IT University of Copenhagen – Bachelor of Science in Data Science

Projects in Data Science - 2nd semester

Group Letter: D

Participants: Hugo Olmo Lorience (huol@itu.dk), Bartosz Pliszka, Nikita Konjushenko, Victor Mathias Wiese Stage, Nagib Mahfuz Prince

Exploratory data analysis

In this preliminary analysis, Group D explored a given sample of a dataset of dermoscopic skin lesion images, including segmentation masks and associated metadata. During the preliminary exploration, we observed variation in image characteristics that could potentially influence later analysis. Many images contained visible body hair, while others were hair-free. Additionally, a number of images included pen markings drawn on the skin surface, whereas others did not. We also noted that a small portion of the images appeared slightly blurry, which could reduce feature clarity and potentially affect the performance of artificial intelligence models. Furthermore, lighting conditions varied across images; while most were adequately illuminated, some exhibited minor inconsistencies that could influence visual interpretation and model accuracy.

In addition to image characteristics, the metadata file provided demographic and clinical information, including patient gender, diagnostic category, anatomical lesion location, parental medical history (mother and father), and lifestyle factors such as smoking and alcohol consumption. The gender distribution was relatively balanced, with approximately 35.9% females, 32.5% males, and 31.6% categorized as NaN (Undefined), indicating no strong gender imbalance within the sample. Information about family history and lifestyle allows us to better understand potential risk factors associated with different lesion types.

Since hair presence and pen markings may introduce visual artifacts that affect model performance, we manually annotated these features to systematically quantify their distribution within the dataset. The annotations are linked further down below in a public GitHub repository.

The next step in the research was the selection of categories for further analysis. The following categories address two fundamental aspects of the dataset: the medical classification of lesions and their anatomical occurrence, thereby offering a more comprehensive overview of the sample. By examining the diagnostic distribution, we gain insight into the prevalence of different types of skin conditions within the dataset. At the same time, analyzing the anatomical distribution allows us to identify patterns related to lesion location across the body. Together, these perspectives contribute to a deeper understanding of the dataset's structure and variability, which is essential for informing subsequent data preprocessing and modeling decisions.

Diagnostics distribution

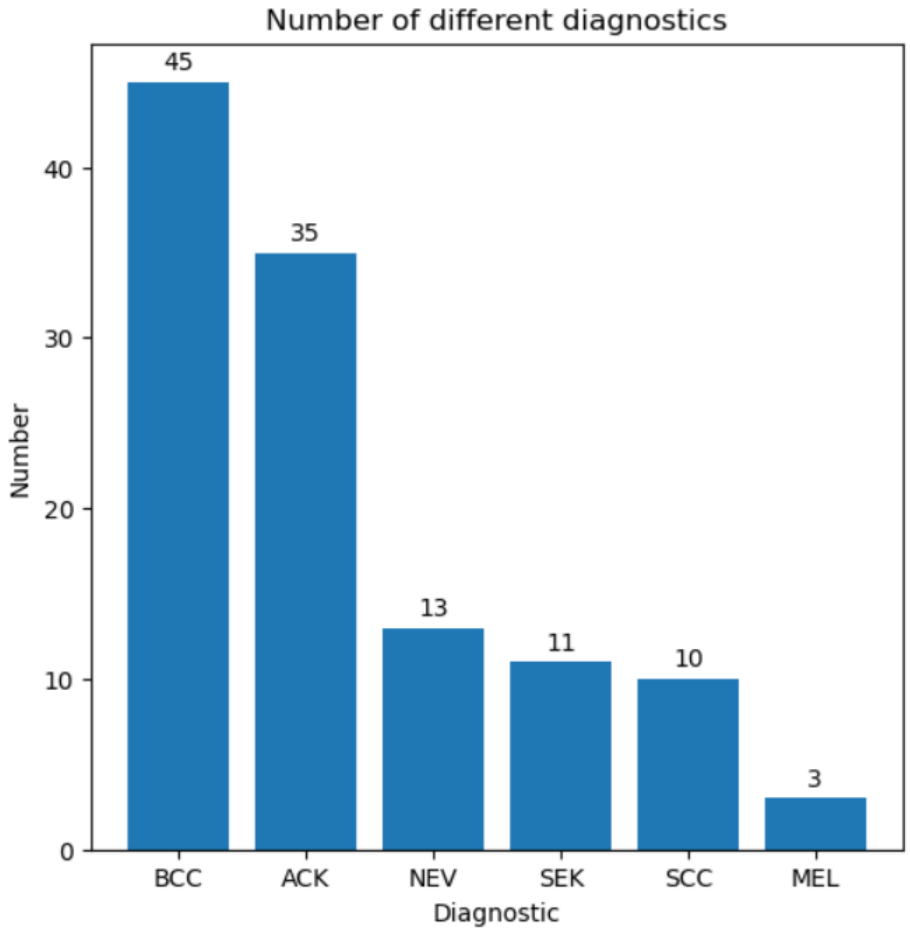

Figure 1. Number of different diagnostics

Figure 1 illustrates the distribution of diagnostic categories among patients in the dataset, representing different types of skin cancer. The most common diagnosis is Basal Cell Carcinoma (BCC), with 45 reported cases. Basal Cell Carcinoma and Actinic Keratosis (ACK) together account for approximately 68% of the dataset, corresponding to 117 cases in total. In contrast, Melanoma (MEL), which is considered one of the most serious forms of skin cancer, represents the smallest proportion, with only 3 cases.

On the whole, the diagnostic distribution indicates that the dataset is dominated by BCC and ACK cases, while more severe conditions such as Melanoma are comparatively rare. This imbalance may have implications for future modeling tasks.

Lesions in certain regions distribution

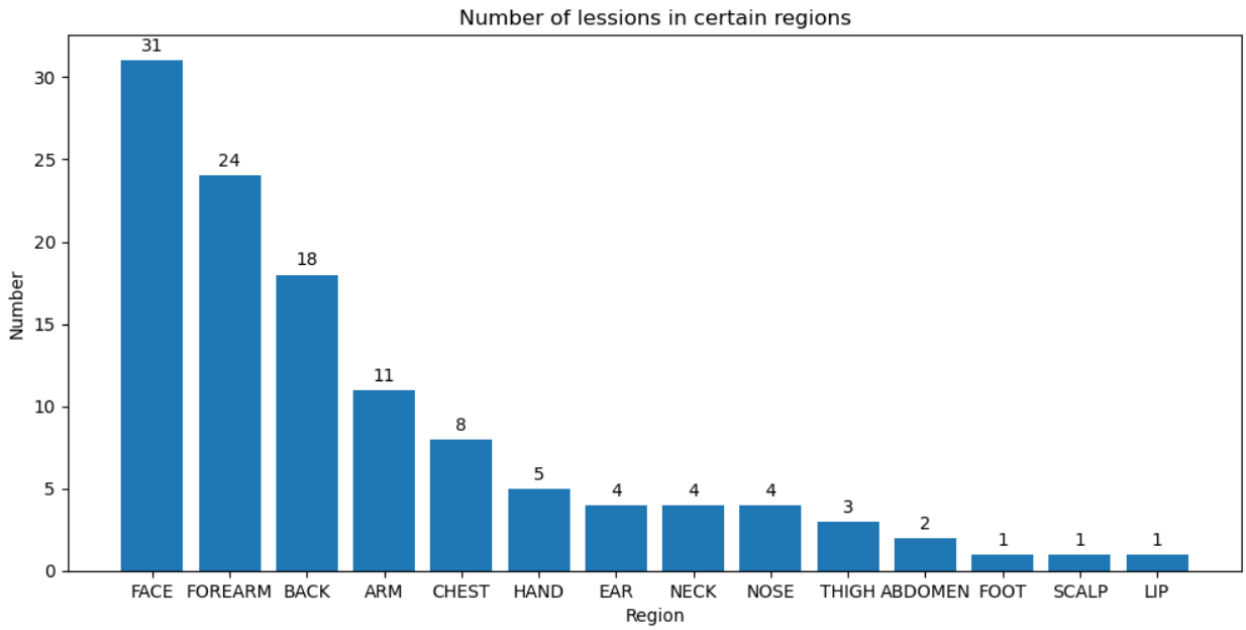

Figure 2. Number of lesions in certain regions

The bar chart (Figure 2) illustrates the distribution of lesions across different body regions. The x-axis represents the body regions, while the y-axis indicates the number of lesions in each region. As shown in Figure 2, lesions are most common on the face (31 cases), followed by the forearm and back. After the top three regions, there is a steep decline in the number of lesions, followed by the arm (11 cases), chest (8 cases), and hand (5 cases). The other regions show relatively low frequencies; the lowest counts were identified in the foot, scalp, and lip, each with a single case.

Overall, the figure highlights an uneven distribution of lesions, where the highest concentration is observed in the face and upper body regions, with vastly fewer cases in the remaining areas.

Conclusion

In summary, the exploratory analysis highlights meaningful variation in both diagnostic categories and anatomical lesion distribution within the sample. The data show a predominance of certain diagnoses, particularly Basal Cell Carcinoma and Actinic Keratosis, as well as a concentration of lesions in the face and upper body regions. Additionally, variation in image quality and metadata characteristics suggests potential factors that may influence future preprocessing and modeling decisions. Overall, this preliminary exploration provides a solid

foundation for subsequent annotation validation, feature extraction, and machine learning model development.

Annotations

Link to GitHub: https://github.com/nkon-no/2026_PDS_GroupD/tree/main

Names and usernames:

- Annotator 1: Nikita Konjushenko (Username: DragTown)
- Annotator 2: Victor Mathias Wiese Stage (Username: StageMan-hub)
- Annotator 3: Hugo Olmo Lorience (Username: hugoolmo63)
- Annotator 4: Bartosz Pliszka (Username: CzescBartek)
- Annotator 5: Nagib Mahfuz Prince (Username: NM-Prince)