

# An Approach to Mammary Cancer Diagnosis Based on Mammography

First Author

zni32@wisc.edu

Second Author

cwei@wisc.edu

Third Author

thirdauthor@wisc.edu

December 9, 2018

## Abstract

The abstract for your project goes here. The length of the abstract should be between 200-250 words. Tips for writing a good abstract can be found at [https://writing.wisc.edu/Handbook/presentations\\_abstracts.html](https://writing.wisc.edu/Handbook/presentations_abstracts.html).

## 1 Introduction

In the America, about 1 in 8 women will develop breast cancer at some point in their lifetime. Over 40,000 women are expected to die in 2018 from this disease. Mammogram screens, which are X-Ray pictures of the breast, are currently the best test for doctors to identify this cancer as early as possible. Sometimes, they are able to identify the disease up to two years before it can be physically felt, which can be crucial in preventing fatality.

Its important to note the difference between two types of mammograms: screening and diagnostic. Mammogram screens are routinely implemented to check for any abnormalities. If an abnormality is found, it calls for further inspection and possibly a follow up diagnostic test. This test could be a diagnostic mammogram or one of the several other alternatives, such as a biopsy. The diagnostic test will result in a benign or malignant decision for the tumor. This project is focuses on screening mammograms, which try to detect and classify any abnormalities that might signify cancer and lead to additional testing.

Despite the benefits from mammogram screens, there are serious downfalls. For starters, nearly 1 in 5 screenings will not detect breast cancer if it is present, result-

ing in a false-negative. Estimates suggest that approximately 10-20% of mammograms without breast cancer will be incorrectly diagnosed with an abnormality, resulting in a false-positive as well. Nearly half of women who get an annual mammogram over a 10-year period will receive an incorrect positive test result because of this high false-positive rate. Not only does this create inefficiencies in the healthcare system, but it also leads to emotional stress and anxiety in women who are incorrectly diagnosed with cancer. Because of these downfalls, many people have questioned whether mammograms should be routinely implemented.

The previous reasons have lead to our motivation in this topic. More effective diagnosis could reduce over-diagnosis, over-treatment, fatality rates, healthcare costs, and financial and emotional burden put on patients. Even marginally small improvements could prove quite meaningful; over 33 million mammograms are performed each year in the United States.

## 2 Related Work

Within the past five years, theres been an increase in research focused on improving the various stages of mammograms via machine learning. From literature, most work falls under two categories: Multi-Stage and End-to-End approaches. The process from screening to diagnosis can be split into three stages: detection, further analysis, and final diagnosis. There has been some, but not much, research on End-to-End approaches, which attempt to tackle all three stages of mammogram assessment.

<https://arxiv.org/pdf/1703.07047.pdf>

Method	Accuracy
Method 1	$70 \pm 3 \%$
Method 2	$76 \pm 3 \%$

Table 1: This is an example of a table.

Most research, however, focuses on addressing a singular stage in this process, such as mass detection. Most similar to our project is the additional research that's been done trying to classify specific regions of interest. Much of this research addresses classification by using deep convolutional neural networks, which is also the method we will be applying. Some of the techniques employed are more advanced and beyond the scope of our knowledge. This research also varies on the resolution of the images used as well as the preprocessing techniques.

High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks (KJ Geras, et al).

### 3 Proposed Method

#### 3.1 KNN

k-nearest neighbors algorithm (KNN) is a non-parametric method used for classification and regression. Here, we will use the KNN to solve the classification problem. In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If  $k = 1$ , then the object is simply assigned to the class of that single nearest neighbor.

## 4 Experiments

#### 4.1 Dataset

The data set is the Mammographic Image Analysis Society (MIAS)(<https://www.kaggle.com/kmader/mias-mammography>) database provided by university of cambridge, which contains 322 images of mammography along with its conditions(see Figure 1).

REFNUM	BG	CLASS	SEVERITY	X	Y	RADIUS
mdb001	G	CIRC	B	535	425	197
mdb002	G	CIRC	B	522	280	69
mdb003	D	NORM				
mdb004	D	NORM				
mdb005	F	CIRC	B	477	133	30
mdb005	F	CIRC	B	500	168	26
mdb006	F	NORM				
mdb007	G	NORM				
mdb008	G	NORM				
mdb009	F	NORM				
mdb010	F	CIRC	B	525	425	33
mdb011	F	NORM				
mdb012	F	CIRC	B	471	458	40
mdb013	G	MISC	B	667	365	31
mdb014	G	NORM				
mdb015	G	CIRC	B	595	864	68

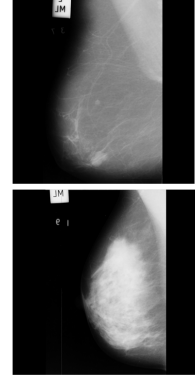


Figure 1: Samples of MIAS dataset.

The first column reference the number of the image. The second column is the character of background tissue: where F means Fatty, G means Fatty-glandular and D for Dense-glandular. The third column is quite important for its the label column of the class of abnormality present.

#### 4.2 Software

Python

#### 4.3 Data preprocessing

The data pre-processing is very important for complex data set. After checking the images of mammography, we noticed several problems.

Firstly, the images are arranged in pairs, where each pair represents the left (even filename numbers) and right mammograms (odd filename numbers) of a single patient. So the first thing is to reverse some of the images to make all of them face the same side.

Secondly, the images are recorded not so well. There are many bright spots and lines in the dark areas of images. We should dealing with these parts very carefully. Generally, we erase the noise on the right and left side of the breast shape by setting the pixels be zero. But images with id 287 and 280 are exceptions. So we designed special functions to pre-process these pictures.

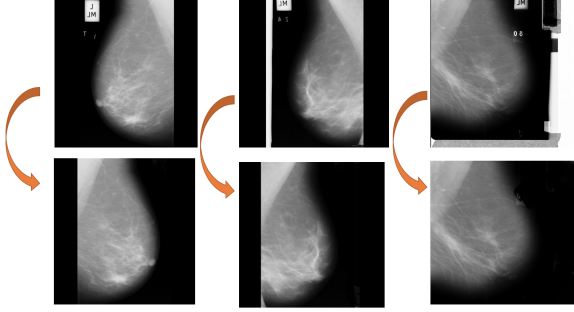


Figure 2: Three samples of preprocessing. (mirroring and noise removing)

Another problem is the high dimension. The images are in shape  $1024 \times 1024$ . This will cost us a lot of time to train the model. So we use the PCA method to deduct the dimension to  $4 \times 1024$  which still save 97% of the information, but only use 1/256 of the original features.

#### 4.4 KNN and random forest

KNN and random forest are two very basic but effective machine learning method. Using KNN and random forest is quite simple for the data has been processed well enough. We simply split the data set into training set(70%) and test set(30%). Since the amount of samples is very small, the number of a certain class may be only 7 or 8, after splitting the original data set into training set and test set, there is no more space to do cross validation.

We tried several coefficients for the two models. After several experiments we find that KNN and random forest both reach an accuracy of 60% to 65%. The accuracy reached the highest when  $n=9$  for KNN, which is 64.65%. While the highest accuracy for random forest is 63.64% when  $n\_estimators = 20$ . It seems KNN has a higher accuracy than random forest, however, if we see it carefully, none of the abnormal class other than CALC is predicted correctly. So basically, its a predictor considering everything normal. As for random forest, though it has a lower accuracy, the recall scores show that this predictor successfully predict a few samples for different classes. I believe with a larger amount of samples, we will get a better

performance. (see Figure 3 and Figure 4)

Accuracy 61.21%				
	precision	recall	f1-score	support
ARCH	0.00	0.00	0.00	9
ASYM	0.00	0.00	0.00	8
CALC	0.00	0.00	0.00	15
CIRC	0.00	0.00	0.00	13
MISC	0.00	0.00	0.00	8
NORM	0.62	0.98	0.76	103
SPIC	0.00	0.00	0.00	9
micro avg	0.61	0.61	0.61	165
macro avg	0.09	0.14	0.11	165
weighted avg	0.39	0.61	0.47	165

Figure 3: The accuracy of raw KNN algorithm.

Accuracy 57.58%				
	precision	recall	f1-score	support
ARCH	0.17	0.11	0.13	9
ASYM	0.00	0.00	0.00	8
CALC	0.17	0.13	0.15	15
CIRC	0.60	0.23	0.33	13
MISC	0.00	0.00	0.00	8
NORM	0.64	0.86	0.74	103
SPIC	0.00	0.00	0.00	9
micro avg	0.58	0.58	0.58	165
macro avg	0.22	0.19	0.19	165
weighted avg	0.47	0.58	0.51	165

Figure 4: The accuracy of raw random forest algorithm.

Its important to emphasize that the cases in this dataset is very complicated. Its harder than the original image processing. The abnormality we are trying to find is located in different places in the images, the breast part locates in different places, and there is even some images with two different types of abnormality. So we need to do more work when we are using KNN and random forest. This will be introduced in the next part.

#### 4.5 Improved KNN and random forest

From the previous result, we can see that the accuracy of the specific classification of the cancer type is extremely bad, which is even all zero (KNN) and about 40% for random forest. We can see that it is not only the problem of the dataset, but also due to the basic structure of the algo-

rithms. From the previous research, we can see that there are some featured problems causing the low accuracy, and in this part, we will try to do some targeted changes to improve the KNN and Random forests. Additionally, to get the more useful result, we reduce the labels into just Normal and Cancer. Then, the problem becomes a Binary classification problem.

As we know from the previous parts, one kind of problems may occur when the position of the cancer core differs. While the whole image is rather big (1024x1024) and the radius of the cores are mostly about 40, the position of the cancer core in the picture has a very vital impact on the accuracy. Thus, we tried to split the whole pic into the small pieces, like small squares. We use the coordinates (x, y) provided in the labels and the median of the radius to extract the cores out of the picture. Then, we randomly pick up squares of the same size from the normal samples. Due to this operation, the effect of the position of cores can be erased because if the cores appear on the picture, they will take the majority of the area. Another benefit of this operation is that the size of the picture has been successfully reduced to 20x20, which is rather smaller than the raw samples. This means the computational efficiency has been improved a lot.

Another problem is the dataset itself. We can see that besides the distribution problem, the whole dataset only contains 330 samples, while some of them are still repetitive. Small size means the outliers may have a big impact on the final result, and also will easily lead to overfitting problems for the models. For the distribution problem, we tried replication way to balance the distribution. Here, for the data size problem, we considered some oversampling techniques. Using elastic distortions, one image can be used to generate many images that are real-world feasible and label preserving. There are several types of elastic distortions, like perspective transforms, rotation, shearing, cropping, and random erasing. With only a few operations, we can produce a large number of samples which is applicable in the model. Here, because the size of the samples is important in our classification, we decide to choose the size preserving ways, rotation and transition. To control the distribution of the dataset, we keep the ratio between the core parts and the normal parts 1:2.

Because the size of the samples is already small enough, we didnt use PCA method to reduce the dimension of the samples this time. We directly apply the KNN

method to the samples. We split the whole samples into training set and test set. Then we use the KNN and random forest to see wether the outcome improved. From the code, we can see that the accuracy of KNN has been increased to 86.15% and the accuracy of the random forest has been increased to the 84.62% which is obviously better compared with former results. (see Figure 5 and Figure 6)

Accuracy 86.15%					
	precision	recall	f1-score	support	
BALA	0.75	0.46	0.57	26	
NORM	0.88	0.96	0.92	104	
avg / total	0.85	0.86	0.85	130	

Figure 5: The accuracy of improved KNN algorithm.

Accuracy 84.62%					
	precision	recall	f1-score	support	
BALA	0.64	0.54	0.58	26	
NORM	0.89	0.92	0.91	104	
avg / total	0.84	0.85	0.84	130	

Figure 6: The accuracy of improved random forest algorithm.

However, the most important part is still a struggle place in this method. The accuracy of the classification of whole mammography is still very low. We tried to use the expectation of the number of negative squares to determine whether the sample is normal or abnormal. But the result is not adorable, which may due to the accuracy of the algorithm being still low. We are still finding how to make it a more applicable method.

## 4.6 CNN

Compared with KNN and random forest, CNN is much better to deal with this kind of problems.

Convolutional Neural Network is really a good method to deal with images. We tried VGG16 in this project (the code and pre-trained model can be

found in <https://www.kaggle.com/kmader/pretrained-vgg16-for-mammography-classification-notebook>). VGG16 is a basic method with totally 16 layers in the model. To use this method, we also balanced the distribution of the dataset since there are 70 percent of samples with class Normal. Then, from the result, we can see that compared with KNN and random forest, CNN is much outstanding. (See Figure 7)

	precision	recall	f1-score	support
ARCH	0.20	0.40	0.27	5
ASYM	0.20	0.33	0.25	3
CALC	0.25	0.88	0.39	8
CIRC	0.33	0.50	0.40	6
MISC	0.15	0.50	0.24	4
NORM	1.00	0.02	0.04	52
SPIC	0.24	0.80	0.36	5
avg / total	0.72	0.24	0.15	83

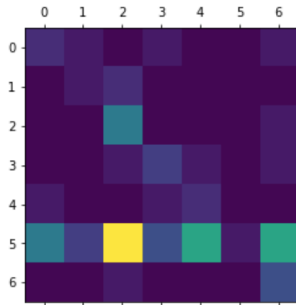


Figure 7: The accuracy of CNN VGG16.

## 5 Results and Discussion

From the result from the experiments (see Table 5), we can see that while treating well processed datasets, KNN and random forest are truly efficient and well-performed. However, when facing the dataset that is not that standard, their limitations are undoubtedly revealed. We can see that KNN and random forest are easily affected by the position and size of the cancer core. And the degree of the gray scale which is determined by the thickness of the fat and muscle also result in the misclassification in the model. Even though the improved KNN and random forest's results seem perfect, the most important part is

still not solved in this report. That is the classification of the whole mammography. We are still work on this part to find a way to accomplish it.

Compared with the KNN and random forest, CNN is much more robust to deal with the noised dataset. While treating the 7 classification problem, we can see that CNN's precision of the classification is much more acceptable than KNN, which is average 72% and its recall ratios for abnormal classes are all high enough. However, it is much more time-consuming than the other two in this problem, while it will use about 380s every epoch and about half an hour all 5 epochs (differs on different computers). Besides, from the outcome of CNN in this problem, we can see that the recall ratio of normal samples are really low. That is because the CNN algorithm class almost all the samples into abnormal ones. This may caused by the distribution balance part. But due to the time limitation, we can't find out why. We will try to fix this problem in the future study.

Method	Precision	Recall
KNN	$39 \pm 3 \%$	$61 \pm 3 \%$
R. F.	$47 \pm 3 \%$	$58 \pm 3 \%$
KNN (Binary)	$85 \pm 3 \%$	$86 \pm 3 \%$
R. F. (Binary)	$84 \pm 3 \%$	$85 \pm 3 \%$
CNN	$72 \pm 3 \%$	$24 \pm 3 \%$

Table 2: The accuracy of the all five experiments.

## 6 Conclusions

Describe your conclusions here. If there are any future directions, you can describe them here, or you can create a new section for future directions.

## 7 Acknowledgements

List acknowledgements if any. For example, if someone provided you a dataset, or you used someone else's resources, this is a good place to acknowledge the help or support you received.

## 8 Contributions

### **Project Report (writing)**

Introduction - Dan

Related Work - Dan

Proposed Method - Nemo, Zheng Ni

Experiments - Nemo, Zheng Ni

Results and Discussion - Dan, Zheng Ni

Conclusions - Dan

Contributions - Dan, Zheng Ni, Chong Wei

### **Computational Tasks**

Preprocessing - Nemo, Zheng Ni

Model Selection- Zheng Ni

Evaluation and Parameter Tuning - Nemo, Zheng Ni

Result evaluating - Nemo, Dan