An Approach to Mammary Cancer Diagnosis Based on Mammography

Zheng Ni zni32@wisc.edu
Nemo Wei cwei48@wisc.edu
Dan Kojis dkojis@wisc.edu

Abstract

Our research goal is to build a model that can correctly classify abnormalities in mammogram screenings. To build the model, 322 mammogram screening images provided by the University of Cambridge were used. Initially, basic kNN and Random Forest approaches were applied. Cross validation was used to with a training and validation set to select the best parameters for our models. We used a 70-30 training to test split for all models discussed. Various steps were taken to preprocess the data and principal component analysis was used for dimension reduction. These results were quite poor. To improve these models, a different image preprocessing approach was taken that takes into account the positioning of the cancer core and the problem was changed to binary classification of either a normal or abnormal mammogram screening. These models performed much better with accuracy ratings around 85%, but they failed to address the original classification problem. Because of this, we turned to an Artificial Neural network approach. Convolutional Neural networks are good with handling images, so tried the VGG16 method, which has 16 layers. The distribution of the dataset was balanced before application. This approach was significantly better than the original kNN and Random Forest approach at handling the noise, with a precision rate of 0.72. However, the recall rate was troublingly low at 0.24 which puts into question the effectiveness of this approach as well.

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.[3] 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.[2]

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, resulting 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 overdiagnosis, overtreatment, 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.[1]

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.[5][6][15]

Most research, however, focuses on addressing a singular stage in this process, such as mass detection.[7][8][13]

Most similar to our project is the additional research thats been done trying to classify specific regions of interest. Much of this research addresses classification by using deep convolutional neural networks[12][11][10], which is also the a method we will 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. The results from this published research ranged from marginally to noticeably more effective at classifying abnormalities when we compared our evaluation metrics.

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.

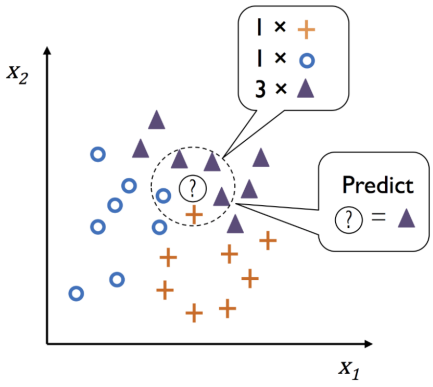

Figure 1. Illustration of kNN for a 3-class problem with k=5.

The K-Nearest Neighbor classifier usually applies either the Euclidean distance or the cosine similarity between the training samples and the test sample. And here we use the Euclidean distance. The Euclidean distance between a training sample and a test sample can be derived as follows:

Let X_i be an input sample with p features $(x_{i1}, x_{i2}, \ldots, x_{ip})$

Let n be the total number of input samples (i = 1, 2, , n)

Let p be the total number of features (j = 1, 2, , p)

The Euclidean distance between sample X_i and $X_t (t = 1, 2, \ldots, n)$ can be defined as

$$d(x_i, x_t) = \sqrt{(x_{i1} - x_{t1})^2 + (x_{i2} - x_{t2})^2 + \ldots + (x_{in} - x_{tn})^2}$$

The nearest sample is determined by the closest distance to the test sample. The K-NN rule is to assign to a test sample the majority category label of its K-Nearest training sample.

In this project, if we apply the kNN directly to the mammographies, the sample is the matrix of mammographies, which can be treated as the following form.

$$< x_1, x_2, x_3, \ldots, x_n >$$

where x_i is 1024×1 vector and n=1024.

3.2. Random forests

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.[9] [4] Compared with common decision, random forests can effectively correct for decision trees' habit of overfitting to their training set.

Bagging The training algorithm for random forests applies the general technique of bootstrap aggregating, or bagging, to tree learners. Given a training set $X = x_1, \ldots, x_n$ with responses $Y = y_1, \ldots, y_n$, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples:

For $b = 1, \ldots, B$:

1.Sample, with replacement, n training examples from X, Y; call these X_b, Y_b.

2.Train a classification or regression tree fb on X_b, Y_b.

After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x':

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x')$$

or by taking the majority vote in the case of classification trees.

This bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias. This means that while the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated. Simply training many trees on a single training set would give strongly correlated trees (or even

the same tree many times, if the training algorithm is deterministic); bootstrap sampling is a way of de-correlating the trees by showing them different training sets.

Additionally, an estimate of the uncertainty of the prediction can be made as the standard deviation of the predictions from all the individual regression trees on x':

$$\sigma = \sqrt{\frac{\sum_{b=1}^{B}(f_b(x') - \hat{f})^2}{B-1}}.$$

The number of samples/trees, B, is a free parameter. Typically, a few hundred to several thousand trees are used, depending on the size and nature of the training set. An optimal number of trees B can be found using cross-validation, or by observing the out-of-bag error: the mean prediction error on each training sample x_i, using only the trees that did not have x_i in their bootstrap sample. The training and test error tend to level off after some number of trees have been fit.

From bagging to random forests The above procedure describes the original bagging algorithm for trees. Random forests differ in only one way from this general scheme: they use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features.

https://en.wikipedia.org/wiki/Random_forest#cite_note-elemstatlearn-3

3.3. ANN

An artificial neural network(ANN) is a network of simple elements called artificial neurons, which receive input, change their internal state (activation) according to that input, and produce output depending on the input and activation. ANN can be divided into a lot of classifications according to the structure of the model (The configurations of VGG models are showed in Figure 2). And what we used in the report is VGG16. VGG16 is a basic method with a total of 16 layers in the model. [14]

4. Experiments

4.1. Dataset

The data set is the Mammographic Image Analysis Society (MIAS)(https://www.kaggle.com/kmader/mias-mammography) database provide by university of cambrige, which contains 322 images of mammography along with its conditions(see Figure 3).

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.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224×224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Figure 2. **ConvNet configurations** (shown in columns). The depth of the configurations increases from the left (A) to the right (E), as more layers are added (the added layers are shown in bold). The convolutional layer parameters are denoted as convreceptive field size-number of channels. The ReLU activation function is not shown for brevity.

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

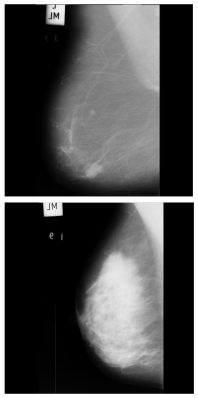

Figure 3. Samples of MIAS dataset.

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

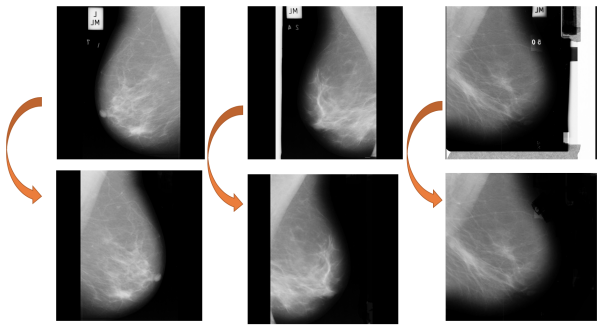

Figure 4. Three samples of preprocessing. (mirroring and noise removing)

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.

Another problem is the high dimension. The images are in shape 1024*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*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 forests are two very basic but effective machine learning method. Using kNN and random forests 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 forests 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 forests 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 5 and Figure 6)

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 5. 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 6. The accuracy of raw random forests 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 algorithms. 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 dif-

fers. 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 forests 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 forests has been increased to the 84.62% which is obviously better compared with former results. (see Figure 7 and Figure 8)

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 7. 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 8. The accuracy of improved random forests 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. ANN

Compared with kNN and random forest, ANN 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, ANN is much outstanding. (See Figure 9)

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 forests 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 forests 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 complish it.

Compared with the kNN and random forest, ANN is much more robust to deal with the noised dataset. While treating the 7 classification problem, we can see that ANN's precision of the classification is much more acceptable than kNN, which is average 72% and its recall ratios for abnor-

	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

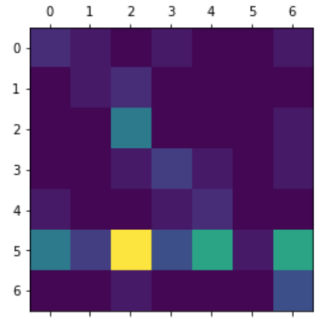

Figure 9. The accuracy of ANN VGG16.

mal 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 ANN in this problem, we can see that the recall ratio of normal samples are really low. That is because the ANN 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 ± 3 %	61 ± 3 %
R. F.	47 ± 3 %	58 ± 3 %
kNN (Binary)	85 ± 3 %	86 ± 3 %
R. F. (Binary)	84 ± 3 %	85 ± 3 %
ANN	72 ± 3 %	24 ± 3 %

Table 1. The accuracy of the all five experiments.

6. Conclusions

The initial goal was to build a model that could effectively diagnose and classify mammogram abnormalities. The results from our ANN model were considerably more effective than the kNN and Random Forest approach, but the results are troubling because of its low recall rate. Compared to past research and literature, these results were not great, as others work has performed noticeably better at times.

However, these results should not be discouraging. Although the model wasnt fantastic and will not be publication worthy, it was a good learning experience with ample room for improvement. More time, experience, and expertise on these topics can eventually lead to better and more meaningful results.

While we focused on the classification of abnormalities in cancer screening, there are still a lot of related avenues for future exploration. While we attempted diagnosis for a binary classification on screening mammograms, future research could extend to studying images that have already been screened and referred to a physical diagnostic test. Additionally, research into diagnostic image testing extends far beyond mammograms and can be applied to an array of health issues. We consider our project a success in the sense that it has deepened our interest in machine learning and public health and will likely encourage a deeper pursuit of these topics, this time with a stronger foundation.

7. Acknowledgements

Our work was conducted for a class project and out of personal interest, and did not receive any funding from outside sources. Our professor, Sebastian Raschka, provided us with much of the foundational material for learning the machine learning techniques applied in this paper.

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 , Dan
Evaluation and Parameter Tuning - Nemo, Zheng Ni
Result evaluating - Nemo, Dan

References

[1] A better future.
[2] Breast cancer detection.
[3] U.s. breast cancer statistics.
[4] I. Barandiaran. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8), 1998.
[5] A. S. Becker, M. Marcon, S. Ghafoor, M. C. Wurnig, T. Frauenfelder, and A. Boss. Deep learning in mammography. *Investigative Radiology*, 52(7):434440, 2017.

[6] G. Carneiro, J. Nascimento, and A. P. Bradley. Unregistered multiview mammogram analysis with pre-trained deep learning models. *Lecture Notes in Computer Science Medical Image Computing and Computer-Assisted Intervention MICCAI 2015*, page 652660, 2015.

[7] N. Dhungel, G. Carneiro, and A. P. Bradley. Automated mass detection in mammograms using cascaded deep learning and random forests. *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2015.

[8] M. G. Ertosun and D. L. Rubin. Probabilistic visual search for masses within mammography images using deep learning. *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2015.

[9] T. K. Ho. Random decision forests. In *Document analysis and recognition, 1995., proceedings of the third international conference on*, volume 1, pages 278–282. IEEE, 1995.

[10] B. Q. Huynh, H. Li, and M. L. Giger. Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. *Journal of Medical Imaging*, 3(3):034501, 2016.

[11] A. Rakhlin, A. Shvets, V. Iglovikov, and A. Kalinin. Deep convolutional neural networks for breast cancer histology image analysis. May 2018.

[12] A. Rampun, B. W. Scotney, P. J. Morrow, and H. Wang. Breast mass classification in mammograms using ensemble convolutional neural networks. *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 2018.

[13] F. Rong, L. Shasha, X. Qingzheng, and L. Kun. A detection algorithm based on convolutional neural network. 2018.

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

[15] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie. Deep multi-instance networks with sparse label assignment for whole mammogram classification. 2016.