As we come close to submitting the capstone project, a project reflecting my accumulation of learning and understanding in ML, I have realized that there is still much to learn to apply to a real-world context. For example, I am building the first Pay-Per-Use dermatology EMR solution for the Indian market, and a feature to triage patients' complaints would be valuable to doctors. But the system(symabMED.in) is not ready; our system would have to run for 12 to 18 months to collect data in the real world (In India) to accumulate relevant metadata and high-resolution images with defective categorization.

A screenshot of a computer

Description automatically generated with medium confidence

So in the meantime, it would be great if I could build a generic framework with data from the HAM10000 data set, and when we are ready, we could provide the actual data to run the first pass and see the results.

Skin cancers are a group of diseases characterized by abnormal growths or sores that won't heal or changes in the size, shape or colour of a mole. The two main types are melanoma (less common, more dangerous) and non-melanoma skin cancers (common, less dangerous). India has a relatively lower incidence of skin cancer than Western countries, primarily because people with darker skin are less susceptible to skin cancer due to their higher levels of melanin, which helps protect against UV radiation damage.

However, regional variations can play a role in the prevalence of skin cancer in India: our system have folwing metadata in addition to other standard data like age, gender and what is included in the HAM dataset

* **Latitude**: Regions closer to the equator receive more intense sunlight, increasing the risk of skin cancer. Hence, regions in the southern part of India may see a slightly higher incidence rate.
* **Altitude**: Higher altitudes receive more UV radiation. Therefore, areas at higher elevations could potentially have a higher incidence of skin cancer.
* **Outdoor occupations**: In regions where more people work outdoors (like farming communities), there are increased sun exposure and a potentially higher risk of skin cancer.
* **Urban vs rural**: Urban areas might see different rates due to lifestyle differences, including more indoor work and potentially more access to healthcare for early detection.
* **Cultural and social practices**: Clothing, sunscreens, and shade can vary by region and influence skin cancer rates. Some traditional attire covers much of the body, providing natural sun protection

For my capstone project, I plan to write a generic framework for metadata cleaning and normalizing, a routine that takes DataFram, Columns property type, and method to identify and perform like oneHotCoding, replacing Null value etc. I will test this routine on HAM10000, the year 2020 dataset.

For Images, write a routine that standardizes the image files and removes the record of the corrupt file ( including metadata). Test this routine on HAM datasets.

This generic framework should produce pre and post-report and visitation of row data and finalize the data set to be used in finding the suitable ML model and test it gainest a test HAM10000 2020 test dataset. Again as part o the project, I will write pre and post-report and visitation. If all works, I will have code that can take any DataFram for image classification, perform data visualization and preparation for ML modelling, go through a selection of Models and provide API that can be used in the real world as a service