Assignment 3.1: Data Summary Section Draft

GitHub: https://github.com/mrusd/AAI-590-Team-8/blob/main/3-Code/1-EDA-Preprocessing-Images-Dataset.ipynb

- What **variables** are present in your **dataset** and what **are their datatypes**?
  - Overall, 3 datasets have been employed in our study.
    - https://www.kaggle.com/datasets/stevepython/osteoporosis-knee-xray-dataset
    - https://www.kaggle.com/datasets/sachinkumar413/osteoporosis-knee-dataset-preprocessed128x256
    - https://www.kaggle.com/datasets/mrmann007/osteoporosis
  - The dataset comprises of 'labeled images' only. Thus, only the target label is available: 'Osteoporosis' vs. 'Normal' (i.e., no presence of osteoporosis).
  - For bookkeeping purposes only, we have created a dataframe containing the following characteristics of the images (none of these will be used as variables in our study, except for the 'encoded labels'):
    - **filepaths:** Paths to the image files.
    - **width:** Width of the images.
    - **height:** Height of the images.
    - **aspect_ratio:** Aspect ratio (width/height) of the images.
    - **zero_pixel_percentage:** Percentage of zero (black) pixels in the images.
    - **mean_pixel_value:** Mean pixel value of the images.
    - **std_pixel_value:** Standard deviation of the pixel values.
    - **file_size:** Size of the image files on disk.
    - **image_format:** Format of the image files (e.g., JPEG, PNG).

        - **original_vs_processed:** Flag indicating whether the image is original or processed.
        - **labels:** Original labels of the images (e.g., 'Osteoporosis', 'Normal').
        - **encoded_labels:** Numerical encoding of the labels.

- What **issues** were present in your **dataset** and **what steps** were taken to handle them (missing data, regularization, etc.)? What is your best guess for the source of these data issues?
    - The dataset does not have any missing data, it's balanced, but small. Issues include the presence of black borders in the images, images of different sizes, and images with different contrasts or brightness.

- **What steps** were taken to handle them (missing data, regularization, etc.)?
    - **Normalization and Resizing:**
        - All images were normalized by dividing the pixel values by 255 to bring them into the range [0, 1].
        - Images were resized to a consistent size, such as 224x224 pixels, suitable for CNN input.
    - **Cropping Black Boundaries:**
        - Images with significant black boundaries were cropped to remove unnecessary black regions before resizing.
    - **Converting Grayscale Images to Single-Channel:**
        - Converted grayscale images saved in RGB mode to single-channel grayscale images.
    - **Texture Enhancement:**

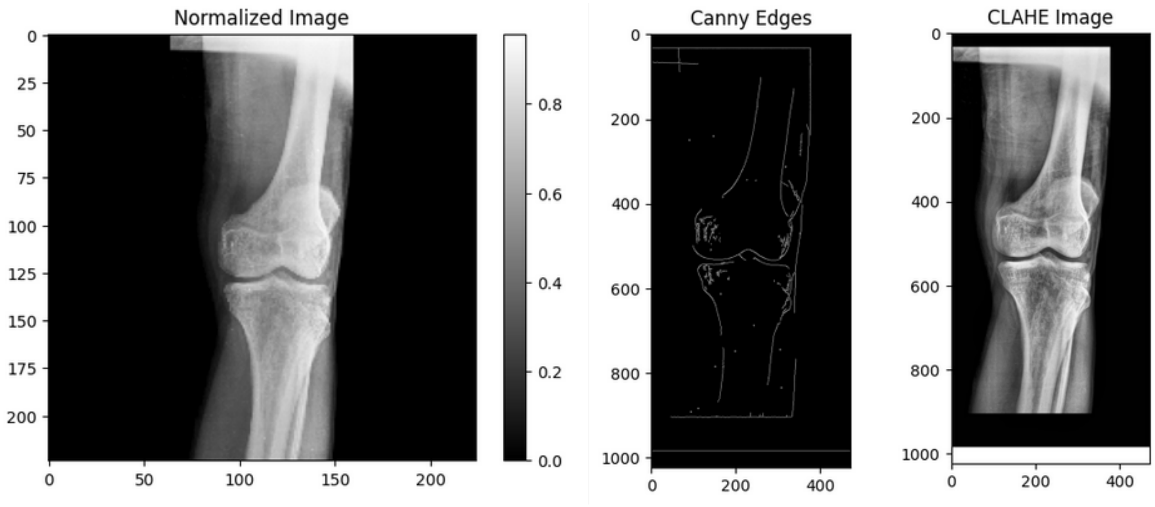
        - Applied techniques like CLAHE (Contrast Limited Adaptive Histogram Equalization) to enhance the texture of bones.

- What is your best guess for the source of these data issues?
    - Differences in imaging equipment and different acquisition settings:
    - Differences in exposure settings (time, radiation dose, and detector sensitivity) can affect image contrast and brightness.
    - Variation in how patients are positioned during the X-ray procedure can lead to varying amounts of black borders in the images.
    - Patient movement during the X-ray can introduce artifacts and affect image quality.
    - Inconsistent manual cropping or framing by technicians can lead to varying image sizes and black borders.
    - Different image formats and compression settings (e.g., JPEG vs. PNG) can introduce artifacts and affect image quality.
    - Different hospitals or clinics may follow varying protocols for capturing and storing X-ray images, leading to inconsistencies in the dataset.
    - Combining historical data with recent data may introduce inconsistencies due to changes in technology and protocols over time.

- How are your **variables** related to your project goal? Do you see any **patterns** in the data that would suggest that they are/are not going to be useful in your machine learning model(s)? Do you need to **transform** or create new variables in order to reach your project goal?
    - Images needed resizing, normalization, and we are studying segmentation, too.
    - Dataset needs augmentation (to account for images alignments, black borders, etc.)

Preprocessing:

- Normalization: To ensure that the pixel values are in a consistent range [0, 1].
- Contrast Enhancement: For those histograms showing that pixel values are clustered in a narrow range, we need to enhance the contrast of the images. Techniques like histogram equalization or adaptive histogram equalization (CLAHE) are used here.
- Resizing: To ensure that all images have the same dimensions.
- Data Augmentation: To increase the variability in the training set to improve generalization.



Learning:

For a Convolutional Neural Network (CNN) tasked with image classification, the primary data input is the images themselves, and the CNN learns directly from the pixel values and patterns within those images. No other variables are used as input features for the CNN model.

- How are your variables related to each other? Are there any **strong correlations**? How might this affect how you build your machine-learning model(s)?
  - Histograms of Pixel Intensities provide insights into the distribution of pixel intensities and can highlight differences between classes.

- Aggregating and comparing histograms by class can reveal patterns in intensity distributions. Statistical features extracted from these histograms could be used for correlation analysis to identify potential relationships with the labels.