Advancing Equity in AI-based Skin Cancer Diagnosis

Dhruv Dewan, Monish Napa, Raghavendra Sunkara, Sanjana Vandanapu

Mentor: Dr. Heng Huang
PhD Student Advisor: Reza Shirkavand
Librarian: Drew Barker

Table of Contents

Abstract

This project addresses disparities in skin cancer detection models, which often underperform on patients with darker skin tones, contributing to inequities in healthcare diagnosis. Standard ResNet-based models, even when fine-tuned on publicly available dermatology datasets, show clear gaps in performance across skin-tone subgroups. To improve representation and generalization, we applied targeted preprocessing techniques and balanced sampling strategies to enhance representation of minority skin tones. Building on these approaches, we incorporated self-supervised learning using the DINO training process on unlabeled dermatology images to learn more domain-aligned feature representations. We experimented with partial fine-tuning, domain-specific pretraining, and group-aware evaluation metrics to enhance fairness and reduce performance disparities under domain shift. Overall, our work demonstrates the importance of equitable training practices, advanced representation learning, and careful dataset curation in developing reliable, generalizable, and fair AI-driven skin cancer diagnostic models.

Introduction

As of July 2025, skin cancer is the most prevalent form of cancer in the United States with approximately 9,500 new diagnoses occurring daily with 1 in 5 Americans likely to develop skin cancer at any point in their lifetimes [1]. When detected early and confined to the skin, the 5-year survival rate is approximately 99%; however this significantly drops to 76% upon regional lymph node presence, and to 35% in cases of distant metastasis making early detection critical for patient survival [2]. The annual cost of treating skin cancer in the United States exceeds $8.9 billion, underscoring both the scale and urgency of the problem [3]. Prognosis varies significantly depending on the type and stage of skin cancer at the time of diagnosis [4].

To address the limitations of traditional diagnostic methods, artificial intelligence (AI) -based systems have emerged as a promising tool for early skin cancer detection. Researchers at Stanford Medicine found that clinicians using AI assistance achieved a sensitivity rate of 81.1% and a specificity rate of 86.1%, compared to 75% and 81.5% respectively for those working without AI support [5]. This shows a meaningful improvement that suggests AI-augmented workflows have strong potential as clinical decision-support tools [5]. However, AI usage in diagnosing skin cancer still carries a serious inequity issue.

The promise of AI-assisted dermatology is undermined by a critical inequity: these models do not perform equally across patient demographics. The five-year survival rates for melanoma are approximately 90% for white patients compared to 66% for black patients, a disparity that reflects the longstanding gaps in both clinical care and the data used to train diagnostic systems [6]. This is largely driven by the severe underrepresentation of darker skin tones in dermatological image datasets, which leads to models that generalize poorly to non-white patient populations. This underrepresentation means that current AI diagnostic tools, despite strong overall performance, are effectively optimized for a narrow subset of the intended patient population, often underperforming for those who need them most [7,8].

Team AID's research investigates how pretraining strategy and dataset diversity jointly influence diagnostic accuracy and equity in AI-based skin cancer classification. The study compares supervised ImageNet pretraining against self-supervised DINO initialization using ResNet-50 as a consistent backbone, and evaluates both full and partial fine-tuning strategies. Domain adaptation is further examined by incorporating HAM10000 through split-training and joint-training approaches, and by assessing whether domain-specific self-supervised pretraining on unlabeled ISIC 2019 images yields more transferable representations. All models are

evaluated on the Diverse Dermatology Image (DDI) dataset using accuracy, precision, recall, and F1 score to assess both diagnostic performance and equity across skin tone groups. This work is guided by the following research questions:

1. Does self-supervised initialization improve performance on a small, diverse dermatology dataset?
2. Does initial labeled training on a larger dermatology dataset help or hurt transfer to diverse skin tones?
3. Can dermatology-domain unlabeled pretraining produce more transferable representations than standard supervised or ImageNet SSL initialization?

Literature Review

Overview of Skin Cancer and Imaging

Skin cancer is defined as the abnormal proliferation of cells within the epidermis and is broadly categorized into three primary types: basal cell carcinoma (BCC), squamous cell carcinoma (SCC), and melanoma. BCC is the most prevalent form of human cancer worldwide, with incidence rates continuing to increase, resulting in a significant disease burden and high healthcare costs [9]. In the United States alone, an estimated 2 million Americans are directly affected by BCC annually, with UV radiation exposure identified as the primary factor [9]. SCC arises from the abnormal proliferation of squamous cells in the outer layer of the epidermis and is the second most common form of skin cancer. Together, BCC and SCC account for approximately 5.4 million new cases in the United States each year, and this number continues to grow [10].

Melanoma, though comparatively rare, carries the most severe clinical prognosis due to its high potential for metastasis while not detected at an early stage. Although invasive

melanoma accounts for approximately 1% of all skin cancer diagnosis, it is responsible for over 75% of skin cancer deaths [11]. The global impact of this disease has grown considerably in recent decades; worldwide melanoma cases rose from approximately 230,000 in 2012 to 325,000 in 2020, which is a 41% increase with the highest incidence recorded in Australia and New Zealand [11]. Stage at diagnosis remains the strongest predictor of patient outcomes, with survival declining dramatically as the disease progresses - emphasizing why early and accurate detection is clinically important [12]. In addition to these primary forms, several rare skin cancers exist, including Merkel cell carcinoma (MCC), Kaposi sarcoma (KS), Actinic keratosis (AK), and skin lymphoma [13]. Although these conditions are less common, they can still pose significant health risks.

Traditional Diagnostic Methods

A variety of both non-invasive and invasive diagnostic methods have been developed in identifying suspicious skin lesions. These methods vary significantly in accuracy, accessibility, and clinical demands, which have driven development in more efficient and scalable alternatives.

Among non-invasive techniques, dermoscopy has emerged as the most extensively validated tool for in vivo lesion evaluation. Dermoscopy allows clinicians to visualize subsurface structures that are not visible to the naked eye, allowing for recognition of pigmentation patterns, vascular morphology, and structural features associated with malignancy [14]. Studies show that dermoscopy combined with visual inspection outperforms visual examination alone for diagnosing both melanoma and BCC [15]. Despite these advantages, dermoscopy proficiency varies considerably across clinical settings; a 2017 survey found that only 35% of U.S. dermatology residents received formal dermoscopy training during residency, limiting its consistent application outside specialist care [16].

Multi-modal spectroscopy is a complementary non-invasive technique that analyzes how polarized light at different wavelengths interacts with skin tissue [14]. Spectral filtering can enhance image contrast and highlight structural abnormalities within lesions that may not be apparent through standard visual inspection [14]. Computational methods have also been integrated into non-invasive workflows with feature extraction algorithms applied to lesion images and classification performed using approaches such as support vector machines (SVMs) [17].

Despite advances in non-invasive imaging, tissue biopsy remains the clinical gold standard for definitive skin cancer diagnosis. Histopathological analysis of biopsied tissue provides reliable confirmation of malignancy; however the procedure carries drawbacks including patient discomfort, procedural cost, and risk of complications such as infection or scarring. The combined limitations of invasive procedures, variable access to dermoscopy expertise, and a global shortage of dermatologists, especially in underserved regions, have motivated extensive research into computational and AI-based tools capable of augmenting clinical workflows and broadening access to early, accurate skin cancer detection.

Bias in Medical AI

As artificial intelligence and machine learning models become increasingly integrated into medical diagnostics, concerns have emerged regarding algorithmic bias and its potential to worsen existing healthcare disparities. Bias can occur at multiple stages of the AI development pipeline, including in data collection, model training, evaluation, and deployment, and when left unaddressed, can lead to substandard clinical decisions that disproportionately harm vulnerable patient groups [18]. Two primary categories of bias are relevant to medical AI systems. Statistical bias refers to cases in which the dataset does not accurately represent the true

distribution of the population, while social bias reflects existing societal inequalities that may result in unequal outcomes for specific groups within the human population [19].

These issues are especially consequential in AI-based skin lesion classification systems. Convolutional neural networks (CNN) models designed to classify skin lesions are predominantly trained on dermascopic image datasets in which lighter skin tones are heavily over represented. The International Skin Imaging Collaboration (ISIC) dataset, one of the most widely used benchmarks in computational dermatology has over 70% of its images depicting light skin tones with fewer than 8% representing dark skin tones [7]. A systematic review of publicly available skin lesion datasets found that the vast majority do not report skin type metadata at all, and among those that do, darker skin tones are considerably underrepresented - raising significant concerns about the generalizability of models trained on these datasets [20].

This imbalance has measurable consequences for model performance across patient populations. Daneshjou et al. (2022) demonstrated that most AI dermatology models have not been assessed on images of diverse skin tones, and that state-of-the art skin lesion classification models performed significantly worse on images of individuals with darker skin tones compared to those with lighter skin tones [8]. These performance gaps persist even when models achieve high aggregate accuracy on standard benchmarks, revealing that models may overfit to majority group characteristics and fail to capture diagnostic patterns in darker skin tones, posing significant clinical risk for underserved populations. This pattern is further illustrated by the performance of DeepDerm and ModelDerm, two leading AI skin cancer classifiers, which achieved AUC scores of 0.94 and 0.88 respectively on standard, predominantly light-skinned datasets. When evaluated on the Diverse Dermatology Images (DDI) dataset - a dataset balanced across skin tones - their scores dropped sharply to 0.65 and 0.56 respectively. Notably, when

DeepDerm was retrained using the diverse dataset, its AUC recovered to 0.74 on dark skin and 0.72 on light skin, demonstrating that performance gaps can be meaningfully narrowed through more representative training data [8]. These findings reveal that even classifiers considered state-of-the-art are not fully equipped to generalize across diverse patient populations, and that performance gaps exist even when models achieve high accuracy on standard benchmarks [8]. Addressing these disparities requires not only the collection of more demographically representative datasets, but also the development of training strategies explicitly designed to improve fairness and generalizability across all skin tone groups.

CNNs for Image Classification

Machine learning has significantly advanced the field of medical image analysis by enabling computers to identify patterns within large datasets that may be difficult for humans to detect [20]. In the context of skin cancer detection, machine learning models are typically trained to analyze dermascopic images of skin lesions and classify them as benign or malignant. Many medical imaging applications rely on supervised learning algorithms, where models are trained using labeled datasets in which each image is associated with a known diagnosis [21]. Although supervised approaches can produce highly accurate models, labeling large medical datasets requires significant time and clinical expertise. To address this limitation, researchers have explored hybrid approaches such as semi-supervised learning, which leverage smaller labeled datasets alongside larger unlabeled datasets to reduce the time and resources required for model training.

One of the most widely used deep learning architectures for image classification is the convolutional neural network (CNN). A standard CNN architecture consists of three primary layer types: convolutional layers, which apply learned filters to the input image to extract

features such as edges, textures, and shapes; pooling layers, which reduce the spatial dimensionality of feature maps while preserving salient information; and fully connected layers, which integrate the extracted features to produce a final classification output [22]. Beyond this basic structure, several additional components are critical to CNN performance. Activation functions such as the Rectified Linear Unit (ReLU) introduce non-linearity into the network, enabling it to learn complex, non-linear patterns in image data that linear models cannot capture [23]. Pooling operations such as max pooling further reduce computational complexity while retaining the most discriminative features. The network is trained through backpropogation, in which prediction errors are propagated backward through the network and used to update weights via optimization algorithms such as stochastic gradient descent (SGD) or Adam, iteratively minimizing the loss function over successive training epochs [24].

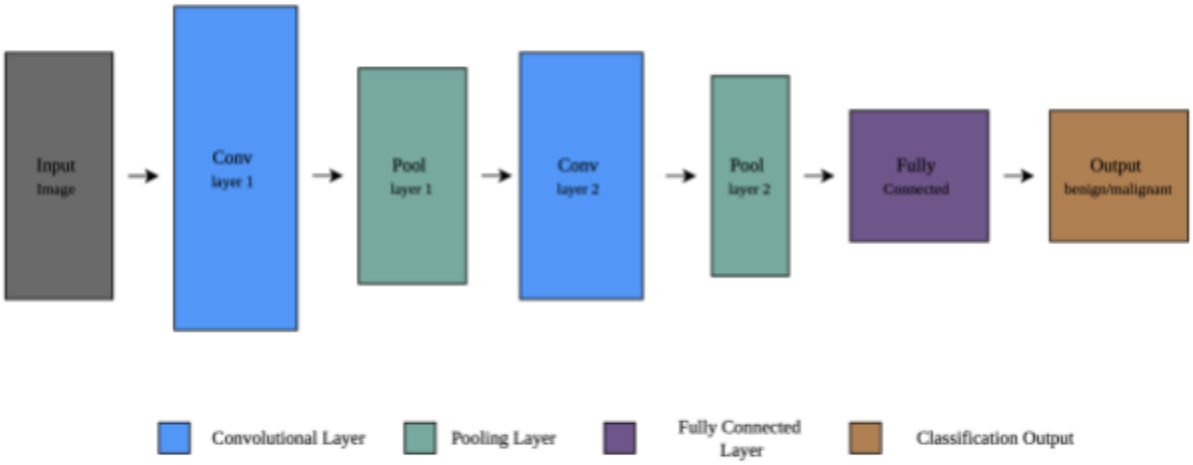

Figure 1. Convolutional Neural Network (CNN) architecture for image classification

Data augmentation is another important technique used during CNN training, particularly in medical imaging contexts where labeled data is limited. By applying transformations such as

rotation, flipping, and scaling to training images, augmentation artificially increases the diversity of the training set, reduces overfitting, and improves the model's ability to generalize to unseen examples [24]. Transfer learning further addresses data scarcity by initializing a CNN with weights pretrained on large general-purpose dataset (most commonly ImageNet) and fine-tuning those weights on a smaller, domain-specific dataset. This approach leverages generalizable low-level features already learned from large-scale datasets, allowing rapid convergence and reducing the risk of overfitting when labeled medical data is scarce [25]. Commonly used pretraining architectures in dermatological image classification include ResNet, VGGNet, and Inception, all of which have been applied to skin cancer classification tasks with strong results [26].

The clinical potential of CNN-based classification was established by Esteva et al., who demonstrated that a deep CNN trained on over 129,000 clinical images achieved diagnostic accuracy comparable to board-certified dermatologists across a wide range of skin disease, including melanoma and carcinoma, without requiring handcrafted features or domain-specific rules [27]. The model was evaluated against biopsy-confirmed cases and matched dermatologists on sensitivity and specificity, establishing CNNs as a credible decision-support tool in clinical dermatology. By leveraging CNN-based approaches, diagnostic capacity could be expanded in areas with limited specialist availability, helping to identify high-risk cases more efficiently.

Self Supervised Representation Learning

Self-supervised representation learning is a powerful approach in machine learning where labeled data is limited. Unlike supervised learning, which relies heavily on annotated datasets, self-supervised learning utilizes the inherent structure of unlabeled data to generate pretext tasks through which a model learns meaningful feature representations without human-provided labels

[28]. These representations can subsequently be fine-tuned for downstream classification or detection tasks. In the context of medical image analysis, self-supervised learning offers an efficient alternative to supervised approaches by reducing dependence on costly and time-consuming manual labeling.

Early self-supervised methods often relied on contrastive learning, which encouraged the model to bring representations of different augmentations of the same image closer in latent space while pushing apart representations of distinct images [29]. One pioneering contrastive framework, SimCLR, outperformed supervised models on the ImageNet benchmark using 100 times fewer labels, though it requires very large batch sizes to perform well which can pose significant computational challenges. To address this, Momentum Contrast (MoCo) introduced a momentum-encoded queue to maintain negative samples without requiring excessively large batches [30]. Non-contrastive alternatives have also explored the Barlow Twins algorithm, which avoids representational collapse through redundancy reduction on the model's outputs rather than through explicit negative pair comparison, and can be trained with batch sizes as low as 128 in contrast to other self-supervised methods which may require batches above 4,000 [31].

The DINO (Self-Distillation with No Labels) framework represents a significant advancement in non-contrastive self-supervised learning. DINO is implemented as a form of self-distillation with no labels, in which a model learns visual representation entirely without labeled data [32]. The framework consists of two networks: a student and a teacher, both of which share the same architecture but have different parameters. The student network is trained using cross-entropy loss to match the outputs of the teacher network, while the weights of the teacher network are updated as an exponential moving average (EMA) of the student's parameters rather than through direct gradient descent [33]. During training, two different

random augmentations of the same input image are passed through the student and teacher networks. Each network produces a feature representation, which is then converted to a probability distribution using a softmax function, and the teacher's outputs are centered before comparison to prevent the model from collapsing to trivial solutions [33]. A key advantage of DINO is its architectural flexibility such that it can be applied to both vision transformers and convolutional backbones such as ResNet-50, making it a suitable option for medical imaging applications where architectural constraints may apply [33].

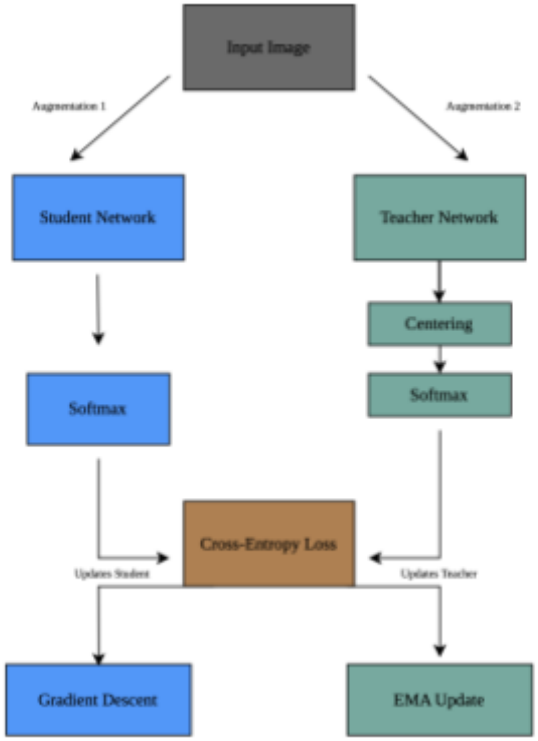

Figure 2. The DINO Self-Distillation Framework

Existing Work On Representation Learning

The applications of self-supervised representation learning to dermatological imaging has grown considerably as researchers have worked to overcome the fundamental limitations of

scarce annotated medical data. Self-supervised learning in computer vision has shown its potential to reduce its reliance on labeled data, though most studies have focused on balanced, large, broad-domain datasets like ImageNet, whereas in real-world medical applications, dataset size is typically limited and class distributions are skewed.

Several studies have directly compared SSL and supervised pretraining strategies for skin lesion classification in limited-data settings. Haggerty and Chandra demonstrated that SSL pretraining on ImageNet via Barlow Twins outperforms supervised pretraining when fine-tuning on a small skin lesion dataset [31]. This advantage arises from enhanced feature reuse, as frozen SSL features outperformed supervised counterparts in linear probing, showing that they capture more transferable and abstract representations [31]. Their study also found that minimal further SSL pretraining on task-specific dermatology data was sufficient to close the performance gap between supervised and self-supervised base models, suggesting that the relevance of pretraining data may matter more than its quantity [31].

More recent work has moved toward applying SSL-based approaches to the clinically motivated problem of total body screening. A study conducted by Useini et al. developed an end-to-end AI pipeline that combined an object detection model to localize lesions with a DINO-based self-supervised model to rank them according to the “ugly duckling” rule - a clinical heuristic which prioritizes lesions that differ from a patient’s other moles [34].

This integrated pipeline achieved approximately 93% average sensitivity for the top-10 AI-identified suspicious lesions, with dermatologists showing 95% agreement with the system’s predictions during clinical validation [34]. Across multiple backbone architectures, DINO consistently outperformed MoCo v2, converging more quickly to high sensitivity, further

reinforcing DINO's suitability as a representation learning framework for dermatological applications [34].

One key question raised about this body of work is whether pretraining on domain specific dermatological data produces more transferable representations than general ImageNet pretraining. Gottfrois et al. examined this directly by pretraining models using SSL on a curated collection of over 240,000 unlabeled dermatological images drawn from both public and private sources, encompassing dermoscopy and clinical imaging modalities [35]. Their results showed that domain-specific SSL pretraining outperformed ImageNet-initialized models across multiple evaluation datasets including DDI, HAM10000, and Fitzpatrick17k, providing evidence that the domain pretraining can meaningfully impact cross-skin tone generalizability. A complementary study found similar results using less curated sources, stating that models pretrained on unannotated dermatology images from online health forums outperformed ImageNet-pretrained counterparts, with top-1 diagnostic accuracy improving from 42.05% to 45.05% where domain-specific unlabeled data replaced ImageNet as the pretraining source [36].

Despite these advances, a consistent equity gap is constant across the existing literature. A systematic review and meta-analysis of AI-based dermatology systems found a pooled AUROC of 0.89 for lighter skin tones compared to 0.82 for darker skin tones, illustrating a consistent performance disparity across studies [37]. This gap continues even in models trained with SSL, suggesting that representation learning alone - without careful considerations of dataset diversity and demographic balance - is insufficient to address equity gaps in clinical AI systems. These findings further motivate the present work, which investigates whether combining domain-specific SSL pretraining with demographically representative training data can both enhance diagnostic accuracy and reduce performance disparities across skin tones.

Methodology

Overview

With this study we aimed to evaluate how different machine learning training techniques and representation learning techniques influence skin cancer classification and diagnostic accuracy of deep learning models under domain shifts. Specifically, we center our approach around the ResNet-50 model and utilize it as a baseline model for our experiments. Initially, we establish a supervised learning baseline by fine-tuning an ImageNet-pretrained ResNet-50 on the Diverse Dermatology Images (DDI) dataset [8]. This baseline provides a reference point for evaluating how different training strategies affect performance on a diverse dermatology dataset.

Building on this baseline, we investigate the impact of domain adaptation techniques involving split training and joint training with initial fine-tuning on labelled datasets such as the HAM10000 [38]. Additionally, we evaluate the impact of representation learning through self-supervised learning (SSL). We employ the DINO framework to initialize a ResNet-50 backbone with self-supervised representations and compare its performance against the supervised ImageNet-pretrained model under identical training conditions. This comparison allows us to evaluate whether SSL initialization improves generalization and diagnostic performance when training data is limited and distributional differences exist between datasets.

Furthermore, we seek to understand the effect of a self-supervised pretraining pipeline that involves data samples in the domain of the downstream task. In this case, we utilize samples of unlabelled dermatology images which enables the model to develop domain specific visual representations of the data distribution prior to being applied to a labelled training dataset for the downstream task of skin cancer classification.

All of the trained models are evaluated on a held-out DDI test set, which no model is ever given prior to evaluation. Performance is measured using standard classification metrics including accuracy, precision, recall, and F1-score, with additional analysis stratified across Fitzpatrick skin tone groups to assess the model fairness and robustness across the diverse populations that this study serves to cover.

The remainder of this section details the datasets used in this study, the preprocessing procedures applied to each dataset, the architecture of the models employed, and the training and evaluation protocols used to conduct our experiments.

Datasets

Throughout the experiments, we primarily employ the use of three datasets: Diverse Dermatology Images (DDI), HAM10000, and ISIC 2019. The DDI dataset is a “pathologically confirmed benchmark dataset with diverse skin tones” created by Stanford University School of Medicine and serves to represent the general demographic and distribution of the population in an equitable fashion which has not been done before [8]. DDI includes 656 image samples that range across 570 distinct patients and every diagnosis is confirmed via pathology reports from biopsy and reviewed by certified dermatologists. There exists a relatively even distribution of images categorized by three main groups within the Fitzpatrick skin types (FST).

Figure 3. DDI samples according to Fitzpatrick skin tones

Within these categories, FST Group V - VI represent darker skin tones and FST Group I - II represent the lighter skin tones generally found in large quantities within publicly available datasets which most classification models are trained on. Following the findings of "Disparities in Dermatology AI Performance On a Diverse, Curated Clinical Image Set", which demonstrated that fine-tuning on diverse data can improve generalization across skin tones, the DDI dataset is utilized for both fine-tuning and evaluation benchmarking [8]. The training, validation, and testing split is described in subsequent sections.

The HAM10000 (Human Against Machine with 10,000 training images) is a widely used dermatology image dataset that consists of dermatoscopic images of common pigmented skin lesions [38]. The dataset was originally compiled from two independent dermatology sources, the ViDIR group in Austria and the Department of Dermatology at the Medical University of Vienna, and contains 10,015 dermatoscopic images which span seven diagnostic categories: Actinic Keratoses and Intraepithelial Carcinoma (AKIEC), Basal Cell Carcinoma (BCC), Benign Keratosis (BKL), Dermatofibroma (DF), Melanocytic nevi (MV), Melanoma (MEL), Vascular skin lesions (VASC). For the purposes of this study, we group these diagnoses into two groups: malignant (AKIEC, BCC, MELl) and benign (BKL, DF, NV, VASC). It is worth noting that although Actinic Keratoses and Intraepithelial Carcinoma is not inherently malignant it is widely

agreed that the lesions can progress into squamous cell carcinoma, which is malignant. As a result, we consider this category as malignant for early detection purposes.

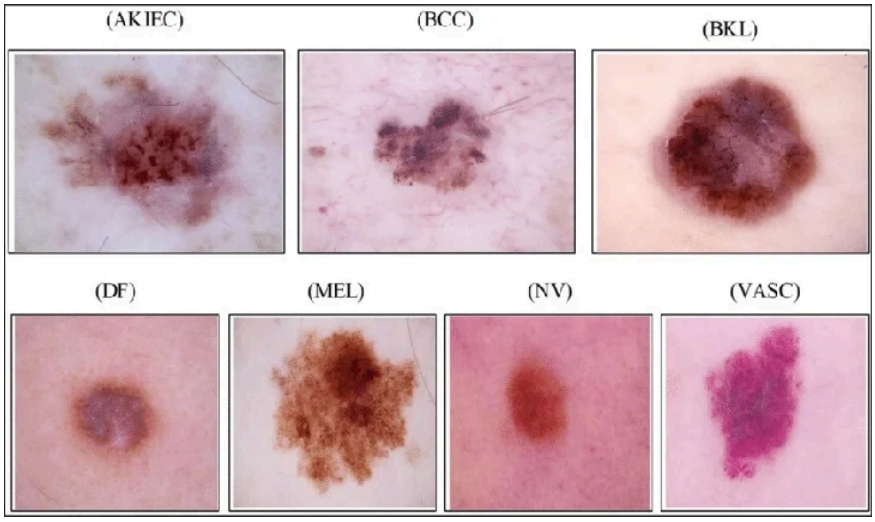

Figure 4. HAM10000 samples according to disease classification [39]

Compared to the DDI dataset, HAM10000 is substantially larger but contains less demographic diversity and primarily represents lighter skin tones, which is a limitation observed in many publicly available dermatology datasets. In this study, HAM10000 is used as an auxiliary dataset for domain adaptation and transfer learning experiments, where models are initially trained on labelled HAM10000 data and limited DDI labelled samples. This setup allows us to investigate whether supervised exposure to a larger dermatology dataset can improve model generalization and classification performance when transferred to a more demographically diverse evaluation dataset.

Finally, we utilize the ISIC 2019 training dataset as the unlabelled training data for our experiments involving self-supervised pretraining with the DINO framework. The dataset contains 25,331 images of skin cancer lesions [38, 40, 41]. Unlike the supervised training

datasets used in this study, the ISIC 2019 dataset is leveraged solely for representation learning, and diagnostic labels are not used during pretraining. This allows the model to learn general visual features and structural patterns present in dermatology images without relying on large amounts of labelled data. By pretraining on a large collection of unlabeled skin lesion images, we aim to capture domain-specific representations that may improve downstream performance when fine-tuned on the DDI dataset and as a result improve generalization of the resulting model.

Data Preprocessing and Augmentation

All images undergo specific preprocessing procedures prior to loading any images into the ResNet model. All images were resized to 256 x 256 and cropped to 224 x 224 to align with the input dimensions of the ResNet-50 architecture. Images were converted to tensors and normalized according to the ImageNet mean and standard deviation: ([0.485, 0.456, 0.406], [0.229, 0.224, 0.225]) [42].

Alongside these preprocessing techniques, we apply certain data augmentations to all image samples to ensure we reduce overfitting and capture relevant skin lesion details. Initially, we resize samples to 256 x 256 and crop to 224 x 224. For training data, we do a random resized crop of the 256 x 256 resized image. This has been shown to help prevent overfitting when training deep learning models over image data [43]. On the other hand, we apply a center crop of the 256 x 256 images on our testing data for evaluation consistency. To increase the robustness and generalization of our trained model, we apply further augmentations to the training images including random horizontal and vertical flips, and random rotations of 90°, 180°, and 270°. Ensuring that our testing transformations remain deterministic, helps to preserve the consistency of our evaluation methods.

Dataset Split	Resize	Crop	Augmentations
Train	256 x 256	224 x 224	Random Flips, Rotations
Val/Test	256 x 256	224 x 224	None

Figure 5. Image transformations according to dataset split

Model Architecture

All experiments conducted in this study utilize the ResNet-50 convolutional neural network architecture as the backbone model for classification of image samples. The architecture was originally introduced by He et al. (2016) as a part of the Residual Network model family. It has become widely adopted in computer vision tasks due to its robust ability to train deep neural networks while mitigating the vanishing gradient problem that is associated with increased depth in neural networks [44]. In this study, the ResNet-50 is the core architecture performing feature extraction across both the supervised and unsupervised training pipelines for consistency in evaluation and comparison between different representation learning strategies.

As the name suggests, the ResNet-50 is composed of 50 layers containing convolutional and fully connected operations organized within residual blocks. These blocks introduce the notion of "skip connections," which allow for the input of a block to be directly added to its output. A residual function, $F(x) = H(x) - x$, is learning instead of a direct mapping, $H(x)$. This results in the formula: $y = F(x) + x$. Here x represents the input feature map and $F(x)$ represents the learning residual mapping. This allows gradients to propagate directly though the identity connections while training, and as a result, reducing the vanishing gradient problem that limited the training of prior deep neural network architectures.

This architecture consists of an initial convolutional layer and is followed by four sequential groups of residual blocks, layer1 through layer4. These progressively learn higher-level visual representations of the data distribution. Earlier layers capture the lower level structure such as edges, textures, and colors. Deeper layers learn abstract semantic representations that are more relevant to the downstream classification task. This structure within the architecture for feature extraction makes the ResNet-50 suitable for medical image analysis where the subtle visual patterns, such as pigmentation irregularities or lesion edges, must be identified in various image conditions, especially in diverse data regimes.

ResNet-50 has been tested for its strong empirical performance across medical imaging benchmarks and dermatological classification tasks. Prior work has demonstrated that these residual networks provide reliable baselines for skin cancer classification, especially within transfer learning scenarios [27]. Furthermore, the architecture is compatible with both supervised pretraining and unsupervised representation learning frameworks, making it a more suitable backbone for the specific experiments within this study.

For implementation, we utilize the ResNet-50 model provided through the PyTorch deep learning framework [45]. This includes the weights pretrained on the ImageNet dataset. This pretraining enables the model to begin the experimental training with general feature representations that capture wide visual patterns, allowing the model to converge more efficiently when fine-tuned on dermatology datasets. In the experiments that involve self-supervised initialization, we load the ResNet-50 backbones pretrained using unlabelled ImageNet data with the DINO self supervised learning framework developed by Meta [32].

The original ResNet-50 architecture is trained for 1000 classes within the ImageNet dataset and thus contains a final fully connected layer outputting a 1000 dimensional vector corresponding to these categories. We modify this final classification layer to a linear layer containing two output neurons, representing the two diagnostic classes relevant to this study: benign and malignant. This modification preserves the pretrained feature extraction from earlier layers and allows us to proceed with binary classification over skin lesion images.

Overall, the ResNet-50 architecture provides a robust and flexible foundation model for the experiments conducted in this study. Its specific residual design allows for the facilitation of training deep convolutional layers and its widespread adaptation for computer vision tasks makes it a strong baseline for method comparison.

Training Pipelines

The training pipelines in this study are structured to systematically evaluate the impact of different representation learning strategies and dataset distributions under domain shift in dermatology image classification. To ensure consistency within evaluation of our methods, we utilize a common ResNet-50 backbone for all experiments while the methods of pretraining fine tuning are altered.

We establish a baseline by training and evaluating the ResNet-50 on the DDI dataset. The backbone is pretrained on the ImageNet dataset. Two forms of pretraining for this backbone are considered for this experiment and applied throughout the rest of the study: supervised pretraining, where labelled images are used, and self-supervised pretraining using the DINO framework, where image representations are learned without the use of labelled data. This allows

us to understand the effect of self-supervised representation learning on downstream diagnosis classification on a diverse dermatology dataset.

To further analyze how pretrained representations are utilized during fine-tuning, we introduce variations in which either all layers of the network are updated or only the higher-level layers (layer4 and the fully connected layer) are trained. This allows us to understand whether preserving pretrained feature representations improves performance when training on relatively small medical datasets such as DDI.

Building on these baseline experiments we extend our analysis to settings for domain adaptation. Specifically, we utilize HAM10000 to serve as a large external set of dermatology training data, and this approach is inspired by and follows from the findings of "Equitable Skin Disease Prediction Using Transfer Learning and Domain Adaptation" [46]. We aim to understand how labelled training on this dataset can help generalize to diverse skin tones. In this setting, we do conduct two experiments: One in which the pretrained backbone model is first trained on HAM10000 and then subsequently fine-tuned on the DDI training data and one in which the HAM10000 and DDI training data is combined and trained jointly.

Finally, we explore the use of self-supervised pretraining on unlabelled dermatology data. In this approach, the ResNet-50 backbone is pretrained using the DINO framework on unlabeled skin lesion images before being fine-tuned on the DDI dataset. This unlabelled dataset is the ISIC 2019 training data. This experiment aims to determine whether domain-specific representation learning improves robustness and fairness compared to standard ImageNet-based initialization which was used for all prior experiments.

Training Configuration

All the models in this study are trained under a consistent set of hyperparameters and optimization procedures that help ensure fair comparison across the different training pipelines. Unless otherwise specified, the same training configuration is applied across all experiments, with only the pretraining strategy and dataset composition varying between them.

Model training is performed using the Adam optimizer, selected for its stability and efficacy in fine-tuning pretrained convolutional neural networks. A learning rate of 1×10^{-4} with a batch size of 32 is used throughout the training. The number of epochs varies per experiment due to consistent early stopping, but a maximum of 100 is enforced within the training scripts.

Early stopping is implemented to prevent overfitting and ensure optimal model performance. Model performance is monitored using the macro-averaged F1-score on the validation set, which provides a balanced evaluation across both classes, which is needed to due the imbalance between classes. Training is terminated if the validation F1-score does not improve for 10 consecutive epochs. The best performing model, defined as the model achieving the highest validation macro F1-score, is saved during training and used for the final evaluation. The loss function used for back propagation during training is cross entropy loss, which is appropriate for the binary classification setting of benign versus malignant lesions.

In experiments that involve partial fine-tuning, we freeze layer4 and the fc, which correspond to the final 3 blocks of 3 convolutional layers each and the fully connected layer. These represent the final 10 layers of the 50 layer convolutional network. This accounts for approximately 14.98 million parameters, which represents 63.6% of the total model parameters. Additionally, all models are trained on NVIDIA A100 GPUs.

Dataset Splits

To ensure consistent and reproducible evaluation across all experiments, each dataset is partitioned into training, validation, and testing subsets according to fixed ratios.

The HAM10000 dataset is divided into a 90:10 split, where 90% of the data is used for training and 10% is reserved for validation. Since HAM10000 is primarily used for pretraining and domain adaptation, no separate test set is defined, and evaluation is conducted on the DDI dataset.

The DDI dataset is split into 80:10:10 proportions for training, validation, and testing, respectively. This split allows for model training and hyperparameter tuning while preserving a held-out test set for final evaluation throughout experiments. The DDI test set ensures fair comparison between different training pipelines.

The ISIC 2019 dataset is used exclusively for self-supervised pretraining and therefore utilizes 100% of the available data without explicit splitting, since no labels or validation metrics are required during this process.

All splits are fixed across experiments for consistency and enable direct comparison of model performance under different training configurations. The number 42 is used as a seed for a majority of computational operations including the random functions from NumPy, CUDA operations from PyTorch, and HPC deterministic flags to ensure reproducibility for the training pipelines.

Results

A series of experiments are made to determine which components improve the baseline and can be used to maximize performance accuracy for the model on diverse data. For each experiment, the overall test loss, accuracy, and F1 score are calculated. The goal is to combine complementary models (supervised + SSL + diverse fine-tuned) into an ensemble that maximizes equitable performance.

Test loss is a statistical measure of how incorrectly the model makes predictions. This result is calculated through the cross-entropy loss function. A lower test loss indicates the model performs better across all test-samples.

Accuracy is the percentage of test samples the model classified correctly, calculated by dividing the number of predictions that match the correct label by the total number of labeled samples.

F1 score is the harmonic mean of precision and recall, which measures how well the model balances to give a more reliable metric for imbalance datasets. It is calculated as 2 * (Precision * Recall) / (Precision + Recall), where Precision is the total number of predicted positives that were actually positive and Recall is the number of actual positives the model successfully predicted. A higher F1 score indicates better performance and suggests that the model can make strong, reliable predictions under imbalance data, avoiding mislabeling negatives and positives. F1 macro score is used to evaluate classes equally by calculating the mean of F1 scores per class, while weighted F1 score calculates F1 based on the frequency of class labels and may skew towards the majority class. Both metrics take the distribution of

classes into account. The score is typically measured from 0 to 1, where a score above 0.5 is considered significant.

In the following experiments, the models are tested on the DDI dataset, consisting of 132 images, where 34 samples are malignant and 98 are benign. Models trained on the ISIC dataset utilize 393 samples, of which 103 are labeled as malignant and 290 are labeled as benign. The HAM10000 dataset consists of 10015 training images, of which 1954 are malignant and 8061 are benign. A 90 - 10 train validation split is used, where the train size is 9013 images and the validation size is 1002.

Experiments

Baseline - Supervised Transfer (ImageNet, Full Fine-Tuning)

This experiment utilizes an ImageNet domain pretrained with standard supervised training and fine tuned on the DDI dataset with all layers of the ResNet model unfrozen. This experiment aims to establish a baseline by using an established approach to train on a diverse dataset.

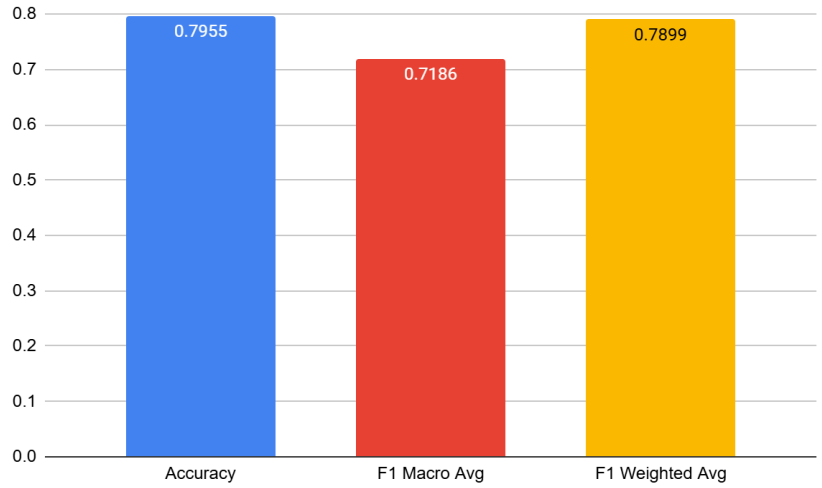

Figure 6a. Fully fine-tuned supervised training performance metrics

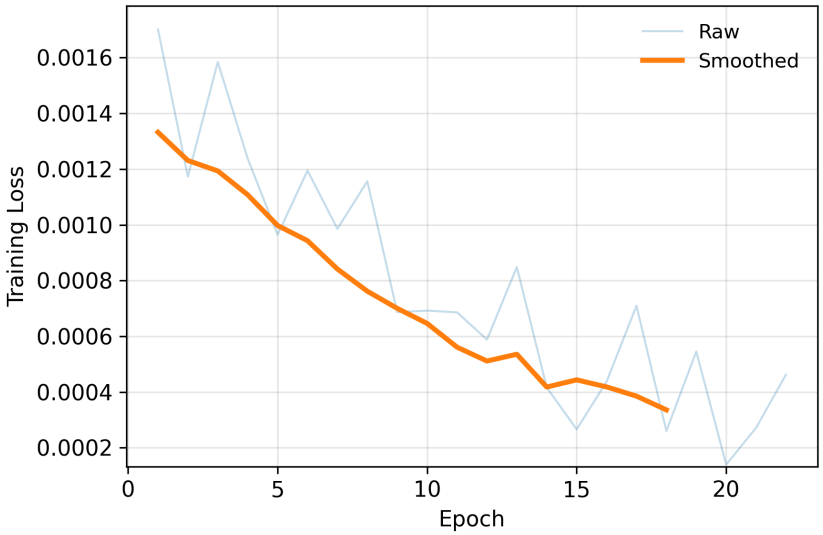

Figure 6b. Loss curves for baseline model. Early stopping occurred after 22 epochs.

It can be observed that the model retains a relatively high accuracy and F1 macro through a supervised pretraining domain.

Baseline - Self-Supervised Transfer (DINO, Full Fine-Tuning)

This experiment utilizes an ImageNet domain pretrained with a DINO model and fine tuned on the DDI dataset with all layers of the ResNet model unfrozen. The goal is to evaluate whether self-supervised representation outperforms the supervised domain. Early stopping was triggered after 27 epochs.

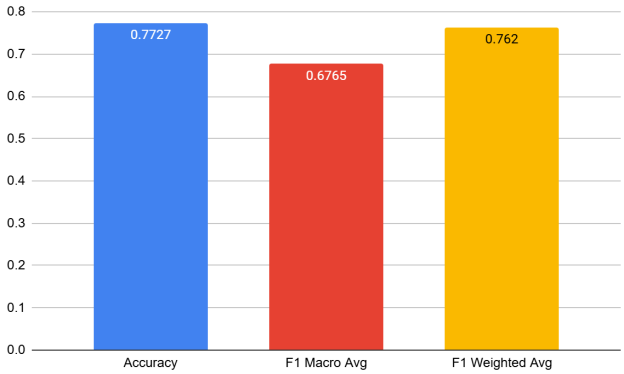

Figure 7a. Fully fine-tuned supervised DINO training performance metrics

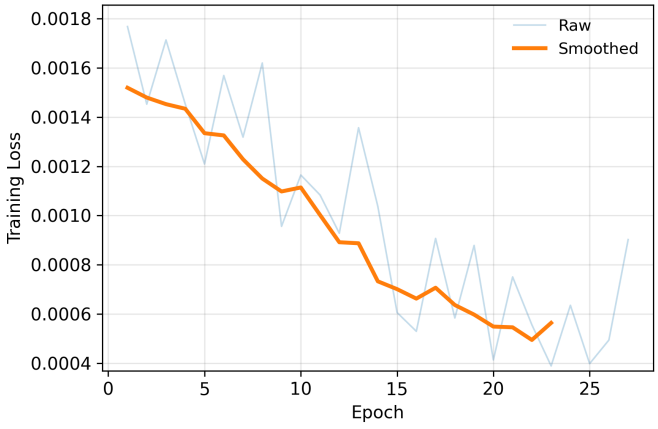

Figure 7b. Loss curves for baseline model with DINO pretraining

It can be observed that self-supervised features did not outperform supervised ImageNet features when all layers are fine-tuned.

Supervised Transfer (ImageNet, Partial Fine-Tuning)

This experiment uses an ImageNet domain pretrained with supervised learning and a ResNet model partially fine-tuned on the DDI dataset with only the last ten layers unfrozen. It is hypothesized that layer freezing has an impact on transfer learning.

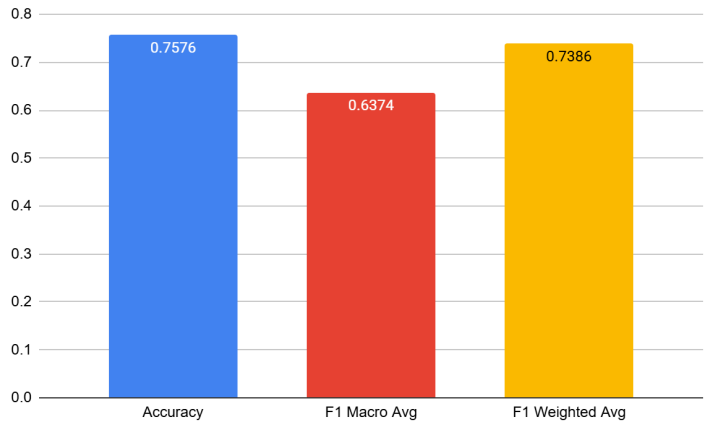

Figure 8a. Partially fine-tuned supervised transfer training performance metrics

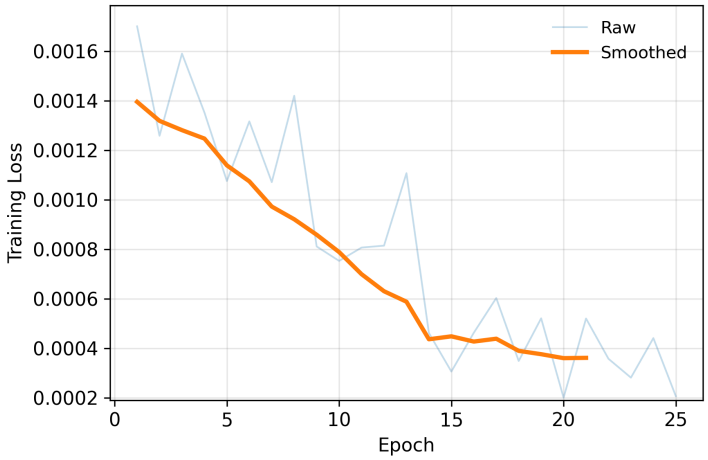

Figure 8b. Loss curve for partially fine-tuned supervised transfer model. Early stopping occurred after 22 epochs.

Based on the results, it is shown that freezing most layers slightly reduces performance, potentially because the model cannot adapt enough to dermatology features.

Self-Supervised Transfer (DINO, Partial Fine-Tuning)

Similar to the last experiment, this method features an ImageNet domain pretrained with DINO through self-supervised learning and a ResNet model fine-tuned on DDI with only the last ten layers unfrozen. It is hypothesized that the addition of self-supervised representation adds more robustness to the model when layers are frozen.

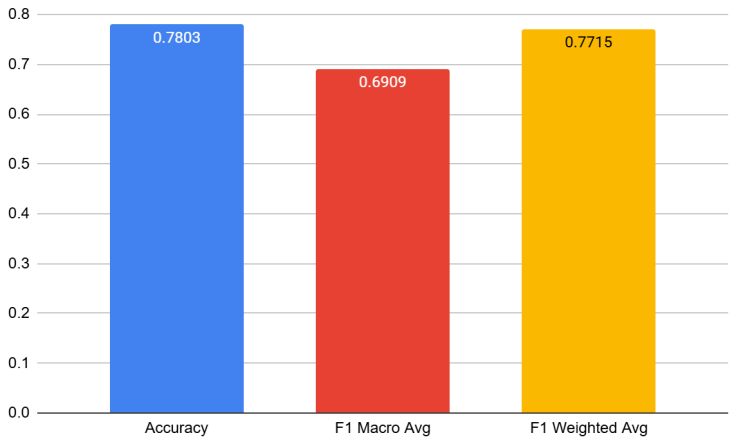

Figure 9a. Partially fine-tuned self-supervised transfer training performance metrics

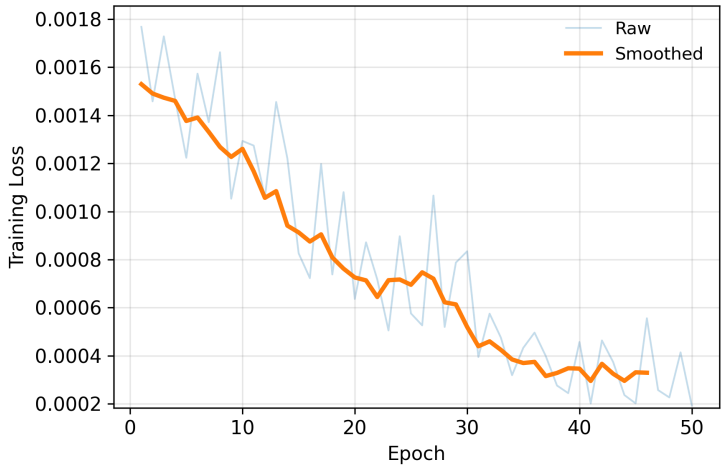

Figure 9b. Loss curve for partially fine-tuned self-supervised transfer model. Early stopping did not occur.

Based on the difference in results, it can be shown that self-supervised learning performance with DINO pretrained improves when layers are frozen. This suggests that some DINO features are more generalizable, benefiting from freezing early layers.

Supervised Domain Adaptation (Split Training)

In this experiment, an ImageNet domain pretrained with supervised learning is used to train a ResNet on both HAM10000 and DDI, where all layers are unfrozen. Through split training, the goal is to evaluate whether intermediate training improves domain adaptation.

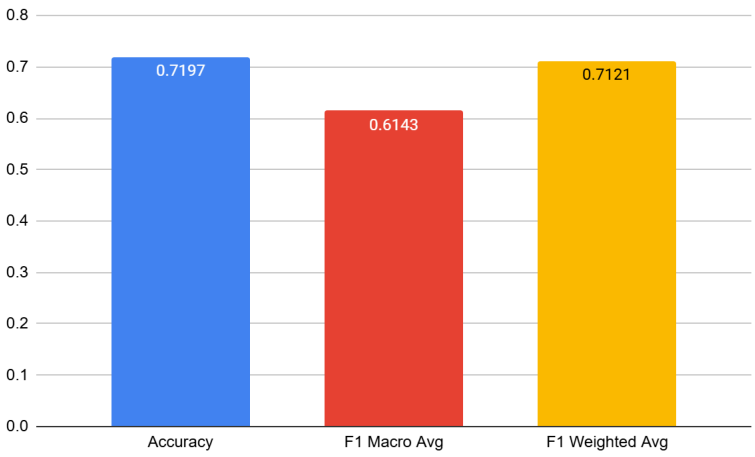

Figure 10a. Split trained model with supervised domain performance metrics.

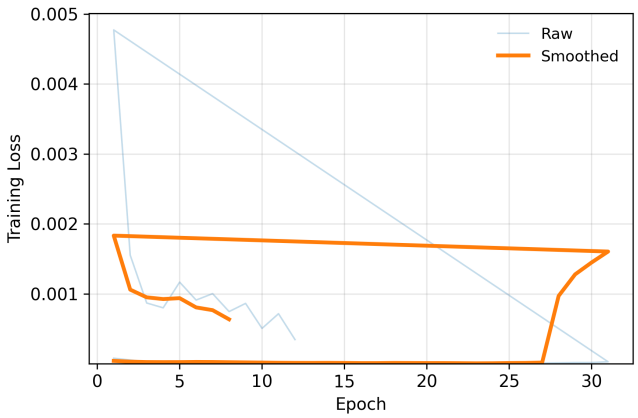

Figure 10b. Loss curve for supervised domain adaptation. Early stopping occurred after 12 epochs.

It is shown that performance drops compared to training directly on DDI, suggesting that split training may downgrade and not achieve domain adaptation on DDI due to a potential difference in the dataset distributions or the introduction of bias from the HAM10000 dataset. According to the loss curve, the model quickly adapts to learned features then plateaus.

Self-Supervised Domain Adaptation (DINO, Split Training)

In this experiment, an ImageNet domain pretrained with self-supervised learning through DINO is used to train a ResNet on both HAM10000 and DDI, with all layers unfrozen. The goal is to examine whether a self-supervised pretrained approach improves domain shifting.

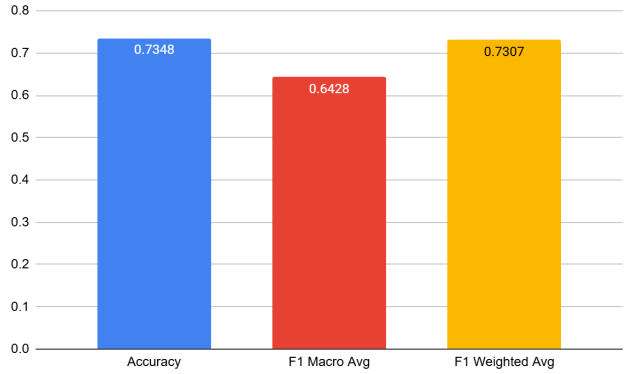

Figure 11a. Split training with self-supervised learning performance metrics

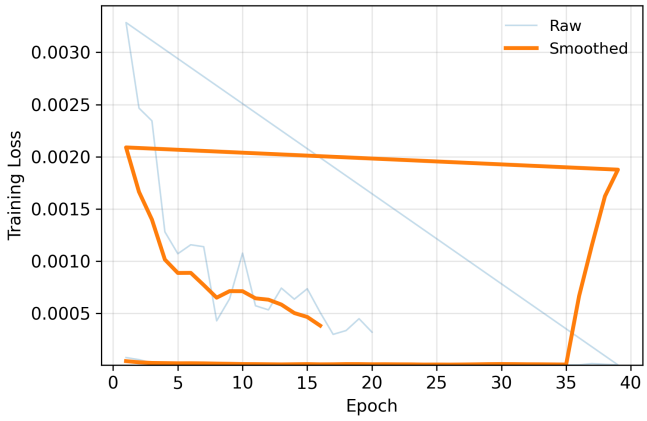

Figure 11b. Loss curve for split training with self-supervised learning. Early stopping occurred after 20 epochs.

While accuracy is lower than the reported baseline, the addition of self-supervised learning representation suggests that SSL is more robust to domain shifts compared to supervised split training.

Supervised Domain Adaptation (Joint Training)

This experiment utilizes a ResNet-50 model with all layers unfrozen, pretrained on a supervised ImageNet domain and fully fine-tuned on both the HAM10000 and DDI dataset combined into a joint dataset. The goal is to examine whether shared features in the dataset can help generalize model performance better.

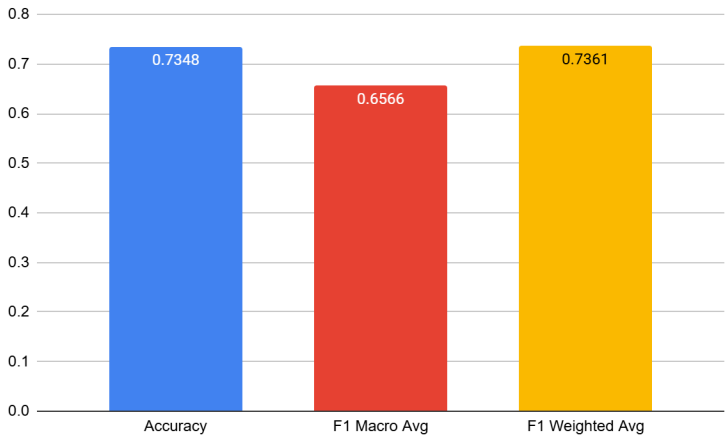

Figure 12a. Supervised joint training performance metrics

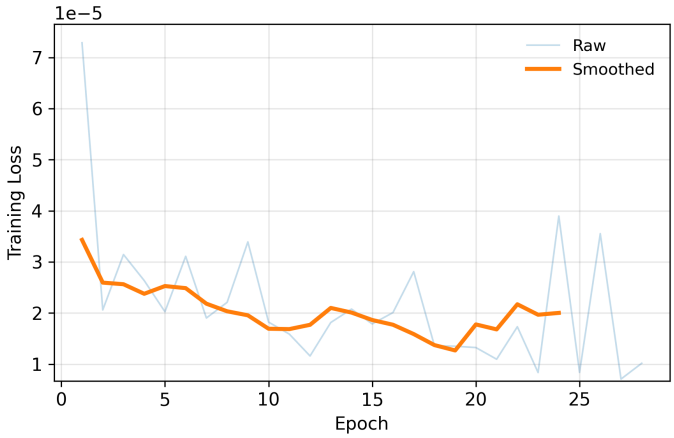

Figure 12b. Loss curve for supervised joint training model. Early stopping occurred after 28 epochs.

Compared to the split training model, the supervised joint dataset produces higher accuracy and F1 macro, suggesting that domain adaptation is retained more when features are shared in a unified set.

Self-Supervised Domain Adaptation (DINO, Joint Training)

Similar to the previous experiment, this experiment utilizes both the HAM10000 and DDI dataset as a joint training set, but with a self-supervised pretrained domain with DINO. Through a self-supervised domain, the goal is to examine whether learned features correlate with the shared features in the joint dataset.

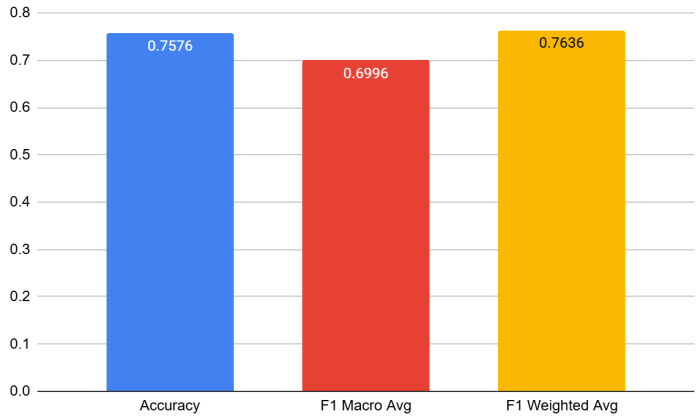

Figure 13a. Self-supervised joint training performance metrics

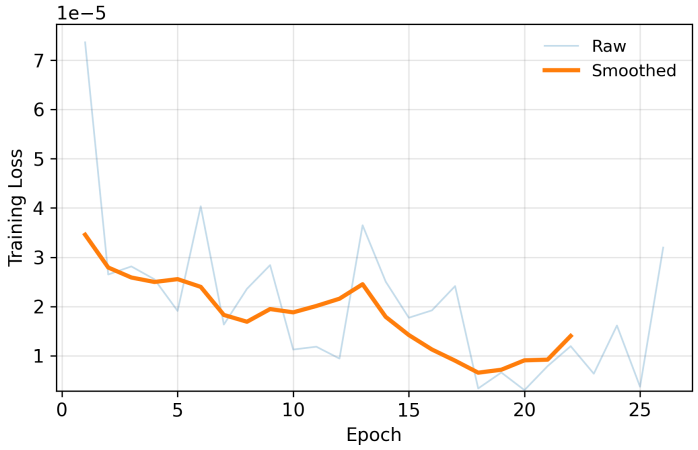

Figure 13b. Loss curve for self-supervised joint training with DINO. Early stopping occurred after 26 epochs.

Similar to the supervised pretrained domain, the self-supervised model with DINO achieves slightly increased overall performance. This suggests that the combination of self-supervised features and joint training supports domain adaptation.

HAM10000 Transfer (Supervised Pretrained, Full Fine-Tuning)

This experiment utilizes a ResNet-50 model with all layers unfrozen, pretrained on a self-supervised ImageNet domain and trained fully on the HAM10000 dataset. It is hypothesized that supervised pretraining on ImageNet will provide strong feature representations and higher overall accuracy, but underform on minority classes due to the disparity between mainstream data and diverse data.

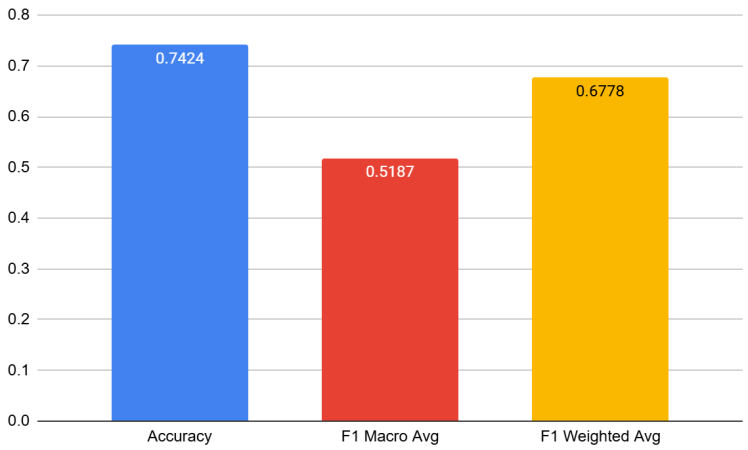

Figure 14a. HAM10000 supervised transfer performance metrics.

Overall, accuracy and F1 macro drop due to the large gaps between HAM10000 and DDI data.

HAM10000 Transfer (DINO Pretrained, Full Fine-Tuning)

This experiment utilizes the same architecture in the previous experiment but fixes a domain pretrained on self-supervised learning through DINO. It is hypothesized that a self-supervised domain will learn more generalized and robust representations and increase class balance.

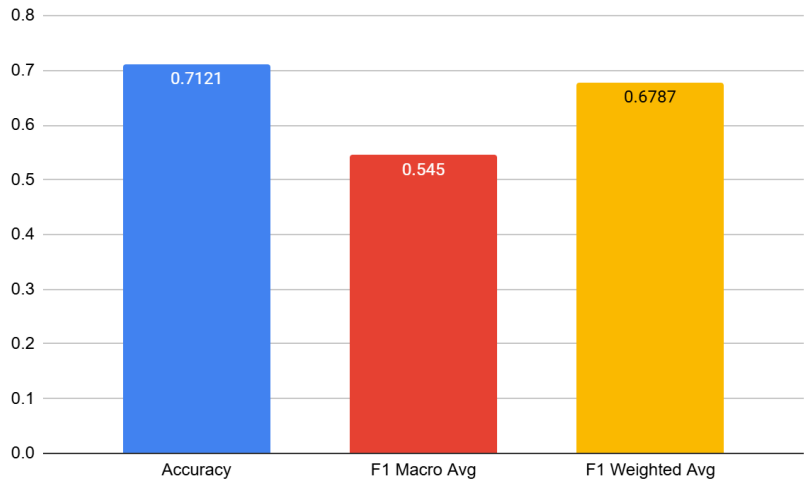

Figure 15a. HAM10000 self-supervised transfer performance metrics.

These results suggest that with the addition of self-supervised pretraining, F1 macro increases due to better minority class handling, but a slight drop in accuracy occurs, representing a tradeoff between overall correctness and class balance.

ISIC Domain Transfer (DINO, Full Fine-Tuning)

In this experiment, a ResNet-50 model is pretrained on the ISIC 2019 dataset, a larger, more generalized dataset with unlabeled data, and trained on the DDI dataset with all layers unfrozen. It is hypothesized that domain-specific pretraining combined with full fine-tuning on a diverse dataset will enable the model to fully adapt to the target dataset and increase macro F1.

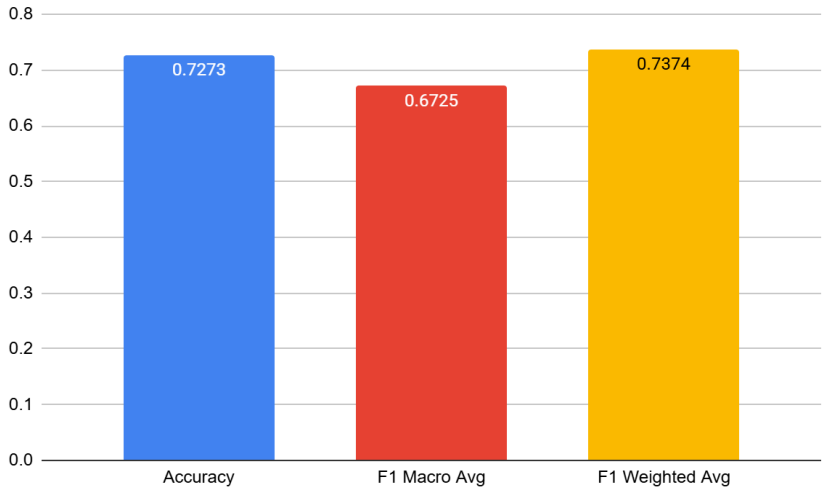

Figure 16a. Unfrozen ISIC domain transfer performance metrics.

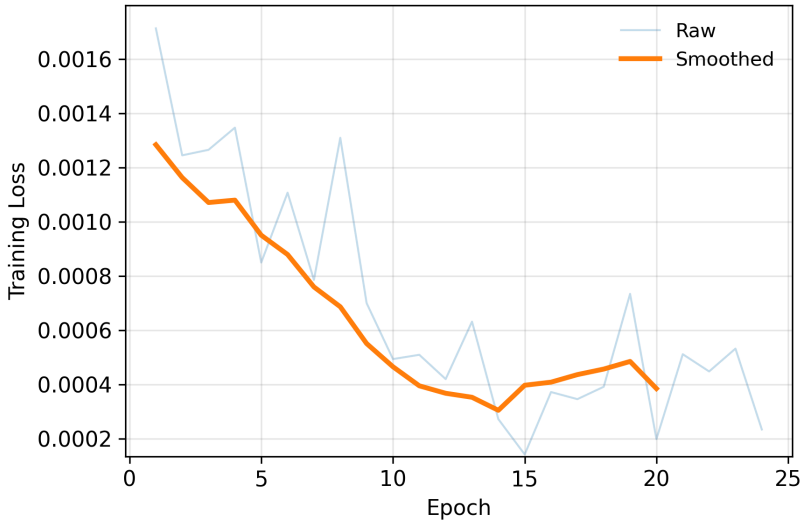

Figure 16b. Loss curve for ISIC domain transfer with fully fine-tuned DINO pertaining. Early stopping occurred after 24 epochs.

Through full fine-tuning, accuracy remains similar to the baseline, but F1 macro slightly drops when compared to pretraining with supervised learning rather than self-supervised.

ISIC Domain Transfer (DINO, Partial Fine-Tuning)

In this experiment, the same architecture as the previous experiment is used but with only the last 10 layers unfrozen. Through retaining most pretrained layers, it is hypothesized that the model will preserve learned representations, leading to higher accuracy, but may limit adaptation to minority classes as opposed to full fine-tuning.

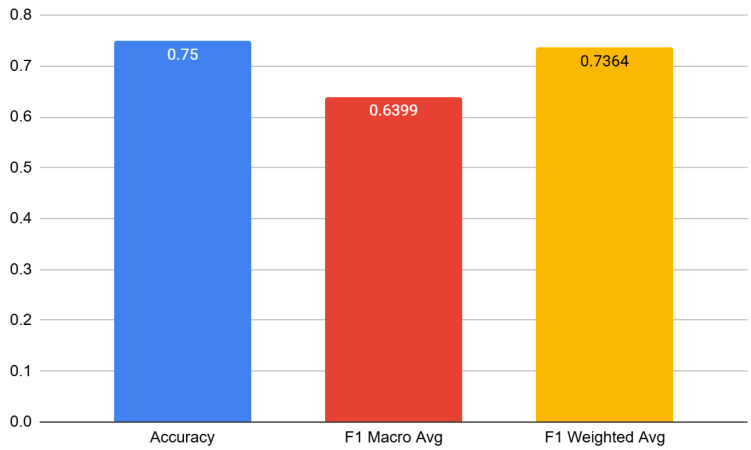

Figure 17a. Frozen ISIC domain transfer performance metrics.

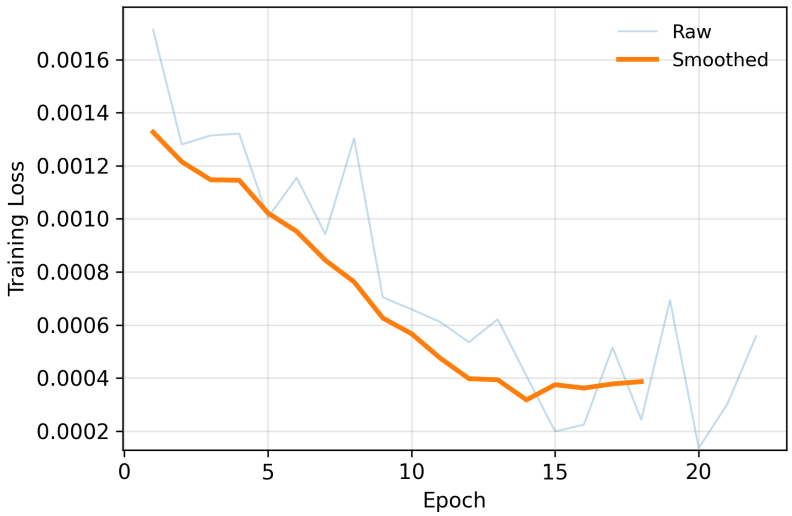

Figure 17b. Loss curve for frozen ISIC domain transfer. Early stopping occurred after 22 epochs.

Through partial fine-tuning, overall accuracy increases due to the presence of learned features, but the drop in F1 macro suggests that class balance decreases when learned representation is lost. However, weighted f1 remains stable, suggesting that the overall performance did not change significantly.

Model Overview

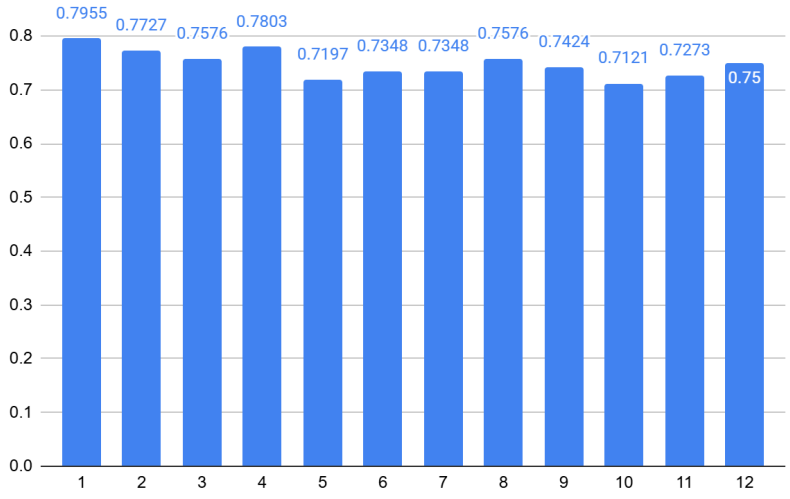

Figure 18. Comparison of all accuracies between all models.

In order of the experiments performed, the accuracies are listed for each model. An overall comparison demonstrates that a self-supervised model pretrained on ImageNet and trained strictly on HAM10000 yields the lowest results, as shown in Figure 15a.

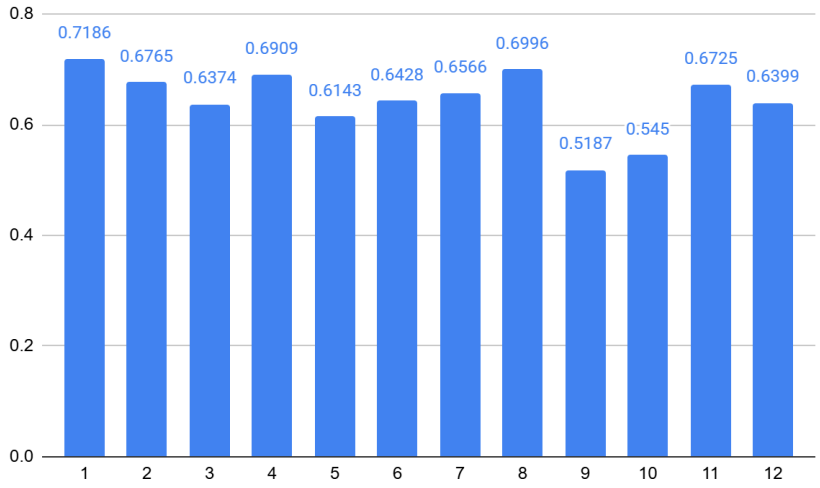

Figure 19. Comparison of F1 macro between all models.

In order of the experiments, the F1 macro scores are listed. The model pretrained with a supervised approach and fully fine tuned strictly on HAM10000 yields the lowest results, as shown in Figure 14a.

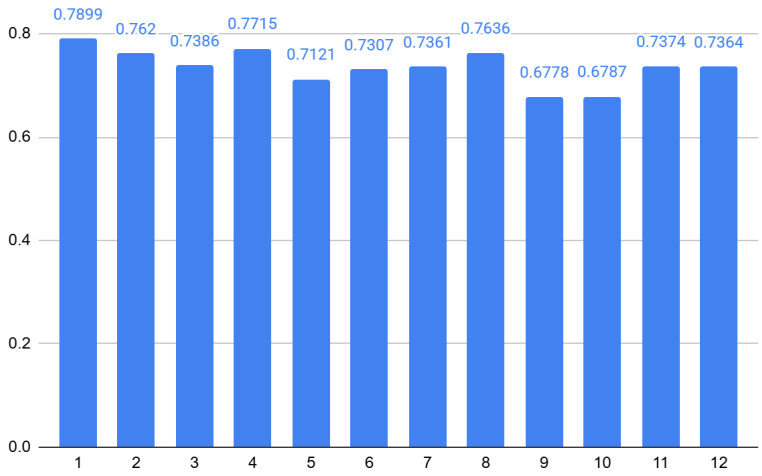

Figure 20. Comparison of weighted F1 between all models.

Finally, the weighted average F1 scores are compared between each model in order of the experiments performed. Similar to the F1 macro, the model with lowest accuracy is strictly fine tuned on HAM10000 and pretrained on ImageNet, as shown in Figure 14a.

In all three charts, it is shown that the baseline supervised model pretrained through ImageNet and trained on DDI yields the highest results in all three categories, suggesting highest overall performance. This is demonstrated in Figure 6a.

Discussion

Key Findings

Across all experiments, several consistent patterns remained regarding changes in domain adaptation, layer freezing, and pretraining approaches and their impact on DDI test performance.

1. **Supervised ImageNet pretraining provides the strongest baseline.**

It is observed that the fully fine-tuned model pretrained on a supervised ImageNet domain and fine-tuned on the DDI dataset achieved the highest overall accuracy and highest macro and weighted F1 score. This suggests that supervised hierarchies learned from the training dataset appearing most similar to the test result in increased accuracy.

2. **Self-supervised DINO features underperform supervised features when fully fine-tuned.**

While performance remains high with a supervised ImageNet pretrained domain, the introduction of self-supervised learning through DINO achieves lower accuracy and F1. This could be explained by the full fine-tuning approach that could possibly be overwritten by the general representations learned during self-supervised pretraining, which would reduce their advantages.

3. **DINO performs best when early layers are frozen.**

When a model is pretrained with DINO and fine-tuned with only the last 10 layers unfrozen, it is observed to outperform supervised models in macro F1 score. This suggests that DINO learns low-level features such as textures, edges, and color patterns that transfer more effectively across skin-tone variations. If these early layers are frozen, the generalizability of the model could prevent overfitting. On the contrary, a supervised transfer model shows a drop in performance with early layers frozen, potentially due to overfitting on natural images from ImageNet that may not transfer to dermatology.

4. **Split training reduces performance and does not achieve strong domain adaptation.**

For both the supervised and self-supervised pretrained models, the performance was reduced when trained first on HAM10000 and then on DDI, suggesting that HAM10000's distribution creates a disparity that biases the model, leading to conflicts with the DDI set. This could be due to the distribution of data in HAM10000, which largely features lighter skin tone images and varying lesion types, while DDI has a more diverse skin tone range.

5. **Joint training performs better than split training.**

When HAM10000 and DDI were combined into a joint dataset, performance was improved for both supervised and self-supervised models. While the dataset distributions are significantly different, the combination of data encourages the model to learn features that simultaneously generalize across both domains and help eliminate domain-specific biases.

6. **Training strictly on HAM10000 performs poorly when transferred to DDI.**

As observed during split and joint training, the HAM10000 dataset provides different learned features from the DDI dataset. Therefore, there is a domain mismatch when shifting from the train to test set, and a severe underrepresentation of darker skin tones. However, DINO is seen to perform better at improving minority-class performance.

7. **ISIC 2019 pretraining does not consistently improve performance on DDI.**

ISIC pretraining and full fine-tuning yields accuracy similar to the baseline but reduces macro F1, suggesting weaker minority-class handling. With partial fine-tuning, accuracy is increased but at the tradeoff of macro F1, further suggesting the ISIC domain preserved features that may not align as well with DDI as ImageNet. Majority classes appear to be favored while class balance reduces, suggesting more bias in class predictions.

Limitations

While this study provides insights into the applicability of representational learning under domain shift, several limitations should be considered when interpreting these results. The primary dataset used within this study is the DDI dataset, which is relatively small (~600

samples) and this limit introduces a large amount of variance within the evaluation metrics. This particularly impacts the macro F1 score, which is sensitive to small changes in model predictions and is used for early stopping and model comparison. Splitting this small dataset into train, test, and validation subsets further limits the ability of each subset to accurately represent the diverse skin lesion distribution that this study aims to make representational improvements upon. This may contribute to the observed discrepancies between validation and test performance across models and limits the reliability of conclusions regarding relative model performance.

Although all the datasets used in this study contain dermatologist verified dermoscopy images, the observed lack of improvement from domain adaptation and in-domain SSL pretraining may be attributed to the distributional differences between the source datasets (HAM10000 and ISIC 2019) and the target dataset (DDI). The DDI dataset contains an intentionally sourced equitable amount of data points per each Fitzpatrick skin tone, which is the source datasets and a majority of publicly available skin cancer datasets lack. Additionally, the quality and framing of the images sourced within the DDI dataset significantly differ from the majority of images within the source datasets. Public datasets such as the HAM10000 include high quality medical images which are cropped to only include the relevant skin lesions, which is not the case with images from DDI. This difference within the content of the images can also explain the lack of transferable representation that the model gained from training on these source datasets through supervised and unsupervised pipelines.

In addition to dataset-level limitations, the effectiveness of self-supervised pretraining in this study is constrained by the scale and composition of the unlabeled data used. While DINO-based pretraining was performed on ISIC 2019, the dataset size (~25k images) is significantly smaller than larger datasets, such as ImageNet, which are typically used to train

robust self-supervised representations. Due to this, the learned feature representations may lack the robustness and generality that is required to transfer effectively to target datasets like DDI. Furthermore, because the ISIC images share similar biases in image quality, lesion framing, and diversity lacking demographic representation, the self-supervised pretraining process may reinforce domain-specific features rather than learn invariant representations across a variety of skin tones and imaging conditions. This could explain why in-domain SSL pretraining failed to outperform standard supervised initialization since the model becomes increasingly specialized to the source domain rather than learning broadly transferable features. Additionally, the training configuration, including augmentation strategies, was tuned to ImageNet. This may further limit the effectiveness of SSL pretraining, as these factors are known to significantly influence the representation quality in self-supervised frameworks.

Initial Experiments

To establish a baseline model, a few experiments were performed. In several trials, an ImageNet pretrained on a ResNet-50 was used to establish a baseline, which serves as an upper bound for all comparisons. In the official experiments, baseline provides a foundation for the DINO self-supervised model pretrained ResNet-50, fine-tuned with the same data and hyperparameters as the pretrained supervised ResNet model. The baseline model achieved an accuracy of about 84.5% on the ISIC dataset, while the DINO-pretrained model outperformed it at 91.1%, suggesting that the self-supervised representation provided a more effective initialization for lesion classification in-domain

When tested on the diverse dataset, the self-supervised feature learning captures more generalizable image representations, particularly texture and structure based patterns that are less dependent on dataset-specific biases which get introduced during supervised ImageNet pretraining. Further experiments utilize around 15-25% of ISIC labeled data combined with diverse fine-tuning strategies to mitigate skin-tone performance gaps, resulting in improvements in out-of-distribution performance on the DDI dataset to around 74–79% accuracy across all tone groups, while ISIC performance remains stable.

Finally, a few experiments involve adding a small subset of Stanford Diverse data in combination with the ISIC training set, which led to substantial gains in cross-tone performance while maintaining high accuracy on the ISIC test set. When trained on 90% ISIC + 10% DDI data, the model achieved around 70–77% accuracy across tone groups on the DDI evaluation set, suggesting no improvement.

While the results were not substantial, they provided insights into potential areas where more advanced methodologies could be made, such as domain alignment, SSL initialization, and joint training. Therefore, we have modified our methodologies to focus on these refined trials.

Future Work

The primary issue that exists is the lack of quality and diversity of skin cancer datasets, which still remain underrepresented. One approach that can be taken to augment skin cancer datasets and improve detection in underrepresented skin tone groups is to utilize synthetic images produced by Generative Adversarial Networks (GANs). GAN models work by learning the underlying data distribution of datasets and can produce novel, verifiable images that make small modifications to existing data, such as changes to texture, lighting, or even the lesion. This

approach can reduce bias and improve model generalization, which decreases the probability of the model architecture overfitting to dominant skin tones.

Another step that can be taken is to improve generalization of models to handle varying quality resolutions of skin cancer/ medical images and improve domain adaptation. In our existing datasets, many images are taken by cellphones or through dermoscopes, which concentrate on the isolated lesions. However, there are several other approaches to taking skin cancer lesion images, such as medical cameras, total body photography, and 3D imaging systems, which will produce varying angles of lesions that our current model may not be able to generalize. Many of the images within a single dataset are captured with the same resolution, which further adds difficulty to generalization for lower resolution images. Future research can involve capturing more types of images from medical centers and training on ImageNet domain with unlabelled medical data from sources such as PADS-UFES-20.

Conclusion

In this work, we investigated the effect of supervised and self-supervised pretraining, domain adaptation, and diverse fine-tuning on melanoma classification performance across skin tones. Our results demonstrate that a standard ImageNet-pretrained ResNet-50 achieves high accuracy on the ISIC dataset, but transfers poorly to out-of-distribution skin tones on the Stanford Diverse Dermatology dataset. By introducing self-supervised DINO initialization we can improve the generalization modestly under full-data conditions, and the introduction of layer manipulation provides higher transfer learning to increase in-class accuracy. Overall, these findings highlight that model fairness cannot be inferred from in-domain accuracy alone, and a more strategic data curation process can contribute as strongly as the decision of model architecture. In particular, it

can be observed that domain alignment provides the strongest indicator of performance, rather than the initialization of self-supervised approaches or transfer techniques. This indicates that the most influential factor is how similar the model's learned feature space is aligned with the diversity, texture patterns, and skin-tone variability of images in the target dataset. Therefore, improving domain alignment through more inclusive training datasets rather than pretraining strategies offers the most reliable path toward better generalization. These results underscore the need for greater diversity of dermatological datasets to ensure fair and equitable performance in real-world clinical applications.

Ultimately, this paper converges on a final equitable pipeline combining (i) self-supervised initialization, (ii) limited diverse fine-tuning, and (iii) fairness-aware training strategies. This approach has the potential to reduce performance disparities across skin tones and serve as a more clinically inclusive foundation for AI-driven melanoma detection.

References

[1] Singson, Ben. "Skin Cancer a Danger of Summer Sun, Experts Say." *Jacksonville Journal-Courier*, 14 July 2025, www.myjournalcourier.com/news/article/skin-cancer-danger-summer-sun-experts-say-20415893.php. Accessed 21 Mar. 2026.

[2] American Academy of Dermatology Association. "Skin Cancer." *American Academy of Dermatology Association*, 22 Apr. 2022, www.aad.org/media/stats-skin-cancer. Accessed 24 Mar. 2026.

[3] CDC. “Health and Economic Benefits of Skin Cancer Interventions.” *National Center for Chronic Disease Prevention and Health Promotion (NCCDPHP)*, 24 May 2024, www.cdc.gov/nccdphp/priorities/skin-cancer.html. Accessed 24 Mar. 2026.

[4] “Survival Rate and Prognosis for Skin Cancer.” *MassiveBio*, 17 Nov. 2025, massivebio.com/survival-rate-and-prognosis-for-skin-cancer-bio/. Accessed 24 Mar. 2026.

[5] Conger, Krista. “AI Improves Accuracy of Skin Cancer Diagnoses in Stanford Medicine-Led Study.” *News Center*, 10 Apr. 2024, med.stanford.edu/news/all-news/2024/04/ai-skin-diagnosis.html.

[6] Balch, Bridget. “Why Are so Many Black Patients Dying of Skin Cancer?” *Association of American Medical Colleges*, 21 July 2022, www.aamc.org/news/why-are-so-many-black-patients-dying-skin-cancer. Accessed 21 Mar. 2026.

[7] Muhammed, Shabu Areez, et al. “A Generative AI Approach for Reducing Skin Tone Bias in Skin Cancer Classification.” ArXiv.org, 2026, arxiv.org/abs/2602.14356.

[8] Daneshjou, Roxana, et al. “Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set.” *Science Advances*, vol. 8, no. 32, 12 Aug. 2022, https://doi.org/10.1126/sciadv.abq6147.

[9] Cameron, M. C., Lee, E., Hibler, B. P., Barker, C. A., Mori, S., Cordova, M., Nehal, K. S., & Rossi, A. M. (2019). Basal cell carcinoma: Epidemiology; pathophysiology; clinical and histological subtypes; and disease associations. *Journal of the American Academy of Dermatology*, *80*(2), 303–317. https://doi.org/10.1016/j.jaad.2018.03.060

[10] Harrison, K. (2024). The accuracy of skin cancer detection rates with the implementation of dermoscopy among dermatology clinicians: A scoping review. *The Journal of Clinical and Aesthetic Dermatology*, *17*(9-10 Suppl 1), S18–S27. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11460753/

[11] Waseh, S., & Lee, J. B. (2023). Advances in melanoma: Epidemiology, diagnosis, and prognosis. *Frontiers in Medicine*, *10*, 1268479. https://doi.org/10.3389/fmed.2023.1268479

[12] Okobi, O. E., Abreo, E., Sams, N. P., Chukwuebuni, O. H., Tweneboa Amoako, L. A., Wiredu, B., Uboh, E. E., Ekechi, V. C., & Okafor, A. A. (n.d.). Trends in melanoma incidence, prevalence, stage at diagnosis, and survival: An analysis of the United States Cancer Statistics (USCS) database. *Cureus*, *16*(10), e70697. https://doi.org/10.7759/cureus.70697

[13] Narayanamurthy, V., Padmapriya, P., Noorasafrin, A., Pooja, B., Hema, K., Firus Khan, A. Y., Nithyakalyani, K., & Samsuri, F. (n.d.). Skin cancer detection using non-invasive techniques. *RSC Advances*, *8*(49), 28095–28130. https://doi.org/10.1039/c8ra04164d

[14] Dinnes, J., Deeks, J. J., Chuchu, N., Matin, R. N., Wong, K. Y., Aldridge, R. B., Durack, A., Gulati, A., Chan, S. A., Johnston, L., Bayliss, S. E., Leonardi‐Bee, J., Takwoingi, Y., Davenport, C., O'Sullivan, C., Tehrani, H., & Williams, H. C. (2018). Visual inspection and dermoscopy, alone or in combination, for diagnosing keratinocyte skin cancers in adults. *The Cochrane Database of Systematic Reviews*, *2018*(12), CD011901. https://doi.org/10.1002/14651858.CD011901.pub2

[15] Davis, S., Piggott, C., Lyon, C., & DeSanto, K. (2020). Effectiveness of dermoscopy in skin cancer diagnosis. *Canadian Family Physician*, *66*(10), 739–740. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7571636/

[16] Fried, L. J., Tan, A., Berry, E. G., Braun, R. P., Curiel-Lewandrowski, C., Curtis, J., Ferris, L. K., Hartman, R. I., Jaimes, N., Kawaoka, J. C., Kim, C. C., Lallas, A., Leachman, S. A., Levin, A., Lucey, P., Marchetti, M. A., Marghoob, A. A., Miller, D., Nelson, K. C., … Liebman, T. N. (2021). Dermoscopy proficiency expectations for us dermatology resident physicians: Results of a modified delphi survey of pigmented lesion experts. *JAMA Dermatology*, *157*(2), 189–197. https://doi.org/10.1001/jamadermatol.2020.5213

[17] Xu, Z., Sheykhahmad, F. R., Ghadimi, N., & Razmjooy, N. (2020). Computer-aided diagnosis of skin cancer based on soft computing techniques. *Open Medicine*, *15*(1), 860–871. https://doi.org/10.1515/med-2020-0131

[18] Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health*, *3*(11), e0000651. https://doi.org/10.1371/journal.pdig.0000651

[19] Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns*, *2*(10), 100347. https://doi.org/10.1016/j.patter.2021.100347

[20] Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine learning from theory to algorithms: An overview. *Journal of Physics: Conference Series*, *1142*, 012012. https://doi.org/10.1088/1742-6596/1142/1/012012

[21] Barragán-Montero, A., Javaid, U., Valdés, G., Nguyen, D., Desbordes, P., Macq, B., Willems, S., Vandewinckele, L., Holmström, M., Löfman, F., Michiels, S., Souris, K.,

Sterpin, E., & Lee, J. A. (2021). Artificial intelligence and machine learning for medical imaging: A technology review. *Physica Medica : PM : An International Journal Devoted to the Applications of Physics to Medicine and Biology : Official Journal of the Italian Association of Biomedical Physics (AIFB)*, *83*, 242–256. https://doi.org/10.1016/j.ejmp.2021.04.016

[22] Chen, C., Mat Isa, N. A., & Liu, X. (2025). A review of convolutional neural network based methods for medical image classification. *Computers in Biology and Medicine*, *185*, 109507. https://doi.org/10.1016/j.compbiomed.2024.109507

[23] Musthafa, M. M., T R, M., V, V. K., & Guluwadi, S. (2024). Enhanced skin cancer diagnosis using optimized CNN architecture and checkpoints for automated dermatological lesion classification. *BMC Medical Imaging*, *24*(1), 201. https://doi.org/10.1186/s12880-024-01356-8

[24] Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: An overview and application in radiology. *Insights into Imaging*, *9*(4), 611–629. https://doi.org/10.1007/s13244-018-0639-9

[25] Cherukuri, S., & Vara Prasad, S. D. (n.d.). EnsembleSkinNet: A transfer learning-based framework for efficient skin cancer detection with explainable AI integration. *Frontiers in Oncology*, *15*, 1699960. https://doi.org/10.3389/fonc.2025.1699960

[26] Wu, Y., Chen, B., Zeng, A., Pan, D., Wang, R., & Zhao, S. (2022). Skin cancer classification with deep learning: A systematic review. *Frontiers in Oncology*, *12*, 893972. https://doi.org/10.3389/fonc.2022.893972

[27] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

[28] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

[29] Huang, S.-C., Pareek, A., Jensen, M., Lungren, M. P., Yeung, S., & Chaudhari, A. S. (2023). Self-supervised learning for medical image classification: A systematic review and implementation guidelines. *NPJ Digital Medicine*, *6*, 74. https://doi.org/10.1038/s41746-023-00811-0

[30] Rajaram, S., Gupta, S., & Medhi, B. (2024). Digital medicine, intelligent medicine, and smart medication system. *Indian Journal of Pharmacology*, *56*(3), 159–161. https://doi.org/10.4103/ijp.ijp_501_24

[31] Haggerty, H., & Chandra, R. (2024). *Self-supervised learning for skin cancer diagnosis with limited training data* (arXiv:2401.00692). arXiv. https://doi.org/10.48550/arXiv.2401.00692

[32] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). *Emerging properties in self-supervised vision transformers* (arXiv:2104.14294). arXiv. https://doi.org/10.48550/arXiv.2104.14294

[33] Doron, M., Moutakanni, T., Chen, Z. S., Moshkov, N., Caron, M., Touvron, H., Bojanowski, P., Pernice, W. M., & Caicedo, J. C. (2023). Unbiased single-cell morphology with self-supervised vision transformers. *bioRxiv*, 2023.06.16.545359. https://doi.org/10.1101/2023.06.16.545359

[34] Useini, V., Tanadini-Lang, S., Lohmeyer, Q., Meboldt, M., Andratschke, N., Braun, R. P., & Barranco García, J. (2024). Automatized self-supervised learning for skin lesion screening. *Scientific Reports*, *14*, 12697. https://doi.org/10.1038/s41598-024-61681-4

[35] Gröger, F., Lionetti, S., Gottfrois, P., Gonzalez-Jimenez, A., Pouly, M., & Navarini, A. A. (2024). 442 towards reliable evaluation benchmarks for digital dermatology. *Journal of Investigative Dermatology*, *144*(12), S304. https://doi.org/10.1016/j.jid.2024.10.452

[36] Shen, Y., Li, H., Sun, C., Ji, H., Zhang, D., Hu, K., Tang, Y., Chen, Y., Wei, Z., & Lv, J. (2024). Optimizing skin disease diagnosis: Harnessing online community data with contrastive learning and clustering techniques. *NPJ Digital Medicine*, *7*, 28. https://doi.org/10.1038/s41746-024-01014-x

[37] Tjiu, J.-W., & Lu, C.-F. (2025). Equity and generalizability of artificial intelligence for skin-lesion diagnosis using clinical, dermoscopic, and smartphone images: A systematic review and meta-analysis. *Medicina*, *61*(12), 2186. https://doi.org/10.3390/medicina61122186

[38] Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data 5, 180161 (2018). https://doi.org/10.1038/sdata.2018.161

[39] Kumar, M., Batra, S., Chakraborty, G., Kaur, P., & Mahajan, M. (2023). An Efficient Deep Learning based Ensemble Approach for Multiclass Skin Cancer Classification Using Medical Imaging. In 2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS) (pp. 417-423).

[40] Hernández-Pérez C, Combalia M, Podlipnik S, Codella NC, Rotemberg V, Halpern AC, Reiter O, Carrera C, Barreiro A, Helba B, Puig S, Vilaplana V, Malvehy J. BCN20000: Dermoscopic lesions in the wild. Scientific Data. 2024 Jun 17;11(1):641.

[41] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, Allan Halpern: "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)", 2017; arXiv:1710.05006.

[42] J. Deng, W. Dong, R. Socher, L. -J. Li, Kai Li and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 248-255, doi: 10.1109/CVPR.2009.5206848.

[43] Shorten, C., Khoshgoftaar, T.M. A survey on Image Data Augmentation for Deep Learning. J Big Data 6, 60 (2019). https://doi.org/10.1186/s40537-019-0197-0

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. (2015). Deep Residual Learning for Image Recognition.

[45] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, & Soumith Chintala. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library.

[46] Dip, S. A., Ibn Arif, K. H., Shuvo, U. A., Khan, I. A., & Meng, N. (2024). Equitable Skin Disease Prediction Using Transfer Learning and Domain Adaptation. Proceedings of the AAAI Symposium Series, 4(1), 259-266. https://doi.org/10.1609/aaaiss.v4i1.31800