Canvas Project Number: 23
Group Members' Names: Ioana Cristescu, Alyssa Mia Taliotis, Jonas Raedler and Alissia Di Maria

Data Description: Our dataset comes from the UCI Machine Learning Repository and includes 782 pediatric patients with tabular clinical data and 1,709 ultrasound images. Each patient is linked to between 1 and 34 image views, creating a one-to-many structure. Following feedback from Milestone 2, we ensured that all ultrasound images are correctly mapped to their corresponding tabular records using the US_Number field. This setup enables a multimodal modeling pipeline that combines both image and clinical data for diagnosis.

Summary of the Data + Data Analysis + Meaningful Insights: The dataset contains a mix of standardized numerical features, such as CRP, WBC_Count, Neutrophil_Percentage, and Appendix_Diameter, and categorical variables like Nausea, Free_Fluids, and Sex. We began by inspecting the distribution of numeric variables using histograms (Figure 1). These plots confirm that the variables have been standardized, and several, most notably CRP, WBC_Count, and Appendix_Diameter, are right-skewed, which aligns with their expected clinical behavior. For example, high CRP levels are typically associated with inflammation, and elevated values are observed in patients with appendicitis.

Categorical features were summarized in Figure 2. We observed strong class imbalance in many binary features. For instance, the majority of patients had "no" for Dysuria, Stool, and Coughing_Pain, while "yes" responses were more common for features like Free_Fluids and Nausea among patients with appendicitis. These patterns helped us identify which categorical features may hold predictive signal, and which may have limited variability. These insights directly shaped how we encoded and prioritized categorical predictors for modeling.

To assess redundancy and relationships across features, we created a correlation matrix for numeric variables (Figure 3) and a Cramér's V heatmap for categorical variables (Figure 4). The correlation matrix confirmed that most numeric predictors are not collinear. The only strong correlation was between the Paediatric Appendicitis Score and Alvarado Score ($r = 0.84$), which was expected given their overlapping clinical inputs. The Cramér's V heatmap similarly showed low redundancy among categorical variables. The highest observed associations were between Nausea and Loss_of_Appetite (0.40) and between Sex and WBC_in_Urine (0.25). These moderate associations will be considered during feature selection, but overall the predictors offer distinct information.

Next, we explored how each feature related to the diagnosis label. Boxplots of numeric predictors by diagnosis (Figure 5) reveal clear differences between the appendicitis and no appendicitis groups. Features like CRP, WBC_Count, Neutrophil_Percentage, and Appendix_Diameter were notably elevated in appendicitis cases, reinforcing their predictive potential. For categorical variables, grouped bar plots (Figure 6) show that predictors like Nausea, Free_Fluids, Appendix_on_US, and Peritonitis differ meaningfully across diagnosis groups. Note that these counts reflect the upsampled dataset used during training, which ensures class balance but may slightly distort raw prevalence.

Finally, to account for confounding, we identified Sex, Age, and BMI as variables that are commonly adjusted for in clinical studies. These features are included in our models to help isolate the true effect of other predictors. Although some unobserved confounders (such as socioeconomic status or access to care) may remain, controlling for these known variables strengthens our modeling approach and increases confidence in our findings.

NB: we also performed some EDA on the images in our dataset. However, upon looking at the images for the appendicitis and no-appendicitis case, we realized that we cannot really make out any clear qualitative differences (likely due to our lack of a medical education), so there weren't any significant insights we felt qualified to draw. Therefore, we do not see any reasons to change our approach in using this data for our task. For sample images of the two cases, please refer to our notebook.

Clean and Labeled Visualizations:

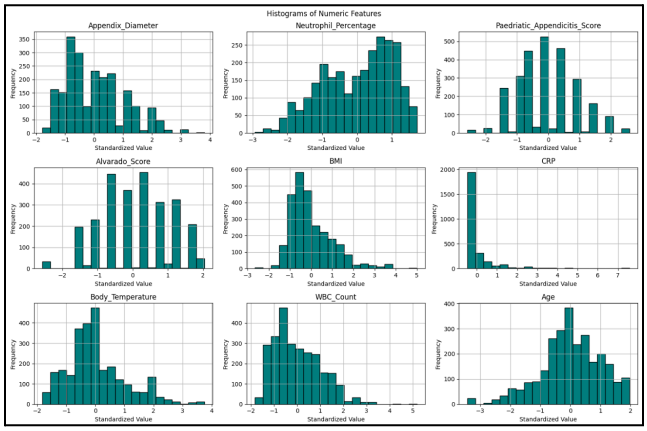

Figure 1

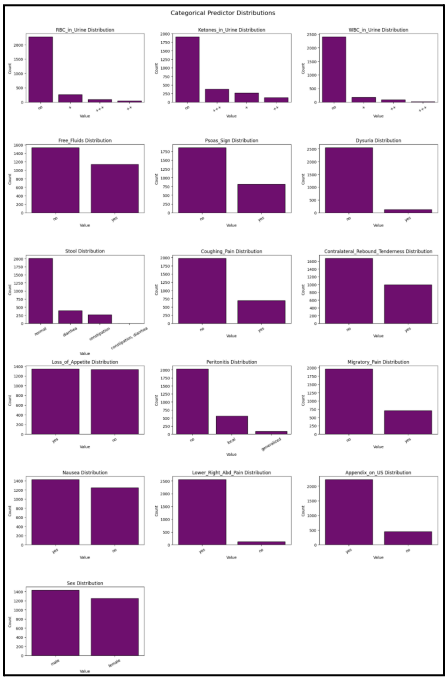

Figure 2

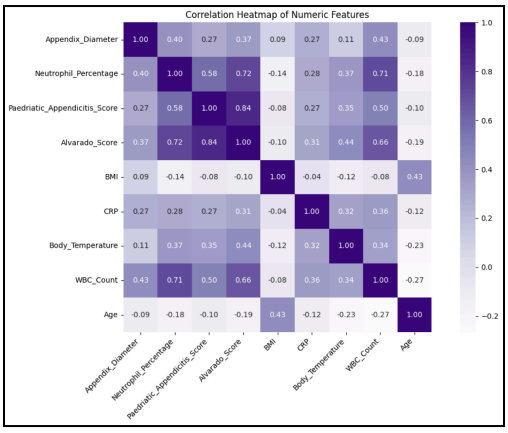

Figure 3

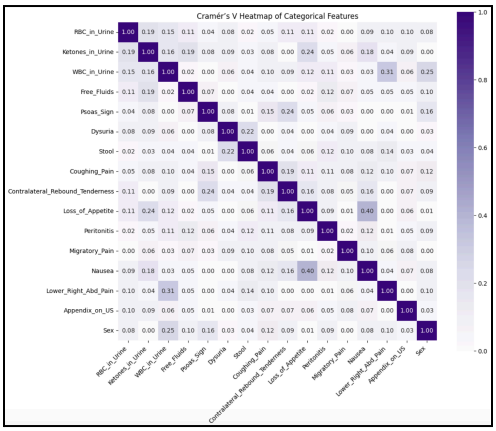

Figure 4

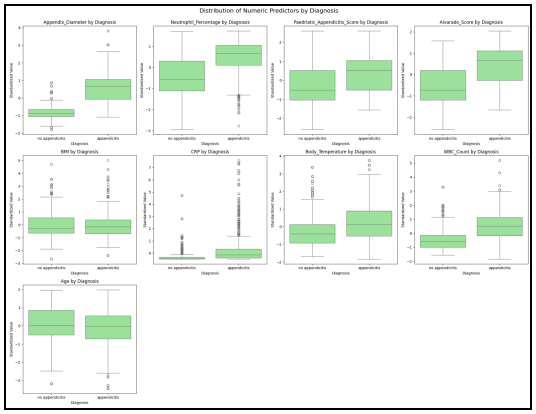

Figure 5

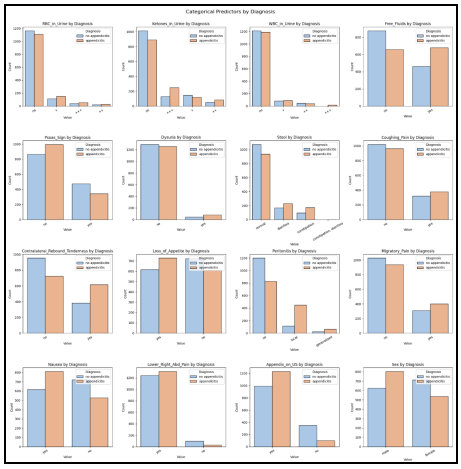

Figure 6

Summary of Findings:

- The dataset is clean, well-labeled, and structured for multimodal modeling.
- All predictors were statistically associated with diagnosis.
- Minimal redundancy across features based on correlation and Cramér's V.
- Several clinical features show strong signal for diagnosis (e.g., Appendix_Diameter, CRP, Peritonitis, WBC_Count).
- Confounders (Sex, Age, BMI) will be included to adjust for bias.

Clear Research Question: How can we build a multimodal model that combines ultrasound images and clinical features to improve diagnostic accuracy and provide interpretable predictions for pediatric appendicitis using CNNs and saliency maps?

Baseline Model or Baseline Model Implementation Plan: We first built a baseline model using only the tabular clinical data. All available predictors were included, with missing values filled in as zero (since the data was already standardized) and missing indicator columns retained. We trained a logistic regression model and evaluated it using an 80/20 stratified train-test split. The model performed well, achieving an overall accuracy of 92% and a ROC-AUC score of 0.98. Precision and recall were balanced across both classes, with slightly higher precision for appendicitis cases and slightly higher recall for non-appendicitis cases. Both classes achieved an F1-score of 0.92, suggesting strong performance without significant class imbalance effects.

Looking at feature importance, the top predictors included Appendix_Diameter, CRP, and Alvarado_Score, as well as missingness indicators like Appendix_Diameter_missing and Appendix_on_US_missing. This suggests that not only are key clinical measurements predictive of appendicitis, but the absence of certain imaging features may also carry diagnostic signal.

In parallel, we developed a baseline image-only model using a simple convolutional neural network. Each grayscale ultrasound image was treated as an individual input. The CNN architecture included two convolutional layers followed by max-pooling, a dense layer, and a final sigmoid output for binary classification. After training, the model achieved a test accuracy of 91.8%, demonstrating strong performance even without access to tabular clinical features.

Our next step will be to build a multimodal model that combines both tabular and image features. Clinical features will be passed through a feedforward neural network, and image features through a CNN encoder. These will be merged into a joint representation before the final classification layer. This

multimodal design aims to take advantage of both data types while enabling interpretability through saliency maps and feature importance analysis.