Machine Learning on Cancer Prediction

UNIVERSITY OF TORONTO

Shuqing Wu, Mahesh Sudhakar, Yuanziwei Li

UNIVERSITY OF TORONTO, DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Introduction

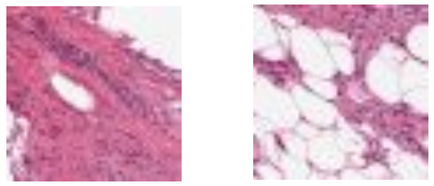
Nowadays, cancer has been prevalent, which is one of the most dangerous diseases that leads to a high mortality rate. In order to detect abnormal issues earlier and offer timely treatment for the patient, we are looking for an approach to enhancing the cancer diagnosis process with this project. The main objective of this project is to scan the image from a part of the human body such as lungs or breast and implement a machine learning algorithm to identify the likelihood of the patient having cancer in that area. We will be taking a set of body scan or microscopic tissue images as inputs, decide how we are going to identify whether the patients have cancer, based on biomedical analysis and label the set with whether they have cancers or not, also train the model with a large amount of labeled images, and finally make use of the algorithm to predict the likelihood of the patient having cancer.

Dataset

"Invasive Ductal Carcinoma (IDC)" is one of the most common subtype of breast cancers. In order to identify breast cancers, pathologists typically focus on the regions containing the IDC. Here we make use of mount slide images of Breast Cancer (BCa) specimens scanned at 40x.

The tissue screenshots are selected from Kaggle Breast Cancer classification datasets. It contains different folders named with patient IDs. Within each folder, we have 2 image folders with label 1 or 0 representing "the screenshot areas have cancer" or "the screenshot areas do not have cancer" respectively. The datasets are collected from a number of patients with cancer. The pathology images containing IDC are labeled as 1 and the ones not containing IDC are labeled as 0.

Below are some samples pathology image <u>without</u> IDC cancer.



Below are some samples pathology image <u>with</u> IDC cancer.

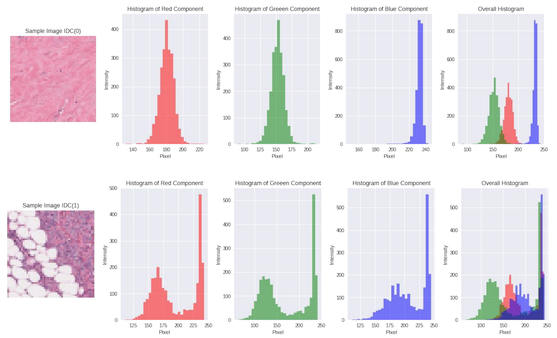
Methods

In this project, the core idea is to train the machine learning model with labeled images to obtain a diagnosis of cancer for the patient. We have applied several image pre-processing steps and machine learning algorithm to build, train, and test the model.

Data Preparation

For datasets without labeling, we will need to manually label the images with cancer or let doctors and medical professionals to manually product the training set label. In this case, we are using the dataset with label, so we will take the dataset directly, and separate them for the purpose of evaluating the model constructed. We build dataset by separating the label 0 images and label 1 images into training, testing, and validation sets.

To extract the information from images accurately, we also apply an image processing algorithm such as histogram equalization to balance the intensity of the images and make sure they are consistent for the training model.



Data Training

Convolutional Neural Network (CNN) is utilized to extract features from the pathology images and to perform image classification. CNN has high accuracy and a less free parameter, which are great for large-scale datasets. We have constructed CancerNet as our model, and we utilize CNN, RELU activation and pooling layer as the 3 layers of our network. After constructing the model, we fit the model on our training set. Then, we run the fitted model on testing and validation set to generate prediction and accuracy results.

Programming Environment

This project will be implemented in Python and will utilize the Python Image Library, Keras, matplotlib, and Tensorflow Library.

Results

We first train the model on 1 epoch using the dataset. The result produced is below:

Loss : 0.4198
Accuracy : 0.8240
Validation loss : 0.4837
Validation accuracy: 0.8149

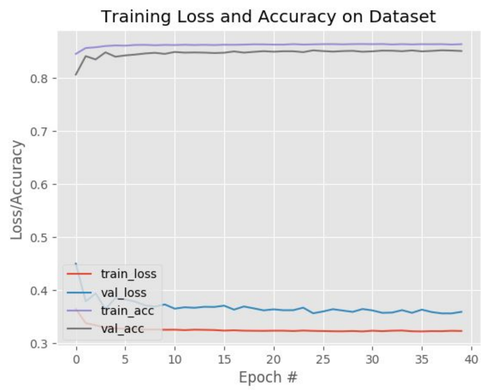
Sensitivity : 0.8140
Specificity : 0.8158

Normalized Confusion Matrix:

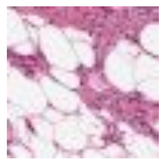
True Positive - 0.8158
False Positive - 0.186
True Negative - 0.814
False Negative - 0.1842

Label	Precision	Recall	F1-score	Support
0	0.92	0.81	0.86	39711
1	0.64	0.82	0.71	15794
Avg/Total	0.84	0.81	0.82	55505

We then trained the model on over 40 epochs using the dataset given. The result graph is below:



Users can also apply the fitted network model on testing one single image, and get the prediction and accuracy data.



Test Image

Loss : 0.4198
Accuracy : 0.8240
Validation loss : 0.4837
Validation accuracy: 0.8149
Sensitivity : 0.8140
Specificity : 0.8158

Conclusions

A novel model for training a large number of labelled pathology images, then to test and validate on the other set of pathology images to predict cancer has been proposed. The solution produces training accuracy, results and labeling. We may also be able to input one image and predict the likelihood of whether that person has breast cancer or not along with the accuracy of the results.

Our solution enhances the application of Machine Learning in medical field, and demonstrates how CNN can be applied in medical imaging. It validates the research results of some current ongoing medical research but also enhance our understanding of medical imaging.

With the model and application, we can enhance it in the future to make it fit into other different dataset, such as images for other parts of the body or for images showing other types of cancer or other types of diseases. As precision enhances, this method will be more and more useful and prevalent.

Bibliography

1. Kaggle breast cancer dataset with labels in each filename, https://www.kaggle.com/paultimothymooney/breast-histopathology-images

2. CancerNet algorithm, by Adrian Rosebrock, https://www.pyimagesearch.com/2019/02/18/breast-cancer-classification-with-keras-and-deep-learning/

3. Kadir T, Gleeson F. 'Lung cancer prediction using machine learning and advanced imaging techniques'. Transl Lung Cancer Res 2018;7(3):304-312. Doi: 0.21037/tlcr.2018.05.15

4. McWilliams A, Tammemagi MC, Mayo JR, et al.Probability of Cancer in Pulmonary Nodules Detected on First Screening CT. N Engl J Med 2013;369:910-9

5. Breast cancer dataset: Kinahan, Paul; Muzi, Mark; Bialecki, Brian; Coombs, Laura. (2017). Data from ACRIN-FLT-Breast. The Cancer Imaging Archive. http://doi.org/10.7937/K9/TCIA.2017.ol20zmxg