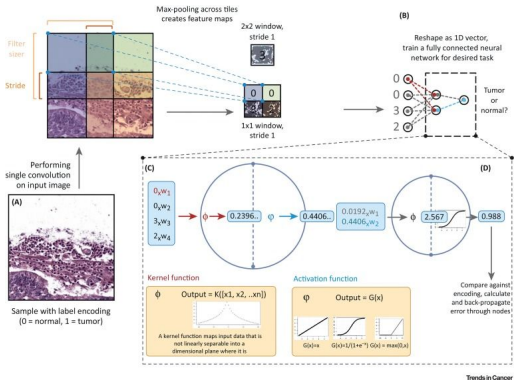
Breast Cancer Classification

Regis University - MSDS696 - Data Science Practicum II
Student: Mary Crawley
Summer 2022

Contents

Problem or Situation

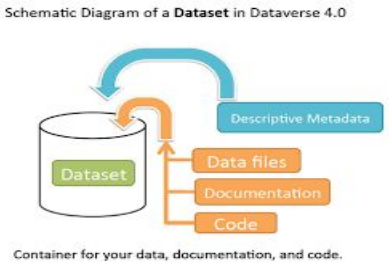
The most common breast cancer is called Invasive Ductal Carcinoma (IDC). Assigning an aggressiveness grade to a whole mount sample, pathologists normally focus on the areas which contain the IDC. This results in a common pre-processing step that aimS for an automatic aggressiveness grading by delineate the exact regions of IDC inside of a whole mount slide.



Research Question

Can a Deep Learning classifier model help increase the accuracy and reduce time in determining whether a histology image is benign or malignant for potential Breast Cancer patients?

The Dataset



Data used: [Kaggle-Breast Cancer Prediction Dataset]

We will be using the IDC_regular dataset (breast cancer histology image dataset) from Kaggle. There is approximately 277,524 patches that are the size of 50x50 extracted from 162 whole mount slide images of breast cancer specimens that are scanned at 40x. The data holds 198,738 test negative and 78,786 test positive with IDC. The dataset is available for the public and can download it here. The size of the data is a minimum of 3.02GB of disk space for this project.

Filenames in this dataset look like this:

8863_idx5_x451_y1451_classo

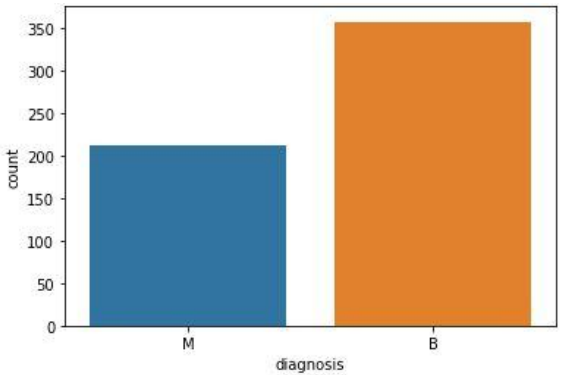
Here, 8863_idx5 is the patient ID, 451 and 1451 are the x- and y- coordinates of the crop, and 0 is the class label (0 denotes absence of IDC).

Methodology

This project will include the use of Keras that will define a CNN (Convolutional Neural Network) to train it on our images. There will be a directory for each patient ID that holds 0 and 1 directories for images. This will be used to split our dataset into training, validation, and testing sets in an 80% for training with 10% of that reserved for validation, and 20% for testing. Using tuples (Python list that can hold a sequence of items) for information about the training, validation, and testing sets.

Results 1: Dropping ID and Unnamed Data

- After removing patient ID’s and data not Malignant or Benign, we have 2 columns remaining.
- There are 212 Malignant and 357 Benign tumors in our data after cleaning.

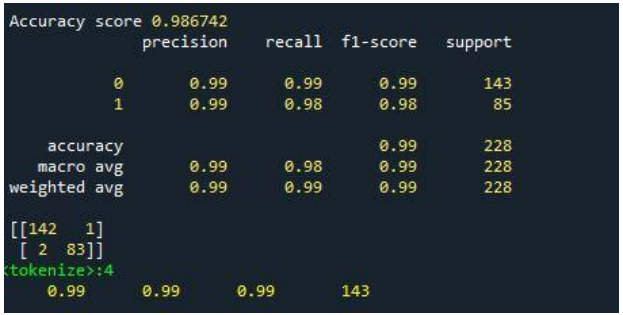


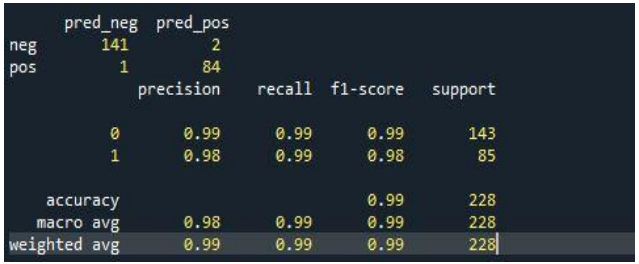
212 Malignant and 357 Benign tumors

Results 2: Hyper Parameter Tuning

Minimizing misclassifications for Malignant ‘M’ and False Negatives ‘FN’.

- Top: After grid searching, ‘M’ is a 2 score (we don’t want FN’s to show as malignant so need it a 1.
- After setting a decision threshold of 0.42, we now have FN to 1.





Results 3: Deep Learning Training

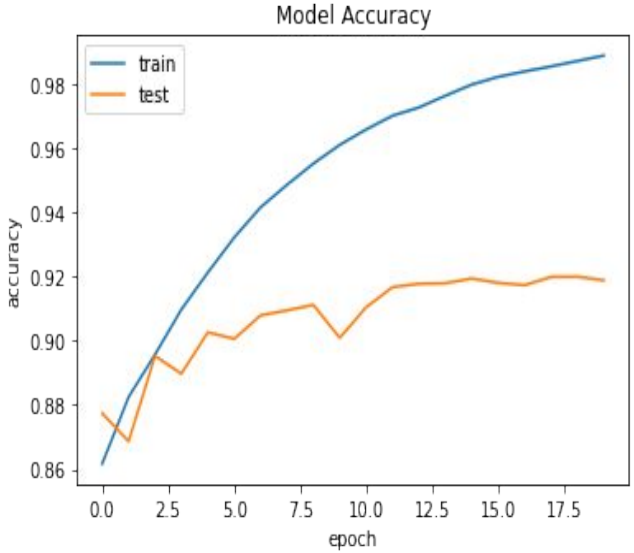
Initialize the training, validation, & testing generators so they can generate in bathes.

- The image shows the output after successfully training our model and computing the confusion matrix and get our raw accuracy

```
Epoch 4/20
3694/3694 - 20s - loss: 0.2445 - accuracy: 0.9097 - val_loss: 0.3116 - val_accuracy: 0.8897
Epoch 5/20
3694/3694 - 20s - loss: 0.2159 - accuracy: 0.9212 - val_loss: 0.3229 - val_accuracy: 0.9026
Epoch 6/20
3694/3694 - 20s - loss: 0.1902 - accuracy: 0.9321 - val_loss: 0.2878 - val_accuracy: 0.9006
Epoch 7/20
3694/3694 - 20s - loss: 0.1675 - accuracy: 0.9415 - val_loss: 0.2969 - val_accuracy: 0.9079
Epoch 8/20
3694/3694 - 20s - loss: 0.1486 - accuracy: 0.9485 - val_loss: 0.3385 - val_accuracy: 0.9094
Epoch 9/20
3694/3694 - 20s - loss: 0.1300 - accuracy: 0.9551 - val_loss: 0.3466 - val_accuracy: 0.9111
Epoch 10/20
3694/3694 - 20s - loss: 0.1153 - accuracy: 0.9609 - val_loss: 0.3609 - val_accuracy: 0.9009
Epoch 11/20
3694/3694 - 20s - loss: 0.1020 - accuracy: 0.9657 - val_loss: 0.3862 - val_accuracy: 0.9104
Epoch 12/20
3694/3694 - 20s - loss: 0.0910 - accuracy: 0.9700 - val_loss: 0.4074 - val_accuracy: 0.9167
Epoch 13/20
3694/3694 - 20s - loss: 0.0821 - accuracy: 0.9726 - val_loss: 0.4166 - val_accuracy: 0.9177
Epoch 14/20
3694/3694 - 20s - loss: 0.0726 - accuracy: 0.9762 - val_loss: 0.4214 - val_accuracy: 0.9179
Epoch 15/20
3694/3694 - 20s - loss: 0.0641 - accuracy: 0.9797 - val_loss: 0.4578 - val_accuracy: 0.9194
Epoch 16/20
3694/3694 - 20s - loss: 0.0577 - accuracy: 0.9821 - val_loss: 0.4861 - val_accuracy: 0.9180
Epoch 17/20
3694/3694 - 20s - loss: 0.0518 - accuracy: 0.9838 - val_loss: 0.5292 - val_accuracy: 0.9173
Epoch 18/20
3694/3694 - 20s - loss: 0.0465 - accuracy: 0.9853 - val_loss: 0.5917 - val_accuracy: 0.9198
Epoch 19/20
3694/3694 - 20s - loss: 0.0424 - accuracy: 0.9870 - val_loss: 0.6124 - val_accuracy: 0.9199
Epoch 20/20
3694/3694 - 20s - loss: 0.0383 - accuracy: 0.9887 - val_loss: 0.6672 - val_accuracy: 0.9188
```

Results 4: Deep Learning Outcome

- Accuracy Plot (Top): Achieved a 98.87% after cleaning our data from missing status of Malignant or Benign and removing ID’s.



Conclusion

The deep learning AI training is a success and distinguishes which images are benign and malignant breast cancer from a combination of small imaging using Deep Learning Python with a 98.87% success rate after cleaning the dataset, and having the False Negatives set to 1 to not show as Malignant by Hyper Parameter Tuning set to 0.42

References

Data Files (image amount is too large to place in my Github):

https://www.kaggle.com/datasets/paultimothymooney/breast-histopathology-images

Github Link:

https://github.com/CrawleyM29/PractiumII-BCC