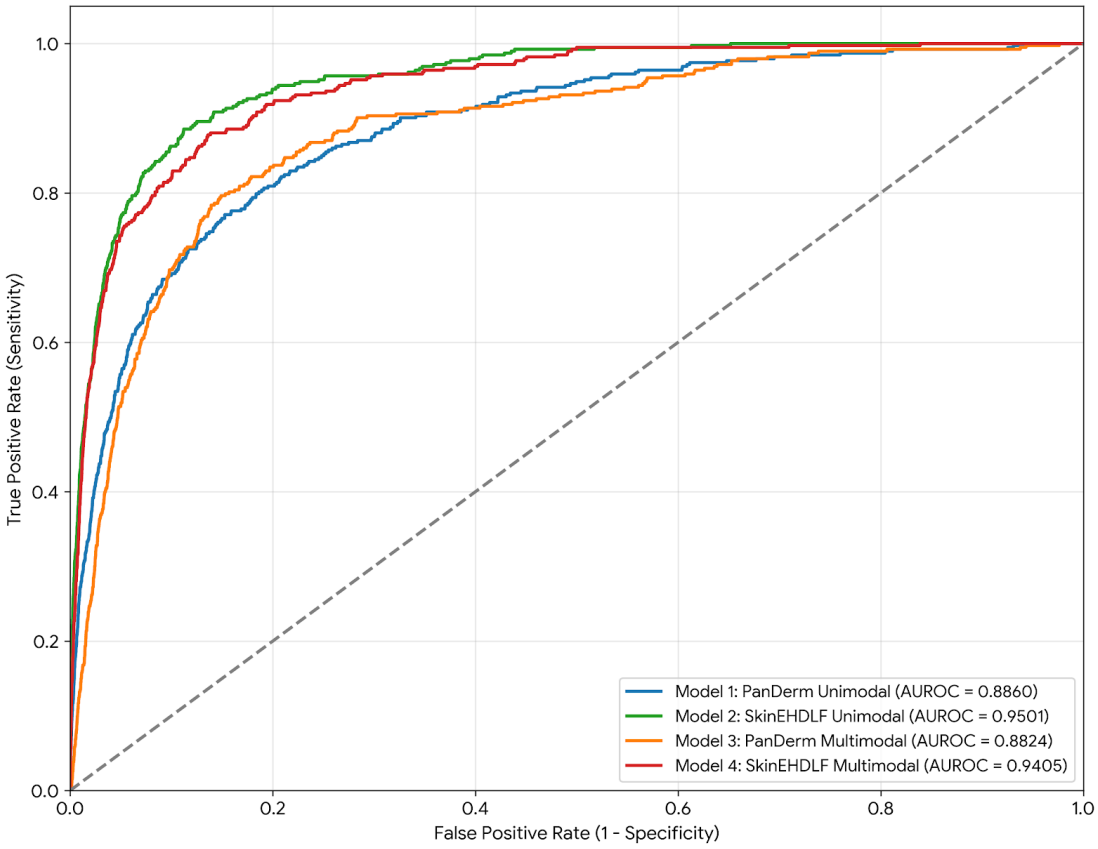


Figure 4.1: Receiver Operating Characteristic Curves by OmniFusion Model

4.4 Phase 4: Pre-Clinical Feasibility Assessment

OmniFusion's viability for pre-clinical deployment was assessed by comparing the performance of its best performing model, Model 2, against thresholds established by the literature for DTC skin cancer screening. Model 2-DS, a zero-shot configuration of Model 2 tested exclusively on out-of-distribution, domain-shifted data, provides insight into how the performance of the model degrades under domain shift and without further training on the target distribution. A DTC skin cancer screening model made available through an app in the Netherlands was analyzed by Smak Gregoor et al. (2023) and provides relevant benchmarks for evaluating the feasibility of OmniFusion's pre-clinical deployment through its reported sensitivity and specificity. In addition, a skin cancer diagnosis reader study by Yan et al. (2025) provides a relevant benchmark for evaluating the feasibility of OmniFusion's pre-clinical deployment in terms of its performance compared to that of high experience clinicians. The results of this evaluation are presented in Table 4.4.

Table 4.4: Evaluation of Model 2 and Model 2-DS Against Established Diagnostic Thresholds

Model	Clinical Standard	Min. Sens[a](%)	Min. Spec[b](%)	Model Sens (%)	Model Spec (%)	Viability
Model 2	Smak Gregoor et al.	>87	>70	88.29	88.79	**Pass**
(Internal)	High Experience Clinician[c]	>78	>78	88.29	88.79	**Pass**
Model 2-DS	Smak Gregoor et al.	>87	>70	91.14	47.38	**Sens Only**
(External)	High Experience Clinician	>78	>78	91.14	47.38	**Sens Only**

[a]Sensitivity; [b]Specificity; [c]According to values derived from the skin cancer diagnosis reader study by Yan et al. (2025). Only accuracy was reported in the original study, so sensitivity and specificity were estimated assuming an equal error rates across classes.
Note: Model 2 metrics utilize the tuned decision threshold ($t = 0.4431$) evaluated on the internal SLICE-3D test set. Model 2-DS performance metrics utilized for comparison are derived from the HIBA dataset evaluation, representing the most stable out-of-distribution clinical scenario.

Model 2 successfully surpassed all established quantitative thresholds for pre-clinical viability. At its optimized decision threshold of 0.4431, Model 2 achieved a sensitivity of 88.29% and a specificity of 88.79%, which both exceed their respective thresholds established by Smak Gregoor et al. (2023) and Yan et al. (2025). Since Model 2-DS was evaluated on multiple datasets of varying domain shift severity, the most stable out-of-distribution clinical scenario was selected for comparison against the established thresholds. Model 2-DS was unable to meet the established thresholds for pre-clinical viability in terms of specificity, achieving a sensitivity of

91.14% but a specificity of only 47.38%. The drop in performance can be largely attributed to zero-shot learning on out-of-distribution data; according to Lilhore et al. (2025), "[SkinEHDLF] requires fine-tuning and adaptation to the particular characteristics of each dataset to achieve optimal performance" during cross-domain evaluation. It should be noted that although Lilhore et al. (2025) present superior performance metrics for SkinEHDLF on out-of-distribution data, contextual information implies that they did not perform a zero-shot evaluation. A breakdown of Model 2-DS performance across all evaluated out-of-distribution datasets is presented in Table 4.5 to provide insight into how the model's performance degrades under varying degrees of domain shift.

Table 4.5: Zero-Shot Domain Shift Performance of Model 2-DS Across External Datasets

Dataset	**Samples**	**AUROC (%)**	**BAcc (%)**	**Sens[a] (%)**	**Spec[b] (%)**	**F1[c] (%)**
HIBA	1,320	79.93	69.26	91.14	47.38	66.89
PAD-UFES-20	436	78.48	61.35	97.83	24.87	38.03
HAM10000	7,989	69.40	59.92	88.01	31.84	45.12
BCN20000	9,926	68.52	53.39	97.25	9.54	41.39
DDI	502	66.16	61.20	79.51	42.89	54.26
Derm7pt	1,371	65.83	54.19	89.62	18.75	34.94
MSKCC	7,221	63.61	58.54	82.62	34.46	49.25
Dermnet	15,684	58.14	53.65	89.92	17.37	28.34
PATCH16[d]	33,165	35.07	38.55	35.61	41.49	44.02

[a]Sensitivity; [b]Specificity; [c]Weighted F1-Score; [d]Histological whole-slide image patches representing extreme domain shift.

Cost effectiveness is another critical consideration for the pre-clinical viability of DTC skin cancer screening models. Smak Gregoor et al. (2023) present Incremental Cost-Effectiveness Ratio (ICER) estimates for DTC skin cancer screening in the Netherlands which are relevant for evaluating the cost-effectiveness of OmniFusion's pre-clinical deployment. ICER is defined as a cross-sectional estimate of the costs for detecting one new skin premalignancy or malignancy compared to the current standard of care, without taking into account the downstream cost savings from early detection. According to Matsumoto et al. (2018), the cost to detect an additional skin (pre)malignancy through total body examination by a dermatologist in the United States is

approximately $2,346, or €1,994 if converted at the current exchange rate. Smak Gregoor et al. (2023) found that the DTC app they evaluated achieved an ICER of €2,567, or $3,021, per additional (pre)malignant lesion detected. Using the ICER table from Smak Gregoor et al. (2023) presented in Figure 4.2, Model 2 achieved an estimated ICER of €1,985, or $2,336, per additional (pre)malignant lesion detected, which is more cost-effective than both the DTC app evaluated by Smak Gregoor et al. (2023) and diagnosis by a dermatologist in the United States. Model 2-DS achieved an inferior estimated ICER of €3,421, or $4,026, per additional (pre)malignant lesion detected.

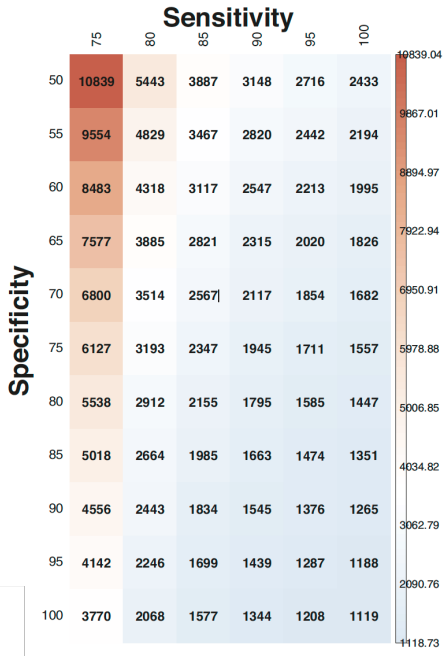


Figure 4.2: Incremental Cost-Effectiveness Ratio Table for DTC Skin Cancer Screening in the Netherlands

Source: Adapted from Smak Gregoor et al. (2023).
Note: ICER values are expressed in Euros.

Sangers et al. (2025) present a qualitative set of preconditions for DTC skin cancer screening model endorsement by clinical experts. Table 4.6 compares Model 2 and Model 2-DS against the qualitative preconditions to provide a more holistic assessment of their pre-clinical viability. While both models arguably meet the evidence-based precondition of having a

representative sample, factors like skin color diversity may be lacking. The requirement for independent verification remains unmet at this stage; because OmniFusion was developed within the context of this study, subsequent validation by a third-party clinical or regulatory body is required to satisfy this precondition. Likewise, the clinical practice integration, liability clarity, and app design preconditions are all currently out of scope for OmniFusion since it is a backend algorithmic architecture that lacks the frontend application layer, security infrastructure, and legal frameworks required for DTC deployment. However, an app similar to the one evaluated by Smak Gregoor et al. (2023) could be developed in the future to facilitate the deployment of OmniFusion and testing by an independent organization, thus meeting the remaining preconditions for clinical endorsement.

Table 4.6: Evaluation of OmniFusion Against Qualitative Pre-Clinical Endorsement Preconditions

Requirement Category	**Specific Precondition**	**Model 2**	**Model 2-DS**
Evidence-Based Verification	Tested by an independent organization	**Unmet**	**Unmet**
	Large representative sample	**Met**[a]	**Met**[b]
Clinical Practice Integration	Appropriate communication of medical info	**N/A**[c]	**N/A**[c]
	Report sharing functionality	**N/A**[c]	**N/A**[c]
	Adequate protection of patient data	**N/A**[c]	**N/A**[c]
	Practitioner involvement in implementation	**N/A**[c]	**N/A**[c]
Liability Clarity	Clarity regarding adverse events	**N/A**[c]	**N/A**[c]
App Design	Accessible and inclusive app design	**N/A**[c]	**N/A**[c]

[a]Evaluated on the SLICE-3D dataset, is sourced from many institutions around the world across an even distribution of ages, sexes, and body sites.
[b]Evaluated across nine diverse international datasets representing varying patient demographics and imaging modalities.
[c]Out of scope.

Chapter 5
Discussion

The experimental results of this study successfully validate OmniFusion as a viable foundation model for pre-clinical skin cancer diagnostics, achieving performance that is on par with the state-of-the-art models it is derived from. Through comparison against pre-clinical benchmarks in the literature, OmniFusion demonstrated that it is capable of performing at a suitable sensitivity and specificity for safe DTC deployment to improve patient triage and outcomes. A phased evaluation of various OmniFusion configurations revealed that the adaptive attention-based feature fusion mechanism used by the SkinEHDLF architecture can achieve superior performance to the vision transformer approach utilized by PanDerm.

Furthermore, the results highlight the complexities inherent to pre-clinical deployment, particularly regarding the handling of multimodal data and out-of-distribution domain shifts. The integration of a LP-FT approach alongside advanced spatial augmentations like CutMix and Mixup ultimately resulted in high performance when testing on the SLICE-3D dataset. However, OmniFusion experienced severe performance degradation when performing zero-shot inference on domain-shifted datasets and varying degrees of success in training convergence.

This chapter contextualizes these findings within the broader landscape of medical AI. The subsequent sections provide insight into the performance of OmniFusion in comparison to the original PanDerm and SkinEHDLF implementations, the architectural advantages of each backbone as implemented in OmniFusion, the reasons for modality collapse, and the challenges of domain shift. Additionally, this chapter discusses model generalizability as evidenced by performance variance across folds, the success of strategies used to address severe class imbalance during model training, and the inherent trade-offs of threshold tuning. The overarching feasibility of DTC implementation and the limitations of this study are also presented to inform future research directions. Finally, the broader implications of this research and potential avenues for future work are discussed.

5.1 Comparison Against Original PanDerm and SkinEHDLF Benchmarks

In order to evaluate the relative efficacy of the OmniFusion framework, it must be compared against the original performance benchmarks for the underlying backbone architectures originally proposed by Lilhore et al. (2025) and Yan et al. (2025). The performance of each OmniFusion model using either a PanDerm or SkinEHDLF backbone closely aligns with the original benchmarks for their respective architectures. Both Model 1 and Model 3 achieve virtually identical AUROC to the original PanDerm implementation, implying that training on multimodal data had negligible benefit for the ViT-Large architecture used in terms of its ability to classify the unimodal SLICE-3D data. In addition, the tight alignment between model performance suggests a superior methodological efficiency of the OmniFusion framework. The original PanDerm implementation underwent hundreds of pretraining epochs on a larger dataset to achieve its performance, yet Model 1 and Model 3 were able to achieve comparable performance with around 40 and 30 training epochs, respectively, including linear probing epochs. The OmniFusion implementation offers a significant reduction in the training time and computational resources required to achieve effectively the same performance. The improved efficiency may be attributed to the use of an LP-FT approach as well as the use of in-place spatial augmentations; specifically regarding the LP-FT approach, Model 3 appears to have benefitted from the linear probing stage as evidenced by a performance jump at the start of its fine-tuning stage visualized in Figure A.11.

Conversely, the comparison between Model 2 and Model 4 against the original SkinEHDLF benchmarks highlights critical discrepancies in evaluation methodology. Lilhore et al. (2025) reported an unusually high AUROC of 99.80% for their model on the SLICE-3D dataset. In contrast, Model 2 and Model 4 achieved highly performant, yet comparatively lower, AUROC of 95.01% and 94.36%, respectively. This performance gap may be attributed to the data augmentation pipeline used in the original SkinEHDLF implementation, which produced a dataset six times larger than the one used in this study. Based on the pipeline described by Lilhore et al. (2025), the original authors appear to have applied geometric and color augmentations to the training data prior to separating it into training, validation, and test sets, which risked augmented versions of the same original images

being present in different sets. This would have led to significant data leakage and overfitting, thereby artificially inflating the performance metrics reported in the original SkinEHDLF paper.

By applying data augmentation techniques strictly in-place during the training phase, the OmniFusion framework ensured that the test sets remained entirely isolated and representative of unseen data. Therefore, although the classification metrics for Model 2 and Model 4 are nominally lower than the original SkinEHDLF benchmarks, they provide a significantly more realistic and clinically reliable estimation of the true diagnostic power of the adaptive attention-based feature fusion architecture. Furthermore, the 95.01% and 94.36% AUROC achieved by Model 2 and Model 4, respectively, confirm that the backbone architecture is highly robust and superior for this specific classification task relative to the ViT-Large approach used by PanDerm. A comparison of OmniFusion models implemented in this study against the original counterparts they were derived from is presented in Table 5.1.

Table 5.1: Comparison of OmniFusion Models Against Original Published Benchmarks

Model	**Backbone**	**Training Data**	**AUROC (%)**	**BAcc (%)**	**Sens (%)**	**Spec (%)**	**F1 (%)**
Original PanDerm[a]	ViT-Large	SLICE-3D + Multimodal	**89.00**	80.00	–	–	92.00
Model 1	PanDerm (ViT-Large)	SLICE-3D	88.41	80.65	74.54	86.75	**92.76**
Model 3	PanDerm (ViT-Large)	SLICE-3D + Supp.	88.83	**81.92**	81.47	82.37	90.12
Original SkinEHDLF[b]	Adaptive Fusion	SLICE-3D (Augmented)	**99.80**	–	**98.30**	**99.00**	**98.70**
Model 2	SkinEHDLF (Adaptive Fusion)	SLICE-3D	95.01	87.86	84.21	91.52	95.47
Model 2 (t = 0.4431)	SkinEHDLF (Adaptive Fusion)	SLICE-3D	95.01	88.54	88.29	88.79	93.96
Model 4	SkinEHDLF (Adaptive Fusion)	SLICE-3D + Supp.	94.36	85.37	78.62	92.13	95.79
Model 4 (t = 0.3884)	SkinEHDLF (Adaptive Fusion)	SLICE-3D + Supp.	94.36	86.96	87.53	86.39	92.57

[a]Derived from Yan et al. (2025) performance on the SLICE-3D test set. Sensitivity and specificity were not explicitly reported for this dataset.
[b]Derived from Lilhore et al. (2025) performance on their augmented SLICE-3D dataset. Balanced Accuracy was not reported (standard accuracy was 98.76%).
Note: OmniFusion models evaluated at the default 0.5 decision threshold unless a tuned threshold (t) is specified. Bold values indicate the best performing metric for each backbone architecture.

5.2 Advantages of PanDerm vs. SkinEHDLF Architectures in OmniFusion

The empirical superiority of the SkinEHDLF backbone architecture over the PanDerm backbone architecture across both unimodal and multimodal experimental configurations can be attributed to the radically different designs of the two backbones. The PanDerm backbone relies exclusively on a ViT-Large encoder, which excels at capturing long-range global contextual

relationships but is inefficient at extracting low-level spatial features within images. Subtle visual cues critical for skin cancer diagnosis, such as jagged lesion borders, are often lost or underrepresented during training. Thus, vision transformers possess inherent limitations within the realm of skin cancer diagnostics.

In contrast, the SkinEHDLF architecture mitigates the bottleneck posed by a single vision transformer encoder through the utilization of a multi-pathway feature extraction framework. Unlike a pure vision transformer, the feature fusion mechanism used by SkinEHDLF constructs a bottom-up spatial hierarchy by routing input images through two different CNNs in parallel. Each CNN processes images layer by layer, first extracting low-level spatial features such as localized edges before deeper layers extract features such as pigment networks. The deepest convolutional layer ultimately aggregates the extracted features to understand the macroscopic structure of the entire image. Simultaneously, a parallel Swin Transformer captures global contextual relationships in a manner similar to the ViT-Large encoder used by PanDerm, excelling at mapping how a lesion relates to the surrounding skin. The three independent pathways produce latent representations that are aggregated within a Feature Fusion Layer that dynamically weights the importance of each representation. The synthesis of features produces a richer, more comprehensive feature map for the classification head of the model than a single vision transformer could provide, as demonstrated by superior performance across all metrics for Model 2 and Model 4 compared to Model 1 and Model 3, respectively.

However, the architectural complexity of SkinEHDLF also introduces significant optimization challenges, specifically regarding training convergence stability. As illustrated by the validation curves for Model 2 and Model 4 in Figures A.10 and A.12, respectively, the SkinEHDLF backbone exhibited highly volatile learning dynamics during the transition between training stages. During the initial 10-epoch linear probing stage, the models' standard validation accuracy spiked while balanced accuracy plummeted, indicating that the frozen backbone and standard cross-entropy loss function were merely allowing the classification head to overfit by guessing the benign majority class. The transition to end-to-end fine-tuning, which introduced a class-weighted cross-entropy loss

function to penalize minority class errors at a 2-to-1 ratio, triggered a massive gradient shock. The models subsequently required the remainder of their 50-epoch training phases to stabilize. Despite this, Lilhore et al. (2025) reported similar accuracy metrics around the same number of epochs and went on to train for twice as many; while the loss curves did appear to level off for Model 2 and Model 4, it is possible that additional training epochs and hyperparameter adjustments may have resulted in further performance improvements.

Unlike the SkinEHDLF implementations, both Model 1 and Model 3 demonstrated steady improvements in validation performance across training epochs using the PanDerm backbone, suggesting that the simpler architecture of PanDerm may be more amenable to optimization and convergence with the hyperparameters used. The loss functions used for each backbone are likely to have played a role as well, with weighting being applied exclusively to the SkinEHDLF backbones to address class imbalance after other strategies resulted in complete model collapse during training. Nevertheless, despite the training convergence difficulties experienced by the SkinEHDLF backbone, higher performance across all metrics during testing confirms that the adaptive feature fusion approach is better suited for the high-variance visual characteristics of skin lesions.

5.3 Modality Collapse and Performance Degradation Under Domain Shifts

For OmniFusion models using the SkinEHDLF backbone, the addition of supplementary data to the training set resulted in marginal performance degradation in terms of AUROC. This phenomenon, known as modality collapse (Chaudhuri et al. 2025), occurs when models trained on multimodal data rely on a subset of the modalities while ignoring others. Considering how models using the PanDerm backbone only experienced marginal improvement in AUROC with the addition of supplementary data, it is possible that modality collapse suppressed potential performance gains that would otherwise be expected from training on a larger, more diverse dataset. Performance stagnation and degradation when adding supplementary data to the training set may be attributed to the proportion of dermatopathology images in the supplementary dataset, which was a much smaller fraction of the original PanDerm authors' complete multimodal dataset. Representing about

40% of the supplementary dataset used in this study, microscopic dermatopathology images are substantially different from the macroscopic 3D TBP images used in the SLICE-3D test set. The dominance of dermatopathology images in the supplementary dataset may have caused the OmniFusion backbones to overfit to features specific to that modality, thereby reducing their ability to properly classify images in the test set. The microscopic features learned from the dermatopathology images were likely irrelevant or even misleading when applied to the SLICE-3D test set, resulting in lower evaluation metrics. Unlike dermatopathology images, the features learned from the macroscopic dermatoscopic and clinical images in the supplementary dataset were more relevant to the SLICE-3D test set, which may explain why the PanDerm backbone was able to achieve a marginal improvement in AUROC with the addition of supplementary data.

The profound incompatibility between 3D TBP and dermatopathology imaging modalities is further corroborated by the zero-shot evaluation performance of Model 2-DS, presented in Table 4.5. Model 2-DS was trained exclusively on the unimodal SLICE-3D dataset and suffered severe performance degradation when evaluated on the PATCH16 subset, far more extreme than the degradation observed with other domain-shifted subsets. The near-identical poor evaluation performance of Model 5 on the SLICE-3D test set provides further evidence of the detrimental impact of dermatopathology images used in the training set, as it was trained exclusively on the supplementary dataset.

Ultimately, these findings confirm that the SkinEHDLF architecture is highly sensitive to modality shifts. While Lilhore et al. (2025) reported superior results relative to Model 2-DS when testing their SkinEHDLF model on the HAM10000 dataset, they indicate that their “strong performance across diverse datasets...requires fine-tuning and adaptation to the particular characteristics of each dataset to achieve optimal performance." Since the authors also do not state that they performed zero-shot inference, it is heavily implied that they trained their pretrained SkinEHDLF model on a portion of the HAM10000 dataset prior to evaluation on the remaining portion of the HAM10000 dataset. If that is how the authors performed their cross-domain performance evaluation, it suggests that fine-tuning a pretrained Model 2 on part of a target dataset

prior to evaluation would have likely resulted in substantially better performance compared to the zero-shot evaluation of Model 2-DS.

5.4 Analysis of Performance Variance Across Folds

A deep learning model's performance variance across cross-validation folds provides crucial insight into its stability and generalizability. High variance in performance metrics can indicate that the model's predictive capability is highly dependent on the specific distribution of the training split, signaling a susceptibility to overfitting and thus poor generalization to unseen data. The SkinEHDLF and PanDerm backbones exhibited differing levels of variance across folds depending on threshold tuning and whether a unimodal or multimodal training dataset was used, as shown in Table A.1. Model 2 demonstrated a tighter range of performance across folds, as evidenced by its smaller standard deviation for all metrics compared to Model 1. This suggests that the SkinEHDLF architecture is more stable and generalizable than the PanDerm architecture when training on the unimodal SLICE-3D dataset. However, Model 4 at its default threshold of 0.5 exhibited mixed results in terms of variance relative to Model 3. Further complicating the analysis, Model 4 exhibited lower variance relative to Model 3 across all metrics only when using a tuned decision threshold of 0.3884. Consequently, the broader implications of their performance variance across folds in the multimodal phase should be considered inconclusive.

5.5 Evaluation of Strategies to Address Extreme Class Imbalance

An analysis of the SkinEHDLF backbone's training and validation curves reveals that despite its overall predictive superiority compared to the PanDerm backbone, training it on highly imbalanced datasets presents a persistent challenge. Following the initial gradient shock at the onset of the fine-tuning stage, the validation curves for the SkinEHDLF models failed to exhibit the continuous, substantial improvement past the early epochs that was demonstrated by the PanDerm models. The early plateau suggests that the strategies implemented to address the extreme class imbalance of the training datasets, namely the use of a class-weighted cross-entropy loss function as well as Mixup and CutMix spatial augmentations, may have had limited impact on driving further

improvements in model performance. However, it is important to note that the implementation of the strategies successfully prevented the complete model collapse that occurred in their absence. The use of the strategies was therefore necessary given the overall hyperparameter configuration, but the SkinEHDLF models ultimately still struggled to continuously extract novel, meaningful patterns from the minority malignant class. While additional training epochs or further hyperparameter adjustments might yield superior performance, extensive empirical testing of alternative configurations proved less successful.

5.6 Threshold Tuning Trade-Offs

The default decision threshold of 0.5 is rarely the optimal threshold in clinical diagnostic applications, as the cost of a false negative is typically vastly higher than the cost of a false positive. This is especially true in the field of cancer diagnostics, where a false negative could lead to delayed treatment and significantly worse patient outcomes. Threshold tuning allows for the recalibration of model predictions to optimize for specific performance metrics that are most relevant to the clinical context. While threshold tuning can improve specific metrics, the extent to which it can do so without severely degrading others is limited by the AUROC of the model. Table A.1 demonstrates the trade-offs involved with threshold calibration for the SkinEHDLF backbones in both unimodal and multimodal training contexts. With significantly higher AUROC values compared to their PanDerm backbone counterparts, Model 2 and Model 4 were able to achieve substantial improvements in sensitivity while experiencing a lesser degree of degradation in specificity. These adjustments placed both metrics within clinically acceptable ranges, balancing sensitivity and specificity while improving overall BAcc.

5.7 Assessment of Pre-Clinical Viability

As analyzed in Section 4.4, the classification performance achieved by Model 2 suggests that it is suitable for pre-clinical deployment. A comparison of its performance against the model studied by Smak Gregoor et al. (2023) as well as the performance of highly experienced clinicians in the reader study conducted by Yan et al. (2025) suggests that Model 2 is capable of performing at

a level that is comparable to or better than existing DTC skin cancer screening models and highly experienced clinicians. Furthermore, Model 2 was shown to be highly effective in terms of its cost per additional (pre)malignancy detected within a population, costing less than a diagnosis by a dermatologist in the United States. Given that a DTC app allows for a diagnosis to be made at any time at no immediate cost to the patient, the impact of early detection on long-term patient and healthcare system costs is likely to be very significant. Likewise, early detection should be expected to substantially improve patient outcomes and optimize the distribution of healthcare resources through more effective patient triage.

Although Model 2 did not meet most of the preconditions established by Sangers et al. (2025) for endorsement from clinical experts, most of the preconditions were not within the scope of this study as they pertain primarily to a nonexistent frontend application layer. However, the remaining preconditions could be easily satisfied through the development of a thoughtful client application as well as testing by an external organization to validate the model's performance and reliability in a pre-clinical setting. While the zero-shot evaluation of Model 2-DS yielded poor results, this degradation is a direct consequence of performing inference on severe cross-modality domain shifts. As discussed in Section 5.3, fine-tuning a pretrained Model 2 on part of a target dataset prior to evaluation would likely remedy the severe degradation in performance when evaluating on domain-shifted datasets.

5.8 Limitations of the Study

While this study successfully demonstrated the architectural superiority of adaptive feature fusion for skin lesion classification as well as its suitability for pre-clinical deployment, several methodological limitations must be acknowledged. First, the exact implementation details of the SkinEHDLF architecture were not provided by the original authors; thus, several assumptions had to be made regarding the design of the feature fusion mechanism and the training procedure. As a result, direct comparison between the results of this study and the results presented by Lilhore et al. (2025) may be imperfect. Second, the unavailability of the majority of the supplementary

dataset used by Yan et al. (2025) necessitated the use of a smaller, less diverse dataset for the multimodal training phase. The disproportionately high representation of microscopic dermatopathology images resulted in degraded model performance when incorporating the available supplementary data into the training sets, obscuring the true potential of multimodal training for both backbones within the OmniFusion framework. Third, computational and hardware constraints necessitated the termination of the fine-tuning stage at 50 epochs. While validation loss curves indicated a plateau, original implementations of the PanDerm and SkinEHDLF backbones were trained for substantially more epochs; it is possible that extended training schedules paired with further hyperparameter adjustments might have yielded performance improvements. Finally, the external evaluation of Model 2 was restricted to zero-shot inference to test raw generalizability. Because no target-domain fine-tuning was performed, the model's performance on diverse, domain-shifted datasets was severely hindered. Further research should prioritize isolating the supplementary data by modality and implementing few-shot fine-tuning to better evaluate cross-domain pre-clinical viability.

5.9 Broader Implications and Future Work

This study illuminated the complex dynamics of multimodal training data and domain shift. While multimodal training can yield positive outcomes and holds immense potential across broader applications as demonstrated by Yan et al. (2025), the inclusion of severely domain-shifted modalities must be carefully considered to prevent modality collapse and degradation in cross-domain generalization. A better approach may be to use OmniFusion as a foundation model that can be fine-tuned on specific modalities for specific applications, rather than training a single model on many modalities at once. Nevertheless, the integration of advanced optimization strategies, specifically the use of an LP-FT pipeline alongside Mixup and CutMix spatial augmentation techniques, proved effective at reducing the number of training epochs required to achieve high performance comparable to the original PanDerm and SkinEHDLF implementations.

Future research should focus on mitigating the impact of severe domain shifts through few-shot fine-tuning on target distributions, isolating training modalities to prevent modality collapse during multimodal training, and replacing the Swin Transformer used in the SkinEHDLF architecture with the ViT-Large encoder used by PanDerm. Further validating the model's performance and reliability in pre-clinical settings through independent verification by external organizations is also essential. Additionally, the specialization of OmniFusion via fine-tuning on training data specifically from dermatology clinics would be a critical next step in preparing the model for real-world deployment. Finally, the creation of an API along with mobile and web applications to facilitate accessibility for clinicians and patients alike is a crucial next step in translating the findings of this research into tangible healthcare impact.

Chapter 6
Conclusion

This study set out to advance the state of the art in computer-aided medical diagnostics by synthesizing the most effective architectural components of leading computer vision models into a cohesive foundation model, OmniFusion, which leverages additional techniques to enhance robustness and generalizability. Through a rigorous four-phase evaluation, this research provided critical insights into the advantages of competing architectural approaches, the complex dynamics of multimodal training data, and the viability of deploying hybrid deep learning models in real-world, pre-clinical settings. The findings of this research also substantiate the strength of the underlying backbone architectures as well as the unique merits of the OmniFusion framework.

Empirical results definitively established the architectural superiority of the adaptive attention-based feature fusion mechanism over the pure vision transformer approach in the context of binary classification on the SLICE-3D dataset, demonstrating its exceptional capability to capture both low-level lesion features and long-range global context. While both architectures achieved exceptional results, statistical testing confirmed that the differences in performance were not due to random chance. Furthermore, the highest-performing OmniFusion model successfully surpassed clinically validated quantitative thresholds for sensitivity, specificity, and cost-effectiveness, rivaling the diagnostic capabilities of highly experienced clinicians. Ultimately, this research demonstrates that scalable, decentralized diagnostic models are not only technically feasible but also possess immense potential to transform early skin cancer detection, optimize healthcare triage and resource allocation, and save lives in the process.

Appendix A

Supplementary Tables and Figures

A.1 Aggregated Performance of OmniFusion Across All Models

The performance metrics of all OmniFusion models are aggregated in Table A.1 to present a comprehensive overview of the results across all experimental phases.

Table A.1: Aggregated Performance of OmniFusion Across All Models

Model	Backbone	Training Data	Threshold	BAcc (%)	Sens (%)	Spec (%)	AUROC (%)	F1 (%)
Model 1	PanDerm	SLICE-3D	0.5000	80.65±5.72	74.54±10.73	86.75±3.95	88.41±3.77	92.76±2.33
Model 2	SkinEHDLF	SLICE-3D	0.5000	87.86±3.98	84.21±8.17	91.52±1.08	95.01±1.73	95.47±0.59
Model 2	SkinEHDLF	SLICE-3D	0.4431	88.54±3.15	88.29±5.89	88.79±1.29	95.01±1.73	93.96±0.72
Model 3	PanDerm	SLICE-3D + Supp.	0.5000	81.92±3.79	81.47±10.28	82.37±6.24	88.83±3.49	90.12±3.83
Model 4	SkinEHDLF	SLICE-3D + Supp.	0.5000	85.37±6.38	78.62±14.55	92.13±2.32	94.36±1.70	95.79±1.25
Model 4	SkinEHDLF	SLICE-3D + Supp.	0.3884	86.96±2.26	87.53±6.19	86.39±3.54	94.36±1.70	92.57±2.04
Model 5	SkinEHDLF	Supp.	0.5000	49.31±2.69	64.47±7.60	34.15±12.48	53.06±2.06	43.23±8.61
Model 2-DS	SkinEHDLF	SLICE-3D	0.5000	49.39±2.66	64.58±7.68	34.20±12.42	53.12±2.02	43.30±8.54

A.2 OmniFusion Model Performance Across Individual Folds

The individual fold performance metrics for each model are detailed in several tables to provide insights into the variability and consistency of model performance across different data splits. Decision thresholds for each model are set at 0.5 unless otherwise specified.

Table A.2: Model 1 (PanDerm Unimodal) Performance Across 10 Folds

Fold	BAcc (%)	Sens (%)	Spec (%)	AUROC (%)	F1 (%)
Fold 1	72.48	56.41	88.55	88.54	93.81
Fold 2	84.83	79.49	90.18	91.22	94.74
Fold 3	84.83	79.49	90.18	91.22	94.74
Fold 4	73.54	69.23	77.85	80.57	87.45
Fold 5	75.70	61.54	89.87	85.85	94.55
Fold 6	80.75	79.49	82.01	89.02	90.02
Fold 7	82.51	77.50	87.52	89.26	93.24
Fold 8	77.34	67.50	87.19	85.22	93.05
Fold 9	84.36	82.50	86.23	89.17	92.50
Fold 10	90.14	92.31	87.97	94.05	93.51
Mean ± SD	**80.65±5.72**	**74.54±10.73**	**86.75±3.95**	**88.41±3.77**	**92.76±2.33**

Table A.3: Model 2 (SkinEHDLF Unimodal) Performance Across 10 Folds

Fold	BAcc (%)	Sens (%)	Spec (%)	AUROC (%)	F1 (%)
Fold 1	78.53	64.10	92.96	92.11	96.24
Fold 2	89.25	87.18	91.31	94.67	95.36
Fold 3	89.22	87.18	91.26	94.75	95.34
Fold 4	86.65	82.05	91.25	93.89	95.32
Fold 5	88.14	84.62	91.66	95.48	95.55
Fold 6	90.23	87.18	93.27	96.49	96.42
Fold 7	87.53	85.00	90.06	94.71	94.67
Fold 8	85.39	80.00	90.79	93.30	95.07
Fold 9	93.71	95.00	92.42	98.02	95.97
Fold 10	89.99	89.74	90.24	96.66	94.77
Mean ± SD	**87.86±3.98**	**84.21±8.17**	**91.52±1.08**	**95.01±1.73**	**95.47±0.59**

Table A.4: Model 2 (SkinEHDLF Unimodal) Performance Across 10 Folds at Decision Threshold 0.4431

Fold	BAcc (%)	Sens (%)	Spec (%)	AUROC (%)	F1 (%)
Fold 1	83.08	76.92	89.24	92.11	94.21
Fold 2	87.81	87.18	88.44	94.67	93.77
Fold 3	88.92	89.74	88.09	94.75	93.57
Fold 4	87.95	87.18	88.72	93.89	93.93
Fold 5	89.47	89.74	89.19	95.48	94.19
Fold 6	91.75	92.31	91.20	96.49	95.30
Fold 7	88.66	90.00	87.33	94.71	93.14
Fold 8	83.82	80.00	87.65	93.30	93.32
Fold 9	92.76	95.00	90.53	98.02	94.93
Fold 10	91.18	94.87	87.48	96.66	93.23
Mean ± SD	**88.54±3.15**	**88.29±5.89**	**88.79±1.29**	**95.01±1.73**	**93.96±0.72**

Table A.5: Model 3 (PanDerm Multimodal) Performance Across 10 Folds

Fold	BAcc (%)	Sens (%)	Spec (%)	AUROC (%)	F1 (%)
Fold 1	84.85	82.05	87.64	89.47	93.32
Fold 2	83.20	89.74	76.65	89.85	86.69
Fold 3	82.02	79.49	84.55	87.46	91.53
Fold 4	80.07	71.79	88.34	85.47	93.70
Fold 5	79.89	89.74	70.03	89.55	82.29
Fold 6	85.79	92.31	79.27	91.55	88.35
Fold 7	79.14	67.50	90.78	91.45	95.06
Fold 8	73.79	65.00	82.59	80.74	90.36
Fold 9	85.76	92.50	79.01	90.42	88.19
Fold 10	84.71	84.62	84.80	92.30	91.68
Mean ± SD	**81.92±3.79**	**81.47±10.28**	**82.37±6.24**	**88.83±3.49**	**90.12±3.83**

Table A.6: Model 4 (SkinEHDLF Multimodal) Performance Across 10 Folds

Fold	BAcc (%)	Sens (%)	Spec (%)	AUROC (%)	F1 (%)
Fold 1	71.01	46.15	95.87	91.15	97.77
Fold 2	87.93	82.05	93.82	94.22	96.71
Fold 3	79.11	64.10	94.11	92.12	96.86
Fold 4	89.22	89.74	88.70	95.59	93.92
Fold 5	87.83	84.62	91.04	94.84	95.21
Fold 6	92.64	94.87	90.40	96.46	94.86
Fold 7	88.36	87.50	89.23	95.84	94.21
Fold 8	81.54	70.00	93.07	94.19	96.30
Fold 9	87.02	82.50	91.53	93.55	95.48
Fold 10	89.09	84.62	93.57	95.60	96.58
Mean ± SD	**85.37±6.38**	**78.62±14.55**	**92.13±2.32**	**94.36±1.70**	**95.79±1.25**

Table A.7: Model 4 (SkinEHDLF Multimodal) Performance Across 10 Folds at Decision Threshold 0.3884

Fold	BAcc (%)	Sens (%)	Spec (%)	AUROC (%)	F1 (%)
Fold 1	82.24	74.36	90.11	91.15	94.70
Fold 2	87.81	84.62	91.00	94.22	95.19
Fold 3	87.11	87.18	87.04	92.12	92.98
Fold 4	88.50	94.87	82.14	95.59	90.10
Fold 5	87.54	89.74	85.34	94.84	92.00
Fold 6	89.92	94.87	84.97	96.46	91.79
Fold 7	87.79	92.50	83.09	95.84	90.67
Fold 8	85.68	82.50	88.85	94.19	94.00
Fold 9	84.42	87.50	81.34	93.55	89.62
Fold 10	88.61	87.18	90.05	95.60	94.67
Mean ± SD	**86.96±2.26**	**87.53±6.19**	**86.39±3.54**	**94.36±1.70**	**92.57±2.04**

Table A.8: Model 5 (SkinEHDLF - Supplementary Only) Performance Across 10 Folds

Fold	BAcc (%)	Sens (%)	Spec (%)	AUROC (%)	F1 (%)
Fold 1	56.14	50.64	61.65	58.28	61.59
Fold 2	50.10	57.01	43.19	52.21	49.90
Fold 3	50.92	55.05	46.78	52.69	52.21
Fold 4	47.97	71.21	24.74	52.41	36.53
Fold 5	48.78	70.87	26.69	53.66	38.36
Fold 6	48.93	67.30	30.56	54.10	41.34
Fold 7	48.19	70.60	25.77	52.96	37.43
Fold 8	46.97	71.53	22.40	51.25	34.25
Fold 9	47.71	64.52	30.89	51.70	41.13
Fold 10	47.41	65.95	28.87	51.30	39.56
Mean ± SD	**49.31±2.69**	**64.47±7.60**	**34.15±12.48**	**53.06±2.06**	**43.23±8.61**

Table A.9: Model 2-DS (SkinEHDLF Unimodal - Domain Shift) Performance Across 10 Folds

Fold	BAcc (%)	Sens (%)	Spec (%)	AUROC (%)	F1 (%)
Fold 1	56.24	50.76	61.71	58.23	61.67
Fold 2	50.05	56.97	43.14	52.29	49.85
Fold 3	50.75	54.87	46.62	52.75	52.05
Fold 4	48.18	71.40	24.95	52.46	36.77
Fold 5	48.95	71.11	26.79	53.72	38.51
Fold 6	49.12	67.47	30.76	54.18	41.56
Fold 7	48.26	70.77	25.76	53.01	37.45
Fold 8	47.13	71.63	22.62	51.40	34.48
Fold 9	47.80	64.68	30.93	51.79	41.19
Fold 10	47.46	66.16	28.75	51.39	39.48
Mean ± SD	**49.39±2.66**	**64.58±7.68**	**34.20±12.42**	**53.12±2.02**	**43.30±8.54**

A.3 Confusion Matrices for OmniFusion Models

Confusion matrices for each OmniFusion model are presented in several figures to illustrate the distribution of true positives, true negatives, false positives, and false negatives, thereby offering granular insight into the classification behavior of each model. Decision thresholds for each model are set at 0.5 unless otherwise specified.

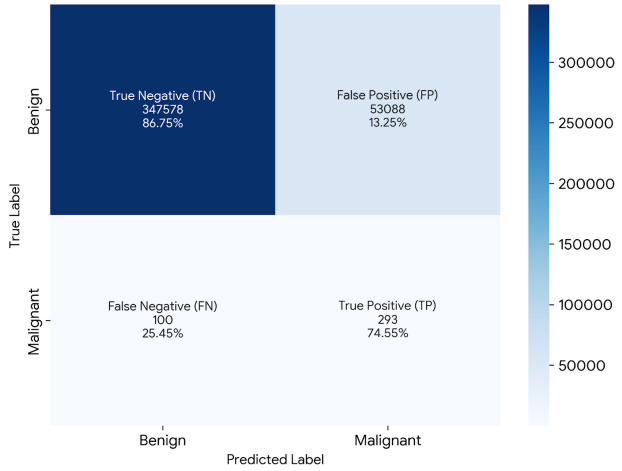


Figure A.1: Model 1 (PanDerm Unimodal) Confusion Matrix

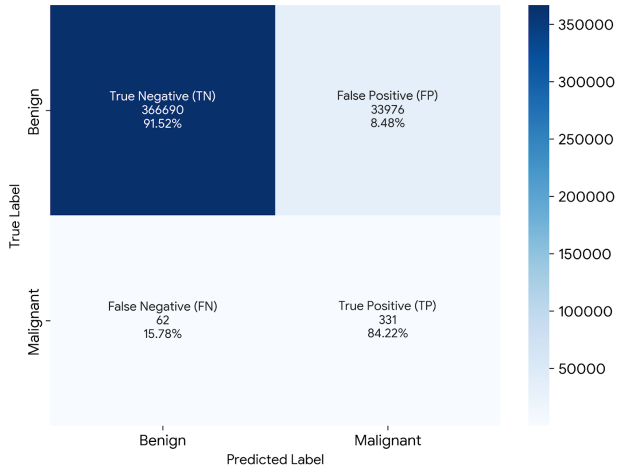


Figure A.2: Model 2 (SkinEHDLF Unimodal) Confusion Matrix

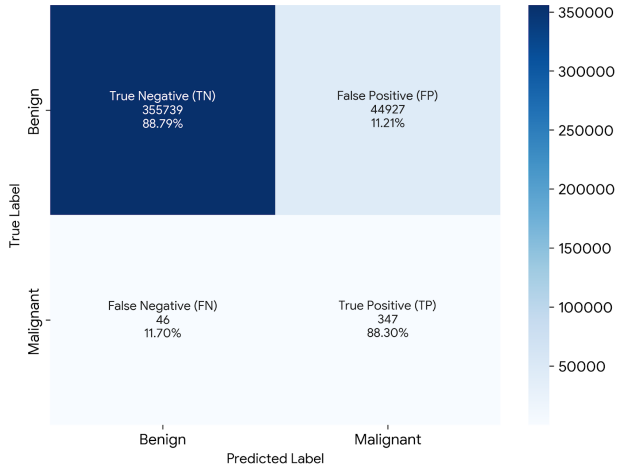


Figure A.3: Model 2 (SkinEHDLF Unimodal) Confusion Matrix at Decision Threshold 0.4431

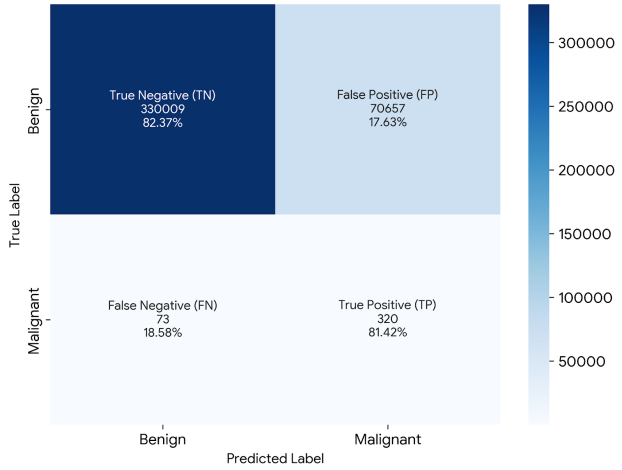


Figure A.4: Model 3 (PanDerm Multimodal) Confusion Matrix

Figure A.5: Model 4 (SkinEHDLF Multimodal) Confusion Matrix

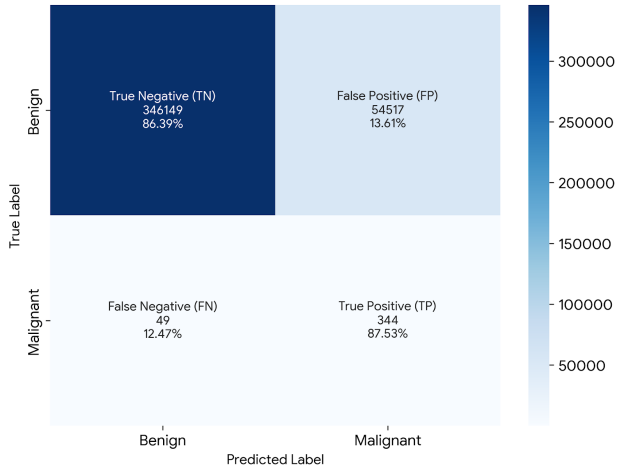


Figure A.6: Model 4 (SkinEHDLF Multimodal) Confusion Matrix at Decision Threshold 0.4431

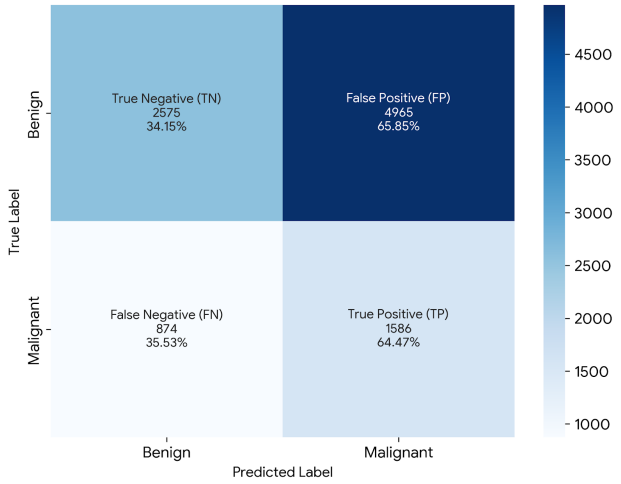


Figure A.7: Model 5 (SkinEHDLF - Supplementary Only) Confusion Matrix

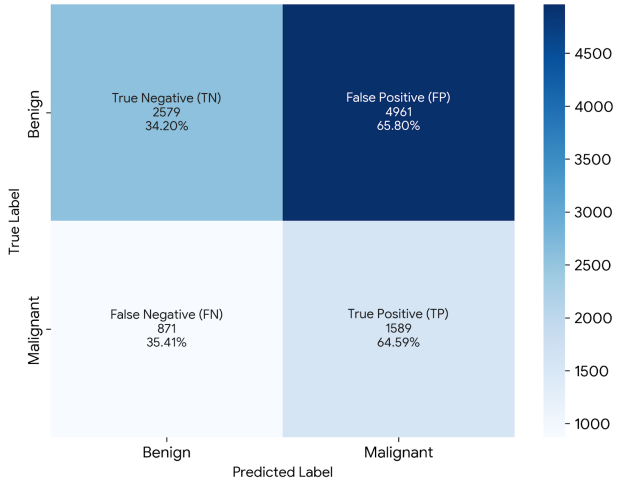


Figure A.8: Model 2-DS (SkinEHDLF Unimodal - Domain Shift) Confusion Matrix

A.4 Training and Validation Curves for OmniFusion Models

The training and validation curves for each OmniFusion model are presented in several figures to illustrate the learning dynamics and convergence behavior during training.

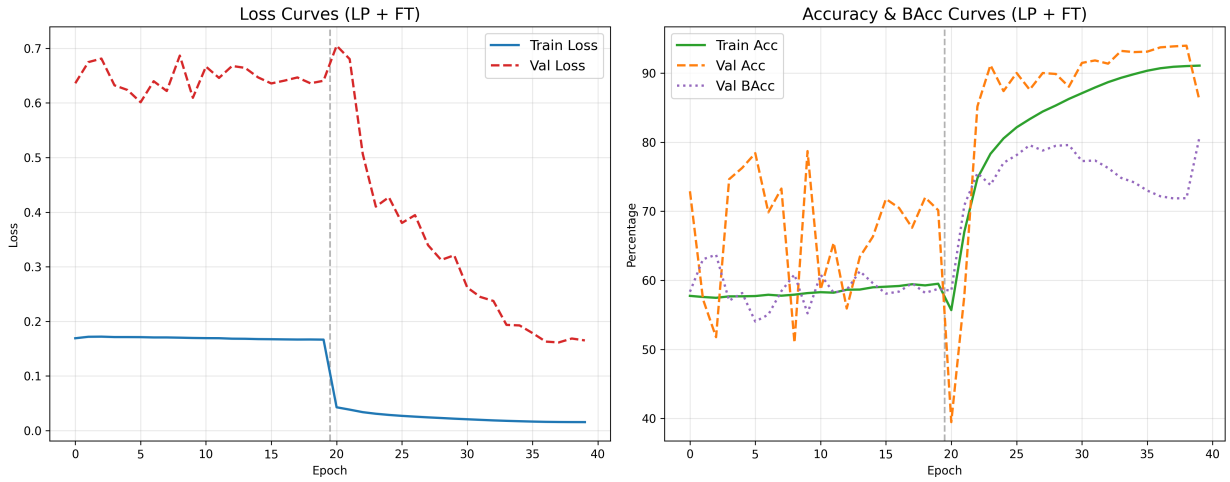


Figure A.9: Model 1 (PanDerm Unimodal) Training and Validation Curves

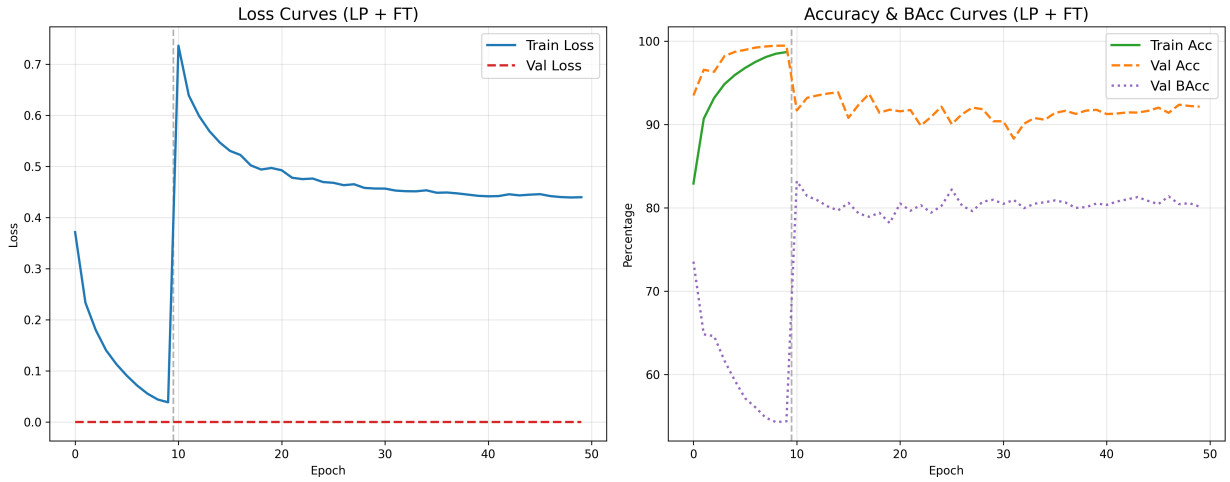


Figure A.10: Model 2 (SkinEHDLF Unimodal) Training and Validation Curves

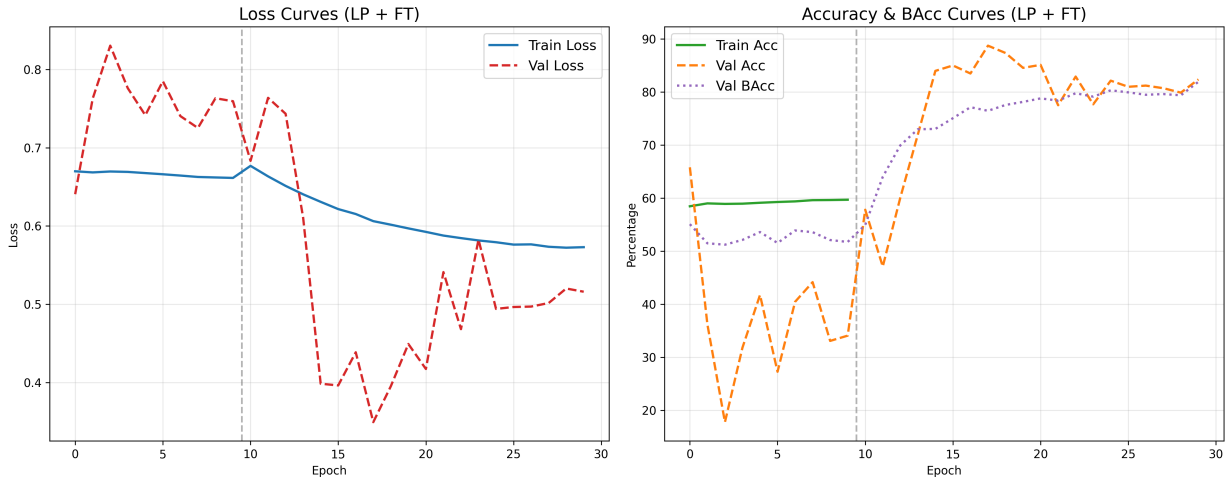


Figure A.11: Model 3 (PanDerm Multimodal) Training and Validation Curves

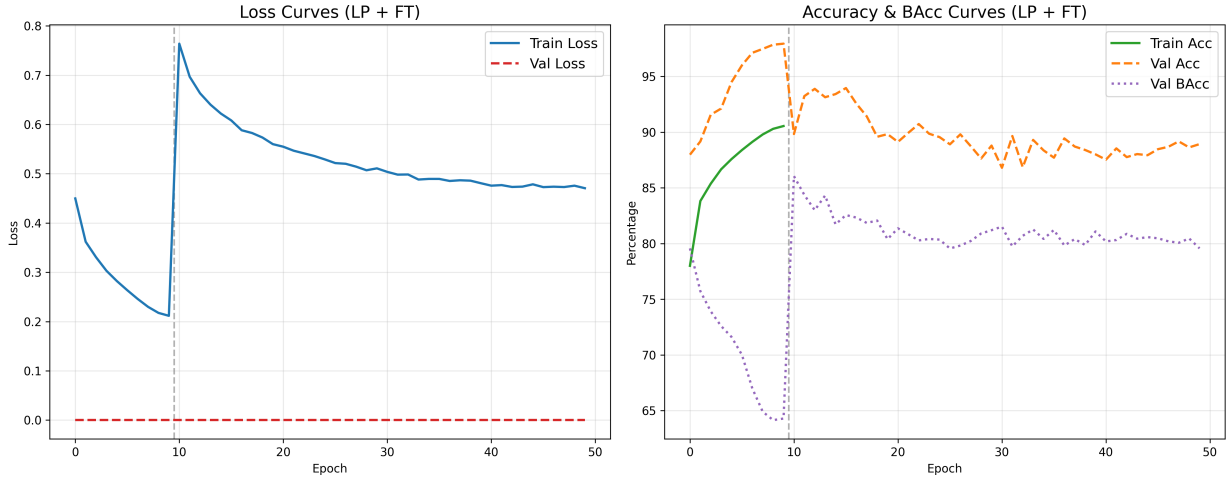


Figure A.12: Model 4 (SkinEHDLF Multimodal) Training and Validation Curves

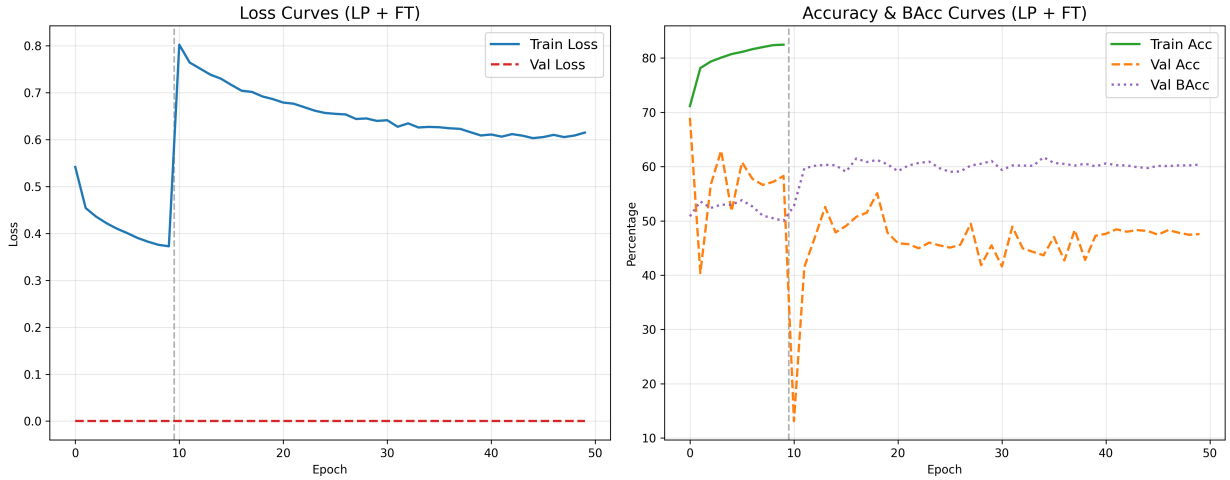


Figure A.13: Model 5 (SkinEHDLF - Supplementary Only) Training and Validation Curves

References

Binder, M., H. Kittler, A. Seeber, A. Steiner, H. Pehamberger, and K. Wolff. 1998. "Epiluminescence microscopy-based classification of pigmented skin lesions using computerized image analysis and an artificial neural network." *Melanoma Research* 8, no. 3 (June): 261–266. https://doi.org/10.1097/00008390-199806000-00009.

Binder, M., A. Steiner, M. Schwarz, S. Knollmayer, K. Wolff, and H. Pehamberger. 1994. "Application of an artificial neural network in epiluminescence microscopy pattern analysis of pigmented skin lesions: A pilot study." *British Journal of Dermatology* 130, no. 4 (April): 460–465. https://doi.org/10.1111/j.1365-2133.1994.tb03378.x.

Chaudhuri, Abhra, Anjan Dutta, Tu Bui, and Serban Georgescu. 2025. *A Closer Look at Multimodal Representation Collapse.* arXiv: 2505.22483 [cs.LG]. https://arxiv.org/abs/2505.22483.

Codella, Noel C., David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, et al. 2018. "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)." *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* (April): 168–172. https://doi.org/10.1109/isbi.2018.8363547.

Codella, Noel C., Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, et al. 2019. *Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC).* arXiv: 1902.03368 [cs.CV]. https://arxiv.org/abs/1902.03368.

Daneshjou, Roxana, Kailas Vodrahalli, Roberto A. Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, et al. 2022. "Disparities in dermatology AI performance on a diverse, curated clinical image set." *Science Advances* 8, no. 32 (August). https://doi.org/10.1126/sciadv.abq6147.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. “ImageNet: A large-scale hierarchical image database.” *2009 IEEE Conference on Computer Vision and Pattern Recognition* (June): 248–255. https://doi.org/10.1109/cvpr.2009.5206848.

Dreiseitl, Stephan, Michael Binder, Krispin Hable, and Harald Kittler. 2009. “Computer versus human diagnosis of melanoma: Evaluation of the feasibility of an automated diagnostic system in a prospective clinical trial.” *Melanoma Research* 19, no. 3 (June): 180–184. https://doi.org/10.1097/cmr.0b013e32832a1e41.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. “Dermatologist-level classification of skin cancer with deep neural networks.” *Nature* 542, no. 7639 (February): 115–118. https://doi.org/10.1038/nature21056.

Goel, Shubham. 2024. *Dermnet.* https://www.kaggle.com/datasets/shubhamgoel27/dermnet/data.

Gutman, David, Noel C. F. Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. 2016. *Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC).* arXiv: 1605.01397 [cs.CV]. https://arxiv.org/abs/1605.01397.

Hospital Italiano de Buenos Aires. 2023. (HIBA Skin Lesions). https://doi.org/10.34970/559884. https://doi.org/10.34970/559884.

Jeyageetha, K., K. Vijayalakshmi, S. Suresh, and A. Bhuvanesh. 2025. “Multi-skin disease classification using Hybrid Deep Learning Model.” *Technology and Health Care* 33, no. 4 (February): 1736–1754. https://doi.org/10.1177/09287329241312628.

Kawahara, Jeremy, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. 2019. “Seven-point checklist and skin lesion classification using multitask multimodal neural nets.” *IEEE Journal of Biomedical and Health Informatics* 23, no. 2 (March): 538–546. https://doi.org/10.1109/jbhi.2018.2824327.

Kharazmi, P., S. Kalia, H. Lui, Z. J. Wang, and T. K. Lee. 2017. "A feature fusion system for basal cell carcinoma detection through data-driven feature learning and patient profile." *Skin Research and Technology* 24, no. 2 (October): 256–264. https://doi.org/10.1111/srt.12422.

Kittler, H., H. Pehamberger, K. Wolff, and M. Binder. 2002. "Diagnostic accuracy of dermoscopy." *The Lancet Oncology* 3, no. 3 (March): 159–165. https://doi.org/10.1016/s1470-2045(02)00679-4.

Kriegsmann, Katharina, Fritjof Lobers, Christiane Zgorzelski, Jörg Kriegsmann, Rolf Rüdiger Meliß, Ulrich Sack, Georg Steinbuss, and Mark Kriegsmann. 2023. (Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections [data]. V. V1). https://doi.org/10.11588/DATA/7QCR8S.

Kumar, Ananya, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. *Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution.* arXiv: 2202.10054 [cs.LG]. https://arxiv.org/abs/2202.10054.

Kurtansky, Nicholas R., Brian M. D'Alessandro, Maura C. Gillis, Brigid Betz-Stablein, Sara E. Cerminara, Rafael Garcia, Marcela Alves Girundi, et al. 2024. "The SLICE-3D dataset: 400,000 skin lesion image crops extracted from 3D TBP for skin cancer detection." *Sci. Data* 11, no. 1 (August): 884. https://doi.org/10.1038/s41597-024-03743-w.

Lilhore, Umesh Kumar, Yogesh Kumar Sharma, Sarita Simaiya, Roobaea Alroobaea, Abdullah M. Baqasah, Majed Alsafyani, and Afnan Alhazmi. 2025. "SkinEHDLF a hybrid deep learning approach for accurate skin cancer classification in complex systems." *Sci. Rep.* 15, no. 1 (April): 14913. https://doi.org/10.1038/s41598-025-98205-7.

Luo, Nan, Xiaojing Zhong, Luxin Su, Zilin Cheng, Wenyi Ma, and Pingsheng Hao. 2023. "Artificial Intelligence-assisted dermatology diagnosis: From unimodal to multimodal." *Computers in Biology and Medicine* 165 (October): 107413. https://doi.org/10.1016/j.compbiomed.2023.107413.

Masood, Ammara, and Adel Ali Al-Jumaily. 2013. “Computer Aided Diagnostic Support System for skin cancer: A review of techniques and algorithms.” *International Journal of Biomedical Imaging* 2013:1–22. https://doi.org/10.1155/2013/323268.

Matsumoto, Martha, Aaron Secrest, Alyce Anderson, Melissa I. Saul, Jonhan Ho, John M. Kirkwood, and Laura K. Ferris. 2018. “Estimating the cost of skin cancer detection by dermatology providers in a large health care system.” *Journal of the American Academy of Dermatology* 78, no. 4 (April). https://doi.org/10.1016/j.jaad.2017.11.033.

Mendonca, Teresa, Pedro M. Ferreira, Jorge S. Marques, Andre R. Marcal, and Jorge Rozeira. 2013. “A dermoscopic image database for research and benchmarking.” *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (July): 5437–5440. https://doi.org/10.1109/embc.2013.6610779.

Natha, Priya, Sivarama Prasad Tera, Ravikumar Chinthaginjala, Safia Obaidur Rab, C. Venkata Narasimhulu, and Tae Hoon Kim. 2025. “Boosting skin cancer diagnosis accuracy with ensemble approach.” *Scientific Reports* 15, no. 1 (January). https://doi.org/10.1038/s41598-024-84864-5.

Pacal, Ishak, Burhanettin Ozdemir, Javanshir Zeynalov, Huseyn Gasimov, and Nurettin Pacal. 2025. “A novel CNN-VIT-based deep learning model for early skin cancer diagnosis.” *Biomedical Signal Processing and Control* 104 (June): 107627. https://doi.org/10.1016/j.bspc.2025.107627.

Pacheco, André G. C., Gustavo R. Lima, Amanda S. Salomão, Breno Krohling, Igor P. Biral, Gabriel G. de Angelo, Fábio C. R. Alves, et al. 2020. “PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones.” *Data in Brief* 32:106221. https://doi.org/10.1016/j.dib.2020.106221.

Pan, Sinno Jialin, and Qiang Yang. 2010. “A survey on Transfer Learning.” *IEEE Transactions on Knowledge and Data Engineering* 22, no. 10 (October): 1345–135 9. https://doi.org/10.1109/tkde.2009.191.

Perez, Carlos H., 0000-0001-5237-4256, 0000-0001-6924-9961, 0000-0002-6998-914X, and 0000-0003-4150-0522. 2023. “BCN20000: Dermoscopic Lesions in the Wild” (December). https://doi.org/10.6084/m9.figshare.24140028.v1. https://figshare.com/articles/journal_contribution/BCN20000_Dermoscopic_Lesions_in_the_Wild/24140028.

Rao, Adrit, Joon-Young Lee, and Oliver Aalami. 2023. “Studying the impact of Augmentations on medical confidence calibration.” *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (October): 2454–2464. https://doi.org/10.1109/iccvw60793.2023.00260.

Sangers, Tobias E., Marlies Wakkee, Folkert Moolenburgh, Tamar Nijsten, and Marjolein Lugtenberg. 2025. “Mobile health apps for skin cancer triage in the general population: A qualitative study on healthcare providers’ perspectives.” *BMC Cancer* 25, no. 1 (May). https://doi.org/10.1186/s12885-025-14244-3.

Sato, Issei, and Akiyoshi Tomihari. 2024. “Understanding linear probing then fine-tuning language models from NTK Perspective.” *Advances in Neural Information Processing Systems,* 139786–139822. https://doi.org/10.52202/079017-4436.

Smak Gregoor, Anna M., Tobias E. Sangers, Lytske J. Bakker, Loes Hollestein, Carin A. Uyl – de Groot, Tamar Nijsten, and Marlies Wakkee. 2023. “An artificial intelligence based app for skin cancer detection evaluated in a population based setting.” *npj Digital Medicine* 6, no. 1 (May). https://doi.org/10.1038/s41746-023-00831-w.

Skin Cancer Facts. 2026, January. https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/.

Tran, Anh T., Tal Zeevi, and Seyedmehdi Payabvash. 2025. “Strategies to improve the robustness and generalizability of deep learning segmentation and classification in neuroimaging.” *BioMedInformatics* 5, no. 2 (April): 20. https://doi.org/10.3390/biomedinformatics5020020.

Tschandl, Philipp, Cliff Rosendahl, and Harald Kittler. 2018. “The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions.” *Scientific Data* 5, no. 1 (August). https://doi.org/10.1038/sdata.2018.161.

Yan, Siyuan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, et al. 2025. “A multimodal vision foundation model for clinical dermatology.” *Nat. Med.* (June). https://doi.org/10.1038/s41591-025-03747-y.

Yap, Jordan, William Yolland, and Philipp Tschandl. 2018. “Multimodal skin lesion classification using Deep Learning.” *Experimental Dermatology* 27, no. 11 (September): 1261–1267. https://doi.org/10.1111/exd.13777.

Yun, Sangdoo, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. 2019. “Cutmix: Regularization strategy to train strong classifiers with localizable features.” *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (October): 6022–6031. https://doi.org/10.1109/iccv.2019.00612.

Zhang, Hongyi, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. 2018. *mixup: Beyond Empirical Risk Minimization.* arXiv: 1710.09412 [cs.LG]. https://arxiv.org/abs/1710.09412.