Project Proposal

Regis University - MSDS696 - Data Science Practicum II

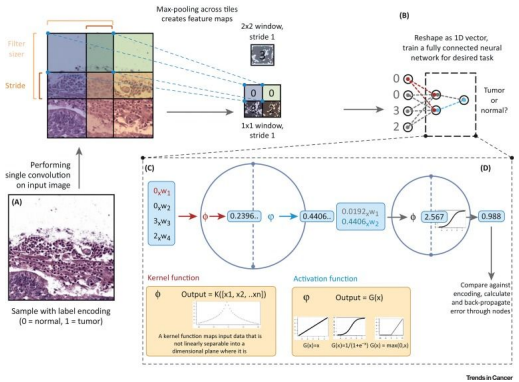
Student: Mary Crawley

Summer 2022

Contents

Problem or Situation

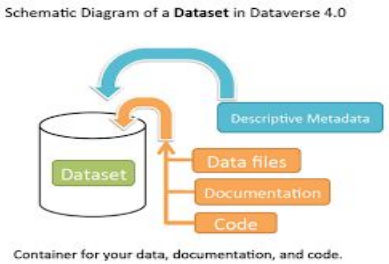
The most common breast cancer is called Invasive Ductal Carcinoma (IDC). Assigning an aggressiveness grade to a whole mount sample, pathologists normally focus on the areas which contain the IDC. This results in a common pre-processing step that aim for an automatic aggressiveness grading by delineate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

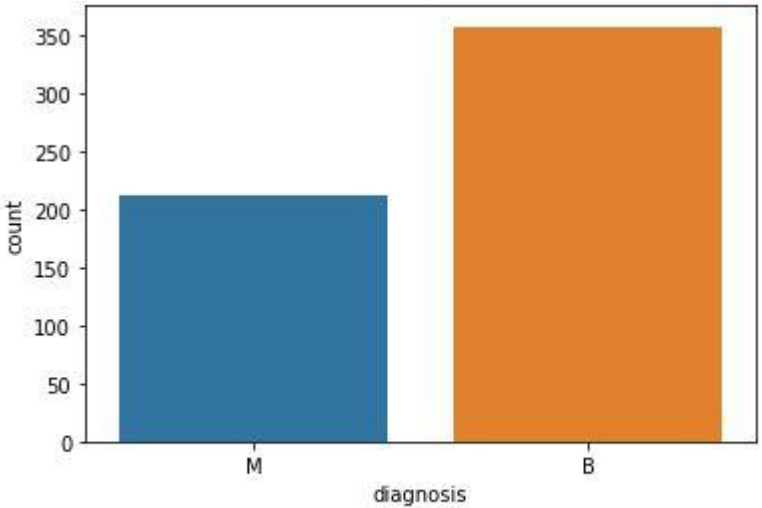
Here, 8863_idx5 is the patient ID, 451 and 1451 are the x- and y- coordinates of the crop, and 0 is the class label (0 denotes absence of IDC).

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 Unamed

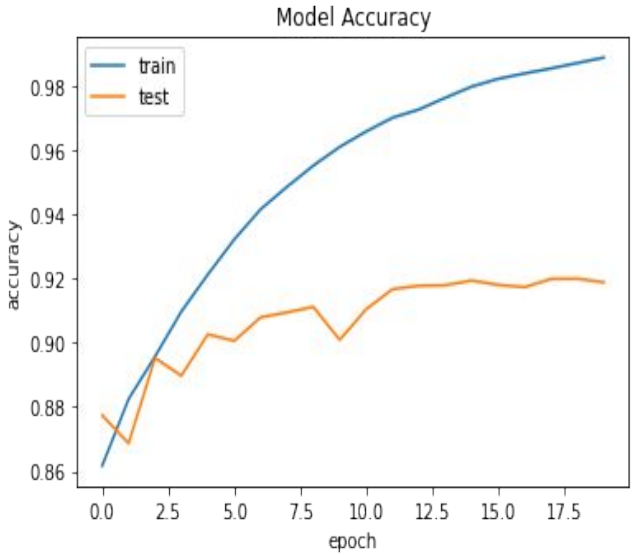
- After removing patient ID’s and those that are not named malignant or benign, we have 2 columns that show the total that are left.
- There are 212 Malignant and 357 Benign tumors in our data.



212 Malignant and 357 Benign tumors

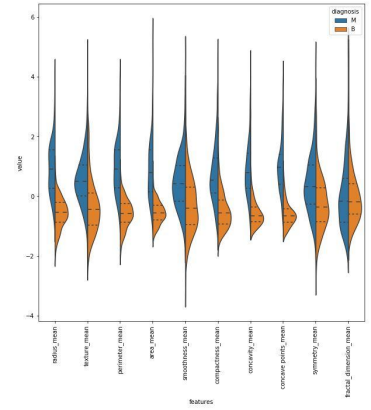
Results 1: Training Data Using Deep Learning

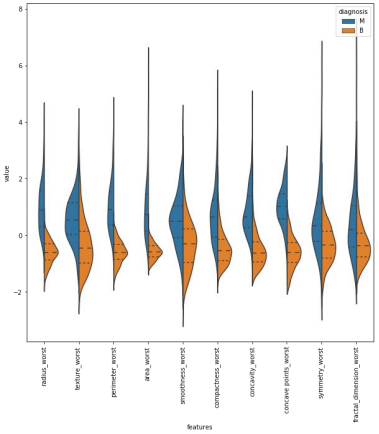
- Accuracy Plot (Top): Achieved a 98.87% after applying the binary-cross-entropy for loss function and Adam optimizer for optimization.



Results 2: Violin Plot

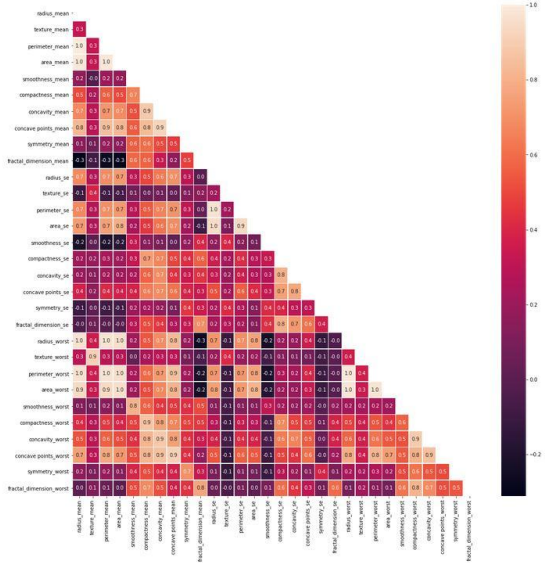
- Top: median of texture_mean for Malignant and Benign looks separated and close to each other for fractal_dimension_mean.
- Bottom: The shape of the violin plot for area_se looks warped and distribution points for benign and malignant are very different. Variance looks highest for fractal_dimension_worst. Concavity_worst and Concave_points_worst look to have a similar data distribution.





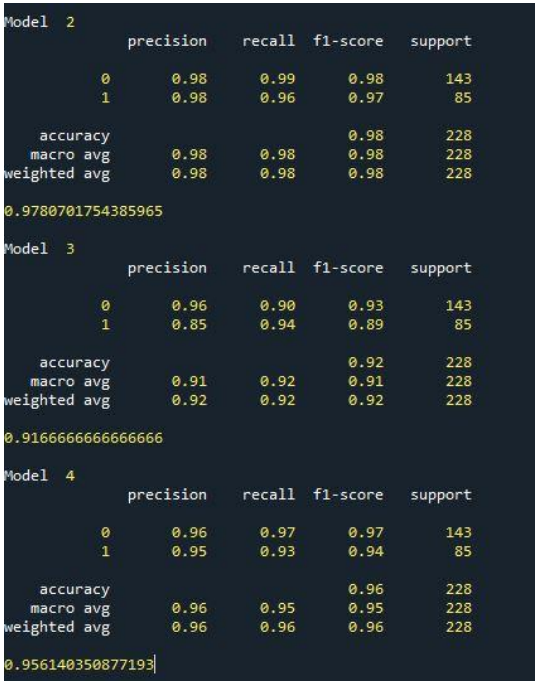
Results 3: Heatmap

- With correlation > 0.8, means, std errors and worst dimension lengths of compactness, concavity and points of the concave are high in correlation to one another.
- Mean, worst dimensions of radius, std errors, area and perimeter of tumors have a correlation of 1.



Results 4: Hyper Parameter Tuning

- After grid searching, accuracy improved a little but the FNs are still 2.
- After custom threshold to increase recall, FN reduced to 1 after setting decision threshold to 0.42



```
Model  2
              precision    recall  f1-score   support

           0       0.98      0.99      0.98       143
           1       0.98      0.96      0.97        85

    accuracy                           0.98       228
   macro avg       0.98      0.98      0.98       228
weighted avg       0.98      0.98      0.98       228

0.9780701754385965

Model  3
              precision    recall  f1-score   support

           0       0.96      0.90      0.93       143
           1       0.85      0.94      0.89        85

    accuracy                           0.92       228
   macro avg       0.91      0.92      0.91       228
weighted avg       0.92      0.92      0.92       228

0.9166666666666666

Model  4
              precision    recall  f1-score   support

           0       0.96      0.97      0.97       143
           1       0.95      0.93      0.94        85

    accuracy                           0.96       228
   macro avg       0.96      0.95      0.95       228
weighted avg       0.96      0.96      0.96       228

0.956140350877193
```

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 and using exploratory data to understand our dataset for better results.

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