Exploratory Data Analysis (EDA) for Melanoma Datasets

1. Introduction

Melanoma is a highly aggressive form of skin cancer, where early detection can significantly improve patient outcomes. In this document, we will outline the exploratory data analysis that was performed on the ISIC 2019 and ISIC 2020 datasets. The original notebook can be found on this link.

As a whole, the aim of our project is to develop a binary classification model that distinguishes between benign and malignant skin lesions using dermoscopic images and structured metadata. A major challenge in this task is the **severe class imbalance** commonly present in medical datasets, where malignant cases are vastly underrepresented. Additionally, the diversity in image quality, anatomical site, and patient demographics adds complexity to the classification problem.

To address these challenges, we conducted a thorough **exploratory data analysis (EDA)** and **dataset enrichment** process. This involved integrating the ISIC 2019 dataset to increase both the size and representativeness of the data, converting multi-class diagnostic labels into a binary format suitable for our classification objective.

Through this process, we also examined **key image properties** and **feature distributions** to better understand the data and inform model design decisions.

2. Metadata Analysis of ISIC 2020 Dataset

2.1 Overview

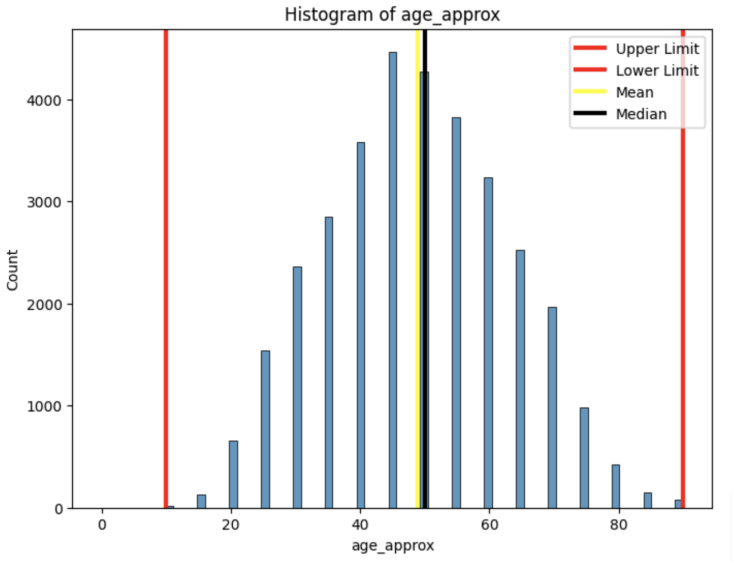
The features of the dataset are as follows

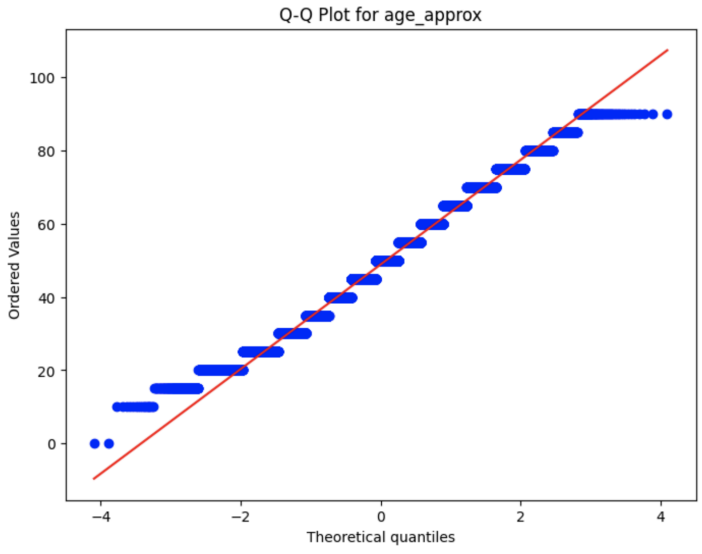
	image_name	patient_id	lesion_id	sex	age_approx	anatom_site_general_challenge	diagnosis	benign_malignant	target
0	ISIC_2637011	IP_7279968	IL_7972535	male	45.0	head/neck	unknown	benign	0
1	ISIC_0015719	IP_3075186	IL_4649854	female	45.0	upper extremity	unknown	benign	0
2	ISIC_0052212	IP_2842074	IL_9087444	female	50.0	lower extremity	nevus	benign	0
3	ISIC_0068279	IP_6890425	IL_4255399	female	45.0	head/neck	unknown	benign	0
4	ISIC_0074268	IP_8723313	IL_6898037	female	55.0	upper extremity	unknown	benign	0

The columns `image_name, patient_id, lesion_id, target` columns will be excluded from the analysis – the first free because they are are unique identifiers and the last one because it is redundant (it is identical to the `benign_malignant` column.

2.2 Numeric Features

The `age` variable appears to follow a normal distribution, with no evident outliers. This is supported by the symmetry in the histogram and the alignment with the reference line in the QQ plot below



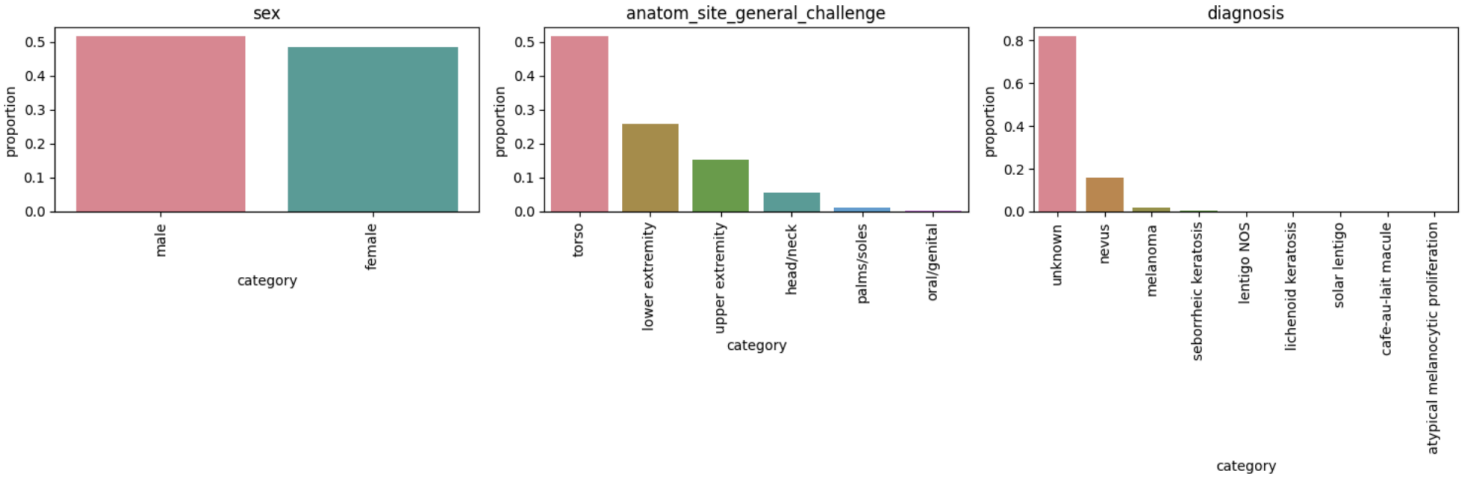


2.3 Categorical Features

Sex**:** The dataset is relatively balanced in terms of sex, with 51.7% male and 48.3% female samples.

Anatomical Site**:** Over half of the lesions are located on the torso (51.7%), followed by the lower extremities (25.8%) and upper extremities (15.3%). Other regions such as the head/neck, palms/soles, and oral/genital areas are underrepresented.

Diagnosis**:** The majority of the diagnoses are labeled as "unknown" (81.9%), which may reflect missing or uncertain clinical information. Among known diagnoses, nevi make up 15.7%, while melanoma comprises only 1.8% of the data. Other conditions such as seborrheic keratosis, lentigo NOS, and various rare skin anomalies are present at very low frequencies.

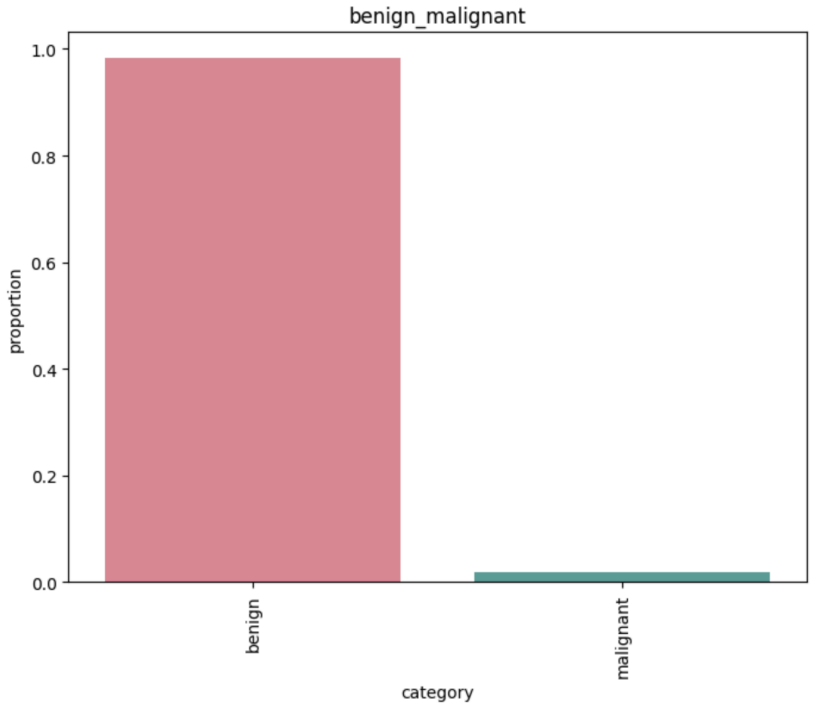



```
sex
male      0.516621
female    0.483379
Name: proportion, dtype: float64
-------------
anatom_site_general_challenge
torso              0.516734
lower extremity    0.258198
upper extremity    0.152857
head/neck          0.056904
palms/soles        0.011503
oral/genital       0.003804
Name: proportion, dtype: float64
-------------
diagnosis
unknown                                0.818813
nevus                                  0.156765
melanoma                               0.017630
seborrheic keratosis                   0.004075
lentigo NOS                            0.001328
lichenoid keratosis                    0.001117
solar lentigo                          0.000211
cafe-au-lait macule                    0.000030
atypical melanocytic proliferation     0.000030
Name: proportion, dtype: float64
-------------
```

2.4 Target Feature

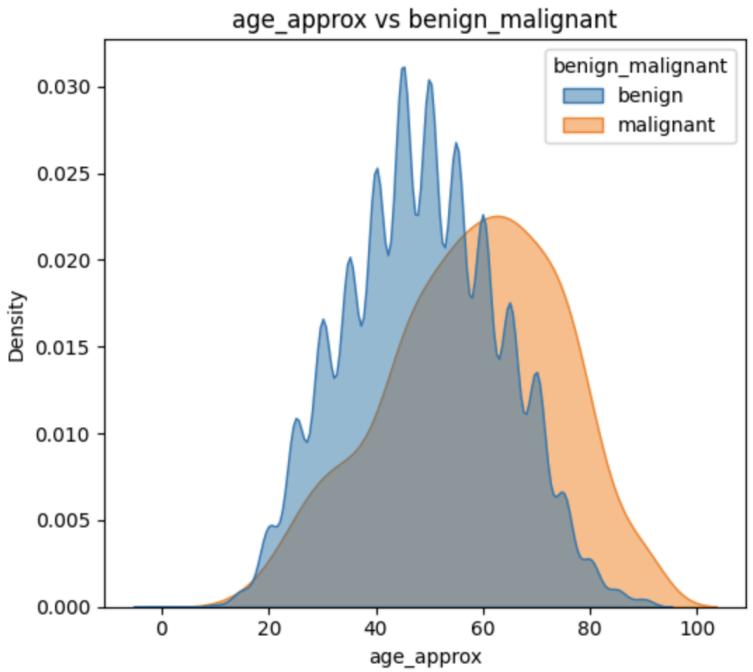
The target feature benign_malignant is highly imbalanced, with only **1.76%** of the samples labeled as malignant. This severe class imbalance requires special attention during model training. Standard loss functions like Cross-Entropy Loss may be suboptimal, as they do not account for the disparity between classes and can lead to biased predictions toward the majority class. In contrast, **Focal Loss** is better suited for this task, as it down-weights easy examples and focuses the model's learning on harder, minority class samples—making it a more effective choice for melanoma classification in this context.

```
benign_malignant
benign       0.98237
malignant    0.01763
Name: proportion, dtype: float64
```

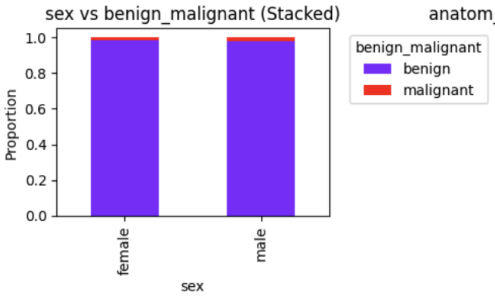


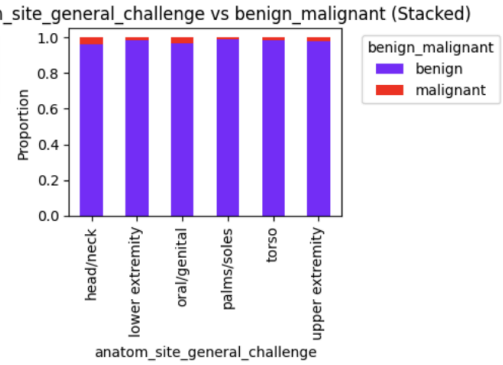
2.5 Relationship Between Features

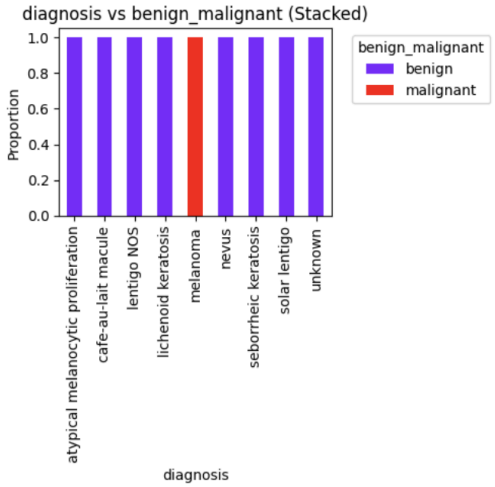
Age appears to have some influence on the target feature, which is expected given that the risk of developing malignant skin conditions generally increases with age. This relationship is visible in the plot below, where malignant cases are more concentrated in the older age ranges.



- Sex and Anatomical Site seem not to have any influence on the target.
- Diagnosis is equivalent to target feature







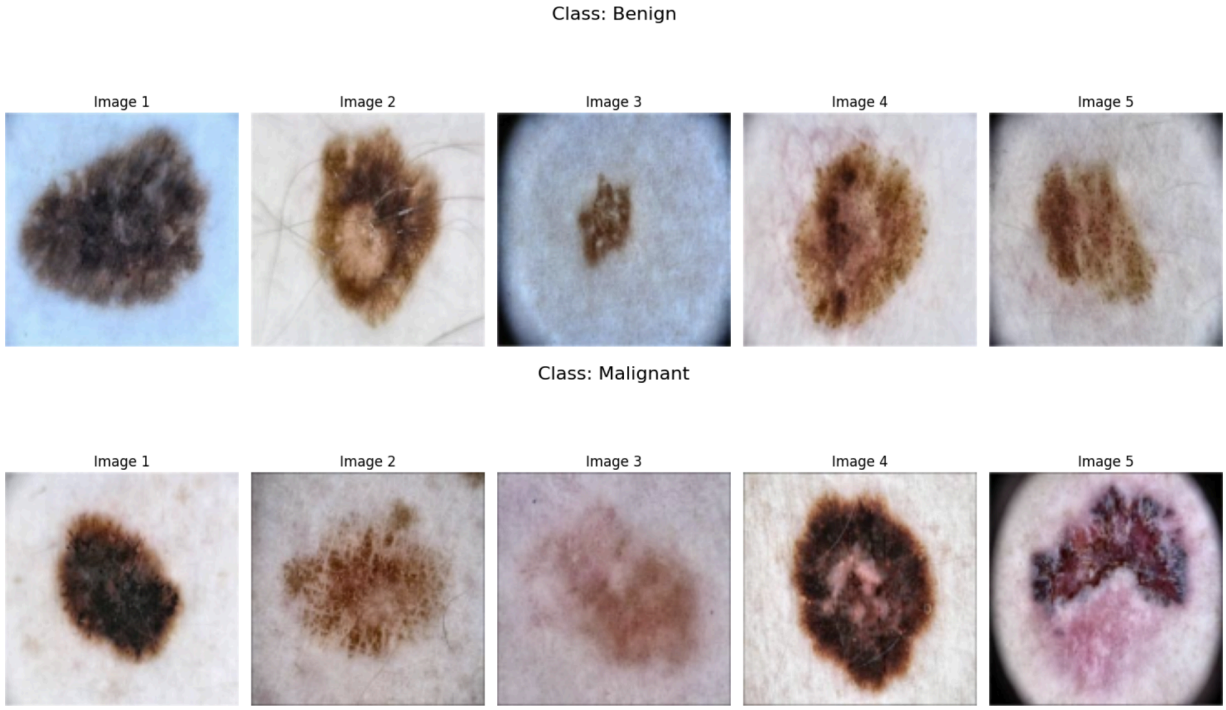
2.6 Conclusion for Metadata Analysis

- `Age` is approximately normally distributed and does not contain significant outliers.
- Among all features, `Age` appears to have the most noticeable influence on the target variable, which aligns with medical knowledge that the risk of malignant skin conditions increases with age.
- The **target feature** (`benign_malignant`) is extremely imbalanced, with malignant cases making up only ~1.76% of the dataset. This imbalance must be addressed during model training to avoid biased predictions.

3. Analysis of combined ISIC 2019 and 2020 datasets

To enhance the performance of our melanoma classification model, we expanded the dataset by incorporating the ISIC 2019 dataset, which provides dermoscopic images along with one-hot encoded diagnostic labels across multiple categories. Since our model is built for **binary classification** (benign vs. malignant), these labels were converted into a single binary column. The downloading of images, extraction of labels, and conversion of the metadata into a unified CSV format—followed by concatenation with the existing dataset—are all handled in the dataset preparation step of our model training pipeline.

The combined dataset now consists of **66,696 images**, which we split into training, validation and testing subsets. Further exploratory data analysis (EDA) is conducted on a **random subset of the training data**, both to prevent data leakage and to reduce computational overhead



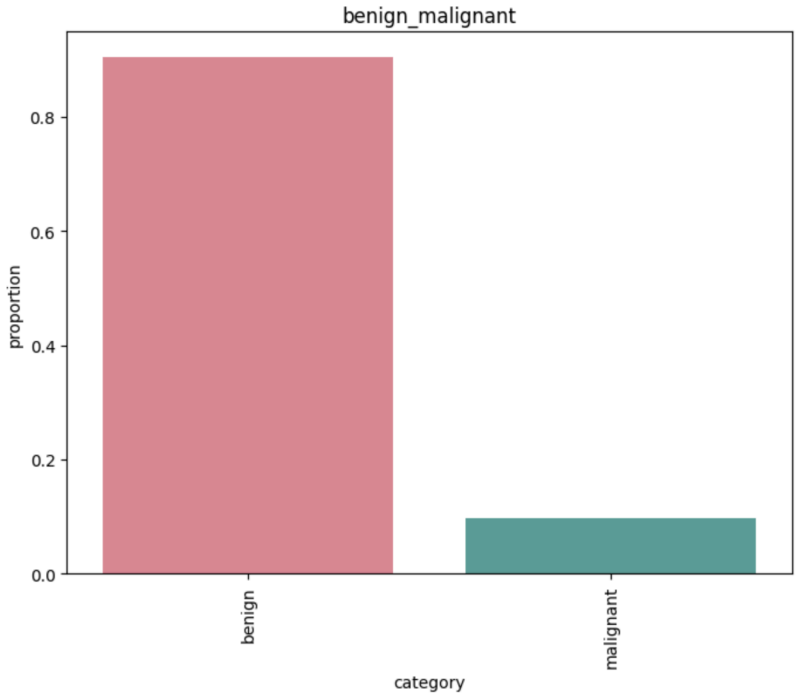
3.1 Overview of combined dataset

```
df.head()
```

	image_name	benign_malignant
0	ISIC_0034321	benign
1	ISIC_0034322	benign
2	ISIC_0034323	benign
3	ISIC_0034324	benign
4	ISIC_0034325	benign

By combining datasets, we not only **doubled the dataset size** (from 33,126 to 66,695 samples), but also **significantly reduced class imbalance**. The proportion of malignant cases increased from approximately **1.7%** to **9%**, reducing the benign-to-malignant ratio from ~98.3:1.7 to ~91:9. This is a substantial improvement that is likely to enhance model performance, particularly in identifying the minority (malignant) class more accurately.

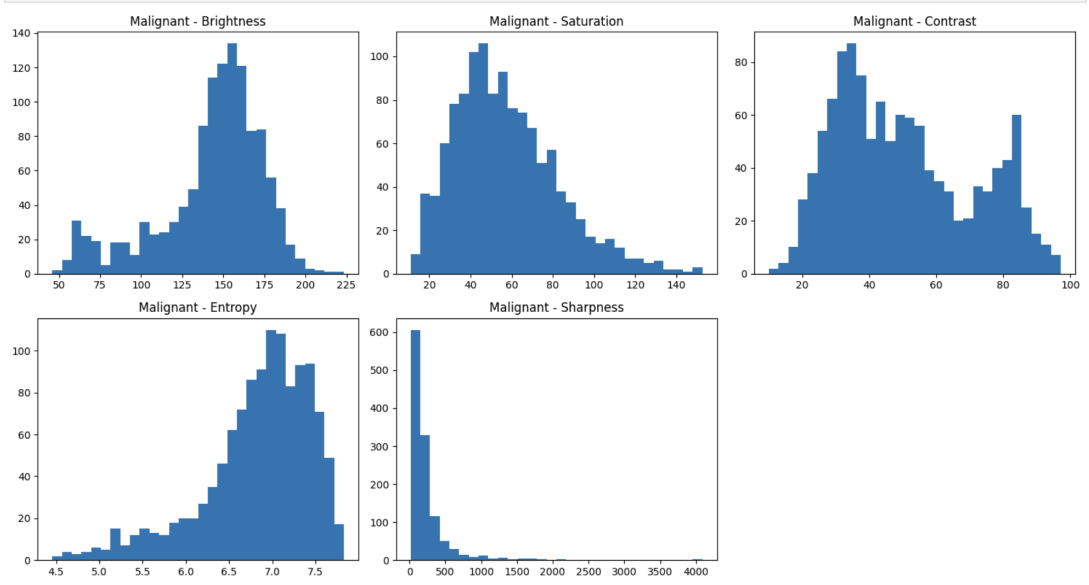
```
benign_malignant
benign       0.903546
malignant    0.096454
Name: proportion, dtype: float64
```

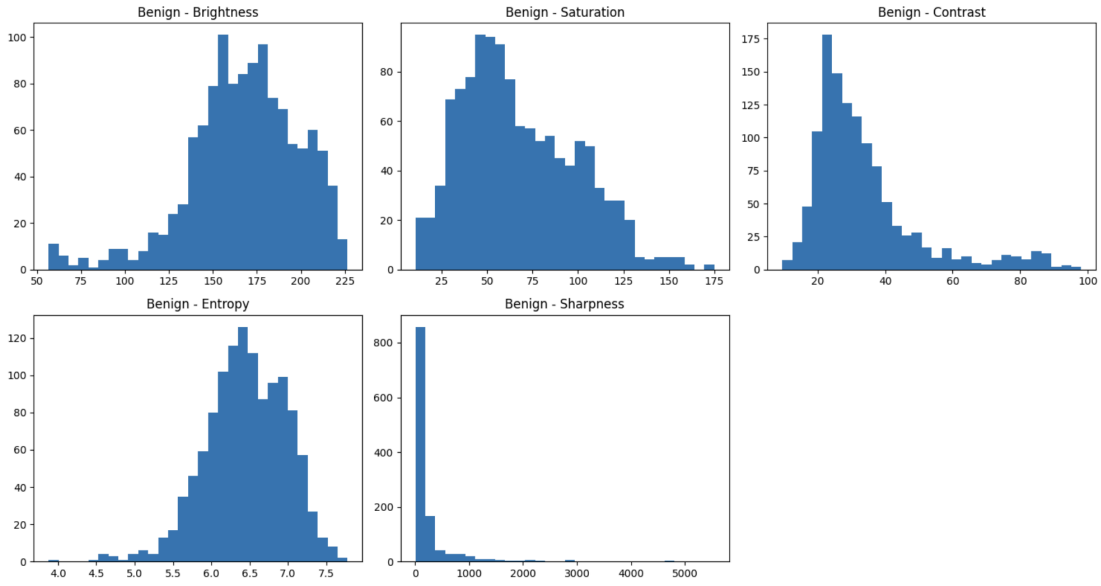


3.2 Image Property Analysis

In this section we will calculate and visualize some basic statistics regarding the image properties:

- Brightness
- Saturation
- Contrast
- Entropy
- Sharpness





	Label	Brightness	Saturation	Contrast	Entropy	Sharpness
0	Malignant	143.095162	57.233710	50.417878	6.831307	251.397841
1	Benign	167.816393	67.617373	33.781292	6.451437	271.714012

According to the histograms and average values displayed above, we can draw the following conclusions regarding the differences between malignant and benign image subsets:

- **Brightness**: Malignant images are on average **darker** than benign ones. This may be due to the darker appearance of melanoma lesions themselves, which reduces overall image brightness.
- **Saturation**: Malignant images exhibit **lower saturation** than benign images. This suggests that benign lesions may have more vivid surrounding skin tones or lighting conditions, whereas melanoma images may appear duller or more uniformly colored.
- **Contrast**: Malignant images show **higher contrast**, indicating a greater difference between lesion and background pixel intensities. This supports the clinical observation that melanoma often stands out more sharply from surrounding tissue.
- **Entropy**: Malignant images have **higher entropy**, reflecting greater texture complexity or variability—consistent with the more irregular and heterogeneous patterns seen in melanoma.
- **Sharpness**: Benign images have **higher sharpness** on average. This could be due to differences in image acquisition quality or lesion borders, as benign lesions often have more regular, well-defined edges.

These insights highlight how basic image statistics can already offer useful distinctions between benign and malignant classes—even before more complex modeling is applied.

4. Summation and Final Conclusion

To improve melanoma classification, we expanded our dataset by incorporating the ISIC 2019 data, converting multi-class labels into a binary format (benign vs. malignant). This increased the dataset size to **66,696 images** and improved class balance, raising the proportion of malignant cases from **1.76% to ~9%**—a meaningful step toward better model performance.

Given the severe class imbalance, standard loss functions like Cross-Entropy may lead to biased predictions. Instead, we recommend using **Focal Loss**, which emphasizes harder, minority class examples and is more suitable for imbalanced medical datasets.

Among structured features, **age** showed the most noticeable influence on the target variable, aligning with medical understanding that melanoma risk increases with age. The age distribution was approximately normal and free of outliers.

Analysis of image-level properties revealed clear differences between benign and malignant lesions:

- **Malignant images** were generally **darker**, **higher in contrast**, and **more texturally complex** (higher entropy).
- **Benign images** were **brighter**, **sharper**, and had higher saturation on average.

These insights show that even basic image statistics can distinguish between lesion types, supporting their use in model development. This foundational work ensures a more robust and fair approach to melanoma classification.

Important note: A more detailed analysis of skin color/type will be conducted in a separate notebook, alongside the fairness evaluation of our model.

5. References

Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data, 5*(180161). https://doi.org/10.1038/sdata.2018.161

Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., & Halpern, A. (2017). Skin lesion analysis toward melanoma detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). *arXiv preprint*. https://arxiv.org/abs/1710.05006

Hernández-Pérez, C., Combalia, M., Podlipnik, S., Codella, N. C., Rotemberg, V., Halpern, A. C., Reiter, O., Carrera, C., Barreiro, A., Helba, B., Puig, S., Vilaplana, V., & Malvehy, J. (2024). BCN20000: Dermoscopic lesions in the wild. *Scientific Data, 11*(1), 641.

Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., Halpern, A., Helba, B., Kittler, H., Kose, K., Langer, S., Liopyris, K., Malvehy, J., Musthaq, S., Nanda, J., ... Soyer, P. (2021). A patient-centric dataset of images and metadata for identifying melanomas using clinical context. *Scientific Data, 8*, 34. https://doi.org/10.1038/s41597-021-00815-z