Dataset Link:
https://www.kaggle.com/datasets/mfazrinizar/skin-cancer-ham10000-raw-and-lesion-segmentation

1 Dataset Overview

Size: 10,015 dermatoscopic images with metadata.

2 Sample Image Grid

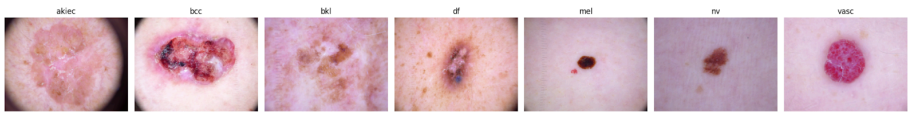

Figure: sample images from each diagnosis class

3 Key Features:

lesion_id: Unique identifier for each lesion.
image_id: Image filename (corresponds to .jpg files).
dx: Diagnosis label (target class).
dx_type: Method of diagnosis (e.g., histo, consensus).
age, sex, localization: Patient demographic and anatomical site information.

4 Data Cleaning Summary:

1. Missing values in age and sex were removed.
2. Checked and removed potential duplicates using lesion_id and visual features.
3. Categorical features (dx, dx_type, sex, localization) were label encoded for modeling.
4. Verified class balance and feature consistency.

5 Class Distribution:

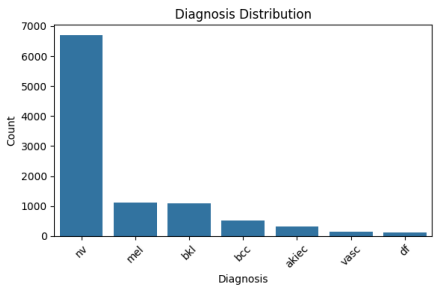

Figure: Class distribution for skin lesion diagnoses.

The dataset is highly imbalanced — "nv" (melanocytic nevi) dominates, while "df" (dermatofibroma) is rare. This affects model training and motivates the use of stratified sampling and performance metrics beyond accuracy.

6 Feature Distribution:

6.1 Age Distribution:

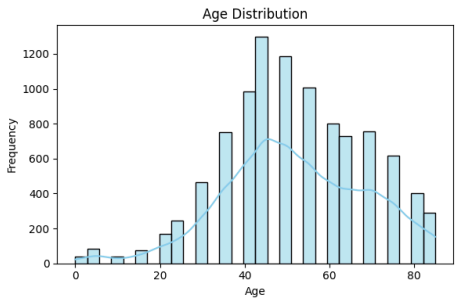

Figure: Histogram of patient ages

It shows a skew towards older individuals, peaking between 45 and 60 years.

6.2 Gender Distribution:

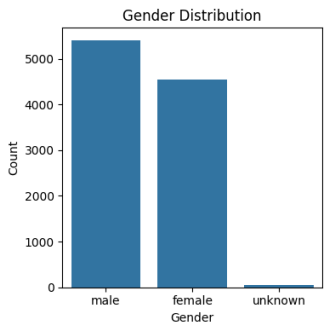

Figure: Bar chart representing gender distribution

Slight male dominance in the dataset, which might introduce bias in prediction.

6.3 Lesion Localization:

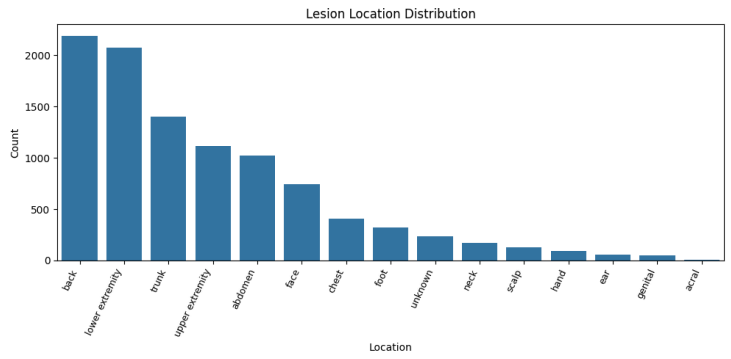

Figure: Barchart representing the distribution of lesion locations across body parts.

Most lesions originate from the back and lower extremities, aligning with dermatological clinical patterns.

7 Clustering and Dimensionality Reduction:

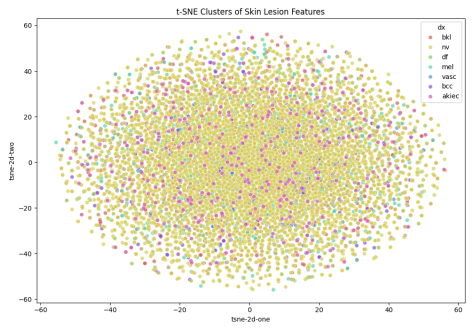

Figure: t-SNE plot of simulated 100-dimensional image features.

Clusters for "nv", "bkl", and "mel" partially overlap, indicating possible confusion in classifier predictions. Hence, deeper feature engineering models along with segmentation will be beneficial.

8 Statistical Analysis Summary:

ANOVA(Analysis of Variance) on age vs diagnosis group: Statistically significant ($p < 0.05$), indicating age varies across classes.

9 External Dataset Comparison:

9.1 Comparison with the ISIC2018 dataset:

i. The accuracy dropped by 3%, which suggests that the model may have overfitted to the original HAM10000 dataset.
ii. Body locations of lesions and the number of samples per class had a gap between the two datasets.