109B – Advanced Data Science Final Report

Florian Hillen - hillen@mit.edu
Nikhil Mallareddy - nikhilmr@mit.edu
Group 21

May 2, 2018

Contents

List of Figures

List of Tables

1 Motivation

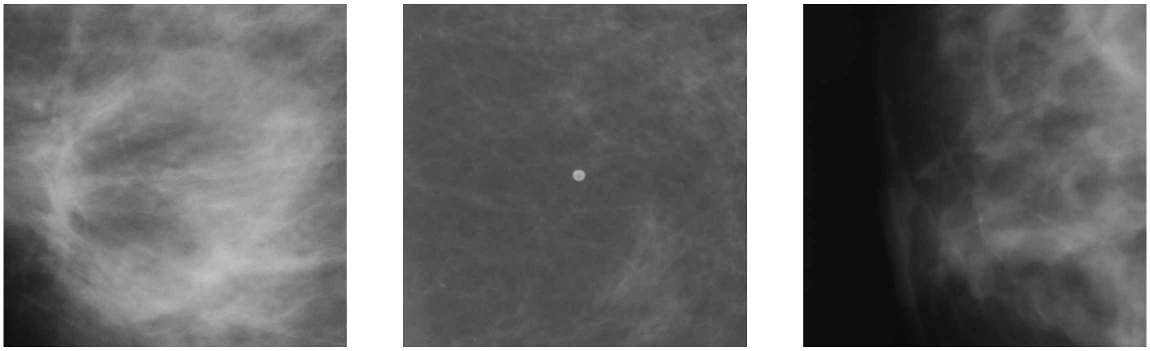
Every year, 12% of women are diagnosed with breast cancer ((McGuire et al., 2015)). In the US alone, 40,000 women die of breast cancer annually ((Cancer.gov, 2018)). Evidence shows that early detection of breast cancer can significantly increase the survival rate of women (Shapiro et al. (1982)).

Mammography, a special X-ray of the woman's breast, is one of the most common diagnostic tools for detecting breast cancer. The images show masses and even calcifications, which are precursors to breast cancer. However, correctly identifying these images can be challenging for radiologists. Moreover, time constraints in assessing the images often result in incorrect diagnosis with detrimental consequences. For instance, a false negative diagnosis, that a case is normal when it is in fact an early form of breast cancer, can decrease the chance of 5-year survival significantly.

In this project, we aim to develop a deep learning algorithm as a support tool for radiologists to decrease false negative and false positive diagnoses, and thereby improve the effectiveness of mammography as a screening tool. By using different architectures and techniques, we aim to match a benchmark set by similar algorithms in the field. Furthermore, by studying literature and interviewing physicians, we want to understand how to effectively deploy such an algorithm in clinical practice.

2 Introduction and Data

A mammogram is a special kind of X-ray that is able to detect calcifications and masses, which are common early-stage signs of breast cancer. For this project, we use the Digital Database for Screening Mammography (DDSM) data set, which is a collection of labelled mammographic images maintained for use by the research community Heath et al. (2000).



(a) Mass (b) Calcification (c) Normal

Figure 1: Three sample images from the DDSM data set

The data set contains 2,620 cases of patients and mammograms with calcifications, masses, and non-pathological findings (see sample images in Fig. 12). The labels for calcification and masses are further specified as "benign", "malignant", "benign without callback" and “unproven” (see Fig. 2). In addition, the images are also categorized on a scale of 1-5 according to the BI-RADS scheme, which is a commonly used method for classifying mammograms in clinical practice. For the purpose of our analysis, we use patches instead of full images, not only for computational feasibility, but also for better performance due to easier feature detection. The data set we use contains 10,713 patches labelled appropriately, the summary statistics of which are shown in Table 1.

	Benign	Benign w/o callback	Malignant	Unproven	Total
Calcification	800	539	797	16	2,152
Mass	1,079	179	1,075	21	2,354
Number of pathological cases					4,506
Number of non-pathological cases					6,207
Total number of patches					10,713

Table 1: Summary statistics of the DDSM patch data set

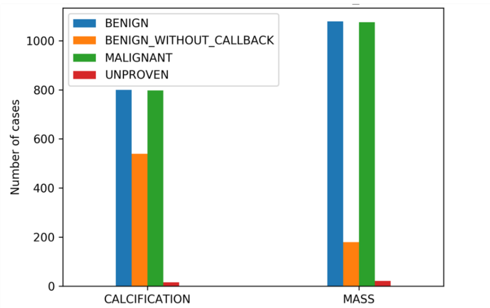


Figure 2: Classes and labels of the DDSM dataset

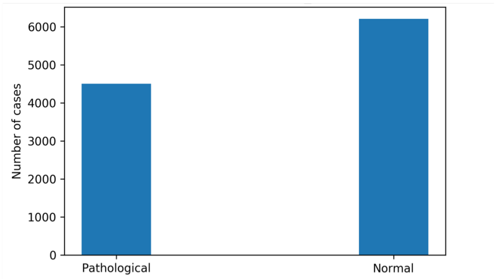


Figure 3: Split of the patches based on pathology

Exploring the data further yields the plots in Fig. 2 and Fig. 3. Overall, there are more cases of masses than of calcification (see Fig. 2). The number of benign and malignant cases for calcification and masses seems to be the same. For both calcification and masses, a few cases have been categorized as 'unproven'. In our analysis, we decide to mark these as pathological as it is not clear if they can be considered healthy patients or not (the number is small and should not have a strong negative impact on our predictive power). As we aim to model a binary classifier, we label all mass and calcification patches as being pathological. In total, we have 4,506 pathological and 6,027 non-pathological patches (see Fig. 3).

3 Literature

Our literature review covered three topics: 1) The state-of-art deep learning architectures for the task of binary classification of images 2) Performance achieved in similar tasks as a benchmark for our algorithm 3) Studies on physicians' performance to understand the clinical implications of such algorithms.

3.1 State of the art architectures

The first architecture we look at is the VGG network (Simonyan and Zisserman (2014)). It is a simple model consisting of a 13 layered convolution neural network using 3x3 filters and 2x2 maxpooling layers. The multiple, smaller-sized kernels as shown in Fig. 4 perform better than a single larger-sized kernel, as the increased depth of the network can enable it to learn more complex features.

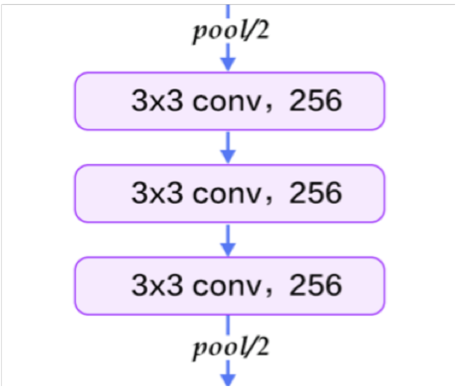


Figure 4: The characteristic 3x3 convolution layers of the VGG

Second, we look at one of the most popular architecture for image classification, which is the Residual Network (ResNet) He et al. (2016). The distinguishing feature in a ResNet is he residual block (see Fig. 5), which allows the ResNet achieve a depth as large as 152 layers. The residual block helps to overcome the problem of vanishing gradients, usually observed in deep conventional networks where the performance degrades with increasing depth.

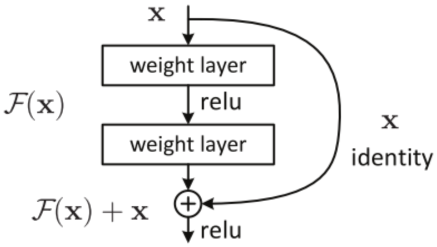
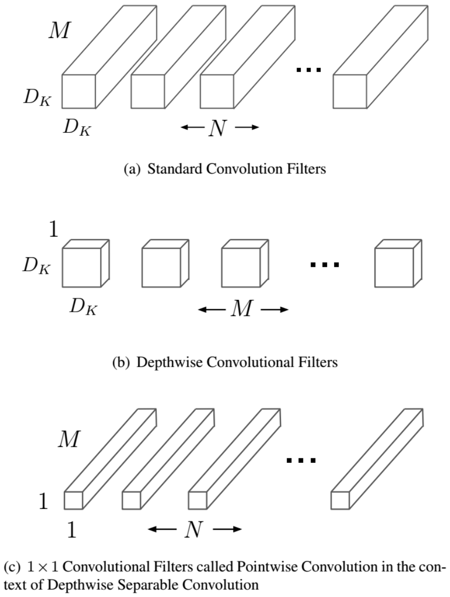

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 5: The residual model of the ResNet architecture

Lastly, we look at the MobileNet, a specialized network for mobile and embedded vision applications (Howard et al. (2017)). It is a simple deep neural network that uses depth-wise separable convolutions, which factorize standard convolutions into one depth-wise and another point-wise convolutions. The advantage is that it results in far fewer parameters and helps build lighter deep neural networks that trade-off accuracy for improved latency (see Fig. 6).


M
D_K
D_K
$\leftarrow N \rightarrow$
...
1
D_K
D_K
$\leftarrow M \rightarrow$
...
M
1
1
$\leftarrow N \rightarrow$
...

(a) Standard Convolution Filters

(b) Depthwise Convolutional Filters

(c) 1×1 Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution

Figure 6: MobileNet Architecture

3.2 Prior art in breast cancer classification

There have been several previous attempts to detect breast cancer using different data sets, algorithms, and classification schemes. For instance, a paper published in 2015 obtained a 85% accuracy for identifying images with a mass, and also localizing 85% of masses in mammograms with an average false positive rate per image of 0.9 Ertosun and Rubin (2015). A more recent publication, Shen (2017), developed an end-to-end training algorithm for a whole-image diagnosis. It deploys a simple convolutional design achieving a per-image AUC score of 0.88 on the DDSM dataset. We adopt this as metric as the benchmark for our algorithm, in addition to an accuracy benchmark of 85%.

3.3 Physician performance

Several high quality studies explored the performance of physicians diagnosing mammograms. A study by Rafferty et al looked at radiologist performance on mammographs from 1,192 patients Rafferty et al. (2013). In a first study, 312 cases (48 cancer cases) were diagnosed by 12 radiologists who recorded if an abnormality which requires a callback was present. This resulted in a sensitivity of 65.5% and a specificity of 84.1%. In a second study, 312 cases (51 cancer cases) were analyzed by 15 radiologists. They obtained additional training and also reported the the type and location of the lesion. This resulted in a sensitivity of 62.7% and a specificity of 86.2%. Another high quality study compared different diagnosis methods such as mammography, ultrasonography (US), and physical examination (PE) using a data set of 27,825 screening sessions Kolb et al. (2002) and compared the results of the three diagnosis methods with the actual biopsy. The results showed a sensitivity of 77.6% and a specificity of 98.8%. However, these scores were not achieved by radiologists using only mammograms and thus do not fit well for a benchmark for this task.

Most relevant as a benchmark for our analysis is the first study by Rafferty et al as the 12 radiologists restricted themselves to binary classification, which is similar to our approach.

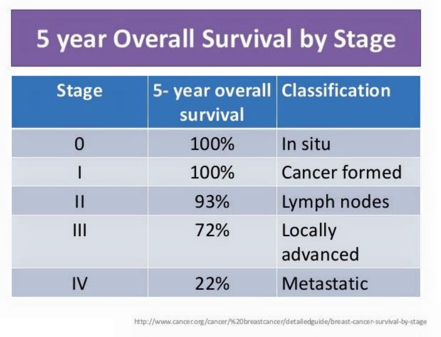
3.4 Clinical relevance

In order to develop an algorithm with actual clinical relevance, further understand of the medical implications of the mammogram diagnosis are necessary. From the discussion above, it appears that radiologists have a significantly higher specificity than sensitivity. This means they have a higher false negative rate compared to the false positive rate. In Table 2, we explain the clinical trade-offs between the two types of errors to analyze implications for clinical practice and algorithm development.

Risks and costs of a diagnostic error	
False positive	• Additional test: Costs and minimal-invasive biopsy • Short-term distress/long-term risk of anxiety
False negative	• 5-year survival rate is strongly impacted by later detection: Decreases from 93% to 72% from stage III to stage II

Table 2: Comparison of risks for type 1 and 2 diagnostic errors

A false positive diagnosis means that the radiologist assesses a normal mammogram of having either a maligned or benign lesion which necessitates a callback. For the patient, this would mean a repeat visit to the clinic, and in most cases further testing through a biopsy. Biopsy for breast cancer detection is minimally invasive and only a small incision is needed. However, there is a range of evidence showing the psychological effects of such false positives. According to a study from 2000, it can lead to short term distress as well as long-term anxiety Aro et al. (2000). On the contrary, a false negative implies that a potentially cancerous case is misinterpreted as healthy. The consequences of this can be very severe because breast cancer, when left untreated progresses in its stages and with each stage a different 5-year life expectancy is associated (see Fig. 7 (Cancer.org (2018))).



5 year Overall Survival by Stage

Stage	5- year overall survival	Classification
0	100%	In situ
I	100%	Cancer formed
II	93%	Lymph nodes
III	72%	Locally advanced
IV	22%	Metastatic

http://www.cancer.org/cancer/%20breastcancer/detailedguide/breast-cancer-survival-by-stage

Figure 7: Survival rate of breast cancer stages

In order to corroborate our observation, we conducted two interviews with a gynecologist in his 2nd year residency (Gyn (2018)) and a radiologist in his 4th year of residency (Rad (2018)). Both interviews took around 30 minutes and physicians were semi-structured in nature. The interview with the gynecologist was aimed at better understanding the medical implications for women, while the radiologist interview helped us to better understand the requirements such algorithm would have.

- Steps involved in the assessment: First, the radiologist has to check the quality of the image, the visibility and the nature of the gland and the tissue. Other structures, such as a Fiebro-Adenom, can be seen in the mammogram as well but for breast cancer detection, calcification and masses are most important (radiologist's interview)
- Usefulness of an algorithm: There is a lot of ambiguity, especially between BI-RADS categories 3 and 4 (probably benign and suspicious abnormality). It would be most helpful to have an algorithm give an estimate of the BI-RADS assessment as these are the categories they have to be judge by in the end. However, even having an assessment of pathological vs. non-pathological can be helpful. It is not very important if it is a calcification or a mass as this can then be seen by the radiologist as well. A heat map would be interesting to better understand what features the algorithm picked up. However, this would be mostly interesting for the purpose of research to see if there are any structures not being paid much attention to earlier. (radiologist's interview)
- Trade-off between false positives/negatives: In general, it is more important to avoid a false negative over a false positive. A mammogram is first followed by a sonography and if this is positive as well, further testing is done via a minimal invasive biopsy. However, in the future it would be great to differentiate further and reduce the false positive in the BI-RADS 3 category. As of now, 98% of patients in this category have to come back every 6, 12, 24 months for a check-up, yet do not have breast cancer. This is a large burden. (gynecologist's and radiologist's interview)

Considering Table 2 and the two interviews, we conclude that a false negative error can have more severe consequences than a false positive error. Thus, we decide to design our algorithm to have a threshold which is more sensitive than specific.

4 Modeling

4.1 Data cleaning

Before building the model, we clean the data set to convert it into the appropriate form. We assign new, binary labels to the images by categorizing all the original mass and calcification labels as 'pathological', and the normal images as 'non-pathological'. Thus, the problem is reduced to a binary classification. Next, we randomly divide the data set into train, validation, and test splits, in approximate proportions of 75:10:15 respectively. While doing so, we ensure that the splits are evenly balanced between the two classes (as evident in Fig. 8).

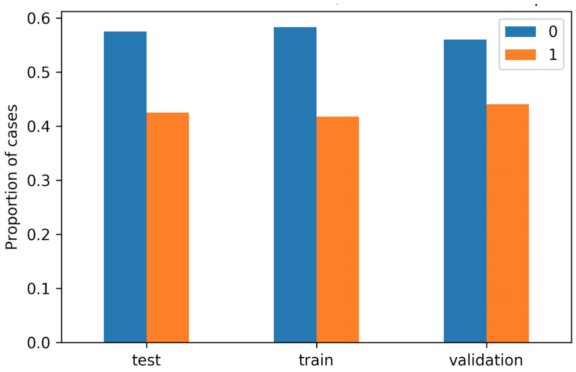


Figure 8: Balance of classes in the train, validation, and test splits

4.2 Performance evaluation

We evaluate the performance of the model using two metrics that are widely-used for classification problems: 1) Accuracy on the test set 2) Area under the Receiver Operating Characteristic curve (AUC). While accuracy refers to the percentage of cases that the model classifies correctly, AUC indicates the ability of the model to discriminate between the two classes.

4.3 Modeling process

The modeling process is an iterative procedure that includes the following steps:

1. Build the model
2. Train the model on the training set for a certain number of epochs
3. Obtain plots of validation accuracy and loss function against epochs and determine where over-fitting occurs
4. Repeat the training process with the appropriate epoch size, using the early stopping regularization technique to avoid over-fitting
5. Compute test accuracy and area under the ROC curve (AUC) and compare it with the benchmark
6. Repeat the process with a different model until the desired performance is achieved.

4.4 Model building

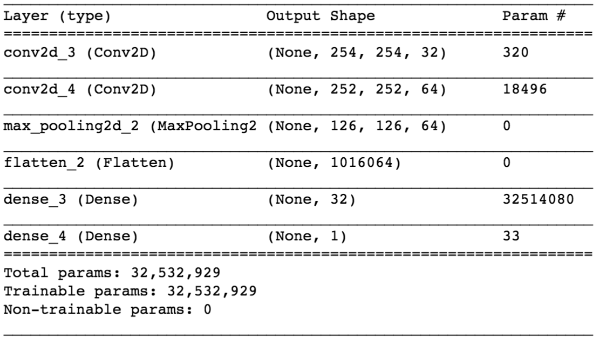
The first step of the modeling process, which is the model building step, is further broken into four sub-steps.They are:

1. Build a baseline model and evaluate performance
2. Train well-known models with different architectures, and choose one that performs the best.

3. Once the model architecture is chosen, perform regularization, pre-training, and data augmentation techniques to improve performance. Choose the model that delivers the best performance as the final model.

4. Tune hyper-parameters on the final model to achieve the desired benchmark.

Thus, we start off the modeling process by building a simple model that serves as a baseline. The purpose of the baseline model is not only to identify unforeseen discrepancies in the data set, but also to provide an estimate of the improvement in performance needed to meet the benchmark.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
conv2d_3 (Conv2D)            (None, 254, 254, 32)      320       
_________________________________________________________________
conv2d_4 (Conv2D)            (None, 252, 252, 64)      18496     
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 126, 126, 64)      0         
_________________________________________________________________
flatten_2 (Flatten)          (None, 1016064)           0         
_________________________________________________________________
dense_3 (Dense)              (None, 32)                32514080  
_________________________________________________________________
dense_4 (Dense)              (None, 1)                 33        
=================================================================
Total params: 32,532,929
Trainable params: 32,532,929
Non-trainable params: 0
_________________________________________________________________
```

Figure 9: Architecture of the baseline model

The architecture of the baseline model includes two 2d convolution layers, with 32 and 64 filters respectively, with one dense layer of 32 nodes on the top (see Fig. 9). Upon evaluating performance of the baseline model on the test set, we found that it delivers an accuracy of 0.759. This is approximately 13 percentage points lower than our benchmark.

As the next step, we implement three well-known image classification models namely VGG16, ResNet50, and MobileNet. We customize these models to our application by tweaking the feed-forward, dense layers in the end to just one layer with 32 nodes, followed immediately by an output layer with sigmoid activation and one node (for binary classification). These models are originally designed to label up to 1000 classes and therefore have wide dense layers (4096 nodes). We cut down the width of these layers so that information in the features is not diluted when passed from 4096 nodes to just one node in the output layer. The models delivered better performance after such changes were made.

Once the final model architecture is chosen, we implement two further techniques to see if they improve performance: 1) Data augmentation 2) Pre-training.

Data augmentation is a regularization technique that involves increasing the size of the training set to reduce variance and enable learning from small data sets. In our analysis, we perform three operations on the input images as a part of the augmentation process: 1) Flip images along a horizontal axis 2) Shift horizontally or vertically within a width range of 0.2 3) Rotate randomly within a 20 degree range.

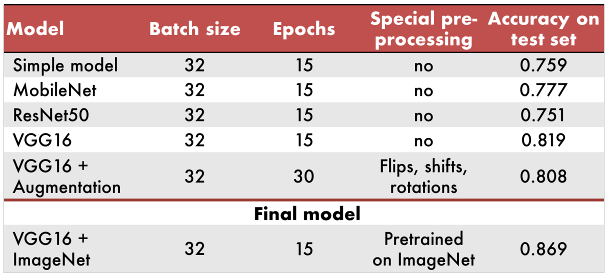
On the other hand, pre-training involves initializing model parameters with values learned from a different data set, instead of random ones. Not only is pre-training expected to speed up learning, but it can also potentially find better local optima during gradient optimization. For our analysis, we pre-train the best model using weights from training on the ImageNet data set.

Lastly, for hyper-parameter tuning on the best model, we vary batch size and learning rates for tweaking the performance to achieve the benchmark.

5 Results and Interpretation

5.1 Performance of different models

The results on performance for the different models tested are shown in Fig. 10.



Model	Batch size	Epochs	Special pre-processing	Accuracy on test set
Simple model	32	15	no	0.759
MobileNet	32	15	no	0.777
ResNet50	32	15	no	0.751
VGG16	32	15	no	0.819
VGG16 + Augmentation	32	30	Flips, shifts, rotations	0.808
Final model				
VGG16 + ImageNet	32	15	Pretrained on ImageNet	0.869

Figure 10: Comparison of performance of different models

At the end of the model building process, we realize that the pre-trained custom VGG16 model outperforms all others in terms of accuracy. The architecture of the model is shown in Fig. 11. It produces an accuracy of 86.9% on the test set and an AUC of 0.933. This is better than our benchmark on both metrics. In Fig. 12b we can see that the model starts to strongly overfit after 6 epochs.

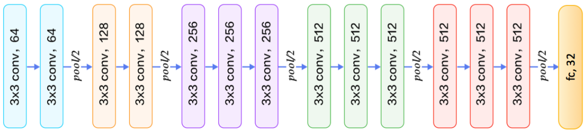
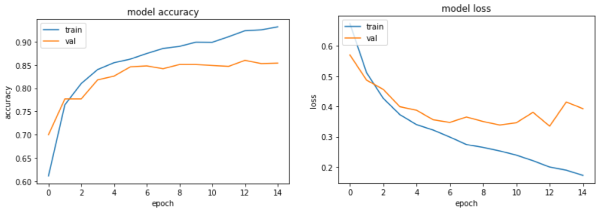


Figure 11: Architecture of the best model



(a) Loss function vs epoch (b) Accuracy vs epoch

Figure 12: Training the best model

Interpretation: Of the three model architectures tested, we see that VGG16 outperforms both ResNet50 and MobileNet. While MobileNet produces lower accuracy by design (for the benefit of training speed), it is surprising that VGG16 outperforms ResNet50. The exact reason for this is not clear as much of deep learning is still based on emperical evidence, we assume that this might be due to the characteristics of the images and that the ResNet50's loss function is stuck on a higher local minimum.

Our second observation is that pre-training delivers a better performance compared to data augmentation. Better performance by pre-training could be because the initial weights might have enabled the model to find a better local minimum of the loss function during the gradient descent process. On the other hand, data augmentation might not have helped as much because of the nature the images, wherein augmentation does not lead to better definition of the features. However, with additional resources it would be interesting to run the model with data augmentation for more epochs since the convergence is slow due to large size of the data.

5.2 ROC Analysis

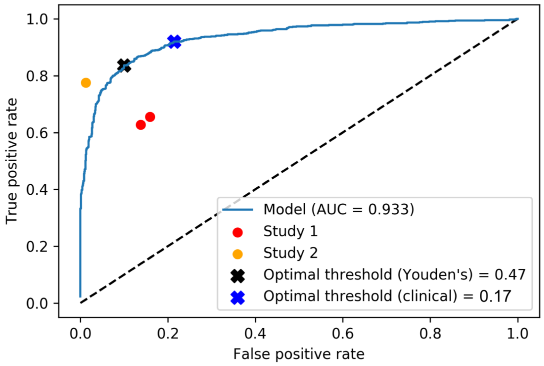


Figure 13: ROC curve of the final model

The ROC curve in Fig. 13 shows that our model has an AUC of 0.93 which is better than our benchmark of an AUC of 0.88 from Shen (2017). Additionally, our model also outperforms radiologists for the comparable task of classifying mammograms as pathological or not. The benchmark of the first study on 12 radiologists on 312 cases with a sensitivity of 65.5% and a specificity of 84.1% Rafferty et al. (2013) was clearly surpassed by our model. For study 2 the physicians did not only use mammography but also other diagnostics, which could be a reason for the better results.

After finalizing the algorithm, we estimate the mathematically optimal mode for the algorithm is, i.e. what the best threshold for the algorithm to declare a mammogram as either pathological or non-pathological. We compute the Youden's J statistic as follows:

$$J = maximumsensitivity(c) + specificity(c) - 1 \tag{1}$$

In different words, this threshold minimizes the error rate of false positive and false negative, taking both as equally important. However, as written in chapter 3.4. from a clinical perspective, reducing false negatives is more important than false positives. Thus, we decided to weigh reducing false negatives twice as important as false positive and calculated the optimal threshold maximizing the cost function = 0.66 * true positive rate + 0.33 * (1-false negative rate). The clinically optimal threshold is 0.17, enabling us to further increase the false positive rate (thereby decreasing the false negative rate) by 10% while increase the false positive rate by 15%.

Conclusively, this model with its well performing accuracy as well as the estimated clinically relevant threshold would be well suited to sufficiently reduce errors, especially false negatives, in the clinical setting.

6 Project Trajectory

As is to be expected, our scope went through several revisions over the course of the project. While some of these changes were a result of hurdles in implementation, others were an outcome of our conversations with physicians and our desire to to make the algorithm more relevant to clinical practice. The major changes were:

- While we started off with the goal of classifying images into multiple categories (benign, malignant, and normal), we eventually shifted our focus to binary classification. This is because classifying a case as normal with higher confidence is more clinically relevant and immediately applicable than multinomial classification. This also made us change the data set we use from CBSI-DDSM, which is an updated, pre-processed version of the DDSM, to the original one.
- Initially, our scope included developing a heat map to help physicians explain the way the algorithm makes classification decisions. However, as we explain further in section 8, our physician interviews reveal that explainability would make sense in the clinical setting only if the heat map identified features that physicians themselves use to assess the BI-RADS category of the image. Therefore, we choose not to pursue this further.
- Finally, we made changes to our implementation plan due to time and resource constraints. We did not focus a lot on hyper-parameter tuning after selecting the best model because the gains we make would probably not improve performance by a lot. Also, we limited the number of iterations on the model once we achieved the desired performance.

7 Conclusions

Our best performing algorithm was a customized VGG16 network which was pre-trained on the ImageNet with an accuracy of 87% and an AUC of 0.933. With that, we met and outperformed our benchmarks. Our clinical analysis shows that sensitivity should be prioritized over specificity in the case of breast cancer. Thus, we choose a clinically optimal classification threshold, which is much lower than the mathematically optimal threshold. This algorithm should, if deployed, will help to significantly reduce the false negative cases of mammograms and increase the chances of 5-year survival.

Strength in this project is the balance between breadth and depth in the scope. Testing different architectures enabled us to identify the best model for the task, without getting into the rabbit hole of testing all different permutations of settings. Once the desired performance is achieved, going into a more thorough analysis of the clinical relevance of such algorithm gives this project ”salience”, an aspect not always considered by researchers dealing with medical applications.

Shortcomings on the other hand were that with more time and computing resources, we could have tried to fine tune our models better, trying different hyper-parameters and building our own network.

8 Future Work

There are several interesting lines of thought we would like to follow in the future. First, instead of a binary classification, it would be interesting to make a categorical classification based on the BI-RADS scores but keeping masses and calcification merged. Expanding on this, it would be great to not only have the BI-RADS score predicted but also if the algorithm can give an explanation why it diagnosed an image with a certain score. Each BI-RADS score has specific requirements which the mammogram has to satisfy. For the algorithm to output these matches of categories would significantly increase the explainability of the algorithm to the doctor, making the results of the ”black-box” more transparent and trusting. The downside is, that we would need additional, very detailed data set containing not only the BI-RADS scores but also these explanations and categories the algorithm could train on.

Furthermore, it would be interesting to incorporate additional information as features into our algorithm. For instance, through our interviews, we know that the tissue density of the woman plays a critical role in the breast cancer assessment. Obtaining this and adding it as a feature could potentially increase the accuracy of the algorithm significantly.

References

Interview with gynecologist, 2nd-year resident, 04/29. 2018.

Interview with radiologist, 4th-year resident, 04/30. 2018.

A. R. Aro, S. P. Absetz, T. M. van Elderen, E. van der Ploeg, and L. T. van der Kamp. False-positive findings in mammography screening induces short-term distress—breast cancer-specific concern prevails longer. *European Journal of Cancer*, 36(9):1089–1097, 2000.

Cancer.gov. https://seer.cancer.gov/statfacts/html/breast.html. 2018.

Cancer.org. https://www.cancer.org/cancer/20breastcancer/detailedguide/breast-cancer-survival-by-stage. 2018.

M. G. Ertosun and D. L. Rubin. Probabilistic visual search for masses within mammography images using deep learning. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pages 1310–1315. IEEE, 2015.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer. The digital database for screening mammography. *Digital mammography*, pages 431–434, 2000.

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

T. M. Kolb, J. Lichy, and J. H. Newhouse. Comparison of the performance of screening mammography, physical examination, and breast us and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. *Radiology*, 225(1):165–175, 2002.

A. McGuire, J. A. Brown, C. Malone, R. McLaughlin, and M. J. Kerin. Effects of age on the detection and management of breast cancer. *Cancers*, 7(2):908–929, 2015.

E. A. Rafferty, J. M. Park, L. E. Philpotts, S. P. Poplack, J. H. Sumkin, E. F. Halpern, and L. T. Niklason. Assessing radiologist performance using combined digital mammography and breast tomosynthesis compared with digital mammography alone: results of a multicenter, multireader trial. *Radiology*, 266(1):104–113, 2013.

S. Shapiro, W. Venet, P. Strax, L. Venet, and R. Roeser. Ten-to fourteen-year effect of screening on breast cancer mortality. *Journal of the National Cancer Institute*, 69(2):349–355, 1982.

L. Shen. End-to-end training for whole image breast cancer diagnosis using an all convolutional design. 2017.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.