Comment on 20 Feb 2024:

1. The dataset is very big (i.e. 4G). Suggest working on a subset of the data.
2. Try to construct balanced training / testing dataset if possible.
3. Do you want to build a computer vision (image processing) model? Or simply using the numeric/textual data? image combine with textual data
4. Do you want to build a prediction model? Which attribute is your target? Diagnostic and picture
5. May I have your team number / member name list?

- Determination of data source for your project and preliminary data analysis

- State the problem statement (i.e. what problem do you want to handle?)

- For problem involving the use of prediction model, specify the target variable for prediction

- State the number of attributes (better to be above 20) and number of rows in the dataset

- Any missing values/outliers? how to handle it?

- Define the training data and testing data

- Is there any performance baseline?

<https://www.kaggle.com/datasets/mahdavi1202/skin-cancer>

About Dataset

Summary description  
Published: 7 Jul 2020

* The skin lesions are: Basal Cell Carcinoma (BCC), Squamous Cell Carcinoma (SCC), Actinic Keratosis (ACK), Seborrheic Keratosis (SEK), Bowen’s disease (BOD), Melanoma (MEL), and Nevus (NEV). As the Bowen’s disease is considered SCC in situ, we clustered them together, which results in six skin lesions in the dataset, three skin cancers (BCC, MEL, and SCC) and three skin disease (ACK, NEV, and SEK)
* All BCC, SCC, and MEL are biopsy-proven. The remaining ones may have clinical diagnosis according to a consensus of a group of dermatologists. In total, approximately 58% of the samples in this dataset are biopsy-proven. This information is described in the metadata.
* The images present in the dataset have different sizes because they are collected using different smartphone devices. All images are available in .png format.
* The metadata associated with each skin lesion is composed of up to 26 features. All features are available in a CSV document in which each line represents a skin lesion and each column a metadata feature.
* In total, there are 1,373 patients, 1,641 skin lesions, and 2,298 images present in the dataset. Each image/sample has a reference to the patient and the skin lesion in the metadata.

Number of instances: 2298

Number of attributes: 26

1) patient\_id: Identifier of the patient under study.

2) lesion\_id: Identifier of the lesion or wound under study in the patient.

3) smoke: Whether the patient has a history of smoking or not.

4) drink: Whether the patient has a history of alcohol consumption or not.

5) background\_father: The history of any diseases or health conditions related to the patient's father, including any history of skin cancer or other diseases that may be related to skin cancer.

6) background\_mother: The history of any diseases or health conditions related to the patient's mother, including any history of skin cancer or other diseases that may be related to skin cancer.

7) age: Age of the patient at the time of examination.

8) pesticide: Whether the patient has been exposed to pesticides or other chemicals.

9) gender: Gender of the patient.

10) skin\_cancer\_history: History of skin cancer in the patient's family.

11) cancer\_history: History of cancer in the patient's family.

12) has\_piped\_water: Indicates whether the location or area of the patient's residence has access to piped water or not.

13) has\_sewage\_system: Indicates whether the location or area of the patient's residence has a proper sewage system or not.

14) fitspatrick: Skin tolerance to sunlight.

15) region: The area of the body where the lesion or wound has been examined.

16) diameter\_1: Primary diameter of the lesion or wound.

17) diameter\_2: Secondary diameter of the lesion or wound.

18) diagnostic: The type of lesion or wound is diagnosed.

19) itch: Whether the lesion or wound has itched or not.

20) grew: Whether the size of the lesion or wound has grown or not.

21) hurt: Whether the lesion or wound has hurt or not.

22) changed: Whether the appearance of the lesion or wound has changed or not.

23) bleed: Whether the lesion or wound has bled or not.

24) elevation: Description of the of the lesion or wound relative to the skin surface of the patient.

25) img\_id: Identifier of the image related to the lesion or wound.

26) biopsed: Whether the lesion or wound has been biopsied or not.

