Identification and Classification of Breast Tissue Abnormalities from Curated Mammography Imaging

Adam Langenbucher[1], **Han Truong**[2], **Yi Li**[2]
[1]Massachusetts Institute of Technology, [2]Harvard University
{alangenb, hantruon, yili537}@mit.edu

Abstract

We implement deep learning algorithms to analyze mammogram images in the Curated Breast Imaging Subset of Digital Database for Screening Mammography for regions of interest segmentation and malignancy diagnosis. First, we identify regions of interest (ROI) from full mammography images to produce segmentation using the state-of-the-art model Mask R-CNN. Second, we use the popular MobileNetV2 and InceptionV3 models to classify abnormalities as benign or malignant for mass and calcification ROI. For segmentation, we achieve an average dice coefficient of 0.456 and a max dice coefficient of 0.903 on mass tissues. For classification, we achieve an accuracy of 0.87 and 0.84, and an area under the curve of 0.74 and 0.84, for malignancy detection for mass and calcification, using the larger patches re-extracted on our own. These methods outperfom most results in the literature using the small patches. Our findings suggest that deep learning algorithms can be utilized for robust malignancy classification during mammogram readings to aid radiologists in their decision making process.

1 Introduction

The current breast cancer screening protocol in the United States is for an expert radiologist to review mammogram imaging (reading) once per patient visit, also known as single reading. In comparison with the European practice of having a secondary expert review for every patient visit (double reading), mammogram-based breast cancer screening in the US has lower accuracy and significantly higher recall rates that could be due to the implementation of single-pass review (Beam, Sullivan, & Layde, 1996). Even though there are benefits of double reading (Domingo et al., 2016), this practice is costly in terms of time and human resources. Having an automated system that could serve as the initial reading in a double reading practice during mammograms screening would cut down costs and increases efficiency in how breast cancer are diagnosed and treated in the US healthcare system.

With expert radiological analysis being an extremely time-intensive step in the mammography screening process, as well as having substantial implications for patients' clinical outcomes, there is a dire need for improved automated tools to aid in the detection and diagnosis of abnormal lesions in medical imaging. In an effort to increase efficiency and accuracy in mammogram readings, computer-aided detection (CAD) systems have been designed to emphasize regions of interest (ROI) to be presented directly to radiologists as well as making benign vs. malignant classifications on their own. However, Kohli A & Jha S (2018) reported that traditional CAD systems have been proved less effective than a truly independent human reader and have high false positive rates, which has a large impact of downstream health care costs (Kohli & Jha, 2018; Harvey et al., 2019).

Thanks to recent breakthroughs in computer vision, deep learning algorithms for automated mammogram screenings have achieved some success in image analysis tasks. Deep learning algorithms outperform traditional CAD systems and is close to exceed single readers' performance. In practice,

the trained neural net models can be used to deploy on existing patient data in real time as well as cross-validate physician's diagnosis to flag potential false negatives that should be called back for a second mammography screening.

In this paper, we aim to apply state-of-the-art neural networks to perform segmentation and classification on mammogram images from the Curated Breast Imaging Subset of Digital Database for Screening Mammography (CBIS-DDSM).

2 Related Work

Convolutional neural networks (CNN) have achieved higher accuracies in image segmentation on the DDSM and inBreast datasets using U-Net, SegNet, Faster R-CNN (Moreira et al., 2012; Ren, He, Girshick, & Sun, 2015; Abdelhafiz, Bi, Ammar, Yang, & Nabavi, 2020; Chen, Chen, Wang, & Chen, 2020) and malignancy detection (Lai, 2021; Lévy & Jain, 2016; Das et al., 2020). However, most image segmentation and malignant vs. benign classification has been done only for mass tissues, but not for calcification tissues. The accuracies achieved range from 0.65 to 0.70.

3 Dataset

The CBIS-DDSM dataset contains 10,239 images from 1,566 patients that are split into a training set and a testing set in an 80-20 fashion and stratified by their Breast Imaging Reporting and Database System Score (BI-RADS) category. The dataset provides full mammograms, ROI segmentation masks for both mass and calcification, which are two types of breast tissue abnormalities. In addition, hand-drawn borders provided by an expert radiologist are also available for a subset of the mammography images. Figures 3 and 9 in Appendices show the full mammograms and example patches and their labels.

4 Methods

The paper focuses on two goals. First, we aim to identify and segment ROI from mammograms using Mask R-CNN. Second, we perform five-class classification (background, calcification-benign, mass-benign, calcification-malignant, and mass-malignant) and binary classification (benign vs. malignant, on mass-only and calcification-only patches) using MobileNetV2 and InceptionV3.

Methods for segmentation and classification are described in this section. All mammograms go through the preprocessing process described in Section 9.1 in Appendices.

4.1 Segmentation

Downsampling. To maintain computational feasibility, we attempt two approaches to downsample the original (4000×4000) to a reasonable resolution for our model to process. The first approach is *ROI patch re-extraction*, where we extract 256×256 patches from each image, with a stride of 128. For each of the four sub-datasets (calcification-training, calcification-test, mass-training, mass-test), normal-tissue and blank-background patches are randomly sampled, such that the resulting dataset contains 50% lesion-bearing, 25% normal-tissue and 25% blank-background images. The advantage of the patch re-extraction approach is that it preserves high-resolution detail of each region. However, the severely constrained spatial context and inability to share information across neighboring patches (each patch is treated as an independent sample) makes segmentation and the satisfactory model convergence extremely difficult. Therefore, we continue with the second approach: *resizing images* to 512×512. This approach yields some success for the mass segmentation, but the calcification segmentation is much more difficult, as the calcification ROI are diffuse regions containing many small ROI invisible at the highly decreased resolution, whereas the mass ROI are larger, more uniform, and can still be detected. Because of these challenges, we only perform segmentation on mass tissues.

We use Mask R-CNN with ResNet-101 as the backbone to perform segmentation on mass images. The model is initialized with the pre-trained weights generated on the COCO image dataset. The model is trained in three stages: head layers, ResNet stages ≥ 4, and finally all layers trained in succession, with a 10-fold decrease in the learning rate between each stage of training.

Cont.

To improve robustness, we employ data augmentation techniques including flipping, contrast and brightness adjustment, noising, cropping and affine transformation. A variety of bounding-box anchor sizes are tested, but anchors of $(16, 32, 64, 128, 256)$ pixels prove to be the most effective.

4.2 Classification

Patch Extraction. Image patches of size 224×224 need to be extracted from the full mammogram images based on provided ROI border annotations by CBIS-DDSM. To perform hyper-parameter tuning, the training set mentioned in Section 3 is further split into training and validation in a 90-10 fashion. Overall, the patches dataset composes of training, validation and testing sets (Table 9). Breakdowns of patients count, abnormalities count, and patches count for each split are shown in Table 11 and Table 9 (Appendices).

We perform transfer learning using MobileNetV2 and InceptionV3 architectures to generate benign vs. malignant classifiers. We also experiment with using either pre-generated ROI annotations (small context) and our re-extracted larger patches (large context) as inputs to the models. We choose MobileNetV2 due to its low computational cost and comparable test performance to other state-of-the-art CNNs on the ImageNet Challenge (Howard et al., 2017). Levy & Jain reported very high test performance (an accuracy of 0.93) on binary mass classification from the same dataset by utilizing the Inception (GoogLeNet) (Lévy & Jain, 2016). Therefore, we also experiment with the InceptionV3 architecture as an effort to replicate their results.

To evaluate the effectiveness of transfer learning, we train a baseline MobileNetV2 model with all weights randomly initialized and a second MobileNetV2 model with pre-trained weights on the ImageNet dataset. The base model is `tf.keras.applications.MobileNetV2`, and we fine-tune the last few convolutional layers.

To reduce overfitting, we also perform varying levels of data augmentation and implement drop out in the final dense layer of our models.

4.3 Evaluation

We use dice coefficient to evaluate the performance of segmentation, and accuracy (ACC) and area under the curve (AUC) to evaluate the performance of classification. Most previous work have reported only ACC, but we report AUC as well because AUC takes into account the sensitivity and specificity, especially if the dataset is not balanced. In the field of medicine, false negative is usually more harmful than false positive. Therefore, we examine how the neural network performs in detecting false negatives and present a confusion matrix in Tables 13 and 14 in Appendices.

5 Results

5.1 Segmentation

Table 1 shows the baseline dice coefficients of image segmentation on the CBIS-DDSM dataset in the literature on mass tissues. The highest dice coefficient is achieved by Vanilla U-Net, which is capable of predicting a precise pixel-wise segmentation map of a full mammogram image because it incorporates more multi-scale spatial context and captures more local and global context.

Table 1: Baseline Dice Similarity Coefficient on Mass Tissues

Architecture	Dice Coefficient
Dilated-Net (Yu, Koltun, & Funkhouser, 2017)	0.799
AUNet (Sun et al., 2020)	0.818
Original U-Net (Ronneberger, Fischer, & Brox, 2015)	0.818
Multi-Scale Adversarial Networks (Chen et al., 2020)	0.822
SegNet (Badrinarayanan, Handa, & Cipolla, 2015)	0.824
Vanilla U-Net (Abdelhafiz et al., 2020)	0.951

In Table 2, we present the dice coefficients on the testing set using Mask R-CNN. “Resizing” refers to resizing the original full-scale images down to 512×512, “patch-extraction” refers to the method

Cont.

of 256×256 ROI generation discussed earlier, "3-stage" denotes training the head layers, ResNet Stage ≥ 4, and all layers successively, while "2-stage" denotes only training the head and all layers.

Table 2: Our Dice Coefficient Using Mask R-CNN

Architecture	Average Dice Coefficient	Max Dice Coefficient
Resizing: 3-stage	**0.456**	0.902
Resizing: 2-stage	0.423	**0.903**
Patch extraction: 3-stage	0.198	0.844
Patch extraction: 2-stage	0.235	0.897

As seen in Table 2, the resizing approach results in a higher dice coefficient than the patch extraction approach. However, comparing our dice coefficients to Table 1, only max dice coefficient (the best-fitting ROI) exceeds the baseline performance, whereas the average dice coefficients are all much lower than the baseline, which is likely due to false-positive ROI detection. More details are discussed in Section 6.

Figure 1 demonstrates the loss for the training and validation sets using the 3-stage resizing method.

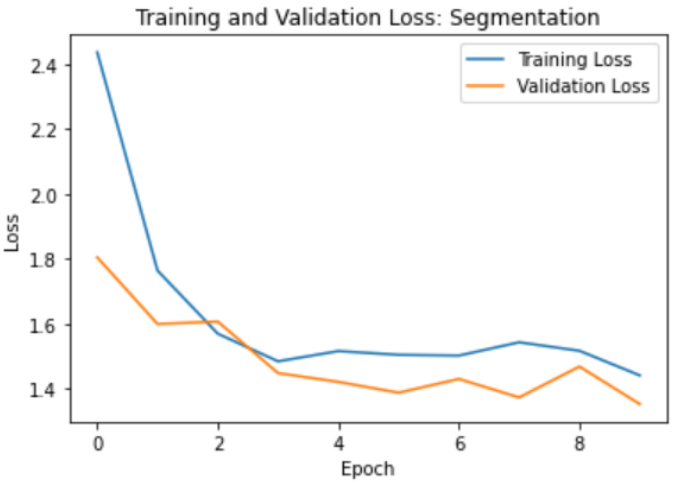


Figure 1: Segmentation Loss by Epoch (Resizing: 3-stage)

5.2 Classification

5.2.1 Multi-Class Classification (Small-Context)

Figures 2 show our training and validation ACC and loss against epoch and training step, if doing five-class classification, including background identification. After experimenting, we find fine-tuning to be helpful for getting both the training and validation accuracy to be above 50% as observed in Table 3. In Figure 2, fine-tuning starts after 30 epochs, which is marked by the green "Start Fine Tuning" vertical line. During fine-tuning, tensorflow's guide only unfreezes the last third of all convolutional layers in MobileNet V2. We go further and perform a second fine-tuning phase, marked as "Start Fine Tuning 2" in red in Figure 2, by unfreezing the last 2/3 of all convolution layers in MobileNetV2 to train for another 50 epochs. This second fine-tuning step allows for a small gain of around 1% in accuracy on training and validation set. After 100 epochs, there is mild overfitting, so we stop training. Since this gain in accuracy is minimal, unfreezing all convolutional layers for a third fine tuning phase seems futile and we focus our efforts on benign vs. malignancy binary classification.

Cont.

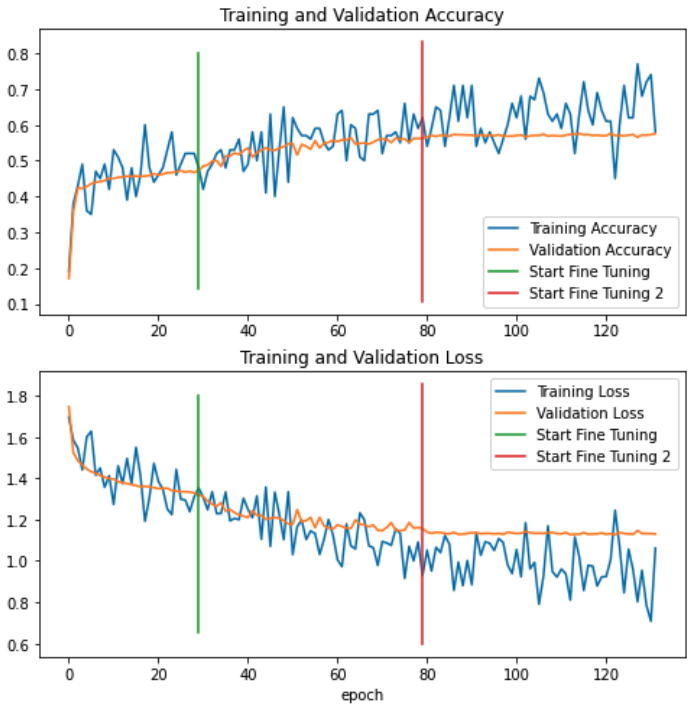


Figure 2: ACC and Loss by Epoch of 5-class Classification by MobileNetV2

Table 3: MobileNetV2 ACC for Five-Class Classification

Learning Style	Accuracy
Randomly initialized weights (baseline)	0.4659
Pre-trained and fine-tuned weights	**0.5685**

5.2.2 Benign vs. Malignancy Binary Classification (Small-Context)

To evaluate our results against most previous work (Table 8) that only perform binary classifications, we train the MobileNetV2 model on patches that contain only either mass or calcification tissues to detect malignancy (binary classification). Table 4 suggests that a pre-trained and fined-tuned MobileNetV2 significantly outperforms the baseline model when the weights are randomly initialized. Using low-level features learned from large datasets such as ImageNet allows the CNN to extract important information that lets the final fully connected layer learn how to detect malignancy in mammogram patches.

Using the pre-generated ROI (small context), our MobileNetV2's ACC is the same as Lai's Medical-CNN for mass, and 2% higher for calcification. Levy's and Jain's AlexNet's ACC is 2% higher than our MobileNetV2 for mass.

Table 4: ACC for Binary Classification (Small Context)

Author	Learning Style	Mass	Calcification
Levy & Jain (Lévy & Jain, 2016)	AlexNet	**0.71**	N/A
Lai (Lai, 2021)	MedicalCNN	0.69	0.69
Randomly initialized weights (baseline)	MobileNetV2	0.59	0.65
Pre-trained and fine-tuned weights	MobileNetV2	0.69	**0.71**

From here on, the results we present are all from the models with pre-trained and fine-tuned weights.

Cont.

5.2.3 Benign vs. Malignancy Binary Classification (Large-Context)

When using small-context patches generated from Tensorflow's pre-made patch extraction code, our model could not get to the level of 90% validation accuracy like obtained by Levy & Jain. We then realize that Levy & Jain used patches that contain much more information than our own patches. We generate our patches using Tensorflow's pre-made code that opt to divide a single abnormality into multiple patches as a way to preserve the original resolution of the abnormality area and skipping any resizing. Levy & Jain also resized the abnormality area so that it would fit into a single patch. In addition, they performed an experiment to show that by encapsulating healthy neighboring tissues around the abnormality into their patches (large context), they obtained a substantial increase in ACC versus just training on patches that only encapsulate the abnormality regions only.

Effectiveness of context. Using larger patches from our own pipeline, the ACC improves by 7% for mass classification, and 2.5% for calcification classification (Table 5). From here on, we use only large-context patches to train our models.

Table 5: Influence of Context

Patch Context	Type of Augmentation	Mass	Calcification
Small	Simple	0.6852	0.7053
Large	Simple	**0.7525**	**0.7319**

Effectiveness of data augmentation. Table 6 compares ACC and AUC using MobileNetV2, when no data augmentation, simple augmentation (random horizontal flip only) and full augmentation (random horizontal and vertical flips, rotation, zoom and translation) are performed, respectively. We conclude that for mass, performing full augmentation largely reduces overfitting (training performance is not shown). In terms of ACC, full augmentation improves the ACC slightly for mass classification (ACC increases by 3%), but hurts for calcification classification (ACC decreases by 8%). For mass classification, AUC is the highest when no augmentation is performed, although there is not much fluctuation when the degree of augmentation varies. Full augmentation results in the highest AUC for calcification classification, however.

Table 6: Influence of Data Augmentation

Type of Augmentation	Mass		Calcification	
	ACC	AUC	ACC	AUC
No augmentation	0.7525	**0.8337**	**0.7319**	0.8211
Simple augmentation	0.7525	0.8310	**0.7319**	0.7883
Full augmentation	**0.7822**	0.8129	0.6522	**0.8559**

Effectiveness of model architecture. Since we also explore InceptionV3, Table 7 compares the ACC and AUC by MobileNetV2 and InceptionV3. Based on ACC given in Table 6, we perform full augmentation for mass classification and simple augmentation for calcification classification. The results show that InceptionV3 performs slightly better for mass, and MobileNetV2 performs slightly better for calcification.

Table 7: Effectiveness of Architecture

Architecture	Mass		Calcification	
	ACC	AUC	ACC	AUC
MobileNetV2	0.7822	0.8129	**0.7319**	**0.7883**
InceptionV3	**0.8020**	**0.8702**	0.6957	0.7561

Lastly, we compare our ACC and AUC with the baseline performance in the literature (Lai, 2021; Lévy & Jain, 2016; Das et al., 2020). Based on ACC given in Table 6 and Table 7, for mass classification, we use InceptionV3 with full augmentation; for calcification classification, we use MobileNetV2 with simple augmentation. We find that for mass classification, our model achieves higher ACC than all other models in Table 8, except Levy & Jain.'s Inception. For calcification classification, our model achieves higher ACC than all other models, except Das et al.'s Xception model and Levy et al.'s

Cont.

Inception. From the confusion matrices in Table 13 and Table 14 in Appendices, we see that false negative rates for both mass and calcification classifications are low.

Table 8: Comparison of Performance on the Testing Set to the Baseline

Authors	Architecture	Evaluation Metric	Mass	Calcification
Lai	MedicalCNN	ACC	0.69	0.69
Levy & Jain	Inception(GoogLeNet)	ACC	**0.92**	N/A
Das et al.	MobileNetV2	ACC	0.64	0.69
	InceptionV3	ACC	0.69	0.65
	Xception	ACC	0.65	**0.77**
Ours	InceptionV3 on mass,	ACC	0.80	0.74
	MobileNetV2 on calcification	AUC	0.87	0.84

6 Discussion

This paper has a few strengths. First, we explore a variety of methods and models from image preprocessing, segmentation to classification. Our best malignant vs. benign classifier with data augmentation achieves higher ACC than most baseline methods in the literature. Meanwhile, the AUC scores are also good for both mass and calcification. Second, we compare the benign vs. malignant binary classification performance for both mass and calcification tissues, while most literature restrict to mass classification. Third, we conduct a comparison of performance between small context and large context and show that small context is indeed a key reason of low ACC, as is seen in literature, although we still do not achieve the ACC as high as Levy & Jain do, because their patch re-extraction method may be different from ours.

Nevertheless, this paper has a few limitations.

Segmentation. As we mentioned in Section 4.1, for the patch re-extraction approach, the spatial context is constrained and patches are not correlated to each other. Thus, it is hard to share information across neighboring patches and our models suffer from convergence issues. For the resizing approach, we see limited results with the mass dataset, but the reduction in resolution removes too much detail to make segmentation of calcification regions feasible.

It is unexpected that the restriction to high-resolution and smaller-context patches do not lead to higher performance for the mass tissues by resizing. From initial observation, this may be due to a large number of positive patches containing only a piece of the lesion mask near the borders, which may make segmentation more challenging. Further improvements include (a) extracting larger-context patches to retain more of the original ROI per image; (b) ensuring better centering of the patch on the ROI centroid to prevent training on small slivers of the ROI edges. Due to the small size of the ROI, resizing calcification images resulted in reduced resolution such that individual calcification is nearly invisible. This implies that successful implementation of a patch-based method is necessary for complete analysis.

In general, the major flaw of our segmentation algorithm may be the false positive rate. In most cases, only one ground-truth ROI exists in the image, which is usually correctly identified by our model, with varying degrees of precision. However, our segmentation often also erroneously identifies additional regions as ROI (with no overlap), driving down the average dice coefficient. The max dice coefficients are more in line with what we might expect for a successful segmentation implementation to look like if the false-positive issue were resolved. It is possible this issue may be mitigated by altering the maximum instance detection limits of our algorithm, or supplementing our dataset with normal negative control images.

Additionally, our segmentation algorithm is trained agnostic to clinical status of ROI. It is possible that training on benign and malignant ROI separately, or introducing separate classes including mixing calcifications and masses (mass-malignant, mass-benign, calcification-malignant, calcification-benign) for the patch-based method, may improve the performance.

Classification. The ACC by ur binary mass classifier is not as high as Levy's and Jain's. Levy & Jain have achieved of 0.93 using Inception with large-context patches and full data augmentation. However, even our InceptionV3 model trained in the exact same scheme is only able to achieve an

Cont.

ACC of 0.80. We are not quite sure about what might cause this discrepancy and have contacted Levy and Jain to share their source code with us.

7 Conclusion

In this paper, we show that deep learning algorithms can be utilized for robust malignancy classification during mammogram readings to aid radiologists in their decision making process. Alternatively, a trained model could potentially serve as the initial reading in a double read practice to drive down costs and miss rate, and improve patients' outcomes. As we are aware, there are major flaws in our segmentation algorithm. If the issues of false positives are resolved and the segmentation algorithm achieves a higher dice coefficient, we can use the cropped ROI by the segmentation algorithm for classification. The performance is expected to outperform the current ACC and AUC. Since there has not been much effort in either image segmentation or classification on calcification tissues, researchers could explore more in calcification tissues. Future work can also include transferring our approach to the inBreast dataset and assess its corresponding performance, and combines deep learning with traditional machine learning methods to see if the ensembl model improves the prediction performance. Moreover, our algorithm could be a component in an ensemble model of machine learning methods to collectively produce clinical diagnosis independent of the physician's opinion.

8 Github Repo

`https://github.com/ryanhantruong/cbis-ddsm-classifier`

References

Abdelhafiz, D., Bi, J., Ammar, R., Yang, C., & Nabavi, S. (2020). Convolutional neural network for automated mass segmentation in mammography. *BMC bioinformatics*, *21*(1), 1–19.

Badrinarayanan, V., Handa, A., & Cipolla, R. (2015). Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. *arXiv preprint arXiv:1505.07293*.

Beam, C. A., Sullivan, D. C., & Layde, P. M. (1996). Effect of human variability on independent double reading in screening mammography. *Academic radiology*, *3*(11), 891–897.

Chen, J., Chen, L., Wang, S., & Chen, P. (2020). A novel multi-scale adversarial networks for precise segmentation of x-ray breast mass. *IEEE Access*, *8*, 103772–103781.

Das, A., Das, H. S., Barman, U., Choudhury, A., Mazumdar, S., & Neog, A. (2020). Detection of breast cancer from mammogram images using deep transfer learning. In *International symposium on signal processing and intelligent recognition systems* (pp. 18–27).

Domingo, L., Hofvind, S., Hubbard, R. A., Román, M., Benkeser, D., Sala, M., & Castells, X. (2016). Cross-national comparison of screening mammography accuracy measures in us, norway, and spain. *European radiology*, *26*(8), 2520–2528.

Harvey, H., Karpati, E., Khara, G., Korkinof, D., Ng, A., Austin, C., ... Kecskemethy, P. (2019). The role of deep learning in breast screening. *Current Breast Cancer Reports*, *11*(1), 17–22.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., ... Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Kohli, A., & Jha, S. (2018). Why cad failed in mammography. *Journal of the American College of Radiology*, *15*(3), 535–537.

Lai, L. (2021, April). *leoll2/medicalcnn: v1.0.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.4700130` doi: 10.5281/zenodo.4700130

Lévy, D., & Jain, A. (2016). Breast mass classification from mammograms using deep convolutional neural networks. *arXiv preprint arXiv:1612.00542*.

Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). Inbreast: toward a full-field digital mammographic database. *Academic radiology*, *19*(2), 236–248.

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*.

Cont.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241).

Sun, H., Li, C., Liu, B., Liu, Z., Wang, M., Zheng, H., … Wang, S. (2020). Aunet: Attention-guided dense-upsampling networks for breast mass segmentation in whole mammograms. *Physics in Medicine & Biology*, *65*(5), 055005.

Yu, F., Koltun, V., & Funkhouser, T. (2017). Dilated residual networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 472–480).

Cont.

9 Appendices

9.1 Image Preprocessing

We perform a series image preprocessing steps, which are useful for image segmentation and classification afterwards. For full mammograms, we crop the borders, normalize the images, and use morphological dilation to smooth the binarized masks of input images. We flip the images if necessary, based on column sum and row sum. To enhance contrast, we employ the contrast limited adaptive histogram equalization (CLAHE) technique. If the image shape is not square, we pad with zeros, and normalize the padded images. For ROI mask images, which share the same dimensions as the full mammograms, we crop the borders, flip horizontally and pad with zeros, if necessary.

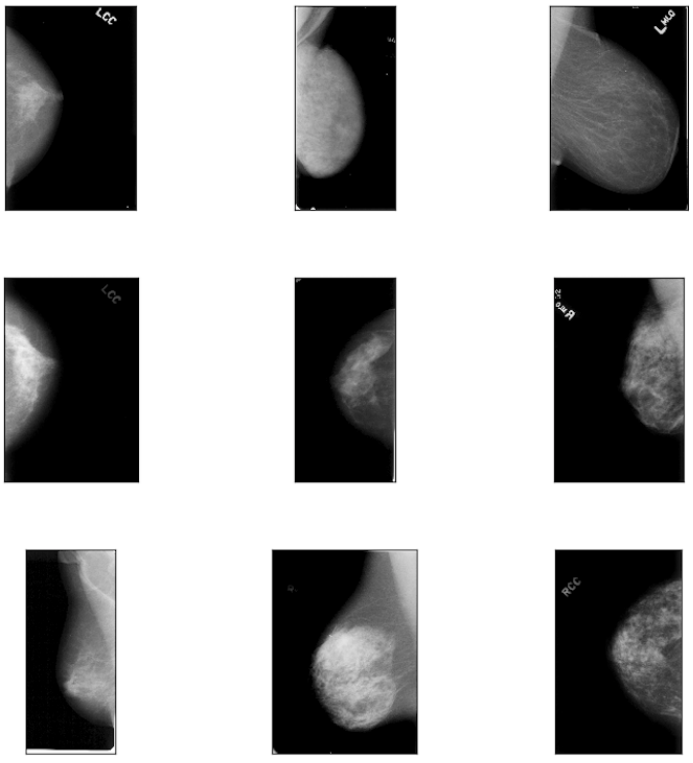


Figure 3: Full Mammograms

Cont.

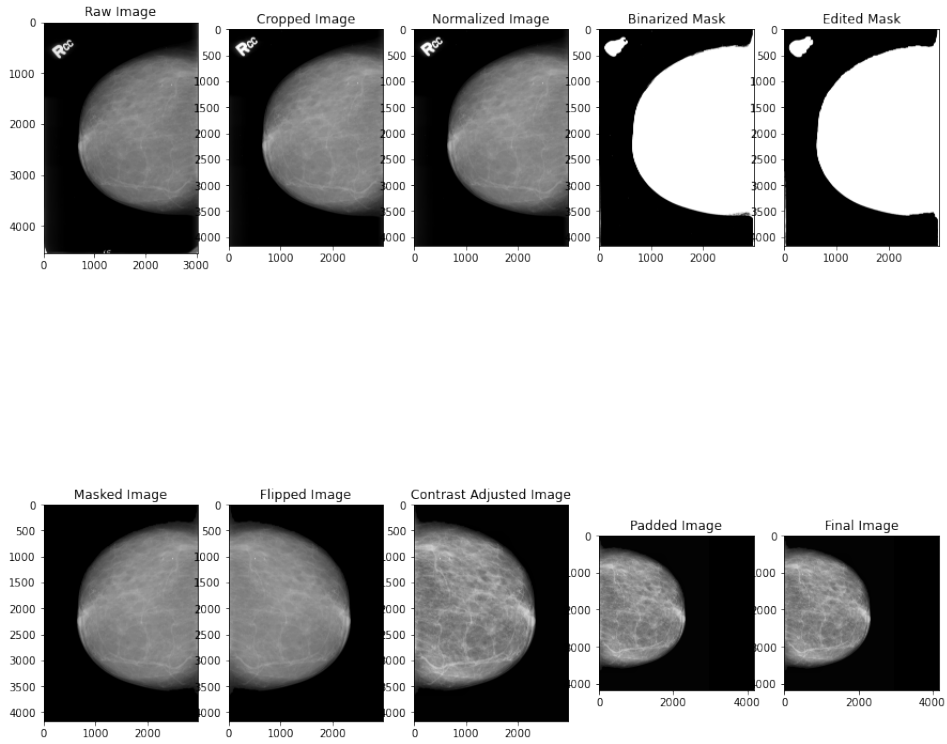


Figure 4: Full mammogram preprocessing

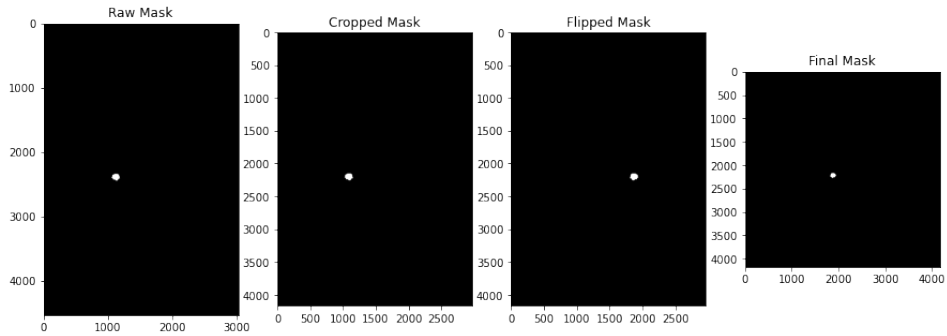


Figure 5: ROI mask preprocessing

Cont.

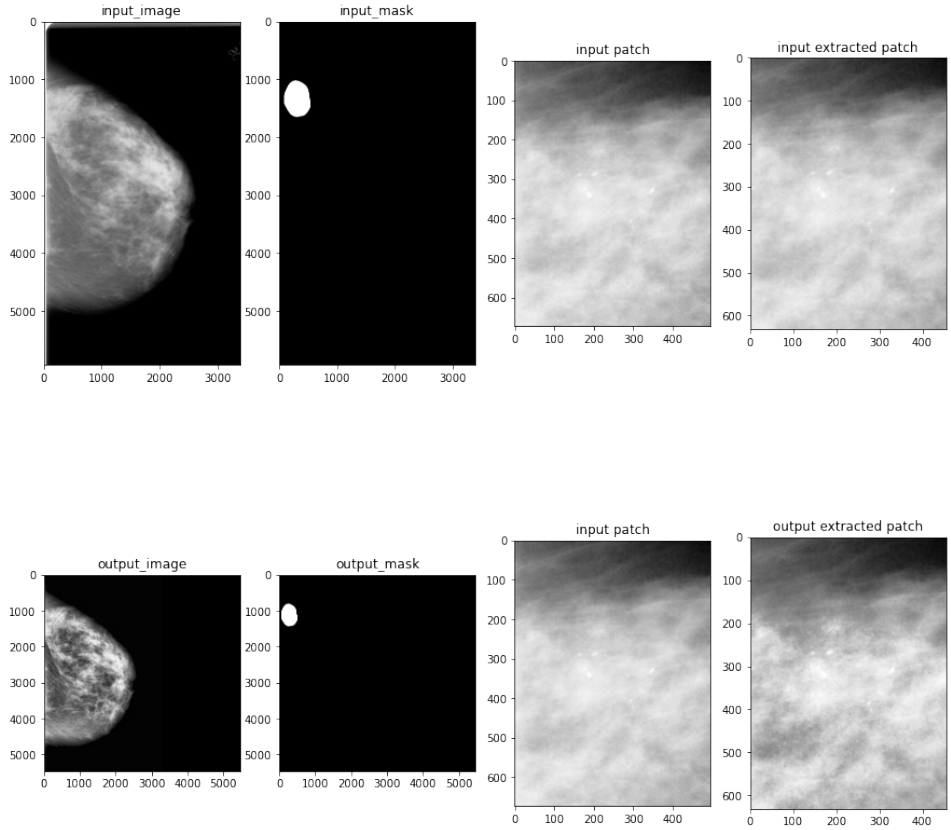
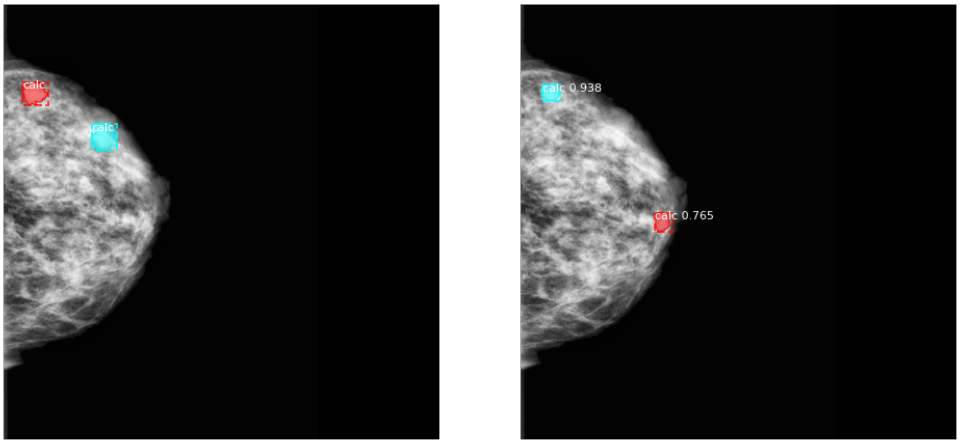


Figure 6: Calcification training image preprocessing

9.2 Segmentation Material



(a) Ground Truth (b) Prediction

Figure 8: Example segmentation result from 3-stage resizing of mass dataset

Cont.

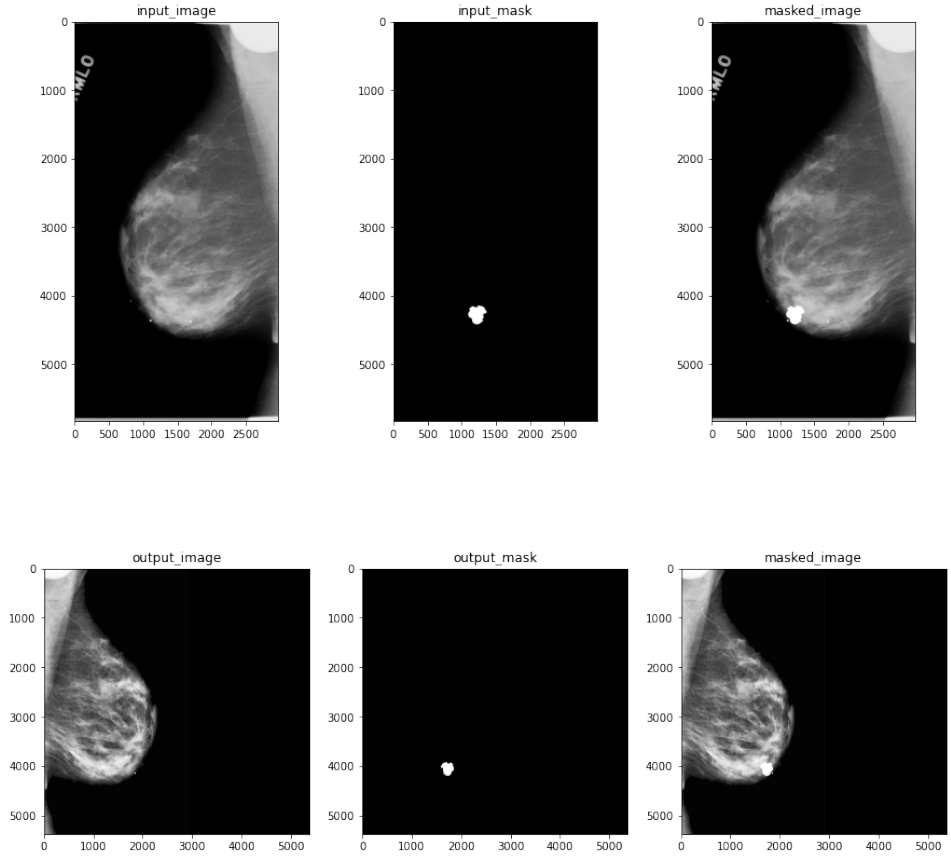


Figure 7: Mass testing image preprocessing

Cont.

9.3 Classification Material

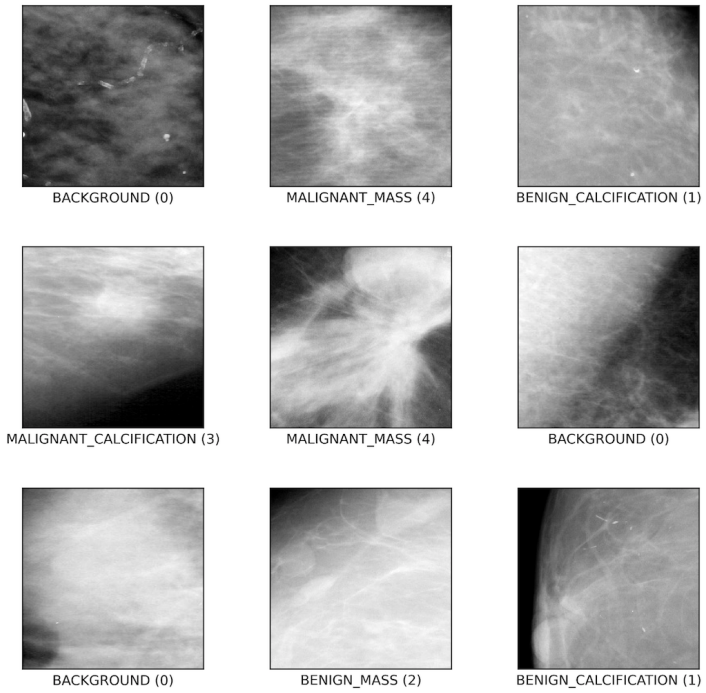


Figure 9: Example Patches and Their Labels

Table 9: Statistics of the Splits of the Dataset

	Whole	Training	Validation	Testing
No. of patients	1566	1197	135	234
% of malignant abnormalities	48.0	48.9	49.6	47.0
No. of mammographies	3103	2386	267	450
% of malignant abnormalities	44.3	44.1	46.5	44.2
No. of abnormalities	3568	2722	308	538
% of malignancies	40.8	41.0	41.2	39.8

Lai developed MedicalCNN and Table 10 shows the performances on CBIS-DDSM dataset using his method (Lai, 2021), where he has tried mass vs. calcification binary classification, benign vs. malignant binary classification and four-class classification (benign-calcification, benign-mass, malignant-calcification and malignant-mass).

Table 10: Baseline Testing ACC by Lai's MedicalCNN

Method	Classification	ACC
CNN from scratch	Mass vs. Calcification	0.91
CNN from scratch	Benign vs. Malignant	0.69
CNN from scratch (composite)	Four-class	0.63
VGG16 transfer learning	Four-class	0.59

Cont.

Table 11: Number of Patches in Each Split (Small Context)

	Background	Benign-calcification	Benign-mass	Malignant-calcification	Malignant-mass
Training	23180	9000	5130	6520	5950
Validation	2600	1110	590	760	520
Testing	4470	1880	990	1220	1210

Table 12: Number of Patches in Each Split (Large Context)

	Benign-calcification	Benign-mass	Malignant-calcification	Malignant-mass
Training	915	626	495	593
Validation	89	57	51	46
Testing	198	232	130	148

Table 13: Test Set Confusion Matrix for Large-Context Binary Calcification

		Ground Truth Benign	Ground Truth Malignant
Prediction	Benign	69	15
Prediction	Malignant	19	35

Table 14: Test Set Confusion Matrix for Large-Context Binary Mass

		Ground Truth Benign	Ground Truth Malignant
Prediction	Benign	181	49
Prediction	Malignant	50	98

The End.