Name	Maksims Petruhins	Jonas Frederik Hansen	Nyasha Margaret Mazimba	Klara Muce Ronnenberg	Jakub Stopiak
Github Username	Petruhinmaksim369-sketch	Jonqz	NyashaMM	KlaraRonne	jakubstopiak
ITU ID	makpe	jfha	nyma	klmr	stop

GitHub repository: https://github.com/Jonqz/2026-PDS-GroupN.git

---

Summary

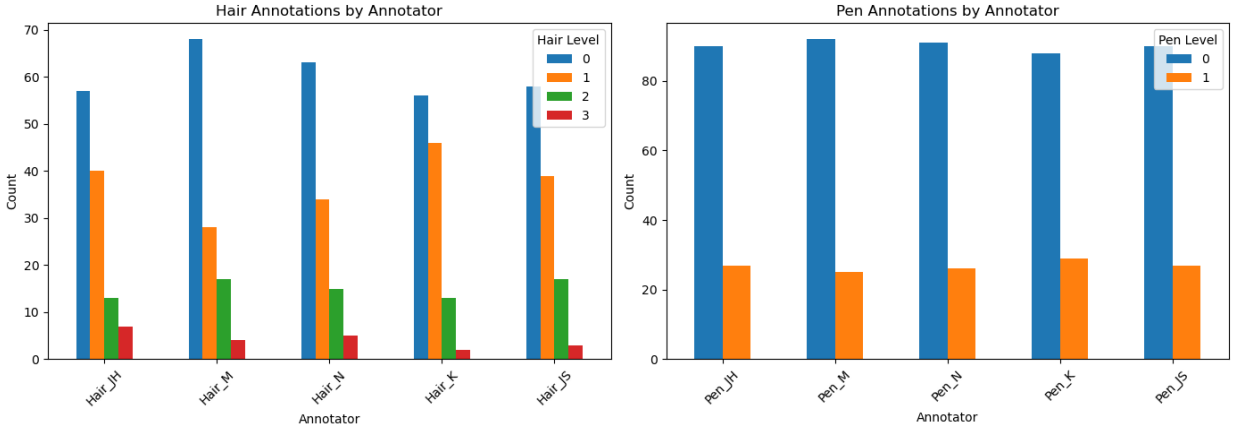
The lack of medical tools and experts in emerging countries and in remote/rural areas, has increased the need for alternative approaches to skin cancer detection. That is the reason we have been introduced to the PAD-UFES-20 dataset, consisting of clinical images of skin lesions collected using different smartphone devices, together with associated patient clinical data. The use of smartphones makes the dataset more realistic and accessible compared to dermoscopic datasets, as smartphones are widely available worldwide.

The dataset was collected from 2018 to 2019 and there are 1373 patients, 1641 skin lesions, 2298 images and 2152 masks. Although the dataset provides extensive patient information (such as smoke or drinking habits, age, family background etc.), our focus in this assignment was on the amount of hair and pen marks in the pictures.

Five group members independently reviewed all images and annotated the amount of hair using a scale from 0 to 3 (0 representing no visible hair and 3 representing a large amount of hair). During this process, we noticed that evaluating hair quantity is somewhat subjective., Our perception varied not only between annotators but also over time, as our internal reference changed after reviewing a larger number of images. In contrast, it was fairly easy to define whether there was a pen mark or not, therefore less open to interpretation.

For hair annotations, the subplot shows how often each hair level was assigned by each annotator, enabling comparison of scoring tendencies and consistency. Similarly, the second subplot presents the distribution of pen annotation levels by annotator, again showing the count of each level as grouped bars. Together, the two plots provide a clear comparative overview of how annotation levels are distributed across annotators for both categories,

helping identify patterns, imbalances, or systematic differences in scoring behavior.



Further, we evaluated the inter-annotator agreement using Cohen's Kappa score which measures the level of agreement between two annotators while correcting agreement that could occur by chance. The score ranges from –1 to 1, where 1 indicates perfect agreement, and 0 indicates agreement equivalent to chance. For the hair annotations, the pairwise Cohen's Kappa score between annotators (JH, M, N K and JS) were:

- JH-M: 0.76
- JH-N: 0.78
- JH-K: 0.78
- M-N: 0.67
- M-K: 0.68
- N-K: 0.75
- JS-M: 0.61
- JS-N: 0.63
- JS-K: 0.66
- JS-JH: 0.68

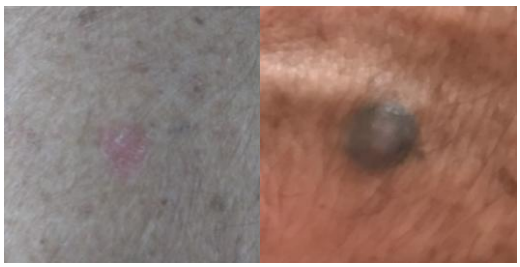
According to commonly used interpretation guidelines, scores between 0.61 and 0.80 reflect substantial agreement. Although we experienced subjectivity during annotation, the relatively high Kappa scores suggest that our evaluations were still fairly consistent across annotators. The slightly lower scores (0.67 and 0.68) may reflect cases where distinguishing between similar levels of hair density, such as 1 and 2 was more challanging.

For the pen mark annotations, the pairwise Cohen's Kappa scores were:

- JH–M:0.80
- JH–N: 0.88
- JH–K: 0.91
- M–N: 0.77
- M–K: 0.81
- N–K: 0.88
- JS-M: 0.80
- JS-N: 0.83
- JS-K: 0.95
- JS-JH: 0.86

These values range from substantial to almost perfect agreement (values above 0.81 are typically interpreted as almost perfect agreement). The higher agreement compared to hair annotations confirms our earlier observation that identifying pen marks was more objective and less open to interpretation, as it was treated as a binary decision. Overall, the Cohen's Kappa analysis quantitatively supports our qualitative findings: while hair annotation introduced some degree of subjectivity, inter-annotator consistency remained high. Pen mark detection, being more clearly defined, achieved even stronger agreement across annotators.

Another challenge during annotation was image quality. Variations in lighting, focus, and resolution (e.g., PAT_1364_1246_447 or PAT_1364_1246_420) sometimes made it difficult to distinguish between light-colored hair and other skin features such as dryness (e.g., PAT_1247_852_178). These ambiguities contributed to disagreements between annotators in certain cases (e.g., PAT_1780_338_999).



We also reflected on the accuracy of the provided segmentation masks. In some instances, the lesion appeared unclear, or the mask did not fully correspond to the visible lesion area (e.g.,

PAT_658_1266_149 or PAT_1464_1612_284). This could potentially affect later feature extraction or model training steps.