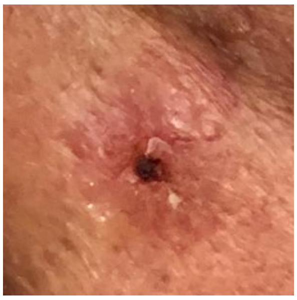
Project Medical Imaging (Projects in data science - Spring 2026)
Group E’s summary
Participants (their github username): Victoria Surdu (VictoriaSrd), Andreea Fedorovici (fedoandreea), Viktória Kapicáková (VikyKapike), Wladimir Lawrow (Wl3-2), Shayan Soltani (shayanst3000)

The data that we explore in this project is derived from PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. Overall, the dataset is pretty diverse, ranging from images with skin lesions that look like regular moles to images with skin lesions that look alarmingly infected.

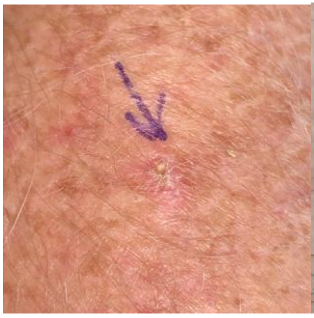
We analyzed 117 images from the dataset and annotated visual structures as either pen marks or hair strands. Pen marks were classified using a binary system (0 or 1) to indicate absence or presence, as their appearance was distinct and consistent. In contrast, hair strands were graded on a scale from 0 to 3 to reflect varying levels of hair density and distribution, requiring a more nuanced assessment.

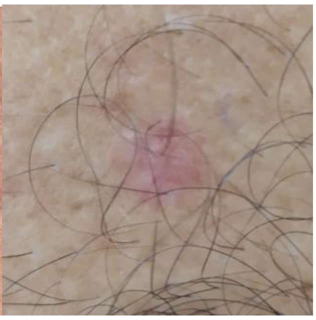
Having been taken with smartphones, there were also some blurry images in the data frame. In this kind of images the color is almost the only distinguishing feature, as in the example:



This can raise real issues for a machine learning model that had training data consisting of high-quality, sharp images because it might be confused and might interpret the blur itself as a feature of a skin regression. On the other hand, it might give a False Negative because it mistook the melanoma for a regular mole, which, in medical diagnostics, is generally considered a much worse failure than a false alarm.

Classification was based on morphological features, including thickness uniformity, edge definition, curvature, continuity, and texture. Pen marks exhibited consistent width, sharp boundaries, and uniform color intensity, whereas hair strands showed natural variation in thickness, tapered ends, and softer, irregular curvature.





For instance, in the first picture, the image was graded 2 or 3 for hair strands by the group members because we had differing views on the extent of visible hair. In the second picture, we marked this image as 0 for pen marks due to the absence of any visible pen strokes, with full consensus among the team.

Looking at the annotations, the averages (average hair annotation = 1.03 and average pen-mark annotation = 0.25) suggest that hair occurs at a noticeable level across the whole sample, whereas pen marks are concentrated in a smaller number of cases. The averages also carry information about annotation confidence and ambiguity. Because five annotators were used, many image-level averages appear in increments (e.g., 0.2, 0.4, 0.6, 0.8), which reflect partial agreement between annotators. The annotators fully agreed on 91.5% of images when it came to pen marks and on 56.4% of images when it came to hair presence. This means that pen-mark identification was generally reliable and reproducible, while hair showed substantially more disagreement, suggesting that it was harder to assess consistently. It is worth mentioning that the binary scale used for pen markings may be a limitation in borderline cases because it forces annotators to choose between only two options (present or absent) and does not allow them to express uncertainty. The averages are especially useful for identifying images that may need review or exclusion. Values near the extremes (e.g., 0 or full agreement) usually indicate clear visual evidence, whereas intermediate averages point to uncertainty or different annotator interpretations.

During the annotation process, we also noticed that lighting and image quality had a strong influence on our judgements. Since the images were taken with smartphones under different conditions, brightness, shadows and blur often varied. In some cases, this uncertainty made it challenging for us to decide how to annotate the image and could also make it difficult to assess whether a lesion is cancerous or not. These observations prove how important consistent image quality is, as small variations in lighting or color can significantly influence interpretation and potentially affect the conclusions.

Building on our annotation results, since we didn't fully agree on the hair presence (only 56.4% agreement), just relying on simple percentages isn't enough. Moving forward in the project, we should use a more robust statistical tool like Cohen's Kappa. This formula will help us measure our real agreement by accounting for lucky guesses and chance.

Furthermore, identifying these pen marks and hair strands is very important to prevent "shortcuts" in our machine learning models. For example, if an AI model sees a pen mark, it might just assume the lesion is dangerous simply because doctors usually circle suspicious moles. We need to handle these artifacts now, so our future AI model learns to evaluate the actual skin lesion, not just the pen marks.

As for the medical aspect of the project, these images succeed in representing the difference between regular moles and melanoma signs on the skin: we can see changes in the surface of a mole like oozing, bleeding, or the appearance of a lump or bump, which indicate the presence of skin cancer.

Our annotation results show that hair was relatively common and more difficult to evaluate consistently, reflected in the lower agreement among annotators. In contrast, pen marks were less frequent and identified with high consistency. We also observed that variations in lighting, blur and color could strongly influence visual interpretation and increase uncertainty in borderline cases. This showed us that identifying and handling these factors is important to reduce ambiguity and support more reliable analysis later in the project.

Our annotations csv: Annotations_GroupE

Our GitHub repository link: https://github.com/VictoriaSrd/2026-PDS-E.git