# Project Proposal:
# Skin Cancer Classification inside a Shiny App using a CNN

ESC403: Introduction to Data Science

Manuel Pfister, 19-659-408
Rahel Eberle, 14-715-080
Carolyne Huang, 22-740-351
Roman Stadler, 18-915-033

31.03.2023

## 1 General Topic

Skin cancer is one of the most common types of cancer globally. Early detection of skin cancer can save lives, and image classification may provide an efficient and objective method of diagnosing skin lesions for clinicians and dermatologists. An example of these lesions is displayed in figure 1.

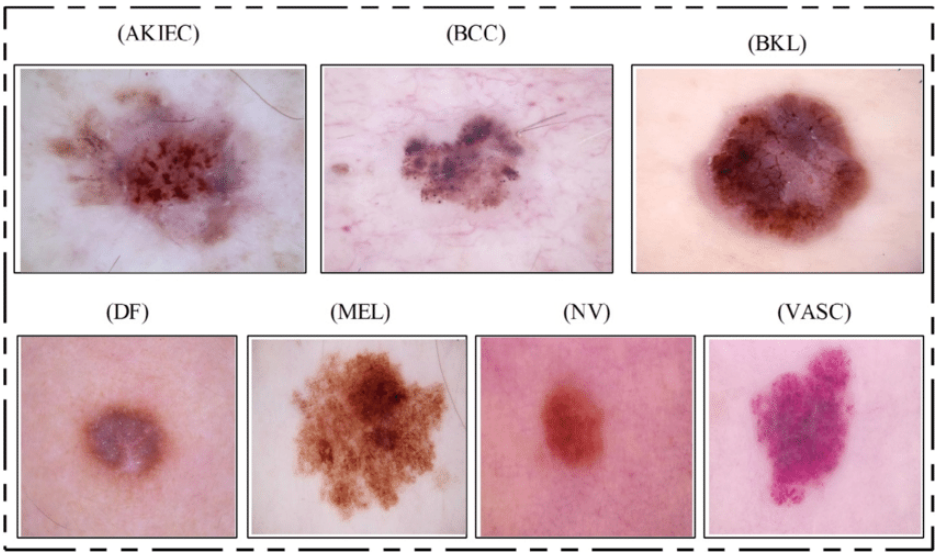


Figure 1: 7 types of skin lesions included in the HAM10000 dataset [1].

## 2 The Data

We are planning to use the HAM10000 dataset, which contains over 10,000 multi-source dermatoscopic images of common pigmented skin lesions. The dataset consists of the following seven different classes of skin 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) [2].

For our project, we will classify these lesions into cancerous (basal cell carcinoma, melanoma) and non-cancerous (benign keratosis-like lesions, dermatofibroma, melanocytic nevi, vascular lesions). We will remove the borderline cases (Actinic keratoses and intraepithelial carcinoma/ Bowen's disease).

# 3 Data Processing

First, we will perform an EDA on the dataset and visualize distributions of the classes, types of skin lesions etc. to gain insight into how the data is structured.

We will remove the lesions categorized as 'Actinic keratoses and intraepithelial carcinoma / Bowen's disease (akiec)' from the dataset because they are precanceroses and therefore borderline cancerous.

Afterwards, the following data-preparation steps will be taken:

1. Removing duplicates: The HAM10000 dataset contains some images that are duplicated, though with different file names. As a result, the same image may be present in both the training and test sets. Removing these duplicates will ensure that the training and test sets are truly independent.
2. Removing irrelevant variables: The metadata associated with each image in the HAM10000 dataset contains many variables, some of which may not be relevant for skin lesion classification.
3. Resizing images: The images in the HAM10000 dataset are of varying sizes, which can be problematic for neural network training. We will resize all images to a consistent size, such as 224 x 224 pixels, which is commonly used for training deep neural networks.
4. Normalizing pixel values: We will scale the pixel values so that they are between 0 and 1, which can help to improve training performance.

# 4 Research Questions and goals

1. Data analysis questions: How accurately can we classify skin lesions into malignant and benign using our model?
2. Practical implications: Create a user friendly interface where pictures of lesions can be uploaded.

# 5 Analysis Techniques

## 5.1 Classification

The image classification will be carried out using a Convolutional Neural Network (CNN). The proposed CNN architecture will consist of multiple convolutional layers, followed by two fully connected layers. We will use batch normalization after each convolutional layer to normalize the input to the activation function. A kernel will be used on the images, and the activation function will be ReLU, though the kernel size and the activation function could be tweaked if the performance of the model is not

satisfactory. The input neuron number is highly dependent on the images, but the output layer will have two neurons for classification (malignant and benign).

The training will be done over many periods using batches of the data, and a potential loss function could be cross-entropy loss. Depending on the model performance, we will include data augmentation techniques such as rotation, flipping, and scaling to increase the diversity of the training data, and consequently to obtain better results. To validate the model, we will split the dataset into training and validation sets, with a ratio of 80:20. We will use the training set to train the model and the validation set to evaluate the model's performance.

After training is complete, we will evaluate the performance of the model on the testing set. We will use standard evaluation metrics such as accuracy, precision, recall, and F1-score to assess the model's performance. We will also use techniques such as confusion matrices and ROC curves to visualize the model's performance.

### 5.2 Deployment

We propose to deploy our machine learning model in a Shiny or Dash framework to enable direct image prediction. This approach will allow users to upload an image and receive an immediate prediction from the model. We believe that this deployment strategy will provide a more user-friendly and efficient experience.

The primary advantage of deploying our model in a Shiny or Dash framework is the ease with which users can interact with the model. Rather than having to manually enter data or code, users can simply upload an image and receive a prediction. This process is much more intuitive and streamlined, which will likely increase the model's usage and adoption.

Additionally, the Shiny or Dash framework provides a user-friendly interface for visualizing and presenting model outputs. This means that we can display the predicted class and probability score in a way that is easy to understand and interpret. Furthermore, the framework provides the flexibility to add other relevant information, such as explanations or contextual data, to enhance the user's understanding of the prediction [3, 4].

## References

[1] Muhammad Khan et al. "Skin Lesion Segmentation and Multiclass Classification Using Deep Learning Features and Improved Moth Flame Optimization". In: *Diagnostics* 11 (Apr. 2021), p. 811. DOI: 10.3390/diagnostics11050811.

[2] Philipp Tschandl. *The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions*. Version V4. 2018. DOI: 10.7910/DVN/DBW86T. URL: https://doi.org/10.7910/DVN/DBW86T.

[3] R Studio. *Installing shiny for python*. URL: https://shiny.rstudio.com/py/docs/install.html.

[4] *Dash Documentation User Guide — Plotly*. URL: https://dash.plotly.com/.