Sri Lanka Institute of Information Technology

IT2011 - Artificial Intelligence and Machine Learning

Year 2 Semester 1 – 2025

Skin Cancer Detection System

Y2.S1.WD.IT.09.02– Project group 175

2025-Y2-S1-MLB-B9G2-05

Final Report

Student Id	Name
IT24102522	Himandi A.H.S.
IT24102555	Weerathunga B.A.
IT24102561	Modarage K.R.
IT24102532	Delpachithra K.N.
IT24102470	Jayawardhana T.N.D.J.

Table of Contents

1. Introduction and Problem Statement

Skin cancer remains one of the most rapidly increasing cancer types globally. Early and accurate detection can save lives, but traditional diagnostic methods rely heavily on human expertise and visual inspection.

This project aims to develop and compare **multiple machine learning and deep learning approaches** to automatically classify **skin lesions** from the **HAM10000 dataset** into benign and malignant categories.

Goal:

To identify the most effective model for classifying skin lesion images and evaluate how traditional ML models perform compared to advanced CNN-based models.

2. Dataset Description

Dataset: HAM10000 (Human Against Machine with 10,000 Training Images)
Source: Kaggle / ISIC archive

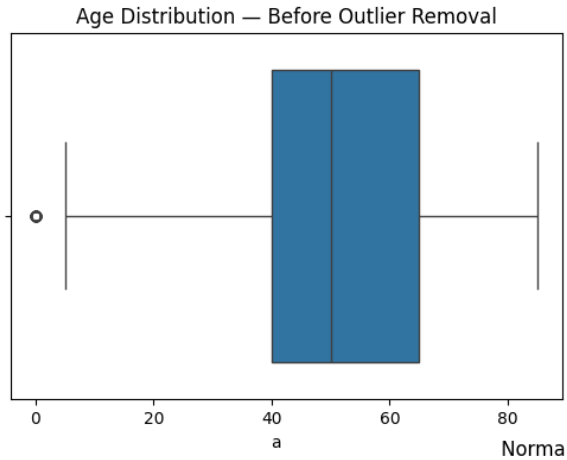
Key details:

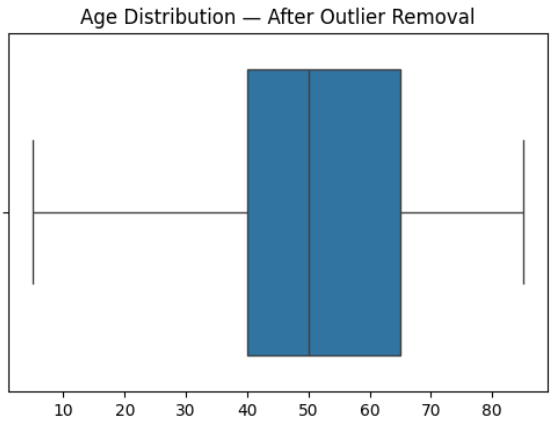
- 10,015 dermoscopic images of pigmented skin lesions
- 7 classes:
    1. Melanocytic nevi (nv)
    2. Melanoma (mel)
    3. Benign keratosis (bkl)
    4. Basal cell carcinoma (bcc)
    5. Actinic keratoses (akiec)
    6. Vascular lesions (vasc)
    7. Dermatofibroma (df)
- Image format: JPEG
- Each image labeled by dermatologists
- Malignant: melanoma, bcc, akiec
- Benign: nv, bkl, vasc, df

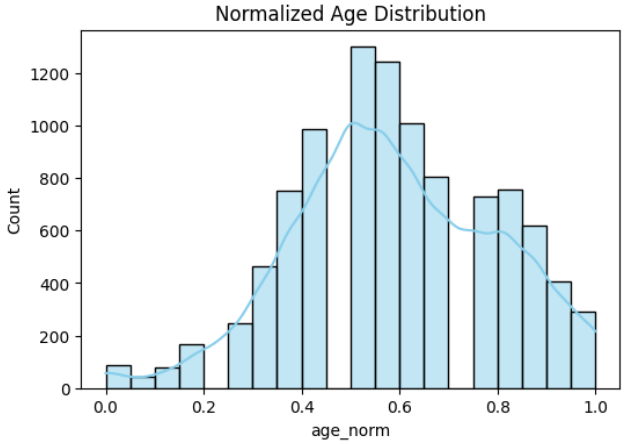
3. Preprocessing & Exploratory Data Analysis (EDA)

1. Missing Values

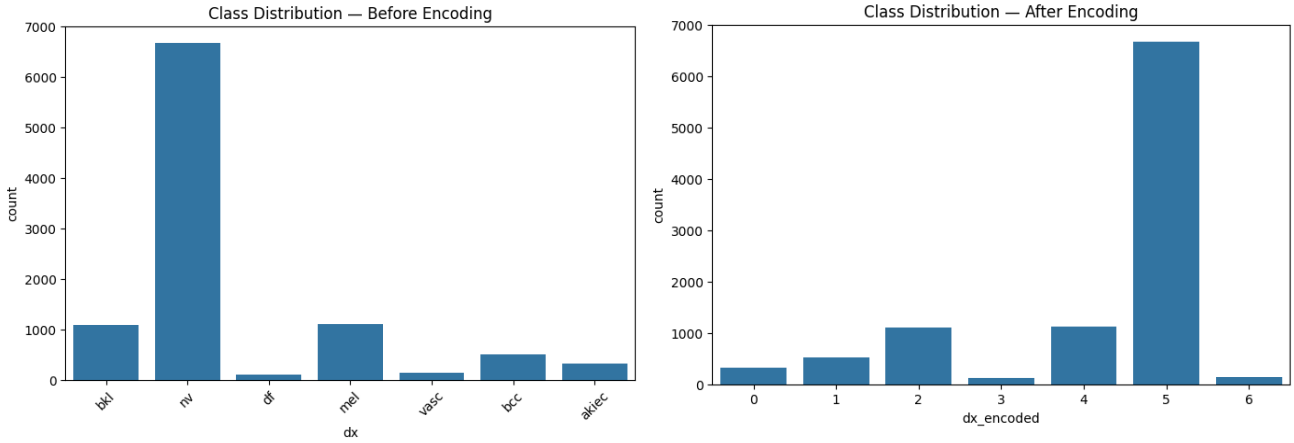
2. Age Outlier removal



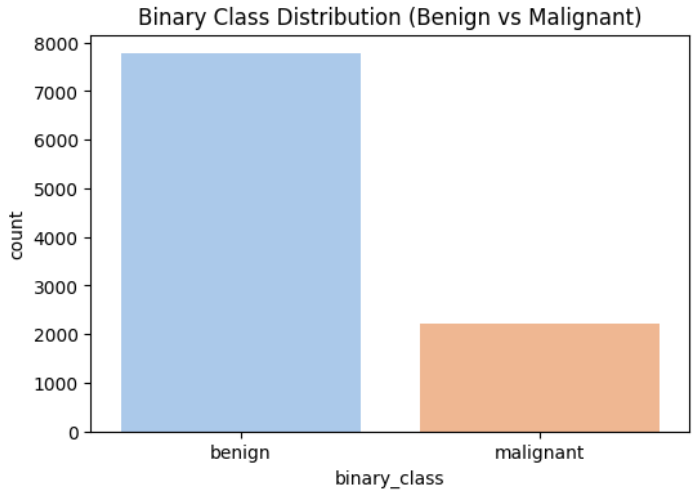




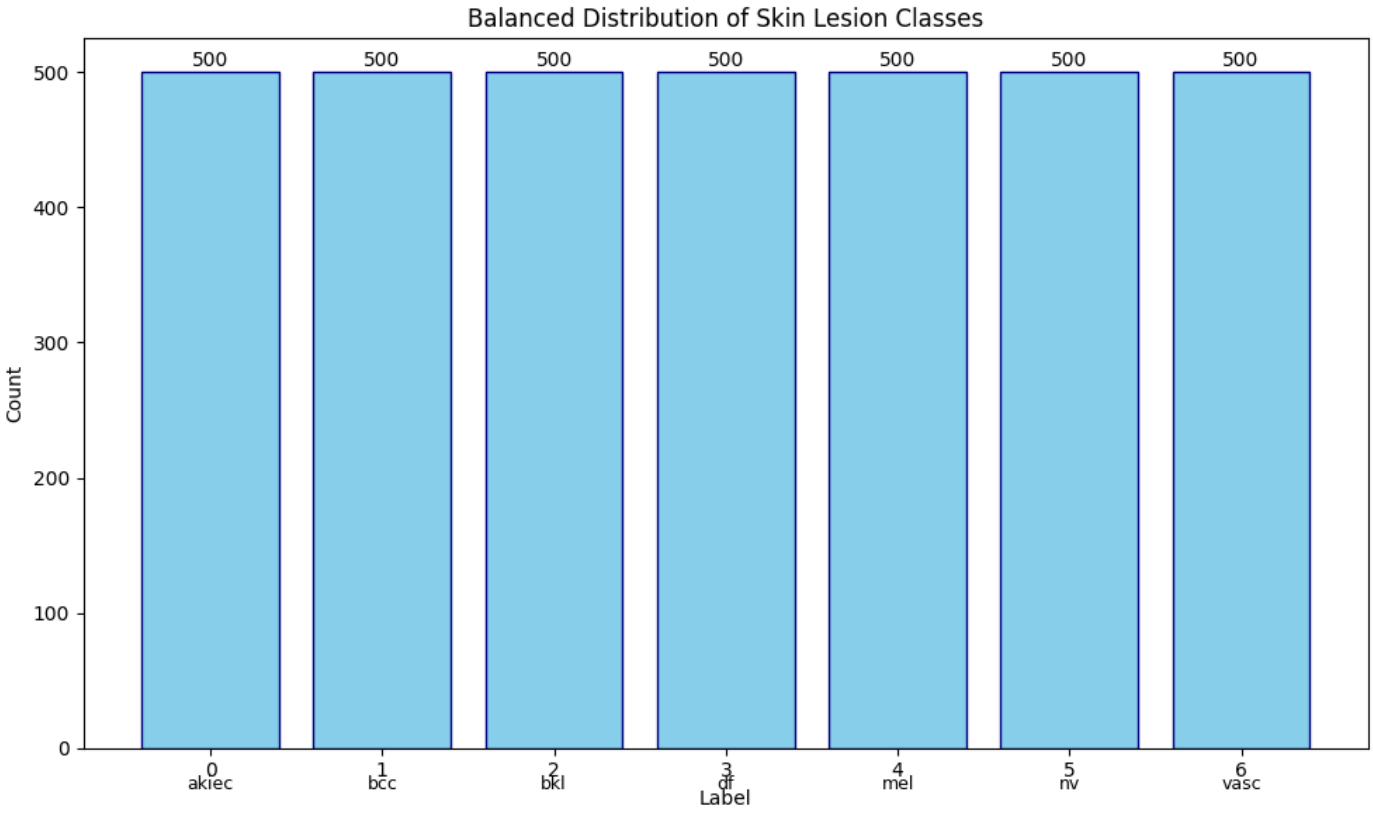
3. Label encoding



4. Feature Engineering: Binary class



5. Balance Classes (500 samples per class)



4. Model Design and Implementation

1. Machine Learning Models

Model	Description
K-Means	Used for unsupervised clustering to group lesion types.
Logistic Regression	Linear classifier to separate benign vs malignant.
Decision Tree	Splits feature hierarchically for interpretability.
Random Forest	Ensemble trees to improve accuracy and reduce overfitting.
Base MLP	Simple neural network with one hidden layer.
Regularized MLP	Adds L2 regularization to prevent overfitting.
Grid Search MLP	Tuned hyperparameters (hidden layers, learning rate).
MLP Ensemble	Combines multiple MLPs for averaged prediction.

2. Deep Learning Models

Model	Description
ResNet50	Pretrained model fine-tuned for lesion classification.
DenseNet121	Feature-rich pretrained CNN known for dense connections.
VGG16	Classic transfer learning model with 16 layers.
Custom CNN (Your Part)	Designed from scratch with tuned hyperparameters.

5. Evaluation and Comparison

Model	Accuracy	Precision	Recall	F1 Score
K-Means	0.2140	0.2863	0.2140	0.2301
Logistic Regression	0.4857	0.5403	0.4954	0.4856
Decision Tree	0.8300	0.8245	0.9680	0.8905
MLP	0.65	0.62	0.59	0.60
CNN	0.99	1.00	0.92	0.96

6. Ethical Considerations and Bias Mitigation

- **Data Bias:** Original dataset was imbalanced; addressed using oversampling.
- **Skin Tone Diversity:** Dataset lacks global skin tone variety—may affect model generalization.
- **Medical Ethics:** Predictions should assist, not replace, professional diagnosis.
- **Transparency:** Used explainable models (Decision Tree, CNN heatmaps) to visualize lesion areas of focus.

7. Reflections and Lessons Learned

- Data preprocessing and balancing had the greatest impact on fairness and accuracy.
- Traditional ML models are faster but less accurate for image-based data.
- Deep learning, especially CNN-based architectures, captures visual features better.
- Hyperparameter tuning (GridSearch, Dropout, Learning Rate) improved model robustness.
- Collaborative learning across classical and deep learning methods enhanced understanding of trade-offs between **interpretability and performance**

8. References

- Tschandl, P. et al. (2018). *HAM10000 Dataset: A Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions.* Scientific Data.
- TensorFlow/Keras Documentation – https://www.tensorflow.org
- Scikit-learn Documentation – https://scikit-learn.org
- Kaggle HAM10000 Dataset – https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000