1 Description

Dataset name	Breast Ultrasound Images Dataset
Link	Kaggle
License	CC0: Public Domain
Medical discipline	Gynecology
Medical procedure	Ultrasound
Multi-class problem	✓
Multi-label problem	✗

The data collected at baseline include breast ultrasound images among women in ages between 25 and 75 years old. This data was collected in 2018. The number of patients is 600 female patients. The dataset consists of 780 images with an average image size of 500 by 500 pixels. The images are in PNG format. The ground truth images are presented with original images. The images are categorized into three classes:

- normal
- benign
- malignant

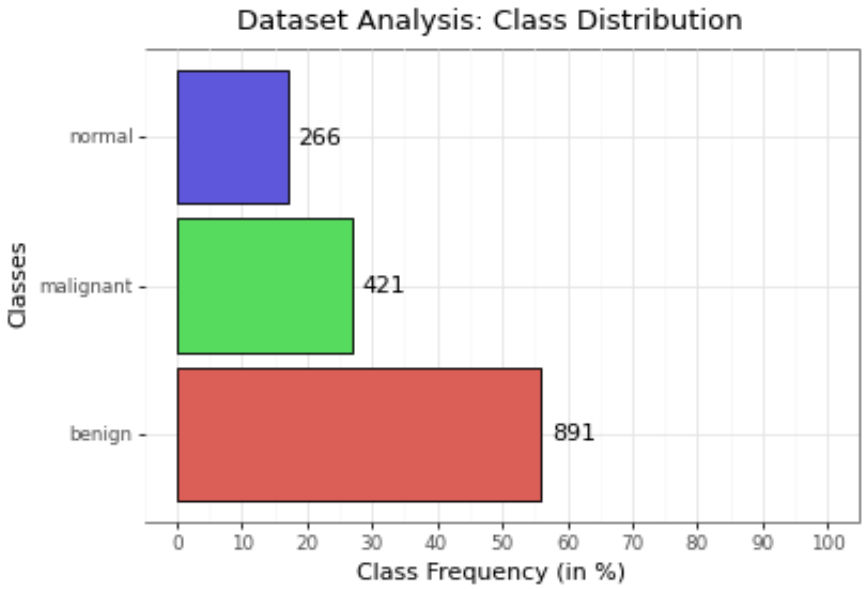


Figure 1: Class distribution: Breast Ultrasound Images Dataset

2 Pre-processing

No pre-processing was done.

3 Training

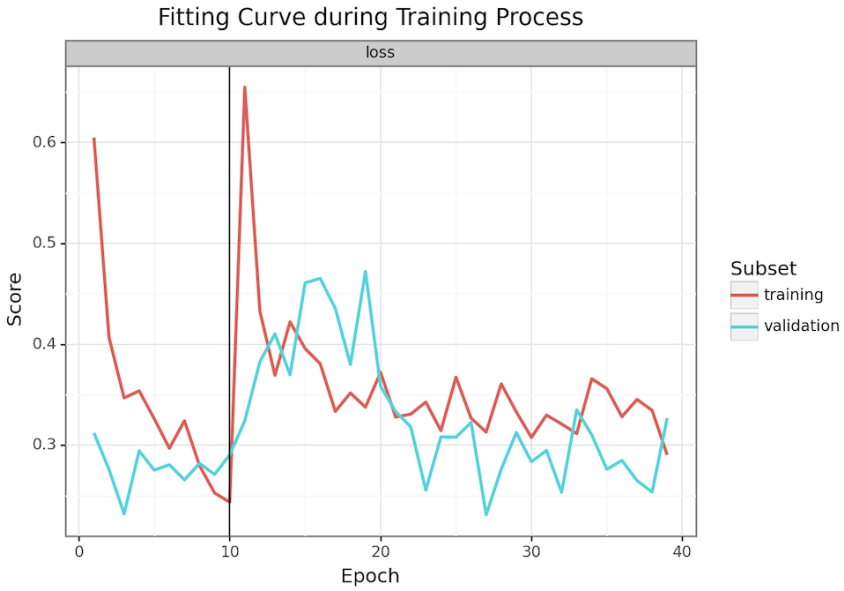


Figure 2: Training: Breast Ultrasound Images Dataset

As usual for transfer learning approaches, training and validation losses peak after the unfreezing of the model's weights in episode 10. The bumpiness of both losses might come from the imbalance within the dataset (see figure 1).

4 Results

Figure 3: Metric overview: Breast Ultrasound Images Dataset

AUC scores show a great performance with more than 90 % throughout all classes.

Precision, formally called *positive predictive value* $PPV = \frac{TP}{TP+FP}$, indicates how many of the classified samples actually belong to this class. Sensitivity, also *true positive rate*, which is defined as $TPR = \frac{TP}{TP+FN}$, expresses how many of a class' samples were correctly identified as such.

For the normal class precision is on a mere 65 %, but sensitivity almost 90 %. That indicates, that many of the samples were classified as normal, often these classifications were not correct.

Contrary to that, the benign class shows a high precision of over 90 % and low sensitivity of 75 %, which means that a relatively low number of the class' samples were classified as such, but these classifications are mostly correct.

ROC curves in figure 4 show fine trends for malignant and normal classes. The benign class' ROC curve shows a plateau, which supports the before-mentioned relation between its precision and sensitivity score.

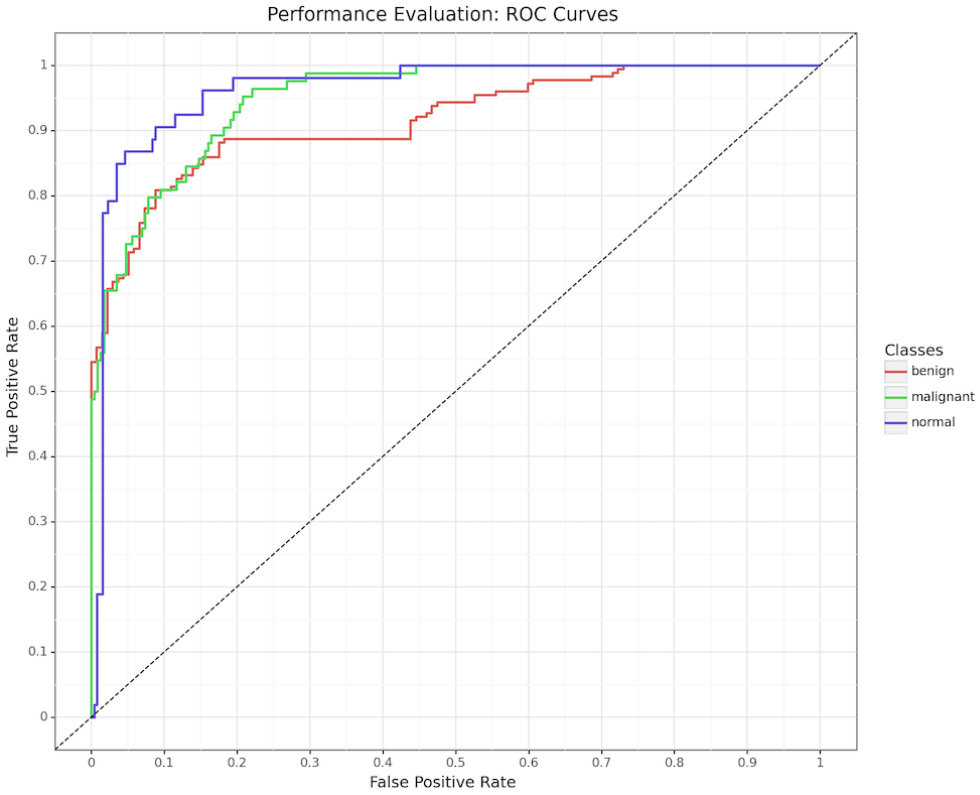


Figure 4: ROC curve: Breast Ultrasound Images Dataset

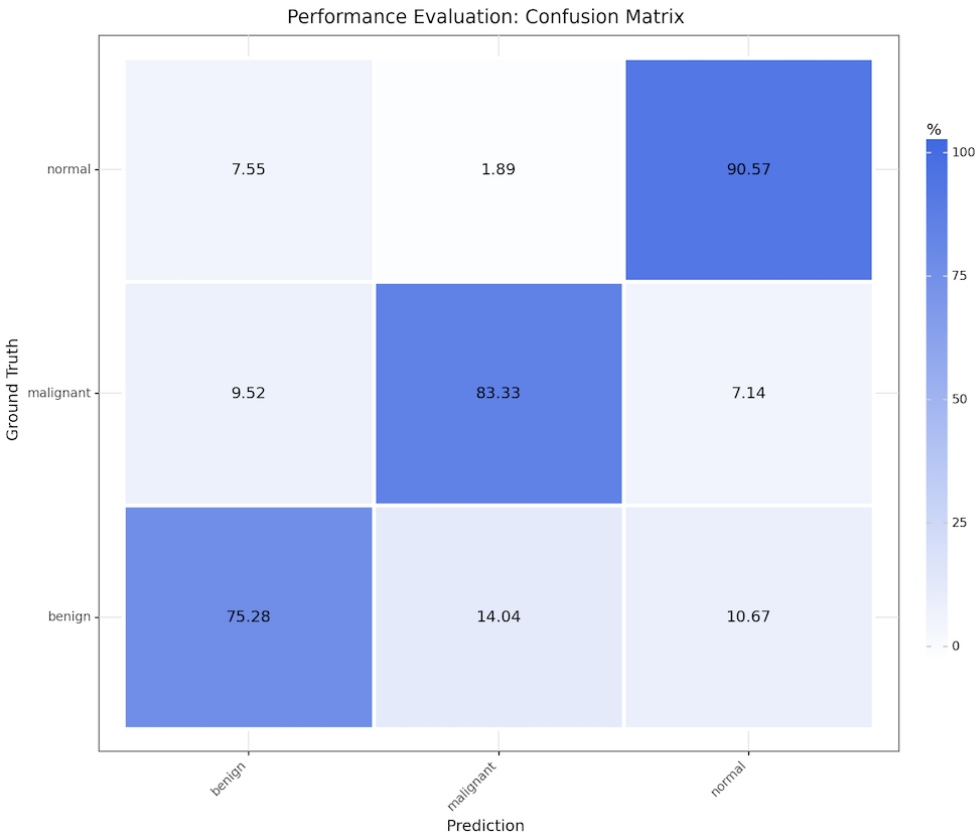


Figure 5: Confusion matrix: Breast Ultrasound Images Dataset

5 XAI

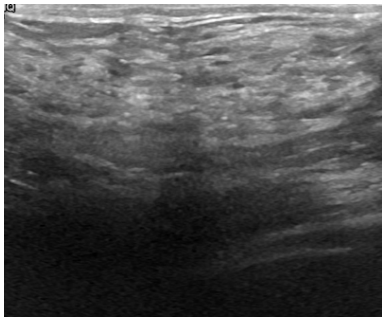
As extensively discussed before, classifications for classes normal and benign seemed unreliable. The three samples rendered here reflect that: Image **benign (32)** was incorrectly classified, which supports the assumption, that many samples were classified as normal, causing the low precision, but high sensitivity. Its Grad-CAM image shows that approximately a quarter of the image is somewhat relevant to the model's decision. The other two samples show a more precise Grad-CAM mark-up. That is curious, because the model's confidence in its decision is with 66 % highest compared to the other two samples. Sample **malignant (113)** for instance, was correctly classified as malignant with a confidence of 52.6 % and its Grad-CAM image precisely marks one small part of the image.

Clearly, classification of normal images was challenging to our model. That may be due to the class imbalance in the dataset (see figure 1). It would be interesting to see, if classification performance could be improved on a dataset with more normal samples.

Image: normal (105)

Class: normal

Classified as: malignant (46.5 %)



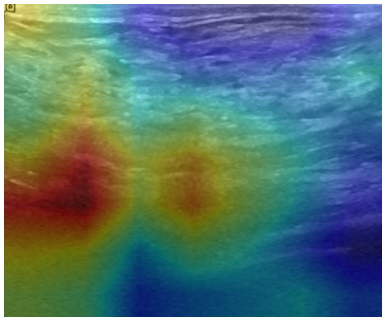
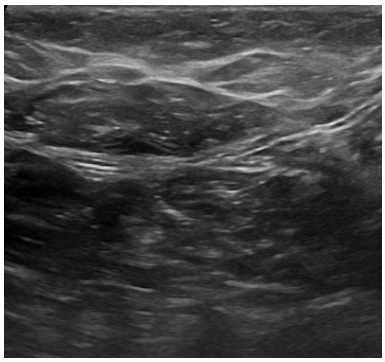


Image: benign (32)

Class: benign

Classified as: normal (66.0 %)



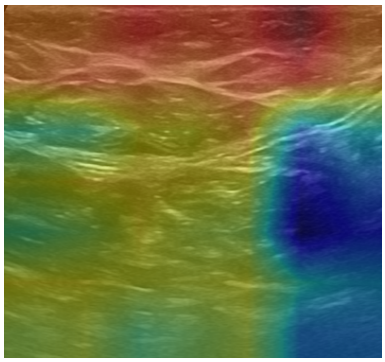
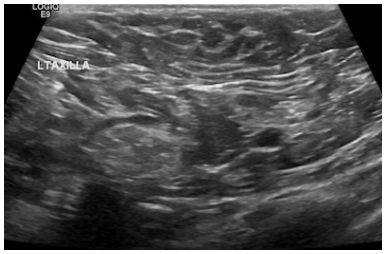
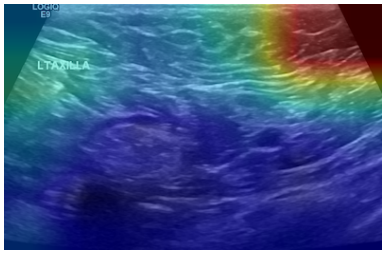


Image: malignant (113)

Class: malignant

Classified as: malignant (52.6 %)





6 Summary

Class imbalance is common within a medical dataset. The overall classification performance for the task at hand is solid. However, the model seems to favor classifying a sample as normal (low precision, high sensitivity) and is cautious in identifying benign samples as such (high precision, low sensitivity). The image sample **benign (32)**, which was considered for explainability supports that. In this instance, the Grad-CAM mark-up for said image highlights big parts of the image as fairly relevant. It would be interesting to observe, how a more balanced dataset would impact Grad-CAM visualizations.