Determining Skin Cancer from a large collection of muli-source dermatoscopic images of pigmented lesions.

Training of neural networks for automated diagnosis of pigmented skin lesions is hampered by the small size and lack of diversity of available dataset of dermatoscopic images. We tackle this problem by releasing the HAM10000 ("Human Against Machine with 10000 training images") dataset. We collected dermatoscopic images from different populations, acquired and stored by different modalities. The final dataset consists of 10015 dermatoscopic images which can serve as a training set for academic machine learning purposes. Cases include a representative collection of all important diagnostic categories in the realm of pigmented lesions: Actinic keratoses and intraepithelial carcinoma / Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses, bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, vasc).

More than 50% of lesions are confirmed through histopathology (histo), the ground truth for the rest of the cases is either follow-up examination (follow\_up), expert consensus (consensus), or confirmation by in-vivo confocal microscopy (confocal). The dataset includes lesions with multiple images, which can be tracked by the lesion\_id-column within the HAM10000\_metadata file.

1. What is the problem you want to solve?  **I am looking to solve two problems. The first would be to see if there is any correlation with the numerical data e.g., relationship between localization and the age of the patient (probably a stretch). The second would be to categorize pigmented lesion into diagnostic categories (dx). The categories are Actinic keratoses and intraepithelial carcinoma / Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses, bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, vasc).**
2. Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn’t have otherwise? **I would believe my client would be research labs or hospitals interested in how the data behaves. Using the machine learning model would assist/provide more confidence in doctors to determine which diagnostic category the lesion would fit. I envision my client after receiving the results of this analysis to collect more data especially if i can conclude correlation within the data.**
3. What data are you going to use for this? How will you acquire this data? **The final dataset consists of 10015 dermatoscopic images which can serve as a training set for academic machine learning purposes.**

**In the Metafile, this is the dx column**

**Cases include a representative collection of all important diagnostic categories in the realm of pigmented lesions: Actinic keratoses and intraepithelial carcinoma / Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses, bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, vasc).**

**Break down of the terminology ( in the metafile this is the dx-type column)**

**50 % of the data are confirmed through Histopathology (histo) - is the examination of tissues from the body under a microscope to spot the signs and characteristics of disease.**

**The remaining are follow-up examination (follow-up) , expert consensus (consensus) or confirmation by in-vivo confocal microscopy (confocal).**

1. In brief, outline your approach to solving this problem (knowing that this might change later). **My approach in solving this problem(s) would be the following:**
2. **The data from the most part seems to be clean/complete e.g., no blanks or missing data therefore, I envision very little data wrangling to be done. I believe I immediately jump into the the engineering \_\_\_\_\_ analysis term.**
3. **For the machine learning portion, the pictures have background information, e.g., each line item points to a picture. Perhaps my model can determine/guess where in the body the cancer is located or guess age/sex.**
4. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck. **The deliverable would be code and a paper.**

**See below for sample of the data e.g., data header**

**lesion\_id image\_id dx dx\_type age sex localization**

**HAM\_0000118 ISIC\_0027419 bkl histo 80 male scalp**

**HAM\_0000118 ISIC\_0025030 bkl histo 80 male scalp**

**HAM\_0002730 ISIC\_0026769 bkl histo 80 male scalp**

20 Questions to ask prior to starting data analysis.

1. Who is the audience that will use the results from the analysis? (board members, sales people, customers, employees, etc)
2. How will the results be used? (make business decision, invest in product category, work with a vendor, identify risks, etc)
3. What questions will the audience have about our analysis? (ability to filter on key segments, look at data across time to identify trends, drill-down into details, etc)
4. How should the questions be prioritized to derive the most value?
5. Identify key stakeholders and get their input on interesting questions
6. Who should be able to access the information? think about confidentiality/ security concerns
7. Who will develop and maintain the report?
8. What information will be on each report?
9. What reports currently exist in another format? What changes might be made to existing reports?
10. What ETLs or stored procedures need to be developed, if any?
11. What database enhancements are required to meet reporting requirements?
12. When will each report be delivered?
13. What is the frequency of updates required for the data? to ensure currency
14. Which data sources are available to work with?
15. Do I have the required permissions or credentials to access the data necessary for analysis?
16. What is size of each data set and how much data will I need to get from each one?
17. How familiar am I with the underlying tables and schema in each database? Do I need to work with anyone else to understand the data structure
18. Do I need all the data for more granular analysis, or do I need a subset to ensure faster performance?
19. Will the data need to be standardized due to disparity?
20. Will I need to analyze data from external sources, which resides outside of my organization’s data?