Distinguishing benign and malignant breast cancer tissues in breast cancer images using Convolutional Neural Network

Meng Ren, mr3847
Biomedical Engineering
Columbia University
New York, United State
mr3847@columbia.edu

Hong Chen, hc3028
Biomedical Engineering
Columbia University
New York, United State
hc3028@columbia.edu

***Abstract*—This project presents a deep learning approach that automatically detects invasive ductal carcinoma (IDC). Breast cancer has been the most prevalent form of cancer in women in the United States. Countless researchers dedicate their time to find the diagnostic modalities and therapeutic cures to this type of disease. Deep learning learns from data and creates a computational modeling of the learning process, which is similar to how human brain works. Deep learning works to extract the useful information and features that organize them into a hierarchical learned representation. Using DL, specifically Convolutional Neural Network, a model was created to distinguish positive IDC against negative IDC. A modified AlexNet was used as the base architecture. 11 total layers were used, including 5 convolutional layers, 4 fully connected layers, 1 flatten layer, and 1 classification layer. Results showed a 81% accuracy of correctly classifying the two classes. AUROC value was calculated to be 0.89. Robustness testing demonstrated that the model can reliably classify IDC at similar accuracy given different training data size. Future work direction can be focused on using deep learning to identify the area and number of necrotic cells or tumor aggressiveness to further evaluate the severity of the breast cancer.**

***Index Terms*—Breast cancer, convolutional neural network, robustness testing**

I. Introduction

Between 1960s to 1970s, there has been a ten fold increase in the number of breast cancer incidence rates, according to the data provided on the World Health Organization (WHO) [1] Breast cancer has been the most common types of cancer present in the female population. The CDC estimated that in 2015, more than 240,000 new cases of female breast cancer were reported [2]. During the same year, more than 40,000 died of the diseases. This is equivalent to 125 diagnosis and 20 deaths per 100,000 women. It is the most prevalent cancer in women by diagnose rate per 100,000 women, and it is the second deadliest cancer after lung and bronchus cancer. The early detection of breast cancer is extremely useful because it can help reduce breast cancer mortality rate for young women significantly. Distinguish of benign and malignant in pathology image is such a challenge because this process is very time consuming, and that there is a constant shortage of pathologists that can work on this tedious tasks. Computer aided diagnosis (CAD), used to help the detection of IDC in breast cancer because it can distinguish subtle imaging characteristics.

Therefore, there is a need to build a deep learning that can be used to detect breast cancer abnormalities. This supervised learning process can help scientists to create a model that can adequately distinguish breast cancer, specifically between positive and negative IDC, which is a condition that represent 80% of all breast cancer.

The dataset used in this project is breast cancer IDC histology slides. This data is readily from Kaggle https://www.kaggle.com/paultimothymooney/breast-histopathology-images/version/1. Histology slides were chosen for this project because traditionally, invasive breast cancer detection is a time consuming and challenging task because it involves a pathologist scanning a large area of benign regions so that areas of malignancy can be determined. To precisely delineate the region of positive IDC can be crucial in determining the tumor aggressiveness, which in turn determines patient outcome.

The histology slides include 162 whole mount images. However, each image is split into patches of size 50x50, totally 277,524 patches, with 198,738 being IDC negative and 78,786 being IDC positive. This huge amount of patch data means that there are a lot of freedom in terms of data split and training. Because no matter what percentage is the training dataset, there will always be a large amount of testing dataset for verification.

Using Convolutional Neural Network and AlexNet, a deep learning architecture was constructed. CNN was chosen because of its proven success track in pattern recognition and computer vision. This involved multiple non-linear ReLU transformations of the data, with the goal of creating more abstract and useful hidden layer representations. The deep neural network includes convolutional, pooling, fully connected, flatten, and classification layers. The output results showed that each image patch can be either classified as positive or negative IDC. The CNN models performance were evaluated using accuracy test and AUROC values.

II. Material and methods

A. CNN Architecture

A Convolutional Neural Network was used for the image classification of the breast cancer images [3]. In this case, AlexNet was used as the base architecture. The advantage of this network include it solely used ReLU as activation function instead of others such as Tanh, which accelerates the processing speed. AlexNet also use dropout instead of regularization to reduce overfitting. This architecture consist of eleven layers. Zero paddings were all enabled for every layers. The first layer is the input, which in the case of the histology patches, is a 55X55X3 image. This image passes through the first convolutional layer with 32 features, with 11x11 kernel and stride of 4. The second convolutional layer has 64 features, with 5x5 kernel and stride of 1. This layer also has a max pooling with 3x3 size and a stride of 2. The third to the fifth layer are again are convolutional layers, with 3x3 kernel with stride of one. The third and fourth layers have 384 feature maps, and the fifth layer has 256 feature maps. The end of the these three convolutional layers are followed by a maximum pooling with 3x3 size and a stride of 2. The final feature map size is 256. Followed by the five convolutional layers is the flatten layer, which resize the data through a fully connected layer with 9216 feature maps each with size 1x1. This network ended with two dense layers with 2048 units, followed by two more dense layers with 512 and 128 layers consecutively. Finally, softmax function was used squeeze the final classification to between 0 and 1, and the class with the highest value is the output classification. See the image on the right for detailed schematics.

Additionally, a Simple CNN network was created for performance comparison. This is created to see if if deeper network will have an effect on the overall accuracy and AUC. The simple network consists of 2 convolutional layers, 1 flatten layer, 1 FC layer, and the same softmax classification layer.

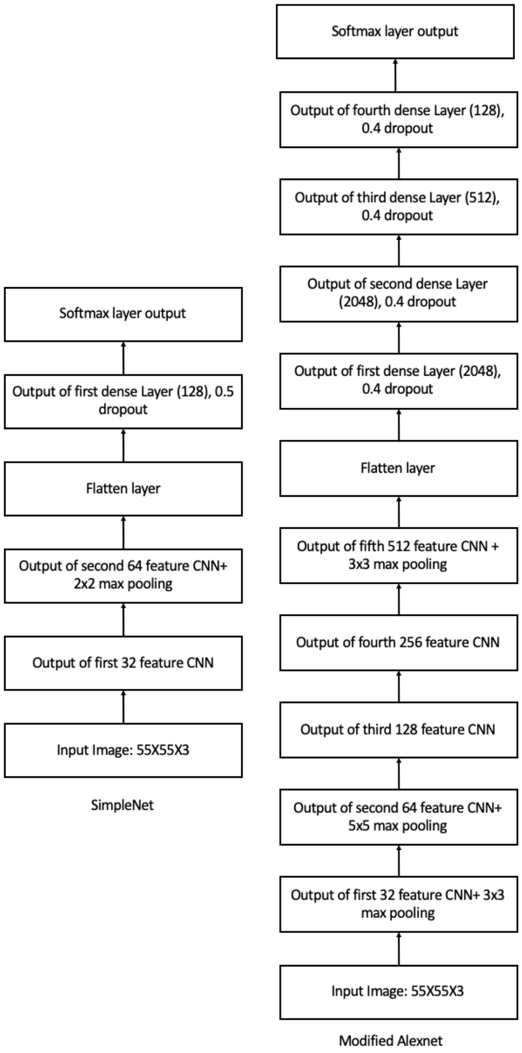
Softmax layer output
Output of first dense Layer (128), 0.5 dropout
Flatten layer
Output of second 64 feature CNN+ 2x2 max pooling
Output of first 32 feature CNN
Input Image: 55X55X3
SimpleNet
Softmax layer output
Output of fourth dense Layer (128), 0.4 dropout
Output of third dense Layer (512), 0.4 dropout
Output of second dense Layer (2048), 0.4 dropout
Output of first dense Layer (2048), 0.4 dropout
Flatten layer
Output of fifth 512 feature CNN + 3x3 max pooling
Output of fourth 256 feature CNN
Output of third 128 feature CNN
Output of second 64 feature CNN+ 5x5 max pooling
Output of first 32 feature CNN+ 3x3 max pooling
Input Image: 55X55X3
Modified Alexnet

Fig. 1. CNN architecture.

B. Dataset Explanation

The pathology image patches were downloaded from the website indicated above in Introduction. Since the images were already in png format, they can be readily read in multiple platforms. So no further image pre-processing was needed. However, the images all contained different naming schemes, which is unnecessary for the purpose of the project. Adding to the fact that there is no csv file that contains the ground truth classification result. But rather the classification is included as the image file name, it was necessary to re-name all the image files and to create a csv file that catalog all the ground truth classification of every single one of the 277,524 patch images. Therefore, a MATLAB script was written to convert all those images into numbered png images, and a classification.csv file was created for performance evaluation purpose such as accuracy and ROC curve. The dataset is split by 20%-80% for testing purpose.

C. Evaluation Metrics

To evaluate the effectiveness of the CNN model, the authors attempted to characterize the total cost of the system in addition to simply defining the accuracy of the system. This means that the target performance parameters are the true positives and false positive that the classifier outputs. To do this, Receiver Operating Characteristic (ROC) plots were created, the Area Under the Curve (AUC) values were also calculated. The AUROC values were easily obtained using Python sklearn libraries using the roc-curve or the auc functions. This is a suitable error metric model for because the results will be binary. In other words, the final classification will output either IDC positive or IDC negative. And ROC curves are typically used in binary classification cases. Additionally, the accuracy of the classification was also calculated for further comparison. The accuracy is calculated by simple division of the correctly classified cases, whether negative or positive, over all cases.

D. Sensitivity and Specificity

The sensitivity, or the true positive rate is calculated as:

$$TPR = \frac{TP}{TP + FN}$$

where TP is true positive, and FN is false negative.
The specificity, or the true negative rate is calculated as:

$$TNR = \frac{TN}{TN + FP}$$

where TN is true negative, and FP is false positive.

E. Precision, Recall, and F1 Score

To further validate the accuracy of the model, precision and recall are calculated. Precision is given by the following formula:

$$precision = \frac{\text{true positive}}{\text{true positive} + \text{ false positive}}$$

Recall is given by the following formula, note that recall is the same as sensitivity.

$$recall = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

After obtaining precision and recall, the F1 score can be calculated using the following formula:

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

F. Robustness Testing

To evaluate the CNN architectures ability to produce a reliable and stable model that adequately captures the difference between malignant and normal breast cancer tissue types, a robustness test was conducted by setting the testing data to 10%, 20% and 30% splits. This robustness test is possible because the dataset is large, as mentioned in introduction, so that there were lots of freedom to choose a high test set percentage while also maintaining a reasonable high amount of training data size.

G. Confusion Matrix

Lastly, a confusion matrix was also constructed for further visualization. This matrix contains information of the number of cases where positive and negative is correctly or incorrectly labeled as positive or negative. Unlike the ROC curve, confusion matrix provides a direct and numerical evaluation of the accuracy of the CNN model.

III. Results and Analysis

Using the modified AlexNet CNN model, the authors tested various different values combinations for the epoch, batch size and learning rate. It was found that the accuracy and AUC values are the most optimal when batch size is 400, epoch is 20, and learning rate is 0.00001. The figure shows the train loss history. The train loss history first rapidly decreased and decayed to around 0.4. No further decrease in the loss was observed.

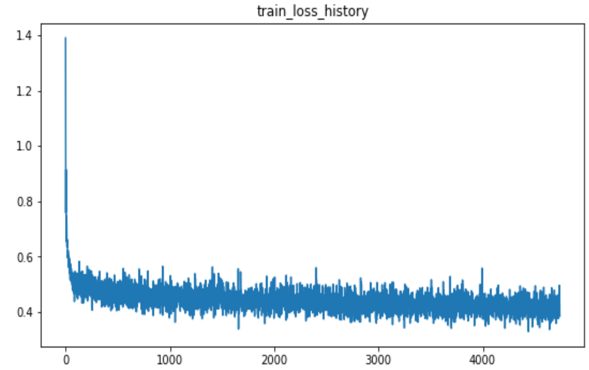

Fig. 2. Train loss history.

With this parameter settings, the accuracy was calculated to be 81%. This value is quite reasonable and very comparable to the other publications that similarly use CNN networks to classify the same dataset images. Additionally, the ROC curve was constructed. From which the AUC values are found to be 0.89 for 20% test split. This is also reasonably adequate and shows a good balance between sensitivity and specificity. This is further verified by the sensitivity and specificity scores calculated using the formula provided in the methods section. The sensitivity is found to be 79% while the specificity is found to be 83% Which is quite balanced because although there is a need to accurately report patients positive conduction, it is almost equally important to reduce the number of false positives as false diagnosis can be traumatic to patients who are healthy all along. Therefore a 79% sensitively and 83% selectivity strikes the balance.

The precision and recall scores are calculated as 83% and 79% respectively. The precision score quantifies the ability of a classifier to not label a negative example as positive. And the current precision demonstrates a good performance. F1 score, the combination of precision and recall via their harmonic mean, is found to be 0.81. This further validates the good performance of the model.

The misclassified images were also plotted for visual confirmation. By visually examining the image, it was clear that the left and right columns exhibit different characteristics. The predicted negatives, i.e. left column all have large sac features, while the predicted positives do not. This might have caused the model to misclassify. In the future, the model can be improved by emphasizing more on the colors. For example, the actual negatives, i.e. top row, have more reddish color while the actual positives have more purplish color.

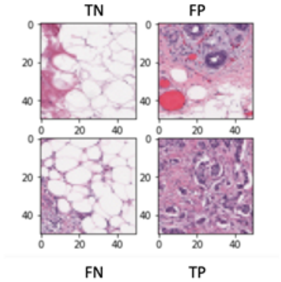

Fig. 3. Images that is classified as true negative, false positive, false negative and true positive.

Confusion matrix for the 20% data split was created to visualize the correct and incorrect classification. It is a extension to the accuracy test. From the matrix, the 81% accuracy can be confirmed, and it was observed that the misclassified cases are split evenly among positive and negative cases. In other words, it is equally likely for a positive to be classified as negative than negative to be classified as positive.

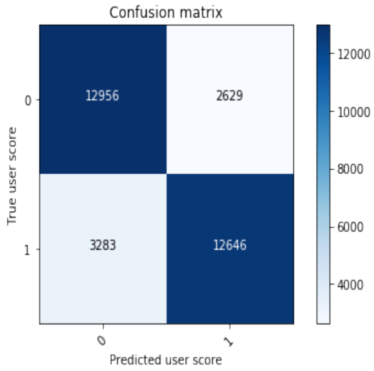

Fig. 4. Confusion matrix.

Robustness testing are conducted using 10%, 20% and 30% testing. The table shows that throughout all different split percentage, the AUROC remained quite constant around 0.89. The Accuracy of the model also remained quite constant, at around 81%. This demonstrates that the modified AlexNet model can reliably classify the IDC pathology images given different amounts of training and testing data. However, it should also be taken note that the existing dataset size is really large. So there might not be a huge difference between different splits since the model may already obtain most of the feature information from the training dataset of the most test splits.Lastly, simple net produced an accuracy of about 0.75 and AUC of 0.8. Both are a lot worse than the results produced by the modified Alexnet, this demonstrates that deeper network may produce better performance.

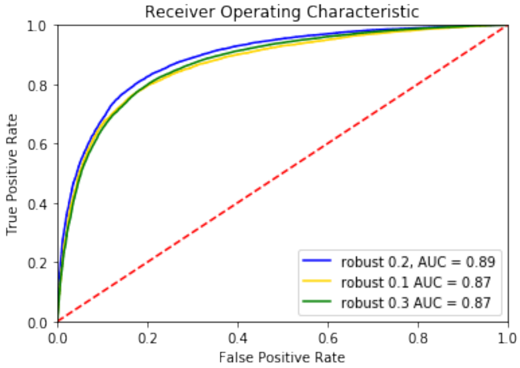

Fig. 5. ROC curve based on different robustness testing.

TABLE I
AUROC, ACCURACY, SENSITIVITY AND SPECIFITY BASED ON DIFFERENT ROBUSTNESS TESTING.

Splite%	**AUROC**	**Accuracy**	**sensitivity**	**specificity**
10%	0.87	0.79	0.74	0.85
20%	0.89	0.81	0.79	0.83
30%	0.87	0.80	0.78	0.82

IV. CONCLUSION

In conclusion, automated detection of invasive ductal carcinoma is a challenging and relevant problem for breast cancer diagnosis. An accurate and reproducible detection of positive IDC is a important step because it can help eliminate the time that pathologists spend on classifying positive and negative IDC regions. An important characteristics of this approach is that CNN is a truly learn from the data model that can perform on operator independent classification tasks.

V. FUTURE WORK

Future development can focus on comparing the CNN models to existing classification methods, such as handcrafted features that measure the luminance, color, and texture of the images. It is also possible to explore CNN with deeper architectures (more neurons and layers) and a validation of larger cohorts. This is supported by the fact that the deeper Alexnet architecture produced better results than the simple net. In addition, it is also interesting to explore if machine learning can be used to identify the area or number of necrotic cells within the pathology slides, this may be used to shed light on the severity of Breast Cancer.

TABLE II
PRECISION, RECALL AND F1 SCORE BASED ON DIFFERENT ROBUSTNESS TESTING.

Splite%	**precision**	**recall**	**F1 score**
10%	0.84	0.74	0.78
20%	0.83	0.79	0.81
30%	0.82	0.78	0.80

References

[1] W. B. Sampaio, E. M. Diniz, A. C. Silva, A. C. D. Paiva, and M. Gattass, Detection of masses in mammogram images using CNN, geostatistic functions and SVM, Computers in Biology and Medicine, vol. 41, no. 8, pp. 653664, 2011.

[2] USCS Data Visualizations. Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, gis.cdc.gov/Cancer/USCS/DataViz.html.

[3] R. Rouhi, M. Jafari, S. Kasaei, and P. Keshavarzian, Benign and malignant breast tumors classification based on region growing and CNN segmentation, Expert Systems with Applications, vol. 42, no. 3, pp. 9901002, 2015.

[4] Cruz-Roa, Angel, et al. Automatic Detection of Invasive Ductal Carcinoma in Whole Slide Images with Convolutional Neural Networks. Medical Imaging 2014: Digital Pathology, 2014, doi:10.1117/12.2043872.