1st semi-annual report

Oz Sam Kilim ozkilim@hotmail.co.uk
Physics PhD.
Supervisors: Prof. Csabai Istvan, Prof. Pollner Peter.

Representation learning of tumor morphology in mammograms

3 projects undertaken during the 2021 autumn semester.

1. Mining for long distance tumor biomarkers in mammograms.

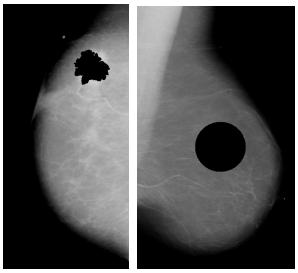
Deep neural networks perform slightly better than expert radiologists at some specific classification tasks. This begs the question: Are models learning feature representations that humans are not? Tumors are located and classified by mammographers in two **distinct** steps:

1. Tumor location search.
2. Tumor investigation with BI-RASDS lexicon (morphology).

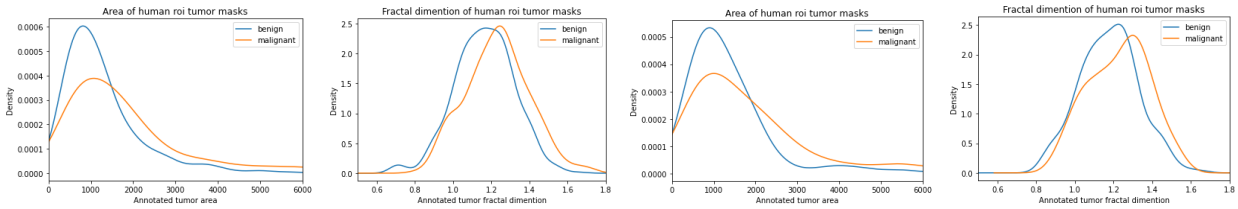
During step 2, out-of-tumor region, long distance information in the mammogram is not used. This project is a data mining investigation to see if long distance signals could be uncovered by a neural network in order to check if this is a factor in the increased performance of models over experts.

Hypothesis: There are visual features outside of labeled tumor region of interest (ROI) that contain clinical information about a distant tumor histology ground truth (Malignant or Benign status). Method: Train a resnet50 neural network on the open source DDSM dataset to classify **whole** mammograms as benign or malignant as a baseline model. Then, cover tumor areas of images to destroy short distance tumor information and see what representations can be learnt and what accuracy the model can achieve. These 3 trained models are compared against each other with the same test set based on:

1. Model classification performance and ROC curves.
2. IOU between thresholded GRADcam CNN attention and ROI annotation.

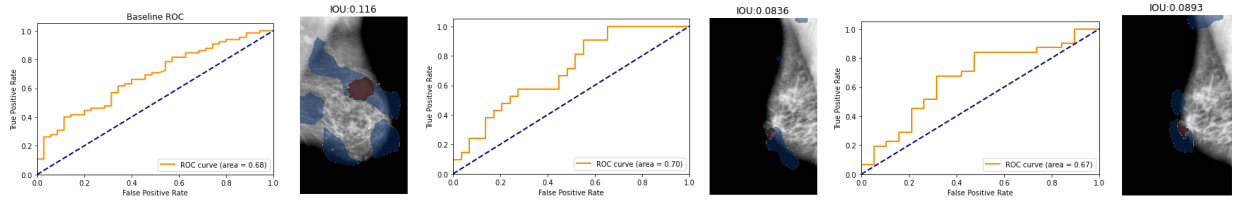


We train the same network on the same images but mask out the tumor in two ways: 1. As a circle of black pixels. (With no nearby tumor information) 2. As the ROI of black pixels annotated by radiologists. (With nearby tumor information).Example of images with ROI (left) and circular masks (right).

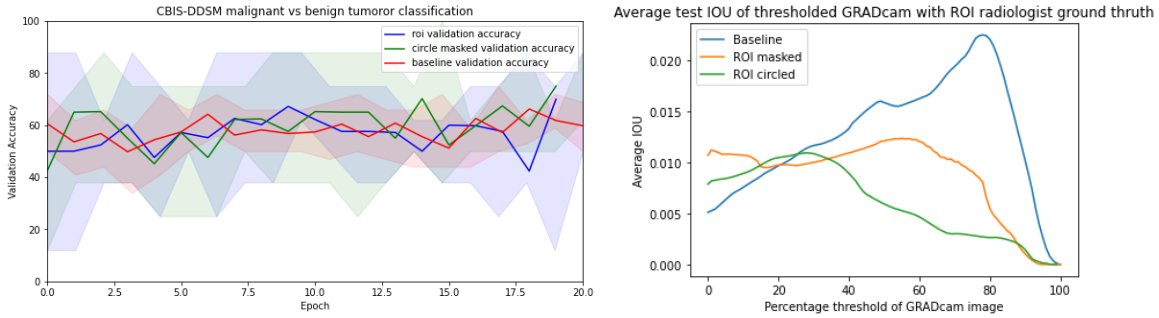


Class frequency analysis of DDSM train (left 2) set and test (right 2) ROI areas and fractal dimensions for malignant and benign classes. Initial search for potentially learnable features from BI-RADS lexicon priors.

Results:



ROC curves for baseline (first 2), ROI masked (second 2) and Circle masked (last 2). An example image of GRADcam : blue, overlap : red, with ground truth : green ROI.



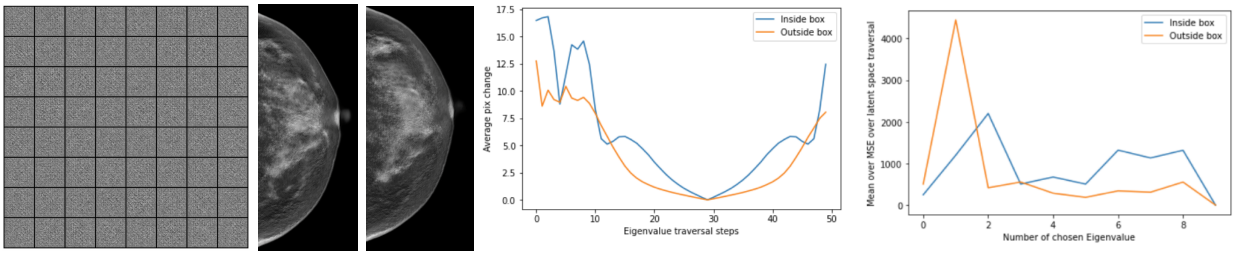
Left: Validation accuracy of each model with accuracy ranges in background highlighted. Right:Average IOU over 50 test images of GRADcam map with ground truth ROI for each trained model as a function of GRADcam threshold. Ablation of learning quality visible with high threshold regime.

Interpretation:

The fact that models trained on DDSM with only the ROI cut outs can achieve up to 93% accuracy indicates that there is **enough** information in the tumor area to retrieve histological ground truth information. The lack of a large amount of data was a bottleneck in this experiment as the baseline accuracy was around 60% on a binary classification problem. If this experiment was recreated with an attention mechanism the baseline results could be greater and a larger ablation (or not) may be able to be observed with and without the tumor edge visible. This experiment could be repeated with the ROI and sub roi annotation blanked out to see the loss of signal as distance is increased from the tumor.

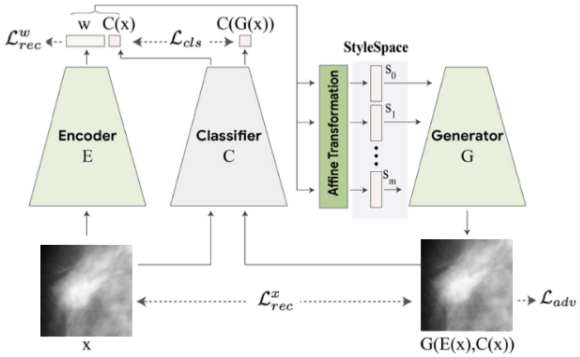
2. Proposing X-RADS lexicon hierarchy for tumor classification.

Hypothesis: The well separated latent space of the StyleGAN will allow disentangling of tumor morphological features that affect state of the art ROI classifier outcomes. This architecture will make it possible to learn the internal representation of a classifier and directly compare the smoothly changing features of tumors to BI-RADS or any X-RADS lexicon. This will be an interpretability tool for future researchers to test their networks on and inspect the learnt feature priorities to see if the networks "internal logic" is consistent with state of the art medical knowledge as well as potentially contribute to an expansion of tumor morphology lexicon.



Left: GAN generated mammograms and Generated images with slight shift in latent space. Right:Changing of pixels inside vs outside ROI for different latent space vectors

Method: A new architecture based on "Explaining in style" combines a pre-trained classifier with a encoder, generator and discriminator to force the latent space of the **generator** to contain **classifier specific** features. This allows for tunable visual single image explanations for classification and global ranking of BI-RADS lexicon features importance in classification decisions.



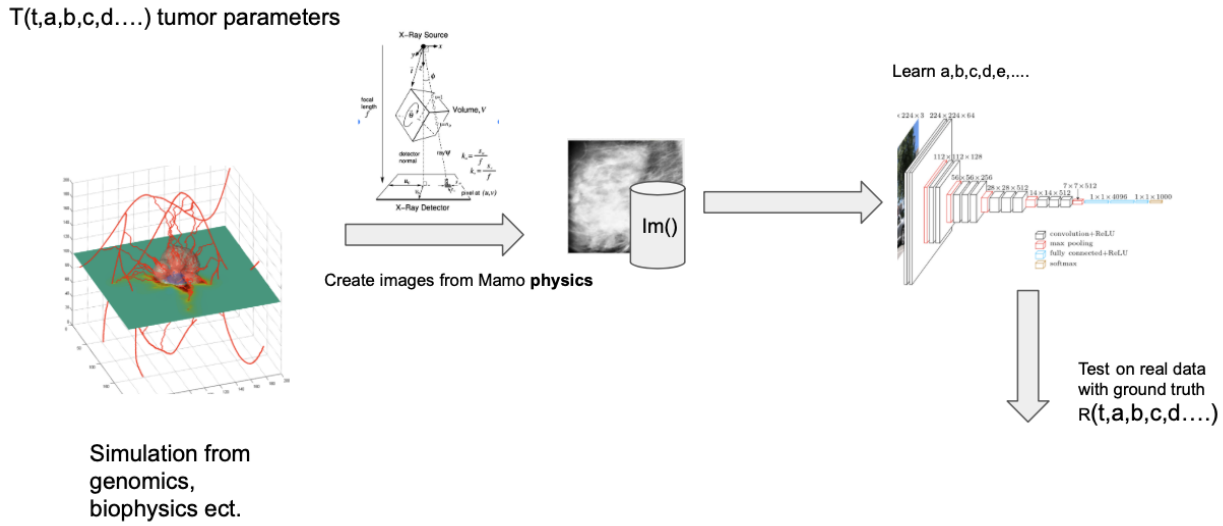
EIS architecture

Pipeline is still in progress and can be followed at: https://github.com/csabaiBio/mammo-styles

3. Tumor morphology parameter learning.

Current networks trained on medical data use histological ground truth data as labels. This is often in the form of a binary histological outcome: Malignant or Benign. Mammographic tumor morphology contains enough signal to accurately extract this ground truth data.

Hypothesis: Useful data that is **not available** from **biopsy** such as **macroscopic tumor growth rate** and **diffusion into external tissue** can only be learnt with some underlying physical model of the tumor. These models exist and offer an in silico image generation mechanism for training deep neural networks to learn such biophysical parameters. Training a neural network on model based synthetic images would allow for extra information for clinicians before biopsy that could guide decision making.
Method: The figure below presents an architecture for parameters learning of simulated tumor X ray images.



Pipeline is still in progress and can be followed at: https://github.com/csabaiBio/tumor-sim

Study activity

5 Modules: Machine learning and data mining.FIZ/3/084 , New results in Machine learning .FIZ/3/092, Clustering with networks. FIZ/3/064E, Pre clinical cancer modeling, Evolutionary game theory.FIZ/3/059E.
Publications: None to date. Expecting submission of: Project 1 and 2.

Source of the data: https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM
These images are free for inclusion in public reports.