Exploratory Data Analysis (EDA) Report – ISIC 2019 Skin Lesion Dataset

1. Dataset Overview

- **Metadata shape:** (25331, 5)
- **GroundTruth shape:** (25331, 10)
- **Merged shape:** (25331, 14)

The merged dataset consists of **25,331 rows and 14 columns**, combining patient metadata with ground truth lesion class labels.

2. Dataset Info

- **Numerical features (10):** age_approx, and 9 binary class labels (MEL, NV, BCC, AK, BKL, DF, VASC, SCC, UNK)
- **Categorical features (4):** image, anatom_site_general, lesion_id, sex

3. Missing Values

- age_approx: 437 missing
- anatom_site_general: 2631 missing
- lesion_id: 2084 missing
- sex: 384 missing
- **Target columns:** No missing values

4. Numerical Summary

- **Age (age_approx):**
  - Mean = 54 years, Std = 18
  - Min = 0, Max = 85
  - Median = 55
- **Target labels:**
  - Values are binary (0/1).
  - Some classes (e.g., NV) dominate, others (DF, VASC) are rare.

5. Categorical Summary

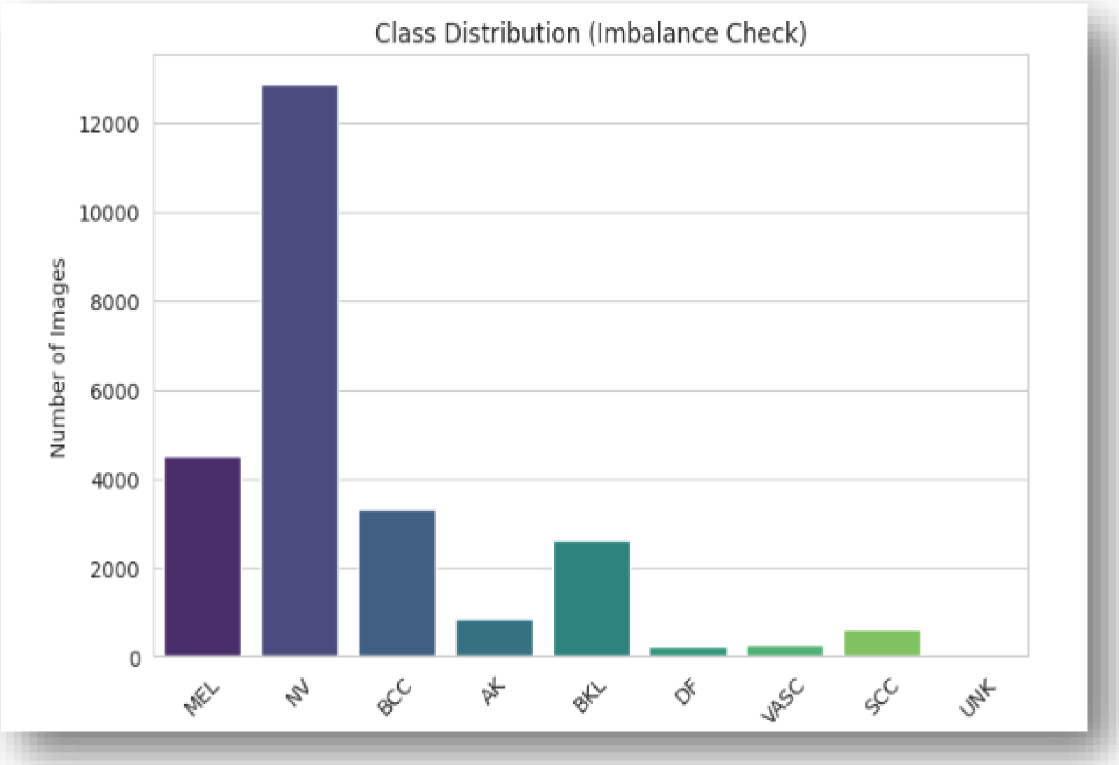
- **Image:** 25,331 unique images
- **Anatomical sites:** 8 unique (most common: *anterior torso* – 6915 cases)
- **Lesion IDs:** 11,847 unique (most common lesion ID: 31 images)
- **Sex:** Male = 13,286 | Female = 11,661

6. Target Classes

Target labels: **MEL, NV, BCC, AK, BKL, DF, VASC, SCC, UNK**

Class Counts

- NV: **12,875**
- MEL: **4,522**
- BCC: **3,323**
- BKL: **2,624**
- AK: **867**
- SCC: **628**
- VASC: **253**
- DF: **239**
- UNK: **0**

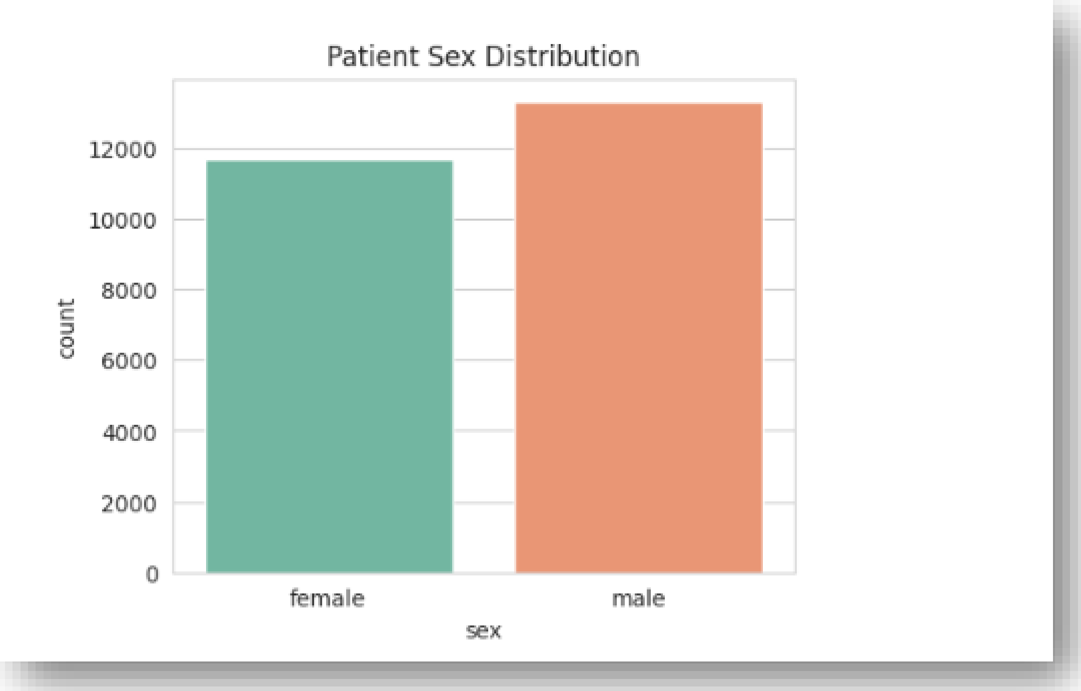


-

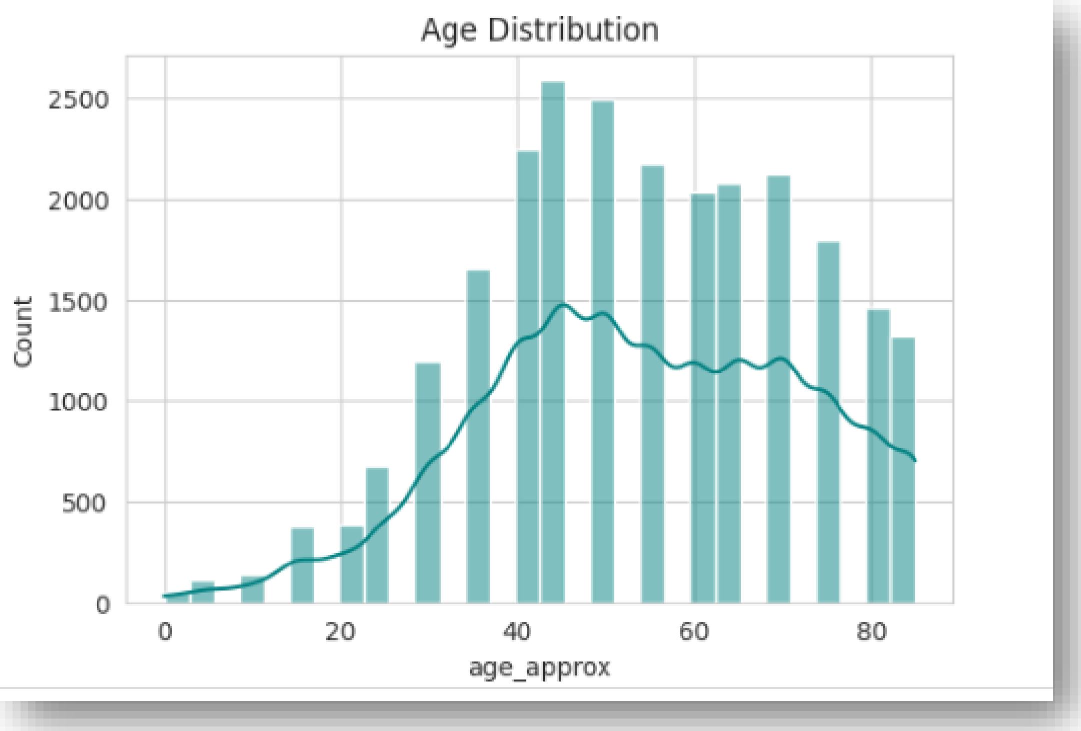
Observation: Dataset is **highly imbalanced**. NV dominates (50%), while rare classes like DF & VASC are <1%.

7. Patient Demographics

- **Sex distribution:** Slightly more males than females.
- **Age distribution:** Peak between 40–70 years.
- **Anatomical sites:** Mostly anterior torso, followed by lower extremity and back.



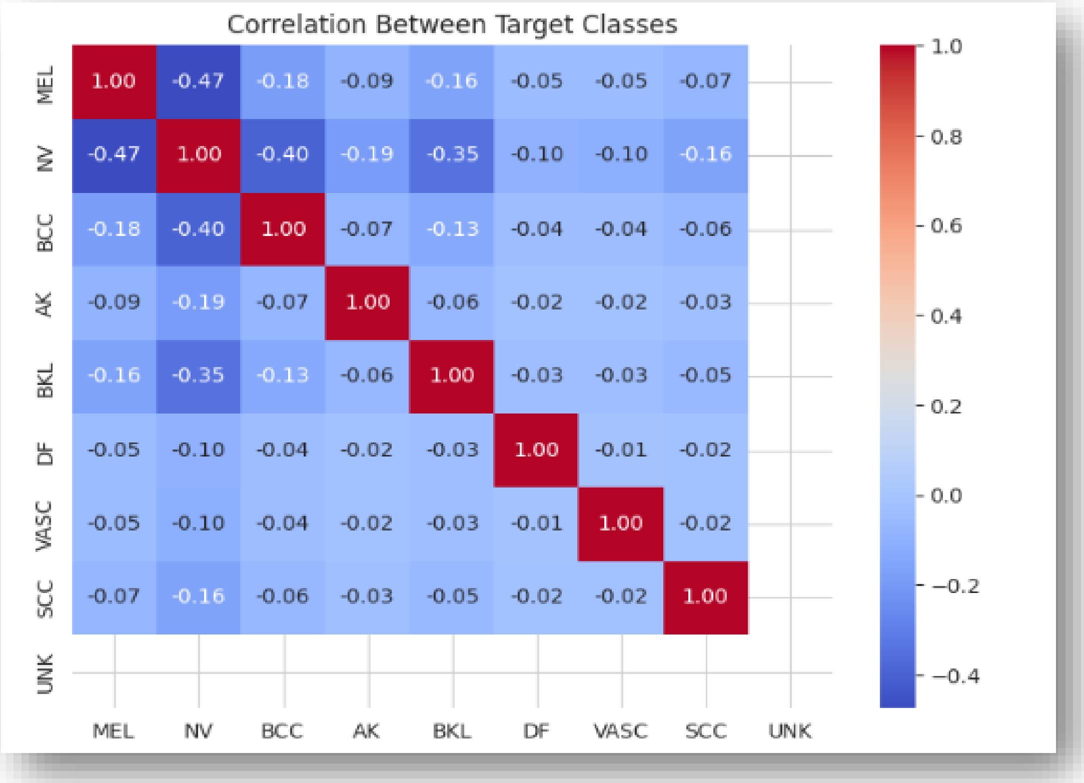
-



-

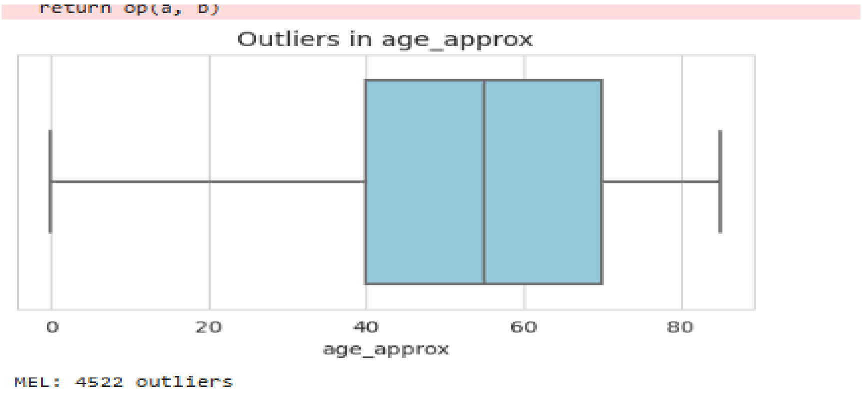
8.Correlation Among Classes

- Heatmap shows low correlation among most classes (since each image usually belongs to only one class).
- Some mild negative correlations exist (since classes are mutually exclusive).



9. Outlier Analysis

- Age has some **outliers (0 years, very low values)** → possible data entry errors.
- Class labels are binary → no numerical outliers.


```
  return op(a, b)
```

```
MEL: 4522 outliers
```

10. Summary of Insights

- Dataset is highly imbalanced (NV dominates).
- Some missing values handled with mean/median/mode.
- Outliers present in age feature.
- Disease prevalence varies across sex and localization.
- Preprocessing steps (balancing, normalization, augmentation) are necessary before model training.