Vietnamese National University HCM

HCMC International University

**School of Computer Science and Engineering**

**AI REPORT**

# **TITLE: Self-Supervised Domain Adaptation for Skin Lesion Classification**

**Course: Artificial Intelligence**

**ID: IT159IU**

*Instructor: Dr. Nguyen Trung Ky*

*MSc. Ly Tu Nga*

# List of Members

| Student Name | Student ID | Member Role |
| --- | --- | --- |
| Trần Nam Anh | ITDSIU23030 | Leader |
| Nguyễn Đức Hải | ITDSIU23006 | Member |

# List of Contents

# List of Figures

# List of Figures

# Abstract

Deep learning models are increasingly used in medical imaging, especially in dermatology. However, the effectiveness of existing models in skin lesion classification is highly dependent on large annotated datasets, which are difficult and costly to obtain. Additionally, models trained on one dataset often fail to generalize to another due to domain shift, which is caused by variations in imaging devices, lighting conditions, and patient skin tones. To address these challenges, this research develops and evaluates an integrated framework that combines **Self-Supervised Learning (SSL)** and **Domain Adaptation (DA)**. We implement a pipeline where a ResNet-50 backbone is first pre-trained using SimCLR on the HAM10000 dataset, followed by fine-tuning using Domain-Adversarial Neural Networks (DANN). Our experiments conduct a rigorous comparison of this Self-Supervised Domain Adaptation (SSDA) model against both a simplified SSL fine-tuning (SSL-FT) approach and a strong, conventionally supervised baseline. The results show that the supervised baseline performed the best *(83.33%)*, beating both the SSL-FT *(70.73%)* and SSDA *(75.67%)* models. This surprising result suggests that for this specific task, the powerful, general-purpose features from ImageNet are more effective than features fine-tuned via SSL on a smaller, domain-specific dataset. Although SSDA worked technically, the simple supervised baseline is still better for this specific dataset.

# 1. Introduction

Deep learning is becoming important in medical analysis. In dermatology, skin lesion classification using Convolutional Neural Networks (CNNs) has shown impressive accuracy in distinguishing between benign and malignant conditions such as melanoma. Several large-scale datasets, such as HAM10000 and ISIC challenges, have enabled significant progress, with models achieving dermatologist-level performance under controlled conditions.

Despite these achievements, the practical, real-world application of these AI models remains limited by two fundamental challenges. The first is a **dependency on large, labeled datasets**. Obtaining expert annotations for medical images is a time-consuming and costly process, leaving vast amounts of unlabeled data in hospitals and public repositories underutilized.

The second critical challenge is the **domain shift problem**. Models trained on data from one clinical environment often experience a significant drop in performance when applied to another due to variations in imaging devices, lighting conditions, or patient demographics. This makes the model unreliable when applied to new data. It also raises fairness issues because a model trained on one patient population may fail to perform effectively on another.

To address these limitations, this project develops and empirically evaluates an integrated **Self-Supervised Domain Adaptation (SSDA)** framework. This approach combines Self-Supervised Learning (SSL) to learn rich visual features from unlabeled data with Domain Adaptation (DA) to enhance the model's robustness against domain shift. The primary goal of this research is to rigorously assess the effectiveness of this advanced framework by comparing its performance on the HAM10000 dataset against strong, conventional supervised baselines. We aim to test if SSDA is actually useful for dermatology images.

# 2. Related Work

This project combines three main topics: supervised learning, SSL, and Domain Adaptation.

## 2.1. Supervised Transfer Learning in Medical Imaging

The dominant paradigm in medical image classification relies on supervised transfer learning (Ahmed et al., 2023; Naji et al., 2022). This approach involves fine-tuning a Convolutional Neural Network (CNN), typically a ResNet pre-trained on the large-scale ImageNet dataset, on a smaller, domain-specific labeled dataset. While this method has achieved dermatologist-level performance on curated datasets like HAM10000, its effectiveness is often limited in real-world scenarios due to two major weaknesses. Firstly, it depends heavily on expert-annotated data, which is expensive and scarce. Secondly, it is highly susceptible to **domain shift**, where performance degrades significantly when the model is applied to images from a new clinical environment with different imaging characteristics (Yadav & Bhat, 2024). This project uses a well-optimized supervised transfer learning model as a strong baseline to benchmark the performance of our proposed methods.

## 2.2. Self-Supervised Learning (SSL)

To address the dependency on labeled data, Self-Supervised Learning (SSL) has emerged as a powerful alternative. SSL aims to learn meaningful visual representations from unlabeled data by solving pretext tasks. In computer vision, contrastive learning has become the leading approach.

The **SimCLR** framework (Chen et al., 2020) is a foundational work in this area. It learns representations by maximizing the agreement between two differently augmented "views" of the same image using a contrastive loss (NT-Xent). The framework's simplicity and effectiveness, relying on a ResNet backbone, a projection head, and strong data augmentations, have demonstrated that an unlabeled model can achieve performance comparable to supervised baselines. In the context of our research, SimCLR provides a powerful mechanism for learning robust, domain-specific features from the unlabeled HAM10000 dataset, forming the initial stage of our advanced frameworks.

## 2.3. Unsupervised Domain Adaptation (DA)

While SSL learns transferable features, it does not explicitly address the domain shift problem. This is the primary goal of Domain Adaptation (DA), which seeks to align the feature distributions between a labeled *source* domain and an unlabeled *target* domain.

A prominent technique in DA is **Domain-Adversarial Neural Networks (DANN)** (Ganin & Lempitsky, 2015). DANN introduces a domain discriminator that is trained to distinguish between features from the source and target domains. Simultaneously, the main feature extractor is trained not only to classify the source data correctly but also to generate features that *confuse* the domain discriminator. This adversarial process, mediated by a Gradient Reversal Layer (GRL), forces the feature extractor to learn domain-invariant representations.

## 2.4. The Research Gap: Integrating SSL and DA

The theoretical connection between SSL and DA is strong; SSL helps reduce the source error ($\epsilon_S$) in the generalization bound, while DA targets the domain divergence term (d(S, T)). Recent works, such as Azizi et al. (2021), have shown promising results by empirically combining these paradigms, for instance, by following SSL pre-training with adversarial alignment. However, few studies combine these methods and compare them strictly against a supervised baseline. This project aims to fill this gap by building and evaluating a full **Self-Supervised Domain Adaptation (SSDA)** pipeline, thereby providing a clear empirical analysis of its practical benefits and limitations.

# 3. Methodology

To address the challenges of labeled data scarcity and domain shift, we designed a comprehensive experimental framework to develop and evaluate three distinct models for skin lesion classification. This section details the data preparation process and the architecture of each of the three methods: our main proposed **SSDA framework**, an **ablation study model (SSL-FT)**, and a **strong supervised baseline**.

## 3.1. Data Preparation and Splitting

Our experiments are conducted on the **HAM10000** dataset, a large collection of dermatoscopic images. The dataset was first processed into a folder-per-class structure, where each subdirectory corresponds to one of the seven lesion types.

To create a consistent environment for all experiments, a definitive **"Golden Data Split"** was established. The entire dataset was partitioned using stratified sampling based on the lesion class, with **60% of the images allocated to a labeled** and the remaining **40% allocated to an unlabeled *target* domain**. The source domain (*src_loader*) is used for supervised training, while the target domain (*tgt_loader*) is used exclusively for evaluation to measure the models' generalization capability.

## 3.2. Method 1: The Proposed SSDA Framework (SimCLR + DANN)

Our primary proposed method is a two-stage Self-Supervised Domain Adaptation (SSDA) pipeline designed to tackle both core challenges simultaneously. The overall pipeline is illustrated in Figure 3.1.

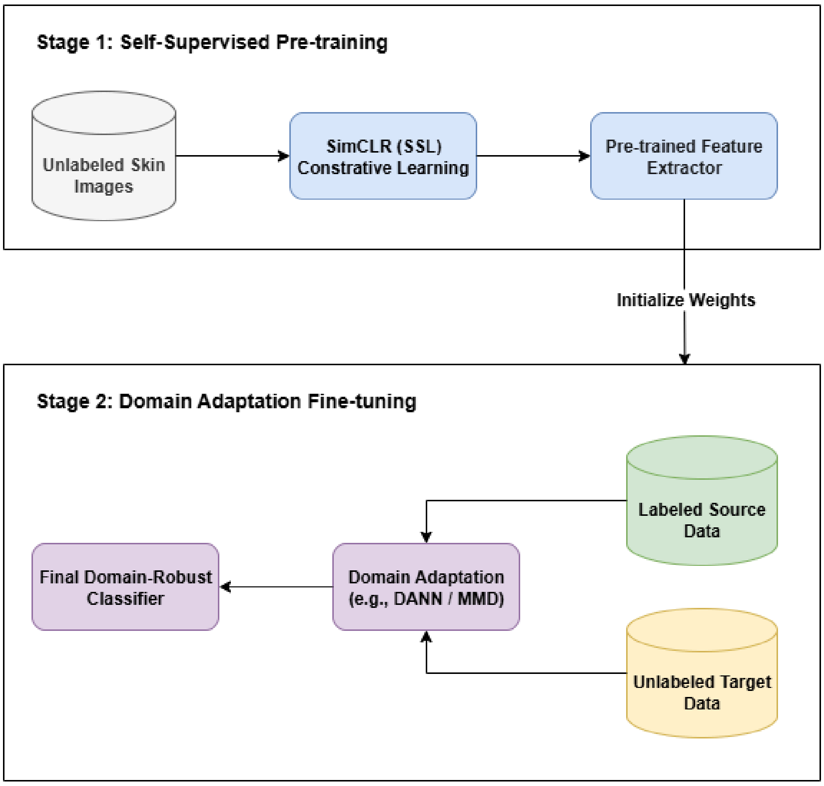


***Figure 3.1: Overall Pipeline of the Proposed SSDA Framework.*** *The framework consists of a self-supervised pre-training stage to learn robust features, followed by a domain adaptation fine-tuning stage to align features across domains.*

### 3.2.1. Stage 1: Self-Supervised Pre-training (SSL)

The first stage aims to learn rich, domain-specific visual representations from all 10,015 unlabeled HAM10000 images, starting from an ImageNet pre-trained ResNet-50 backbone. We adopt the **SimCLR** framework for this purpose, as depicted in Figure 3.2.

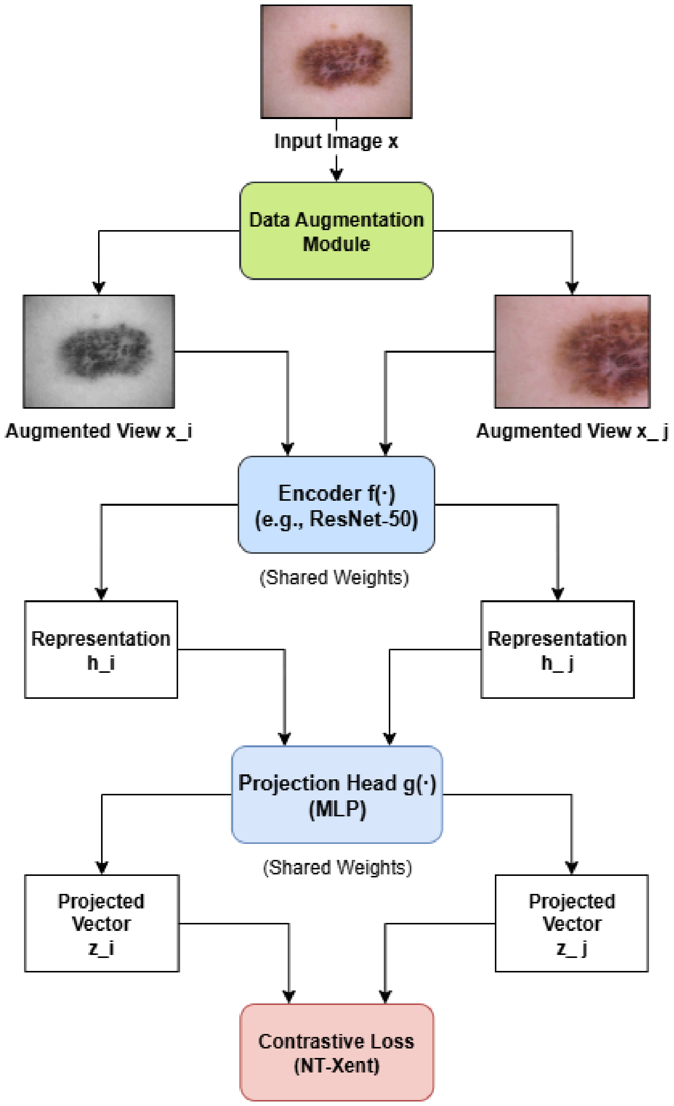


***Figure 3.2: SimCLR Architecture for SSL Pretraining.*** *An input image is transformed into two augmented views. The encoder and projection head map these views into a latent space where the NT-Xent contrastive loss is applied.*

The key components of this stage are:

- **Encoder f(·):** A ResNet-50 backbone that extracts a feature representation h from an augmented image view.
- **Projection Head g(·):** A 2-layer MLP that projects the representation h into a vector z in the contrastive learning space.
- **Contrastive Loss (NT-Xent):** This loss function maximizes the cosine similarity between the projection vectors ($z\_i$, $z\_j$) of two views from the same source image (a

positive pair) while minimizing their similarity to all other image projections in the batch (negative pairs).

After 50 epochs of pre-training, the projection head g(·) is discarded, and the fine-tuned encoder f(·) becomes the pre-trained feature extractor for the next stage.

### 3.2.2. Stage 2: Domain-Adversarial Fine-tuning (DA)

The second stage fine-tunes the pre-trained encoder using the **DANN** architecture, illustrated in Figure 3.3. The goal is to train a classifier that performs well on the target domain by making the feature distributions of the source and target domains indistinguishable.

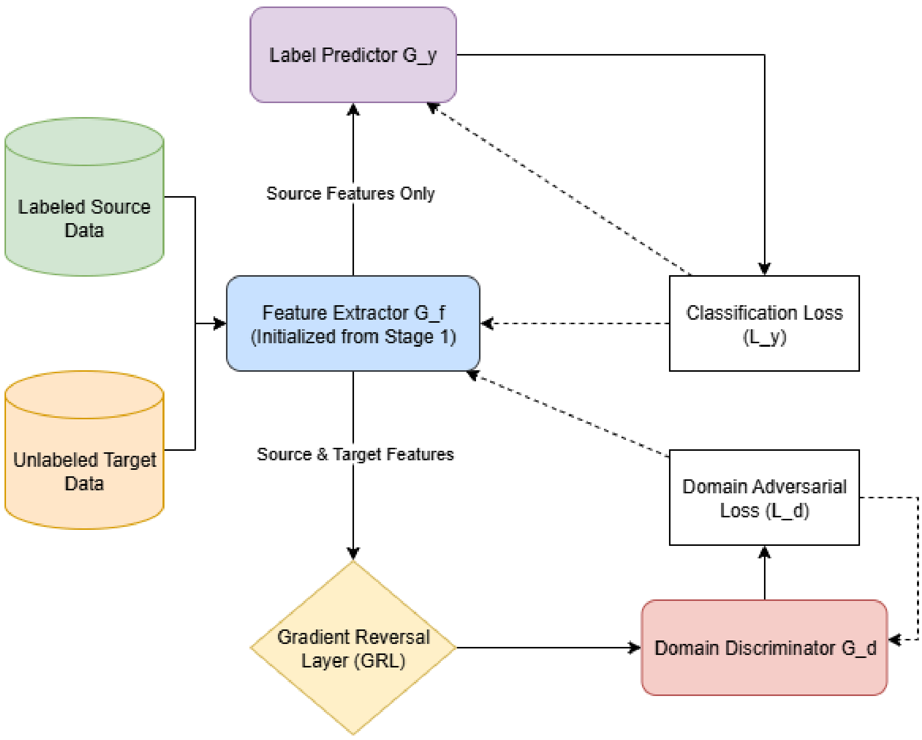


***Figure 3.3: DANN-Based Domain Adaptation Architecture.*** *The Feature Extractor is trained to minimize the classification loss while simultaneously maximizing the domain adversarial loss via the Gradient Reversal Layer (GRL).*

The architecture consists of three components:

1. **Feature Extractor G_f:** The encoder from Stage 1. It is fine-tuned to both minimize classification loss on source data and fool the Domain Discriminator.
2. **Label Predictor G_y:** A classifier trained only on features from the labeled source data to predict the correct lesion class. Its performance is measured by the **Classification Loss (L_y)**, which is a weighted cross-entropy loss to handle class imbalance.
3. **Domain Discriminator G_d:** A classifier trained to distinguish between features from the source and target domains. Its performance is measured by the **Domain Adversarial Loss (L_d)**.
4. **Gradient Reversal Layer (GRL):** During backpropagation, the GRL reverses the sign of the gradient flowing from the Domain Discriminator to the Feature Extractor. This forces the extractor to learn features that the discriminator cannot differentiate, thus aligning the two domains.

The combined objective function is $L = L_y + \lambda * L_d$, where λ is a dynamic parameter that controls the strength of the domain adaptation.

### 3.3. Method 2: The Ablation Study Model (SSL-FT)

To isolate and measure the impact of the SSL pre-training stage, we implemented an "ablation study" model. The SSL-FT method follows the same procedure as the Baseline (described next), but with one critical difference: instead of initializing the encoder with standard ImageNet weights, it is initialized with the powerful, domain-specific weights from our **SimCLR pre-training stage (Stage 1)**. It does not use any DANN components.

### 3.4. Method 3: The Supervised Baseline

To establish a strong performance benchmark, we implemented a conventional supervised transfer learning model. This model consists of a standard **ResNet-50 encoder pre-trained on ImageNet** and a classifier. The entire model is fine-tuned end-to-end on the labeled **source domain** data only. To ensure a fair comparison, it is trained using the same optimized configuration as our other models, including the same number of epochs, SGD optimizer, learning rate scheduler, and weighted cross-entropy loss.

# 4. Implementation and Results

This section details the implementation of our experimental framework and presents the empirical results obtained from evaluating the three distinct models: the Supervised Baseline, our proposed SSDA model, and the SSL-FT ablation study model.

## 4.1. Implementation Details

All experiments were conducted on the Kaggle platform using a single NVIDIA Tesla T4 GPU. The framework was implemented using the **PyTorch** library, with **Torchvision** for models and data transformations, **Scikit-learn** for performance metrics, and **Seaborn/Matplotlib** for visualization.

The key hyperparameters were standardized across all fine-tuning experiments to ensure a fair comparison:

- **Image Size:** All images were resized to 224 x 224 pixels.
- **Batch Size:** A batch size of 32 was used for all fine-tuning stages.
- **Optimizer:** SGD with a momentum of 0.9 and weight decay of 5e-4.
- **Learning Rate:** The initial learning rate for the classifier head was 1e-3, and 1e-4 for the encoder backbone.
- **Scheduler:** A StepLR scheduler was used, reducing the learning rate by a factor of 0.1 every 10 epochs.
- **Epochs:** The SimCLR pre-training was run for 50 epochs, while all supervised fine-tuning experiments were run for 20 epochs.
- **Loss Function:** Weighted Cross-Entropy Loss was used for classification, with weights calculated to balance the classes in the source dataset.

The complete source code for this project, including the training pipeline and the demonstration notebook for result verification, is publicly available on our GitHub repository.

- **GitHub Link:** https://github.com/LineLuLan/SSDA-Model-for-Skin-Lesion-Classification
- **Kaggle Link:** https://www.kaggle.com/code/lineizumi/demonstration
- **Dataset Link:** https://www.kaggle.com/datasets/lineizumi/it159-final-checkpoints

## 4.2. Experimental Results

The three models were trained and evaluated on the same "Golden Data Split" (60% source, 40% target) to ensure reproducibility. The primary metric for comparison is the overall Top-1 Accuracy on the target domain.

### 4.2.1. Summary of Performance

The final performance of the three models on the target set is summarized in the table below. The results reveal a surprising outcome: the strong supervised baseline significantly outperformed both SSL-based approaches.

| Model | Pre-training Strategy | Fine-tuning Method | Final Target Accuracy (%) |
|---|---|---|---|
| **Supervised Baseline** | ImageNet Supervised | Supervised FT | **83.33%** |
| **SSDA (Proposed)** | ImageNet -> SimCLR | DANN | **75.67%** |
| **SSL-FT** | ImageNet -> SimCLR | Supervised FT | **70.73%** |

*Table 4: Summary of Final Model Performance on the Target Domain.*

### 4.2.2. Detailed Analysis of the Supervised Baseline

The baseline model, fine-tuned directly from ImageNet weights, achieved the highest accuracy of **83.33%**. The detailed performance per class is shown in its classification report and confusion matrix below. The model demonstrates strong and balanced performance across all seven classes, with particularly high F1-scores for the majority class ***nv (0.9139)*** and the minority class ***vasc (0.8947)***.

```
==================================================
BASELINE MODEL RESULTS
   - Final Accuracy on Target: 83.33%
==================================================

Classification Report:
               precision    recall  f1-score   support

       akiec      0.6320    0.6031    0.6172       131
         bcc      0.7354    0.7961    0.7646       206
         bkl      0.7095    0.6773    0.6930       440
          df      0.7222    0.5652    0.6341        46
         mel      0.5767    0.6323    0.6032       446
          nv      0.9180    0.9098    0.9139      2682
        vasc      0.8947    0.8947    0.8947        57

    accuracy                          0.8333      4008
   macro avg      0.7412    0.7255    0.7315      4008
weighted avg      0.8358    0.8333    0.8342      4008
```

*Figure 4.1a: Classification Report for the Supervised Baseline Model.*

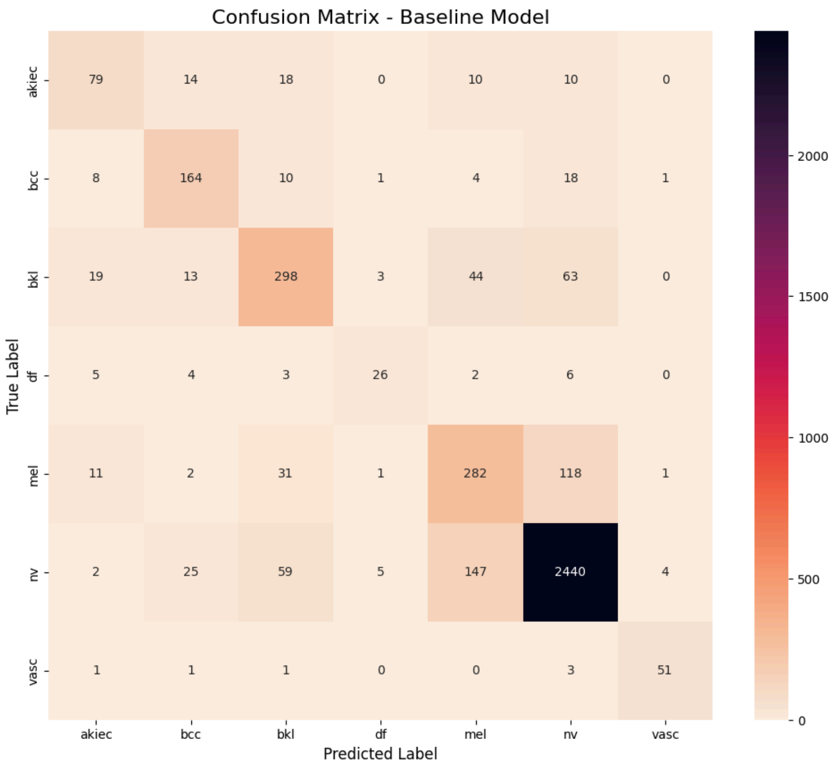


*Figure 4.1b: Confusion Matrix for the Supervised Baseline Model.*

### 4.2.3. Detailed Analysis of the SSDA Model (Proposed Method)

Our main proposed SSDA framework achieved a final accuracy of **75.67%**. While lower than the baseline, the training logs confirmed that the domain adaptation mechanism was technically successful, with the domain loss (*L_dom*) consistently stabilizing around the theoretical optimum of **0.63**. This indicates that the feature extractor was effectively learning domain-invariant features.

The classification report shows that while the model struggled with some classes like ***df*** and ***mel***, it maintained a very high recall for the majority class, ***nv (0.9113)***.

```
==================================================
SSDA MODEL RESULTS
   - Final Accuracy on Target: 75.67%
==================================================

Classification Report:
               precision    recall  f1-score   support

       akiec      0.3800    0.4351    0.4057       131
         bcc      0.6087    0.6796    0.6422       206
         bkl      0.4653    0.5182    0.4903       440
          df      0.3043    0.1522    0.2029        46
         mel      0.4565    0.3296    0.3828       446
          nv      0.8807    0.9113    0.8957      2682
        vasc      0.5556    0.1754    0.2667        57

    accuracy                          0.7567      4008
   macro avg      0.5216    0.4573    0.4695      4008
weighted avg      0.7463    0.7567    0.7482      4008
```

***Figure 4.2a: Classification Report for the SSDA Model.***

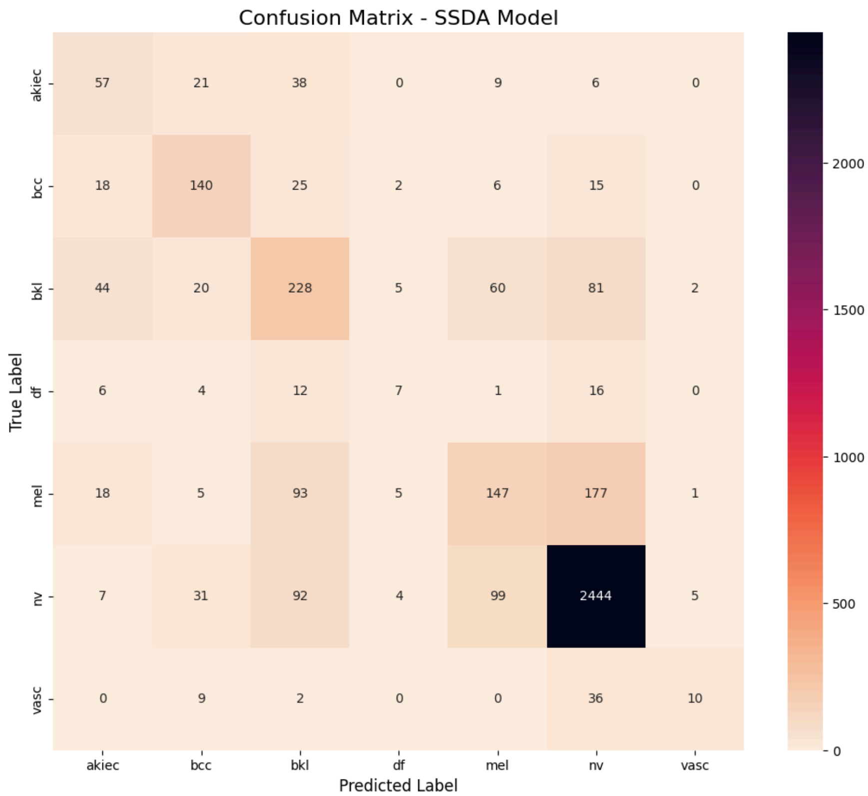


*Figure 4.2b: Confusion Matrix for the SSDA Model.*

### 4.2.4. Detailed Analysis of the SSL-FT Model (Ablation Study)

The SSL-FT model, which isolated the effect of SimCLR pre-training without domain adaptation, achieved the lowest accuracy of **70.73%**. This result is crucial as it allows us to analyze the individual contributions of the SSL and DA components. Comparing this to the SSDA model's performance (**75.67%**) suggests that the DANN component provided a significant performance uplift of nearly **5%**.

```
============================================================
PROPOSED METHOD (SSL-FT) RESULTS
   - Final Accuracy on Target: 70.73%
============================================================

Classification Report:
               precision    recall  f1-score   support

       akiec     0.4247    0.7099    0.5314       131
         bcc     0.5764    0.8058    0.6721       206
         bkl     0.5050    0.5750    0.5377       440
          df     0.3810    0.6957    0.4923        46
         mel     0.3673    0.6390    0.4664       446
          nv     0.9583    0.7289    0.8280      2682
        vasc     0.5100    0.8947    0.6497        57

    accuracy                         0.7073      4008
   macro avg     0.5318    0.7213    0.5968      4008
weighted avg     0.7927    0.7073    0.7318      4008
```

*Figure 4.3a: Classification Report for the SSL-FT Model.*

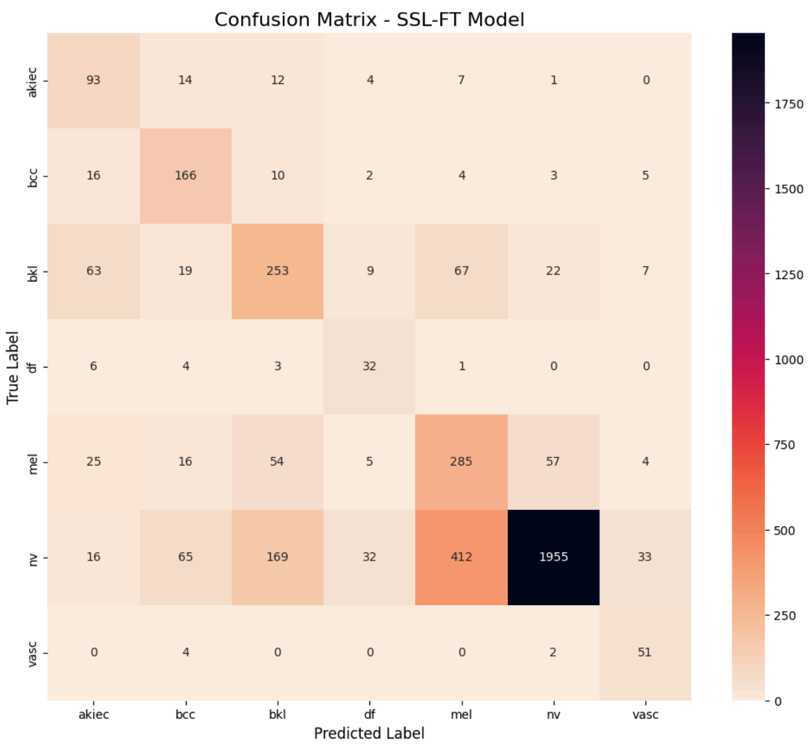


*Figure 4.3b: Confusion Matrix for the SSL-FT Model.*

# 5. Evaluation and Conclusion

This section provides a detailed evaluation of the experimental results presented in ***Section 4***. We analyze the comparative performance of the three models, discuss the key findings, and conclude with the main takeaways and potential directions for future work.

## 5.1. Evaluation and Discussion

The primary objective of this project was to evaluate the effectiveness of a Self-Supervised Domain Adaptation (SSDA) framework against a strong supervised baseline. The results were surprising.

### 5.1.1. Finding 1: The Limited but Evident Impact of Domain Adaptation

A direct comparison between our main proposed method, SSDA (**75.67% accuracy**), and the ablation study model, SSL-FT (**70.73% accuracy**), reveals a notable performance increase of nearly **5%**. The only difference between these two models is the inclusion of the DANN component in the SSDA framework. This suggests a significant domain shift between our source and target data. We also found that the DANN architecture was **technically successful** in its primary objective: aligning the feature distributions and mitigating the negative effects of this domain shift, thereby improving the model's generalization performance. The training logs, which showed the domain loss stabilizing around the theoretical optimum of **0.63**, further substantiate this conclusion.

### 5.1.2. Finding 2: The Unexpected Superiority of the Supervised Baseline

The most critical finding of our research is the superior performance of the supervised baseline. Achieving a final accuracy of **83.33%**, it outperformed both the SSL-FT (**70.73%**) and the full SSDA framework (**75.67%**). This indicates that despite the technical success of our SSL and DA implementations, the foundational features learned through standard transfer learning from ImageNet were ultimately more effective for this specific task. We propose two primary hypotheses to explain this phenomenon:

- **First, regarding feature robustness,** the ResNet-50 model pre-trained on the 1.2 million images of the ImageNet dataset has learned an incredibly rich and general-purpose set of visual features. For classifying dermatological images, which are still natural, this powerful, generalist foundation proved to be more robust than the specialist features developed through SSL on a smaller dataset.

- **Second, regarding the adaptation process,** we hypothesize that the process of fine-tuning the ImageNet-pretrained encoder with SimCLR on the relatively small HAM10000 dataset (**10,015 images**) may have led to a form of "catastrophic forgetting." In its effort to specialize in skin lesion features, the model may have inadvertently discarded some of the powerful, general features learned from ImageNet that were crucial for distinguishing between visually similar classes like ***bcc*** and ***nv***.

This shows an important lesson: advanced, complex architectures do not always guarantee superior performance. A well-optimized, simpler baseline can be remarkably difficult to surpass, especially when it is built upon a strong pre-trained foundation like ImageNet.

## 5.2. Conclusion

This project successfully developed and rigorously evaluated a Self-Supervised Domain Adaptation (SSDA) framework for skin lesion classification. We showed that SSL and DA worked individually, but putting them together didn't beat the baseline.

The key conclusion of this study is that for the HAM10000 dataset, the quality and generalizability of features obtained from large-scale pre-training on ImageNet are the most dominant factors for achieving high classification accuracy. The benefits of in-domain SSL pre-training on a small dataset were outweighed by the potential loss of this powerful initial feature representation. Our work highlights the importance of establishing strong baselines and provides a valuable empirical data point on the limitations of applying complex SSL/DA techniques in data-constrained medical imaging scenarios.

## 5.3. Future Work

Based on our findings, several promising directions for future research are identified:

1. **Exploring Higher Image Resolution:** As demonstrated by Azizi et al. (2021), increasing the image resolution from 224x224 to 448x448, and batch size from 32 to 64 or more could provide the model with more fine-grained details, potentially improving its ability to distinguish between similar lesions.

2. **Investigating Alternative Backbones:** Employing more modern and powerful architectures, such as EfficientNet or Vision Transformers (ViT), could lead to the extraction of more discriminative features.
3. **Applying Pseudo-Labeling:** A promising semi-supervised technique would be to use the strong baseline model to generate high-confidence "pseudo-labels" for the target set, and then retrain the model on the combined labeled data.

# References


- Ahmed, S., et al. (2023). *Human Skin Diseases Detection and Classification using CNN.*

- Azizi, S., et al. (2021). *Big Self-Supervised Models Advance Medical Image Classification. CVPR.*

- Ben-David, S., Blitzer, J., Crammer, K., & Pereira, F. (2010). *A theory of learning from different domains. Machine Learning, 79*(1-2), 151–175.

- Berthelot, D., et al. (2021). *Semi-Supervised Models are Strong Unsupervised Domain Adaptation Learners. NeurIPS.*

- Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). *A Simple Framework for Contrastive Learning of Visual Representations (SimCLR). ICML.*

- Ganin, Y., & Lempitsky, V. (2015). *Unsupervised Domain Adaptation by Backpropagation (DANN). ICML.*

- He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). *Momentum Contrast for Unsupervised Visual Representation Learning (MoCo). CVPR.*

- Li, J., et al. (2021). *Speech SimCLR: Combining Contrastive and Reconstruction Objectives for Self-Supervised Speech Representation Learning.*

- Yadav, R., & Bhat, P. (2024). *A Systematic Literature Survey on Skin Disease Detection and Classification using Machine Learning and Deep Learning.*