SKIN DISEASE DETECTION USING DEEP LEARNING

Jacky Li
Student# 1011271678
jakkii.li@mail.utoronto.ca

Jordan Cui
Student# 1011026916
jordan.cui@mail.utoronto.ca

Lawrence Ding
Student# 1011439025
larryzm.ding@mail.utoronto.ca

Soham Shorey
Student# 1010845169
soham.shorey@mail.utoronto.ca

ABSTRACT

This project presents a deep-learning based approach for multi-class classification of dermatological conditions, including melanoma, actinic keratosis, and seven other forms of benign and malignant lesions by analyzing skin images. Early and accurate detection of skin diseases can significantly improve patient outcomes, particularly for malignant conditions such as melanoma. A combined dataset of 24,738 images from four public sources (BCN20000 (International Skin Imaging Collaboration (ISIC), 2018), HAM10000 (Tschandl et al., 2018), PAD-UFES-20 (Mahdavi, 2022), and DERM12345 (Yilmaz et al., 2024)) was constructed. Labels were standardized through preprocessing, classes were balanced through undersampling and data augmentation, and all images were resized to 512×512.

The best-performing model uses a pretrained Swin Transformer Base backbone followed by a two-layer fully connected classifier. Class imbalance was addressed with a combination of oversampling, WeightedRandomSampler, Mixup/CutMix augmentations, and Class-Balanced Focal Loss with medical importance weighting. A baseline model using ResNet18 feature extraction and KNN classifier achieved 56.3% accuracy.

The proposed model achieved 89.1% test accuracy, 91.3% precision, 88.5% recall, and 89.8% macro F1-score, substantially outperforming the baseline. These results demonstrate the potential of transformer-based architectures as a supportive diagnostic tool. Future work includes further improving minority class performance, incorporating additional data sources, and validating the model on external datasets to assess clinical applicability. —-Total Pages: 9

1 INTRODUCTION

This report outlines a deep learning model to classify dermatological conditions, including melanoma, keratosis, basal cell carcinoma, and actinic keratosis from skin images. This model allows patients to detect dangerous skin conditions like melanoma early and assists healthcare professionals in their diagnosis.

Skin conditions are among the most widespread and varied health conditions, affecting over one-third of the global population (Li et al., 2024). In Canada, the incidence of skin conditions like melanoma has increased by more than 2% annually since the 1980s (Canadian Dermatology Association, 2025). This rise strains healthcare systems worldwide. In 2013 alone, 84 million Americans sought medical attention for a skin-related issue (American Academy of Dermatology, 2025). Our skin classifier could alleviate this burden on healthcare infrastructure by encouraging patients to seek medical attention before their condition deteriorates and by assisting in diagnosis.

Deep learning models, such as convolutional neural networks (CNNs) and vision transformers (ViTs), are highly effective for skin disease classification from images because of their ability to extract spatial patterns. CNN-based and transformer skin disease classifiers have already been implemented (see Section 3) with promising results for physicians and patients.

2 Illustration/Figure

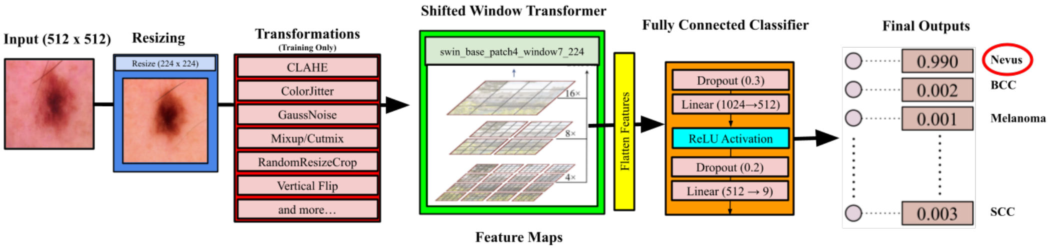


Figure 1: Overview of our model architecture using a Swin Transformer. The model processes image patches through shifted window attention mechanisms for hierarchical feature learning.

3 Background/Related Work

Machine learning models for skin disease detection have been widely studied and are now available to consumers through websites and mobile applications.

Several studies focus on detecting skin cancer from images. For example, a 2017 study compared the performance of two convolutional neural network (CNN) models trained on over 120,000 images to distinguish between keratinocyte carcinomas and benign seborrheic keratoses, as well as malignant melanomas and benign nevi with the diagnostic accuracy of 21 board-certified dermatologists (Esteva et al., 2017). Unlike earlier models limited to standardized dermoscopy and histology images, this model was trained to handle dermoscopy and smartphone-quality images, improving usability for everyday users.

Another study used the Microsoft ResNet-152 CNN model, trained on 19,398 images, to assess the probability of 12 distinct skin conditions and to differentiate between benign and malignant cases. The model achieved AUC scores ranging from 0.82 to 0.96, comparable to healthcare professionals (Han et al., 2018). Variations of ResNet have been implemented frequently in studies. For example, a 2025 study uses a self-reliant residual network (SR-ResNet) to identify melanoma (Radhika et al., 2025). It uses Zoutendijk's Method, a nonlinear optimizer that selects optimal step size and direction to stabilize convergence and prevent gradient vanishing and overfitting. The model achieved an accuracy of 94.79%, precision of 93.09%, and recall of 96.62%, outperforming traditional deep learning models.

Beyond skin cancer, studies have used models to classify a wider range of skin diseases. For instance, a 2019 study evaluated five different models for classifying three types of skin diseases, with CNNs achieving high 90s precision and recall scores (Bhadula et al., 2019). These studies underscore the utility of machine learning, particularly CNNs, in skin disease detection.

Machine learning models that detect skin conditions are readily available on websites and mobile applications. DermascanAI uses an EfficientNetB0-based network, predicting nine common skin conditions with 87.2% accuracy (Vadivelraju et al., 2025). SkinVision detected premalignant and malignant skin lesions with 86.9% sensitivity and 70.4% specificity (Sangers et al., 2022).

These studies and existing successful models helped inform our choices of model architecture.

4 Data Processing

We combined and preprocessed four public skin lesion datasets to create a dataset that covers various skin conditions, skin types, and imaging methods.

4.1 Data Sources

Our data was sourced from four datasets: BCN20000 (International Skin Imaging Collaboration (ISIC), 2018), HAM10000 (Tschandl et al., 2018), PAD-UFES-20 (Mahdavi, 2022), and DERM12345 (Yilmaz et al., 2024).

Table 1: Statistics on Data Sources. Datasets contain different image sizes, class counts, and formats.

Dataset	**Classes**	**Images**	**Size**	**Format/Extraction Process**	**Source(s)**
BCN20000	8	18946	256×256	Extracted BCN20000 images from ISIC archive using metadata CSV file	ISIC archive
HAM10000	7	10015	450×600	Downloaded from Kaggle and stored in 2 separate folders which were combined	Medical University of Vienna, Austria, & the skin cancer practice of Cliff Rosendahl in Queensland
PAD-UFES-20	6	2298	Varies	Downloaded from Kaggle and stored in 3 folders which were combined	Federal University of Espírito Santo (UFES) in Brazil
DERM12345	40 (subclasses)	12345	256×256	Extracted DERM12345 images from ISIC archive using CSV file	ISIC archive

4.2 Data Cleaning & Preprocessing

The unification of these four datasets required several preprocessing steps.

The BCN20000 (International Skin Imaging Collaboration (ISIC), 2018) and DERM12345 (Yilmaz et al., 2024) datasets were not publicly separated from the broader ISIC archive, so we used their respective CSV files to isolate and download their labelled images. The HAM10000 (Tschandl et al., 2018) and PAD-UFES-20 (Mahdavi, 2022) datasets were downloaded from Kaggle, with their respective folders being organized into unified directories.

Data labels were extracted from the diagnosis_3 field in BCN2000 and DERM12345, the dx field in HAM10000, and the diagnostic field in the PAD-UFES-20 CSV. We standardized all labels into nine unified classes: nevus, melanoma, actinic (precancerous) keratosis, basal cell carcinoma (BCC), squamous cell carcinoma (SCC), lentigo, vascular lesion, dermatofibroma, and (benign) keratosis. The images labelled under each class were placed into corresponding folders using a label mapping system. Nevus images were undersampled to prevent severe dataset imbalance and other labels were ignored. Table 2 shows the size of our dataset after preprocessing.

Table 2: Dataset size after preprocessing.

Class	**Nevus**	**Melanoma**	**BCC**	**Keratosis**	**Actinic Keratosis**
Count	7500	6149	4613	2982	1473
Class	**SCC**	**Dermatofibroma**	**Lentigo**	**Vascula Lesions**	
Count	862	463	374	322	
Total: 24738					

Before consolidation, we also resized all images to 512×512 using the PIL.Image library for image size flexibility in different models. For instance, our final model inputs 224×224 images.

The final dataset contains 9 classes of 512×512 images in different formats, including dermatoscopic imaging and phone photos, ensuring the robustness of our model to different types of data.

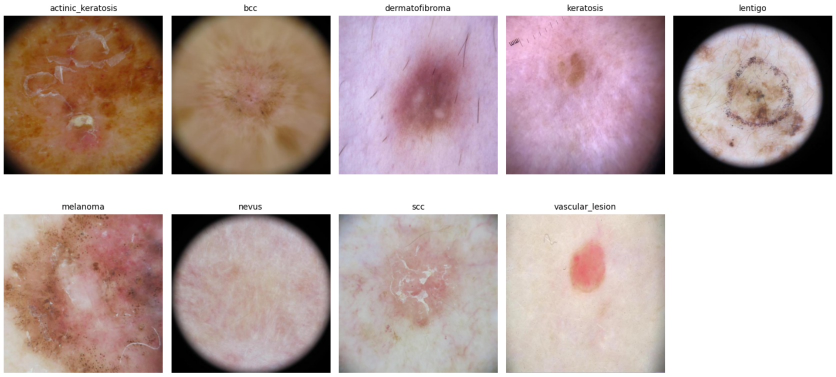


Figure 2: Sample resized and labelled images from each of the 9 classes in the combined dataset.

We split the processed dataset into 70% training, 15% validation, and 15% test sets to ensure robust model training and evaluation. Each split maintains the same class distribution as visualized in Figure 3. However, the dataset remains imbalanced, with the *nevus* and *melanoma* classes accounting for 55% of the data.

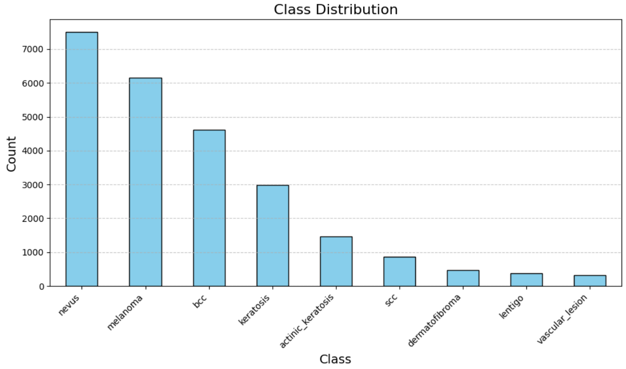


Figure 3: Class Distribution of Combined Dataset

5 Architecture

An illustration of the model architecture is shown in Figure 1.

5.1 Model Architecture

Our best model uses the Swin Transformer Base backbone (swin_base_patch4_window7_224) pretrained on ImageNet. A vision transformer model was chosen due to its higher modeling capacity and lower inductive biases than CNNs. This specific vision transformer processes 4x4 non-overlapping image patches through multi-head self-attention within 7x7 shifted windows. These patches are merged to form multi-scale feature maps. The Swin backbones' pooled feature vector is passed to a 2-layer fully connected linear classifier head with ReLU activation and dropout layers to prevent overfitting. The hidden layer has 512 units and the final layer has 9 units for the 9 skin disease classes. The combined model is extremely large compared to CNN counterparts, with almost 90 million parameters.

5.2 Training Techniques

This class imbalance was addressed using various data sampling and augmentation techniques. During training, we applied 29 different image transformations, including ColorJitter, RandomResizedCrop, Vertical Flip and CLAHE (a dermatology focused contrast normalization) with more transformations applied to minority and challenging classes. Additionally, we used WeightedRandomSampler and oversampling of minority classes to balance the classes seen in training and Mixup/CutMix for 80% of the training epochs which blend both images and labels together to improve performance metrics and smooth decision boundaries. These techniques boosted the model's metrics and robustness by addressing the class imbalance.

During the Mixup/CutMix epochs, the model uses soft-target cross-entropy loss, which directly compares the predicted probability distribution to the soft labels from the augmentation (alpha_mixup = 0.3, alpha_cutmix = 1.0, label_smoothing =0.1). After 80% of epochs are completed, the model uses Class-Balanced Focal Loss (γ = 2.0) to reward classifying difficult samples. The class weights inside this loss are computed from the effective number of samples formula (β = 0.9999) and are further scaled by medical importance factors, which give higher priority to more dangerous conditions shown in Table 3.

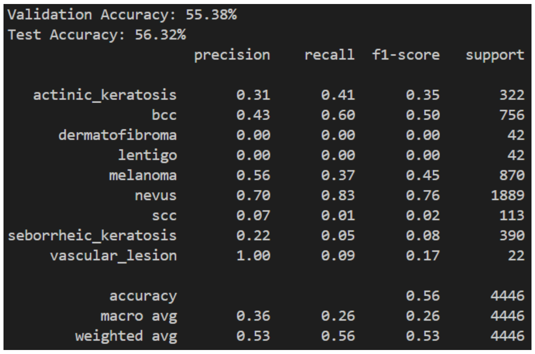
The model was trained for 200 epochs with a batch size of 16 using the AdamW optimizer with a learning rate of 1e-4, weight decay of 1e-4, betas of 0.9 and 0.999. The learning rate schedule combined 3 stages: a linear warmup for the first 5% of epochs, followed by cosine annealing with warm restarts (restart period $\approx$ 12.5% of remaining epochs, ReduceLROnPlateau stage triggered by stagnation in macro F1 over 8 epochs. Early stopping was applied if macro F1 did not improve for 35 epochs.

Table 3: Medical Importance Factors per Class

Class Name	Medical Importance Weight	Notes
nevus	1.0	Most common, baseline weight
melanoma	3.0	Life-threatening malignancy
bcc	2.0	Malignant but less aggressive
keratosis	2.2	Challenging to distinguish from actinic keratosis
actinic_keratosis	2.5	Pre-cancerous lesion
scc	2.5	Malignant, clinically important
dermatofibroma	1.8	Benign, important for differential diagnosis
lentigo	2.3	Age-related, challenging morphology
vascular_lesion	1.8	Specific diagnosis, moderate importance

6 Baseline Model

For the baseline model, a hybrid approach combining pre-trained feature extraction with a simple classifier was selected. Direct use of simple algorithms such as support vector machines (SVM) or k-nearest neighbors (KNN) on raw images was deemed unsuitable as a baseline due to the complexity of skin disease classification and the high-dimensional nature of image data, which would likely result in near-chance performance. Instead, a pre-trained ResNet18 model was employed for feature extraction. All input images were rescaled from 512 × 512 to 224 × 224 pixels to match the ResNet18 input requirements. The model's fully connected layers were removed, retaining only the convolutional and average pooling layers, which output a fixed-length feature vector for each image. These feature vectors were then used as input to a KNN classifier implemented in scikit-learn with the n_neighbors parameter set to 8. This baseline achieved a test accuracy of 56.32%, surpassing the random-chance baseline of 11.11%. Although the resulting F1-scores indicate limited suitability for real-world application, the approach provides a reproducible and reasonable benchmark for evaluating the performance of more complex neural network architectures.



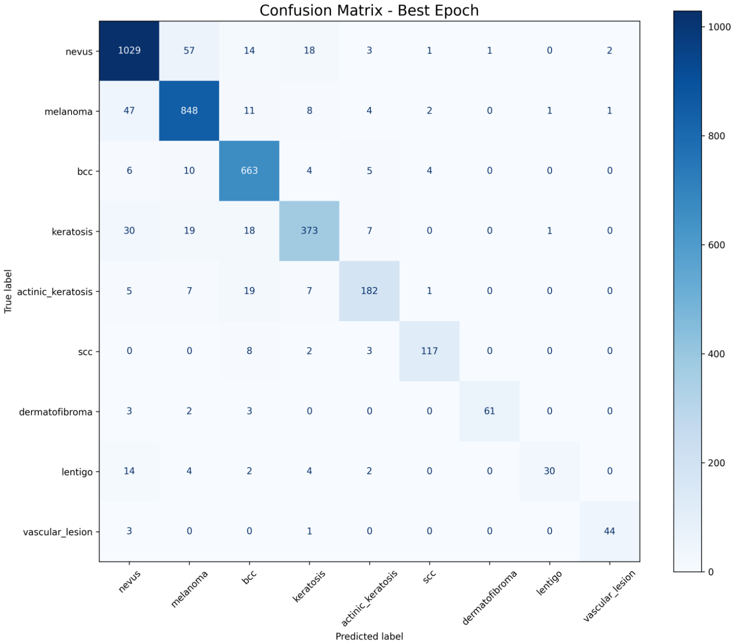
```
Validation Accuracy: 55.38%
Test Accuracy: 56.32%
                     precision    recall  f1-score   support

   actinic_keratosis       0.31      0.41      0.35       322
                 bcc       0.43      0.60      0.50       756
      dermatofibroma       0.00      0.00      0.00        42
             lentigo       0.00      0.00      0.00        42
            melanoma       0.56      0.37      0.45       870
               nevus       0.70      0.83      0.76      1889
                 scc       0.07      0.01      0.02       113
seborrheic_keratosis       0.22      0.05      0.08       390
     vascular_lesion       1.00      0.09      0.17        22

            accuracy                           0.56      4446
           macro avg       0.36      0.26      0.26      4446
        weighted avg       0.53      0.56      0.53      4446
```

Figure 4: Baseline model performance metrics. This model struggles with the imbalanced dataset, with higher F1-scores for the more common classes, but extremely low for the less common classes.

7 Quantitative and Qualitative Results

7.1 Quantitative Results

Our final Swin Transformer Base skin detection model achieved a best validation accuracy of 90.2%, F1 of 88.0%, precision of 92.0% and recall of 85.4% on the 9 classes. Training metrics and losses are not directly comparable to validation results due to the use of a WeightedRandomSampler and extensive data augmentation, which altered the training distribution.



	nevus	melanoma	bcc	keratosis	actinic_keratosis	scc	dermatofibroma	lentigo	vascular_lesion
nevus	1029	57	14	18	3	1	1	0	2
melanoma	47	848	11	8	4	2	0	1	1
bcc	6	10	663	4	5	4	0	0	0
keratosis	30	19	18	373	7	0	0	1	0
actinic_keratosis	5	7	19	7	182	1	0	0	0
scc	0	0	8	2	3	117	0	0	0
dermatofibroma	3	2	3	0	0	0	61	0	0
lentigo	14	4	2	4	2	0	0	30	0
vascular_lesion	3	0	0	1	0	0	0	0	44

Figure 5: Confusion matrix of our best model checkpoint evaluated on the validation data set.

The final model was trained over 170 epochs with frequent checkpoints. Additional augmentations and loss function adjustments were introduced around the 60th epoch, causing the training loss spikes seen in Figure 6 but improving validation performance. The loss function started as soft cross-entropy with heavy augmentation for 80% of epochs, then switched to class-balanced focal loss with reduced augmentation for convergence.

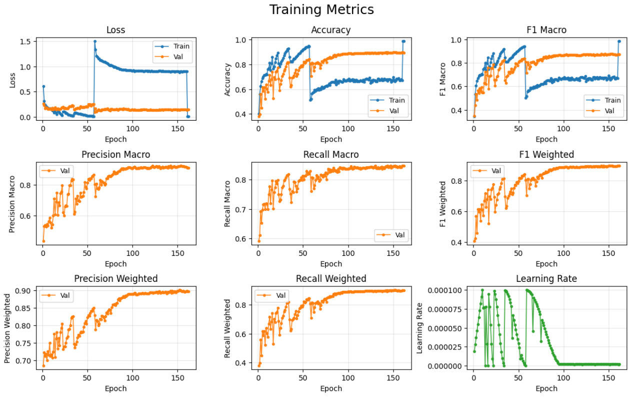
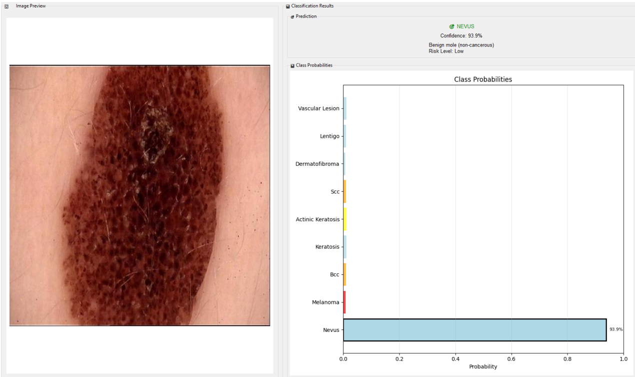


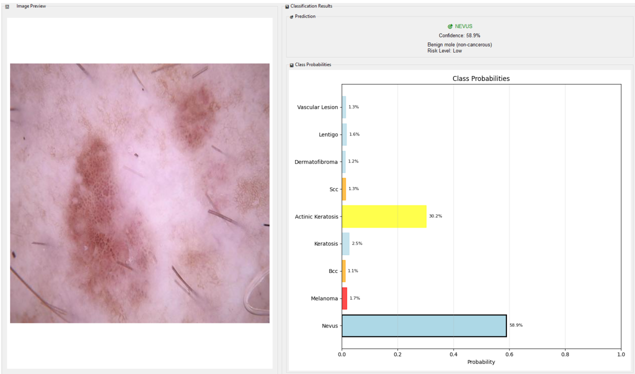
Figure 6: Training Metrics. From the top left graphs show loss, accuracy, macro F1, macro precision, macro recall, weighted f1, weighted precision, weighted recall, and learning rate. Horizontal axis shows epochs, with a total of 170. Blue represents train set while orange represents validation.

7.2 Qualitative Results



NEVUS
Confidence: 93.9%
Benign mole (non-cancerous)
Risk Level: Low

Figure 7: Model output on image with true label nevus, a majority class. The model correctly predicts nevus with 93.9% confidence, demonstrating strong performance on well-represented classes.



NEVUS
Confidence: 58.9%
Benign mole (non-cancerous)
Risk Level: Low

Figure 8: Model output on image with true label actinic keratosis, a minority class. The model incorrectly predicts nevus with 58.9% confidence, reflecting uncertainty on rare conditions.

We selected Figure 7 and Figure 8 to illustrate how class imbalance affects model performance. Figure 7 shows strong performance on nevus (30.4% of training data, 7,500 samples) with 93.9% confidence, aligning with our 88.6% precision results. Conversely, Figure 8 shows lower confidence (58.9%) and incorrect classification for actinic keratosis (only 1,473 samples), contributing to the 79.2% recall for this class. The model appropriately exhibits uncertainty rather than overconfident predictions when training data is limited.

8 Evaluation on Unseen Data

The test dataset created in the data preprocessing stage was strictly separated from training and validation, ensuring unbiased evaluation of model generalization. No test images were used for hyperparameter tuning or architecture decisions. The model achieved 89.1% accuracy, 91.3% precision, 88.4% recall, and 89.77% macro F1 score on test data. These metrics are slightly lower than validation results, indicating minimal overfitting. Our regularization techniques successfully improved generalization, demonstrating that the model effectively distinguishes between nine skin conditions better than our baseline ResNet18-KNN approach (56.3% accuracy).

Table 4: Per Class Precision, Recall, and F1 scores.

Metric	**nevus**	**melanoma**	**bcc**	**keratosis**	**actinic_keratosis**	**scc**	**dermatofibroma**	**lentigo**	**vascular_lesion**
Precision	0.886	0.882	0.892	0.891	0.893	0.928	0.929	0.918	1.000
Recall	0.902	0.910	0.942	0.785	0.792	0.899	0.929	0.804	1.000
F1 Score	0.894	0.896	0.916	0.835	0.839	0.913	0.929	0.857	1.000

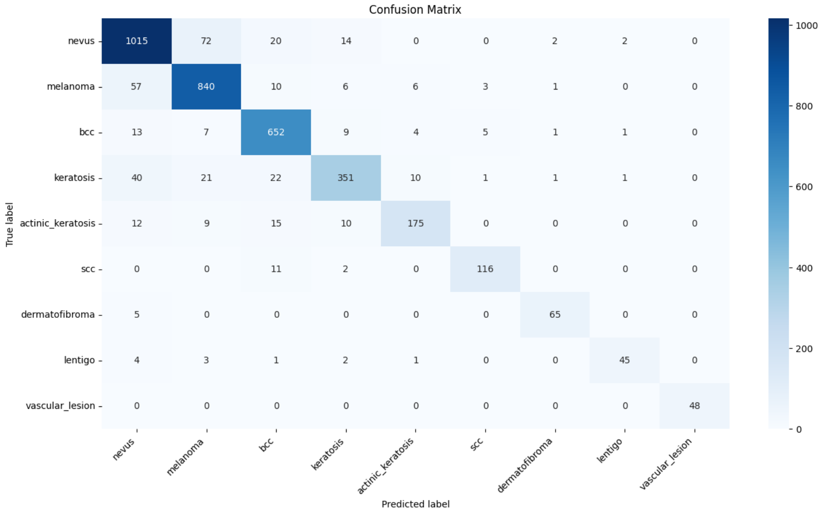


Figure 9: Confusion matrix of our best model checkpoint evaluated on our held out test data set.

9 Discussion

From a clinical perspective, our model's 89.1% test accuracy represents a promising advancement in automated skin disease detection. This performance compares favorably to the reported accuracy of 74.1% for dermatologists in similar classification tasks (Winkler et al., 2023). However, these comparisons must be interpreted carefully, as our model was evaluated on a specific set of nine skin conditions, whereas practicing dermatologists routinely diagnose hundreds of distinct dermatological conditions in clinical practice.

The model's ability to process both high-quality dermatoscopic images and smartphone-captured photos represents a significant advantage for patient accessibility. Many existing systems require specialized dermoscopy equipment, limiting their applicability in primary care settings or for patient self-screening. Our approach democratizes access to preliminary skin disease assessment, potentially enabling earlier detection and treatment initiation.

Despite achieving high overall accuracy, the model's performance varies considerably across different disease classes. The confusion matrix reveals that certain conditions, particularly those with lower representation in the training data, remain challenging to classify accurately. For instance, while the model achieves excellent performance on common conditions like nevus and melanoma, it struggles more with rare conditions such as vascular lesions and dermatofibroma. This performance disparity reflects the inherent challenge of learning from imbalanced datasets, even with the extensive class balancing techniques we employed.

The medical importance weighting system we implemented represents a novel contribution to the field. By assigning higher penalties to misclassifications of more clinically significant conditions (such as melanoma and squamous cell carcinoma), we aligned the model's learning objectives with clinical priorities. This approach resulted in improved detection rates for malignant conditions, though at the cost of slightly reduced overall accuracy for benign lesions.

Our data augmentation strategy, combining traditional geometric transformations with dermatology-specific techniques like CLAHE contrast enhancement, proved effective in improving model robustness. The use of Mixup and CutMix augmentation during training helped the model learn more generalizable features, as evidenced by the relatively small gap between validation and test performance. This suggests that our regularization techniques successfully prevented overfitting despite the model's large parameter count.

One significant limitation of our current approach is the absence of ”healthy skin” samples in the training data. This forces the model to always predict a disease condition, which could lead to false positives when applied to images of normal skin in real-world scenarios. Future iterations should incorporate negative samples to improve the model's ability to distinguish between diseased and healthy tissue. Additionally, the dataset does not include any information about patient demographics or skin type, limiting our ability to assess potential biases in model performance across different populations.

10 Ethical Considerations

Our skin disease classification model raises important ethical considerations regarding healthcare AI deployment. Training datasets are inherently biased toward lighter skin tones, reflecting healthcare disparities and unequal documentation in darker-skinned populations (Alipour et al., 2024). This bias could result in reduced model performance for patients with darker skin, potentially reinforcing existing healthcare inequalities. Additionally, false negatives in melanoma detection could delay critical treatment, while false positives may cause unnecessary patient anxiety. The model should serve as a diagnostic aid rather than a replacement for professional medical evaluation.

11 Conclusion

This project successfully developed a deep learning system for skin disease classification achieving 89.1% accuracy on nine dermatological conditions. The Swin Transformer-based architecture, combined with comprehensive data augmentation and class balancing techniques, demonstrated significant improvements over the baseline ResNet18-KNN approach. The model shows promise as a preliminary screening tool, particularly for early detection of malignant conditions like melanoma.

Key contributions include the unification of four major dermatological datasets, implementation of medical importance weighting in the loss function, and achievement of performance levels comparable to dermatologists. However, limitations remain, including class imbalance effects on minority classes, absence of healthy skin samples, and potential dataset bias toward certain skin types.

Future work should focus on incorporating healthy skin data, improving minority class performance through advanced augmentation techniques, and validating the model on external datasets from different populations and imaging conditions. Additionally, developing uncertainty quantification methods could enhance clinical applicability by providing confidence measures for predictions.

References

N. Alipour, T. Burke, and J. Courtney. Skin type diversity in skin lesion datasets: A review. *Current Dermatology Reports*, 13:198–210, August 2024. doi: 10.1007/s13671-024-00440-0.

American Academy of Dermatology. Burden of skin disease, 2025. URL `https://www.aad.org/member/clinical-quality/clinical-care/bsd`. Accessed: Jul. 9th, 2025.

S. Bhadula, S. Sharma, P. Juyal, and C. Kulshrestha. Machine learning algorithms based skin disease detection. *International Journal of Innovative Technology and Exploring Engineering*, 9(2):4044–4049, 2019.

Canadian Dermatology Association. Skin conditions, 2025. URL `https://dermatology.ca/public-patients/diseases-conditions/skin-conditions/`. Accessed: Jul. 9th, 2025.

A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118, 2017.

S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *Journal of Investigative Dermatology*, 138(7):1529–1538, 2018.

International Skin Imaging Collaboration (ISIC). Isic archive - collection 249, 2018. URL `https://api.isic-archive.com/collections/249/`.

Q. Li, M. T. Patrick, S. Sreeskandarajan, J. Kang, J. M. Kahlenberg, J. E. Gudjonsson, Z. He, and L. C. Tsoi. Large-scale epidemiological analysis of common skin diseases to identify shared and unique comorbidities and demographic factors. *Frontiers in Immunology*, 14, January 2024. URL `https://pmc.ncbi.nlm.nih.gov/articles/PMC10800546/`.

M. Mahdavi. Skin cancer dataset, 2022. URL `https://www.kaggle.com/datasets/mahdavi1202/skin-cancer`.

V. Radhika, A. Muthuchudar, and M. Lingaraj. Self-reliant residual network based deep learning framework for melanoma skin disease detection. *Journal of Theoretical and Applied Information Technology*, 103(10), May 2025.

T. Sangers, S. Reeder, S. van der Vet, S. Jhingoer, A. Mooyaart, D. Siegel, and T. Nijsten. Validation of a market-approved artificial intelligence mobile health app for skin cancer screening. *Dermatology*, 238(4):649–656, 2022.

P. Tschandl, C. Rosendahl, and H. Kittler. Ham10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions, 2018. URL `https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000`.

C. Vadivelraju, K. Bhargavi, P. Rajesh, K. Srinivas, and M. V. N. K. Prasad. Dermascan ai: Deep learning system for preliminary diagnosis of dermatological manifestations. *International Journal of Creative Research Thoughts*, 13(5):406–413, May 2025.

J. K. Winkler, C. Fink, F. Toberer, A. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, W. Stolz, and H. A. Haenssle. Assessment of diagnostic performance of dermatologists cooperating with a convolutional neural network in a prospective clinical study. *JAMA Dermatology*, 159(6):621–621, June 2023. doi: https://doi.org/10.1001/jamadermatol.2023.0905.

A. Yilmaz et al. Derm12345: A large, multisource dermatoscopic skin lesion dataset with 40 subclasses. *Scientific Data*, 11:1302, 2024. doi: 10.1038/s41597-024-04104-3.