Target Variable

The target variable in this dataset is “target,” a binary label indicating whether a lesion is benign (0) or malignant (1). This variable is central to the classification task, as the goal is to predict whether a given lesion is malignant based on the provided image and metadata. Early detection of malignancy is critical for timely intervention in skin cancer, which can significantly improve patient outcomes. As a result, this variable is the primary focus of the predictive model, driving the binary classification task.

Predictors

The predictors in this dataset are divided into two categories: image data and metadata. In the application I am building, users will input an image and some personal information. Any metadata not available to the users will not be used as predictors.

a) Image Data

Image: Cropped lesion images (HDF5 format).

Why: Images are essential because skin cancer detection relies on visual patterns such as shape, size, color irregularities, and texture. A deep learning model can process these patterns to identify malignant lesions.

b) Metadata

Patient-level data:

- age_approx: Approximate age of the patient at the time of imaging.
- sex: Sex of the patient.

Reason: Age and sex are important factors in predicting skin cancer risk. Incidence rates of invasive melanoma vary across age and gender groups. In men younger than 50, melanoma rates have declined by 1% per year and stabilized in older men. However, women under 50 experience higher rates of melanoma than their male counterparts, possibly due to trends like indoor tanning use. In contrast, men over 50 have a higher rate of melanoma than women of the same age group, potentially due to occupational and recreational UV exposure. This makes age and sex critical features for identifying skin cancer risk.

Lesion Characteristics:

- anatom_site_general: The general location of the lesion on the body.
- clin_size_long_diam_mm: The maximum diameter of the lesion in millimeters.

Reason: The anatomical site and size of lesions can be readily assessed by users and are important factors for risk assessment. Lesions in sun-exposed areas, for example, might have a higher likelihood of being malignant.

Unavailable Predictors (Not used in the model):

Features that require medical devices or professional analysis, such as tbp_lv_deltaL, tbp_lv_area_perim_ratio, mel_thick_mm, etc., will not be included as predictors, as these details are unavailable to users.

Exploration of the Dataset

a) Variables and Data Types

The dataset includes both structured and unstructured data (images and metadata). Below is an overview of the variables:

Target Variable:

- target: Binary, {0: benign, 1: malignant}.

Patient Demographics:

- age_approx: Numeric (Integer).
- sex: Categorical (String).

Lesion Characteristics:

- anatom_site_general: Categorical (String).
- clin_size_long_diam_mm: Numeric (Float).
- Image Data:

train-image.hdf5: These files contain cropped lesion images.

b) General Dataset Statistics

Train Metadata (train-metadata.csv):

- Row Count: 401064
- Column Count: 55\
- Format CSV

Train Images (train-image.hdf5):

- Image Count: 401064
- Image Format: The images in the dataset are stored as byte arrays within an HDF5 file, rather than typical image files like JPEG or PNG.
- Image size: 15x15 mm cropped field of view.

Metadata Columns Used as Predictors

Columns	ColumnDescription	Data Type
age_approx	Approximate age of the patient at the time of imaging	Integer
sex	Sex of the patient (e.g., Male, Female)	String
anatom_site_general	Anatomical location of the lesion (e.g., upper arm, lower leg)	String
clin_size_long_diam_mm	Maximum diameter of the lesion in millimeters	Float