Skin Disease Classification

ML Pipeline Report

Early Detection with Human-in-the-Loop Validation

Data Preprocessing & Feature Extraction Pipeline

Report Generated	2026-02-19 16:43:01
Pipeline Start	2026-02-19 16:37:37
Pipeline End	2026-02-19 16:43:01
Total Duration	0:05:23
Python Version	3.11.6
Platform	win32

1. Dataset Overview

Dataset 1 (Kaggle)	1494 images
Dataset 2 (Kaggle)	27153 images
Total Raw Images	28647 images
Number of Classes	15
Image Size (target)	128 x 128 pixels
Colour Space	RGB

Raw Class Distribution

Disease Class	Image Count
1. Eczema 1677	1677
10. Warts Molluscum and other Viral Infections - 2103	2103
2. Melanoma 15.75k	3140
3. Atopic Dermatitis - 1.25k	1257
4. Basal Cell Carcinoma (BCC) 3323	3323
5. Melanocytic Nevi (NV) - 7970	7970
6. Benign Keratosis-like Lesions (BKL) 2624	2079
7. Psoriasis pictures Lichen Planus and related diseases - 2k	2055
8. Seborrheic Keratoses and other Benign Tumors - 1.8k	1847
9. Tinea Ringworm Candidiasis and other Fungal Infections - 1.7k	1702
acne	300
eksim	297
herpes	300
panu	297
rosacea	300
TOTAL	28647

Imbalance Ratio (raw) 26.84 : 1

2. Exploratory Data Analysis

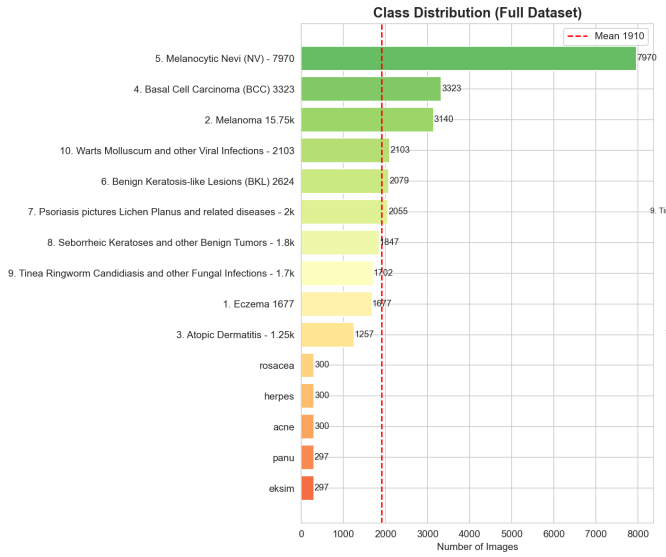
2.1 Data Quality Check

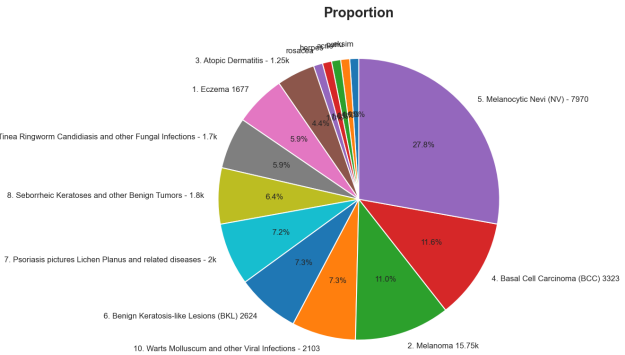
Corrupt / Unreadable	0
Duplicate Images	625

2.2 Image Dimensions

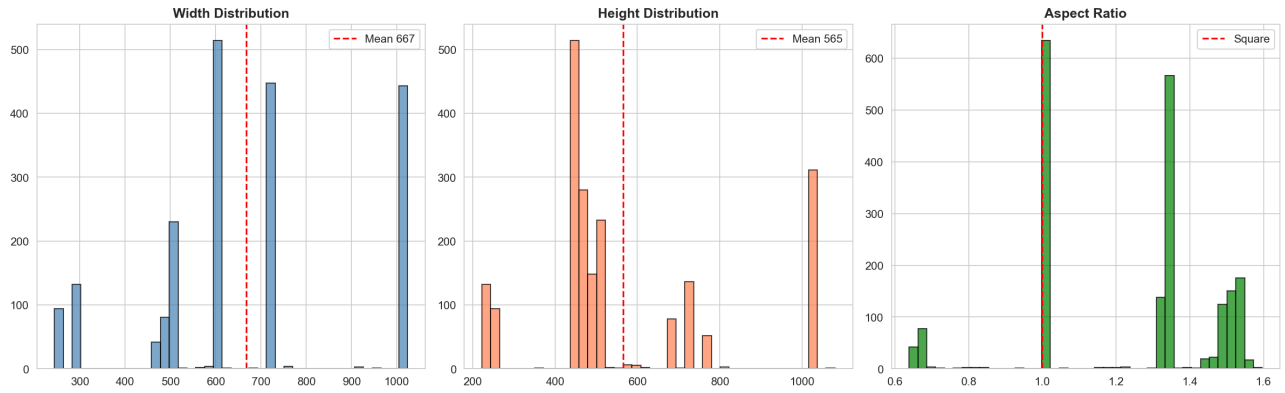
Width Range	244 - 1024 px (mean: 667)
Height Range	222 - 1080 px (mean: 565)

2.3 Class Distribution

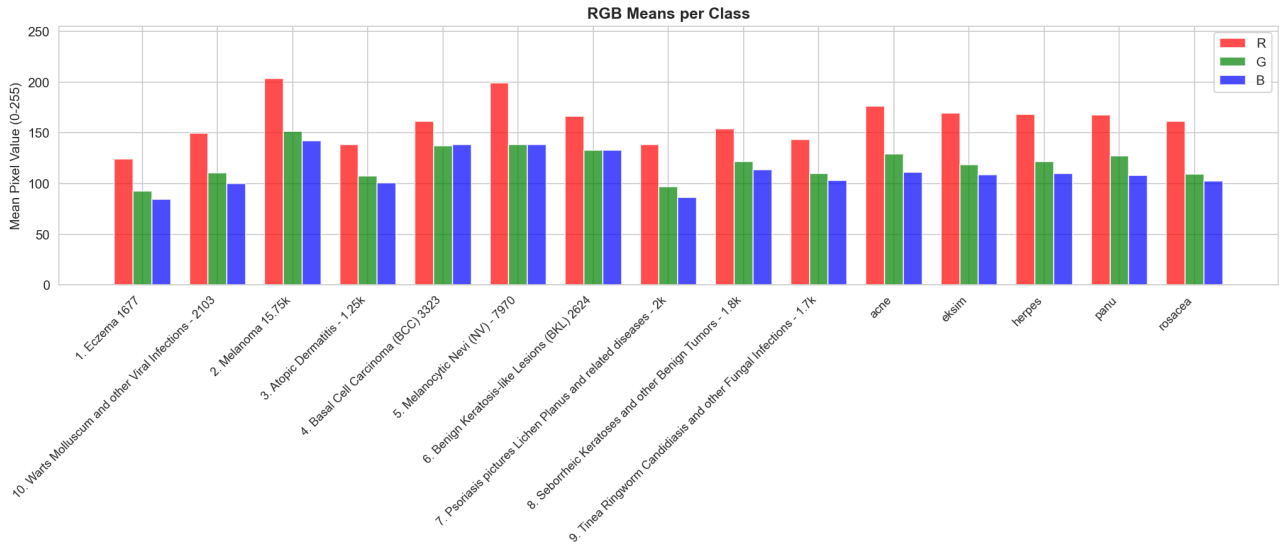




2.4 Image Dimension Analysis



2.5 Pixel Intensity per Class



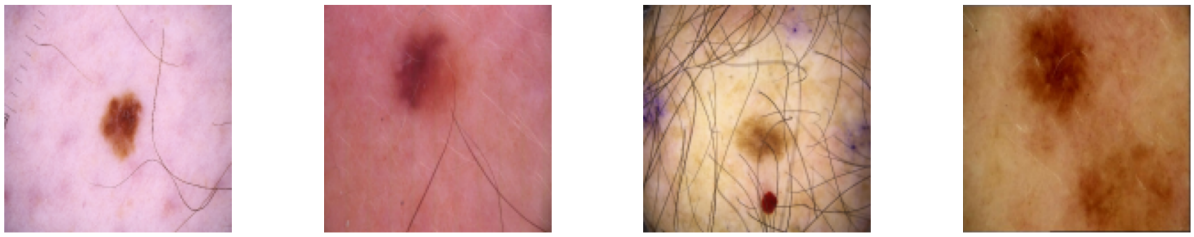
2.6 Sample Images per Class

Sample Images per Class

1. Eczema 1677
(1677)

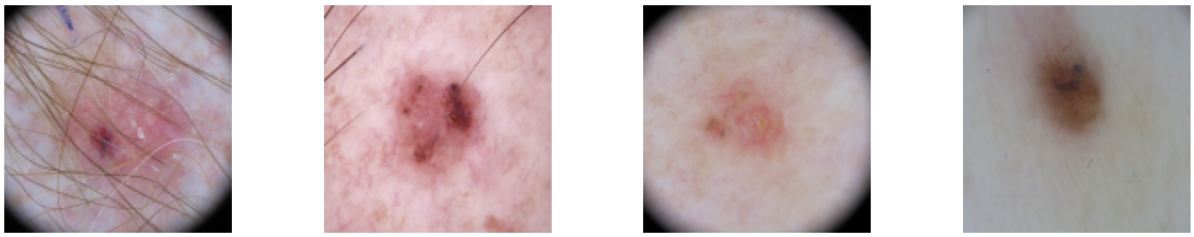
10. Warts Molluscum and other Viral Infections - 2103
(2103)

2. Melanoma 15.75k
(3140)

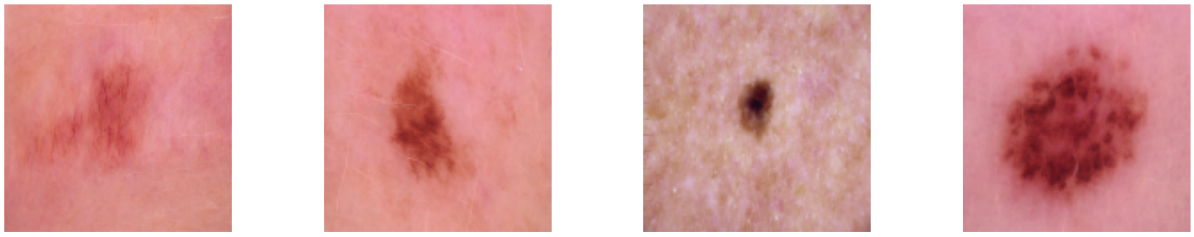


3. Atopic Dermatitis - 1.25k
(1257)

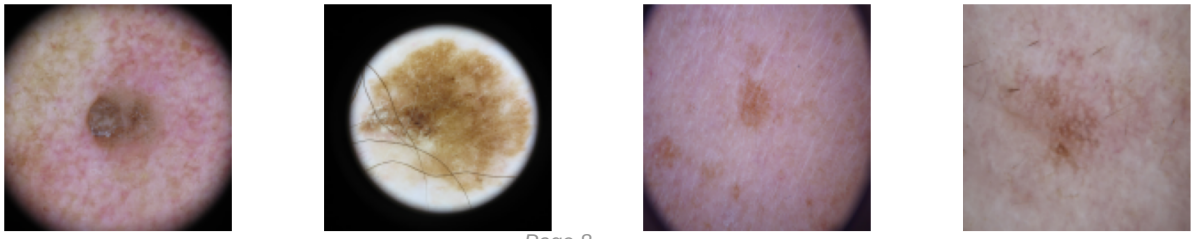
4. Basal Cell Carcinoma (BCC) 3323
(3323)



5. Melanocytic Nevi (NV) - 7970
(7970)

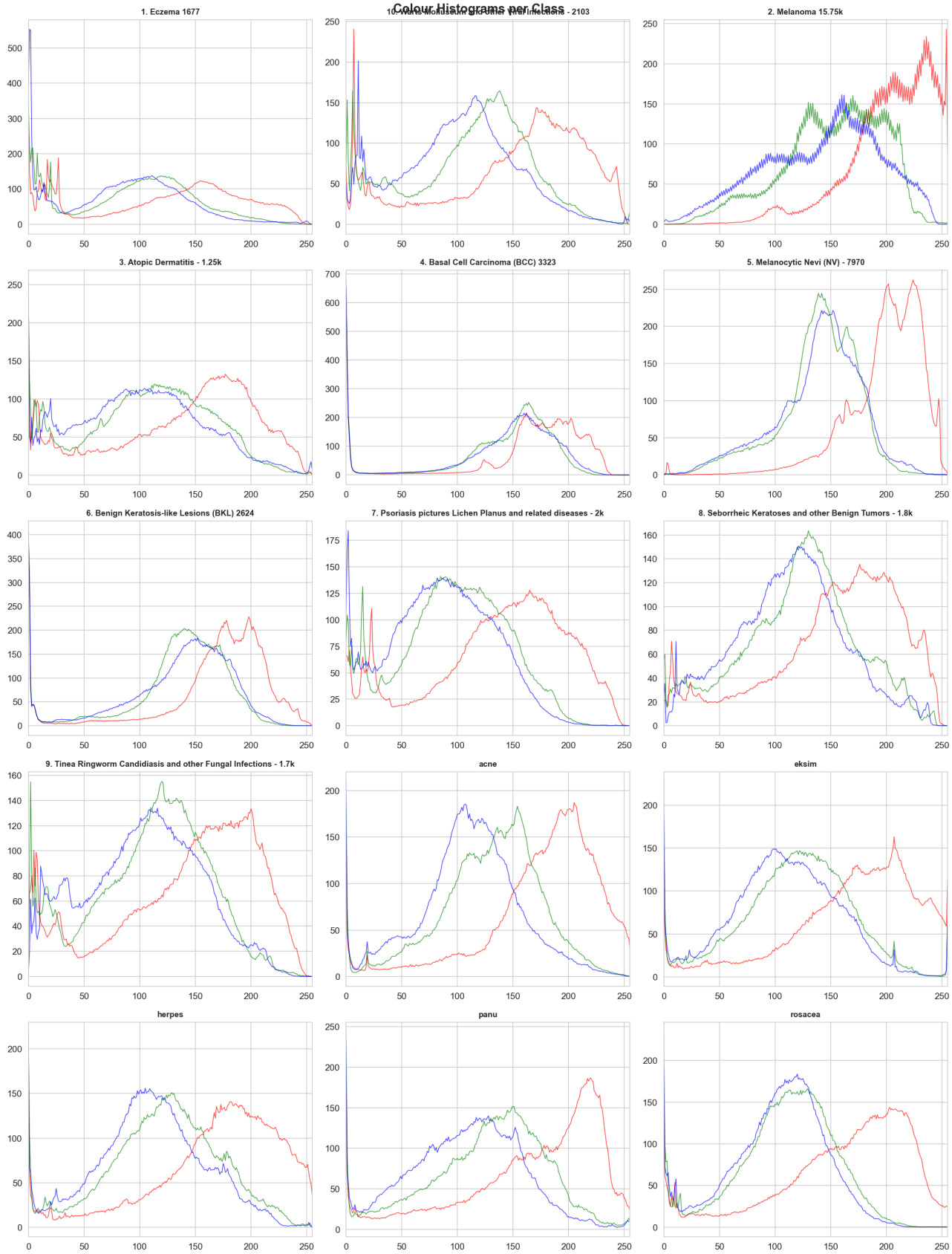


6. Benign Keratosis-like Lesions (BKL) 2624
(2079)

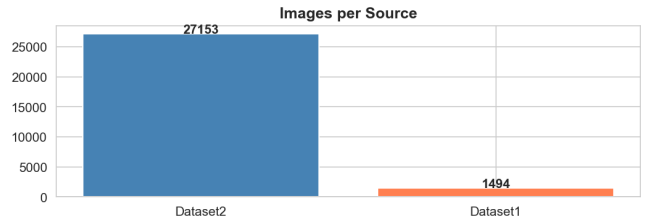


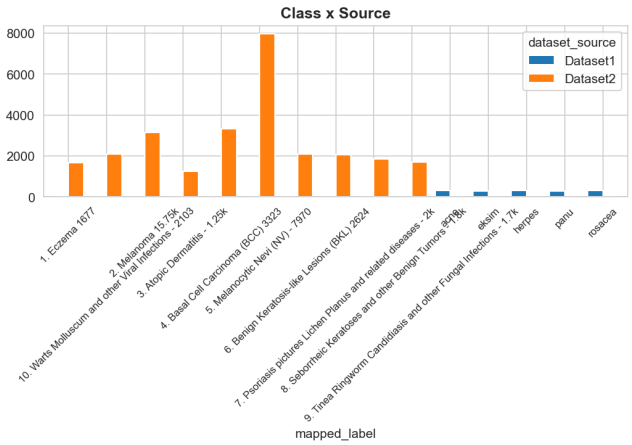
7. Psoriasis pictures Lichen Planus and related diseases - 2k
(2055)

2.7 Colour Histograms per Class



2.8 Dataset Source Comparison





2.9 Pixel Statistics Table

Class	R Mean	G Mean	B Mean
1. Eczema 1677	124.35	92.76	84.46
10. Warts Molluscum and other Viral Infections - 2103	149.69	110.88	100.28
2. Melanoma 15.75k	203.47	151.58	142.45
3. Atopic Dermatitis - 1.25k	138.23	107.37	100.98
4. Basal Cell Carcinoma (BCC) 3323	161.6	137.23	138.78
5. Melanocytic Nevi (NV) - 7970	199.08	138.78	138.53
6. Benign Keratosis-like Lesions (BKL) 2624	166.08	132.9	132.73
7. Psoriasis pictures Lichen Planus and related diseases - 2k	138.24	97.17	86.25
8. Seborrheic Keratoses and other Benign Tumors - 1.8k	153.87	121.56	113.56
9. Tinea Ringworm Candidiasis and other Fungal Infections - 1.7k	143.47	110.04	103.25
acne	176.58	129.15	111.04
eksim	169.57	118.73	108.86
herpes	167.95	121.74	109.97
panu	167.84	127.22	107.95
rosacea	161.22	109.46	102.69

3. Data Cleaning

Corrupt and duplicate images were identified during EDA and removed to prevent errors during training and data leakage during evaluation.

Metric	Count
Images Before Cleaning	28647
Corrupt Images Removed	0
Duplicate Images Removed	625
Images After Cleaning	28022

4. Smart Sampling

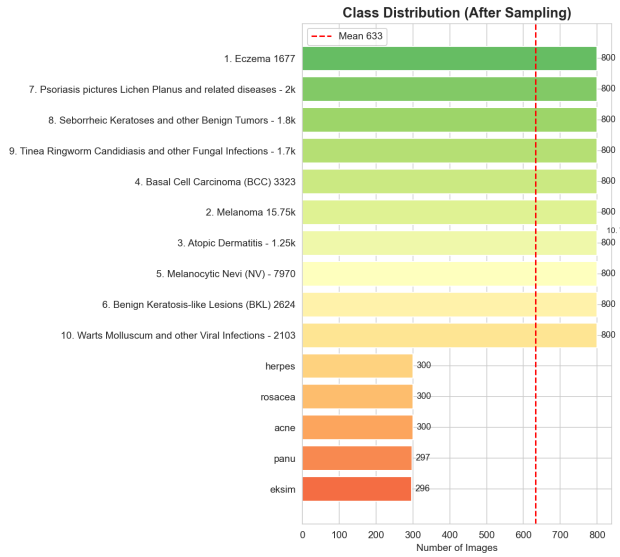
To reduce processing time, classes with more than 800 images were randomly down-sampled to 800. For traditional ML (SVM, RF, KNN), 500-800 images per class is sufficient. This reduced the dataset by 66.1% with minimal impact on accuracy.

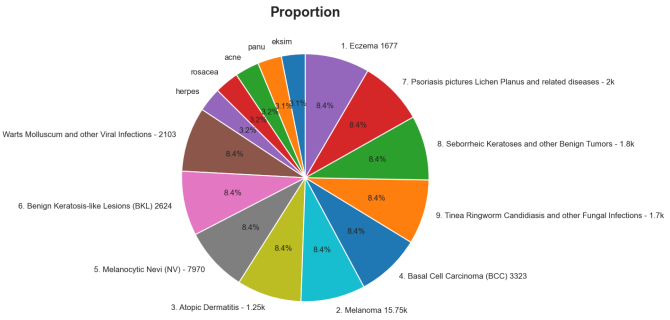
Metric	Value
Images Before Sampling	28022
Cap per Class	800
Images After Sampling	9493
Reduction	66.1%
Imbalance Ratio (after)	2.7 : 1

Per-Class Sampling Results

Class	After Sampling
1. Eczema 1677	800
10. Warts Molluscum and other Viral Infections - 2103	800
2. Melanoma 15.75k	800
3. Atopic Dermatitis - 1.25k	800
4. Basal Cell Carcinoma (BCC) 3323	800
5. Melanocytic Nevi (NV) - 7970	800
6. Benign Keratosis-like Lesions (BKL) 2624	800
7. Psoriasis pictures Lichen Planus and related diseases - 2k	800
8. Seborrheic Keratoses and other Benign Tumors - 1.8k	800
9. Tinea Ringworm Candidiasis and other Fungal Infections - 1.7k	800
acne	300
eksim	296
herpes	300
panu	297
rosacea	300

Class Distribution After Sampling





5. Image Preprocessing

All images were resized to 128 x 128 pixels using INTER_AREA interpolation (best for down-scaling). Pixel values were normalised from 0-255 to 0.0-1.0 by dividing by 255 (Min-Max normalisation). This ensures all features contribute equally during model training.

Before vs After Preprocessing

Metric	Before	After
Total Images	9493	9493
Failed / Skipped	-	0
Image Size	Variable (mixed sizes)	128 x 128 x 3
Pixel Range	0 - 255 (uint8)	0.0 - 1.0 (float32)
Array Shape	N/A	(9493, 128, 128, 3)
Memory Usage	N/A	1779.9 MB

6. Data Augmentation

Minority classes were augmented to 800 images each using random transformations: horizontal flips, rotations (+/-20 degrees), brightness adjustment (0.7-1.3x), contrast adjustment (0.8-1.2x), Gaussian noise, and random crop-zoom. This creates a perfectly balanced dataset.

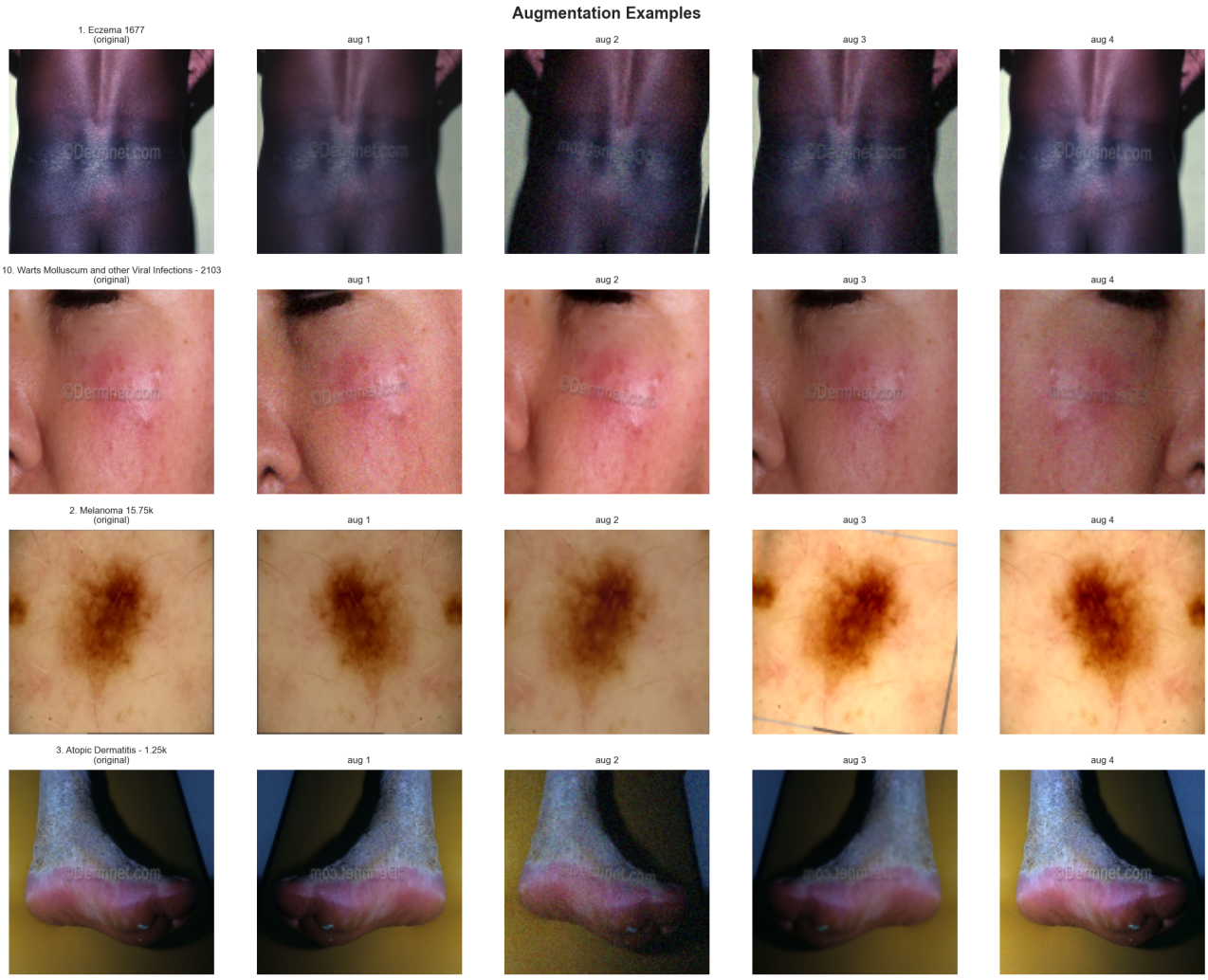
Before vs After Augmentation

Metric	Before	After
Total Images	9493	12000
Target per Class	-	800
Images Added	-	2507
Class Balance	Imbalanced	Perfectly Balanced (1:1)

Per-Class Augmentation Detail

Class	Before	After	Added
1. Eczema 1677	800	800	0
10. Warts Molluscum and other Viral Infections - 2103	800	800	0
2. Melanoma 15.75k	800	800	0
3. Atopic Dermatitis - 1.25k	800	800	0
4. Basal Cell Carcinoma (BCC) 3323	800	800	0
5. Melanocytic Nevi (NV) - 7970	800	800	0
6. Benign Keratosis-like Lesions (BKL) 2624	800	800	0
7. Psoriasis pictures Lichen Planus and related diseases - 1	800	800	0
8. Seborrheic Keratoses and other Benign Tumors - 1.8k	800	800	0
9. Tinea Ringworm Candidiasis and other Fungal Infections -	800	800	0
acne	300	800	500
eksim	296	800	504
herpes	300	800	500
panu	297	800	503
rosacea	300	800	500
TOTAL	9493	12000	2507

Augmentation Examples



7. Feature Extraction

Five complementary feature extraction methods were combined (feature fusion) to create a rich representation of each image. HOG captures shape/edges, LBP captures local texture, colour histograms capture colour distribution, GLCM captures texture regularity, and statistical features capture global properties.

Feature Type	Dimensions	Captures
HOG	1764	Shape & edges
LBP	26	Local texture
Colour Histogram	96	Colour distribution
GLCM	10	Texture regularity
Statistical	21	Global pixel stats
TOTAL	1917	Combined feature vector

Feature Matrix Shape (12000, 1917)

Memory Usage 87.8 MB

Failed Extractions 0

8. Train-Test Split & Feature Scaling

Split Configuration

Split Ratio 80 / 20

Stratified Yes (class proportions preserved)

Random State 42

Set	Samples	Percentage
Training	9600	80%
Test	2400	20%
Total	12000	100%

Feature Scaling (StandardScaler)

StandardScaler transforms features to zero mean and unit variance. CRITICAL: the scaler was fitted on training data ONLY, then applied to test data. Fitting on test data would cause data leakage.

Metric	Before Scaling	After Scaling
Mean	0.1951	-0.0
Std Dev	0.4989	1.0

Class Distribution in Splits

Class	Train %	Test %
1. Eczema 1677	6.7%	6.7%
10. Warts Molluscum and other Viral Infections - 2103	6.7%	6.7%
2. Melanoma 15.75k	6.7%	6.7%
3. Atopic Dermatitis - 1.25k	6.7%	6.7%
4. Basal Cell Carcinoma (BCC) 3323	6.7%	6.7%
5. Melanocytic Nevi (NV) - 7970	6.7%	6.7%
6. Benign Keratosis-like Lesions (BKL) 2624	6.7%	6.7%
7. Psoriasis pictures Lichen Planus and related diseases - 2k	6.7%	6.7%
8. Seborrheic Keratoses and other Benign Tumors - 1.8k	6.7%	6.7%
9. Tinea Ringworm Candidiasis and other Fungal Infections - 1.7	6.7%	6.7%
acne	6.7%	6.7%
eksim	6.7%	6.7%
herpes	6.7%	6.7%
panu	6.7%	6.7%
rosacea	6.7%	6.7%

9. Class Weights

Class weights were computed using sklearn's 'balanced' strategy: weight = n_samples / (n_classes x n_class_samples). Higher weights penalise misclassification of rarer classes.

Class	Weight
1. Eczema 1677	1.0
10. Warts Molluscum and other Viral Infections - 2103	1.0
2. Melanoma 15.75k	1.0
3. Atopic Dermatitis - 1.25k	1.0
4. Basal Cell Carcinoma (BCC) 3323	1.0
5. Melanocytic Nevi (NV) - 7970	1.0
6. Benign Keratosis-like Lesions (BKL) 2624	1.0
7. Psoriasis pictures Lichen Planus and related diseases - 2k	1.0
8. Seborrheic Keratoses and other Benign Tumors - 1.8k	1.0
9. Tinea Ringworm Candidiasis and other Fungal Infections - 1.7k	1.0
acne	1.0
eksim	1.0
herpes	1.0
panu	1.0
rosacea	1.0

10. Output Files

File	Description
ml_ready_dataset.pkl	Complete bundle (all objects)
X_train.npy	Training features (scaled)
X_test.npy	Test features (scaled)
y_train.npy	Training labels (encoded)
y_test.npy	Test labels (encoded)
label_encoder.joblib	Class name to integer mapping
scaler.joblib	Fitted StandardScaler object
pipeline_summary_report.pdf	This report

Output Directory C:\Users\Sharon Shalom\dsproj\dsproj\skin-disease-classification\data\proce

11. Complete Pipeline Summary

Data Flow Through Pipeline

Pipeline Stage	Images	Change
1. Raw Dataset (Combined)	28647	-
2. After Cleaning	28022	-625
3. After Smart Sampling	9493	-18529
4. After Preprocessing	9493	-0 (failed)
5. After Augmentation	12000	+2507

Final Dataset Specifications

Specification	Value
Training Samples	9600
Test Samples	2400
Feature Vector Dimension	1917
Number of Classes	15
Class Balance	Balanced (1:1)
Feature Scaling	StandardScaler (mean=0, std=1)
Total Processing Time	0:05:23

12. Configuration Parameters

Parameter	Value
augmentation_target	800
augmented_path	data\augmented
color_hist_bins	32
color_space	RGB
combined_path	data\combined
dataset1_path	data\raw\dataset1\train
dataset2_path	data\raw\dataset2\IMG_CLASSES
eda_output_path	outputs\eda_plots
glcm_angles	[0, 0.7853981633974483, 1.5707963267948966, 2.356194...
glcm_distances	[1, 3]
hog_cells_per_block	(2, 2)
hog_orientations	9
hog_pixels_per_cell	(16, 16)
image_extensions	{'.png', '.bmp', '.jpg', '.tiff', '.jpeg'}
lbp_n_points	24
lbp_radius	3
max_eda_samples	300
max_per_class	800
min_per_class	250
model_output_path	outputs\models
processed_path	data\processed
random_state	42
target_size	(128, 128)
test_size	0.2

13. Recommended Next Steps

1. Train SVM (RBF Kernel)

Best for high-dimensional feature spaces. Expected accuracy: 85-92%.

2. Train Random Forest

Good accuracy with built-in feature importance for interpretability.

3. Train KNN

Simple distance-based baseline. May struggle with high dimensions.

4. Train Decision Tree

Most interpretable model. Use as explainability baseline.

5. Model Evaluation

Compare using Accuracy, Precision, Recall, F1-Score, Confusion Matrix.

6. Select Best Model

Choose based on weighted F1-Score (handles class importance).

7. Human-in-the-Loop Integration

Present probability-based predictions for clinician validation.

8. Clinical Deployment

Package as decision-support tool integrated into clinical workflow.

Pipeline Completed Successfully

Dataset is ML-Ready

Total Processing Time: 0:05:23

Report generated: 2026-02-19 16:43:03

Total pages: 23