Preliminary Assignment: Data Exploration Summary

Group P

February 26, 2026

1 Group Information

- **GitHub Repository:** `https://github.com/alexandrup2015-rgb/2026-PDS-P`
- **Team Members:**

Name	GitHub Username	Annotator ID	Email
Bartosz Kaluski	bakal3110	bark	bakal@itu.dk
Dan Bivol	Dan-PythonMaster	dbiv	dbiv@itu.dk
Alexandru Poputoaia-Ungureanu	alexandrup2015-rgb	apop	apop@itu.dk
Laura Jakubiak	laurajakubiak	lauja	lauja@itu.dk
Sergiu Cristian Baba	chrisscifi	sbab	sbab@itu.dk

2 Dataset Overview

Images used in this assignment were obtained from the PAD-UFES-20 dataset (Pacheco et al., 2020). A specific subset of 116 images, including their metadata, assigned to our group(Group P) was extracted from the dataset. Images present close-up views of varying skin lesions.

3 Data Exploration & Observations

3.1 Key Statistics and Relationships

Several variables demonstrate clear interdependencies within the data set, particularly regarding symptoms linked to lesion severity. We observed that growth, elevation, bleeding, and pain frequently co-occur. This suggests that as a skin lesion progresses, patients are more likely to exhibit multiple clinical indicators simultaneously.

Anatomical location also emerged as an interesting diagnostic indicator within our specific subset. Notably, 100% of the lesions located on the nose in Group P were diagnosed as Basal Cell Carcinoma (BCC). This perfect correlation suggests that for our group, the nose represents a high-risk site where persistent lesions are almost exclusively malignant.

A significant relationship was also identified between clinical symptoms and the decision to perform a biopsy. Lesions presenting with high-risk features—such as bleeding, pain, elevation, or recent change—showed a higher frequency of biopsy. Age also appears to be a contributing factor; older patients were more likely to undergo biopsy, reflecting increased risk with age.

3.2 Diagnostic Distribution and Real-Life Comparison

We analyzed the diagnostic distribution of our 116 samples. In the table below, we compare our confirmed cancer cases (57 total) against global dermatological statistics.

Diagnostic Type	Count	% of Group P	% of Cancer Types	Global Average
Basal Cell Carcinoma (BCC)	45	38.8%	78.9%	75–80%
Squamous Cell Carcinoma (SCC)	10	8.6%	17.5%	15–20%
Melanoma (MEL)	2	1.7%	3.5%	1–4%

Table 1: Confirmed Cancer Diagnoses in Data Set vs. Global Average

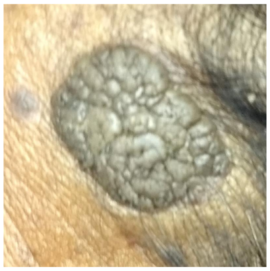
4 Annotations Summary

We manually labeled images for hair density (0–3) and pen marks (0/1).

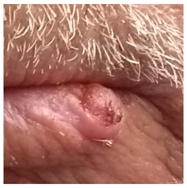
- **Hair Density:** Full consensus was reached on **44.8%** of images.This low percentage could be linked to what ”a little” vs”a lot” of hair means to each of us.
- **Pen Marks:** High consensus of **92.2%**. Approximately 25% of images contained markings.

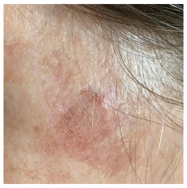
4.1 Notable Samples

Hair Density Conflict: In sample PAT_1286_1000_517, one member (sbab) rated it as 3, while others recorded a 0. This demonstrates how lighting and zoom can make fine hairs look like significant density to some and non-existent to others.

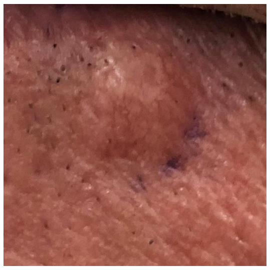


Consensus Outliers: Samples PAT_679_1286_677 and PAT_1767_3340_959 had the highest variation. The ambiguity of the lesion borders often led to different interpretations of the surrounding features.





Marker Detection: Faint blue ink in sample PAT_16_24_691 led to a 3-vs-2 split decision. Such artifacts are easily missed when focusing on the biological characteristics of the lesion.



5 Conclusion

Our manual annotation process for hair density and surgical markings exposed the inherent subjectivity of visual data analysis. While artificial features like pen marks reached a somewhat common recognition (92.2%), biological features like hair density became less consistent (44.8%). This discrepancy shows the necessity of clear annotation rubrics and multi-observer validation in medical imaging tasks.