Big Data Analytics Symposium - Fall 2019

Analytics Project: Early Detection of Breast Cancer Using Spark MLlib and its Analysis

Team Members:

- Ishan Khanka
- Shashank Lochan
- Anisha Salunkhe

Abstract:

Breast cancer is causing widespread deaths and is one of the most common causes of cancers encountered around the world. These deaths can be reduced by early detection of cancerous cells. Unfortunately, not all physicians are experts in distinguishing between the benign and malignant tumors and the classification of tumor cells may take up to two days.

We used the distributed computation framework of Hadoop - Apache Spark for quicker analytics and faster model generations to aid in early detection of breast cancer. The analytic deals with the diagnosis of breast cancer by analyzing survey images of breast cancer cells and raw text data consisting of characteristics of the cell nuclei present in the image.

Motivation

Who are the users of this analytic?

- Cancer Research Laboratories
- Breast Cancer Pathologists
- Healthcare Professionals

Who will benefit from this analytic?

- Breast Cancer Patients
- Hospitals
- Research Institutes

Why is this analytic important?

- It is estimated that 41,400 deaths (40,920 women and 480 men) from breast cancer have occurred in the year 2018. And on an average it takes two days to classify the tumor cells.
- This analytic will reduce the time for classifying the tumor cell and additionally find out the most efficient magnification for the image and important features to look out in a cell.
- The number of published papers which utilize spark as a solution are rare.
- Detection of cancer in early stages increases the chance for a successful treatment.

Goodness

What steps were taken to assess the ‘goodness’ of the analytic?

- Our domain expert, Ms. Nikita Meghani (MS in Biology, New York University) with her expertise in cancer research, helped us build our machine learning model and validated the entire process from a clinical perspective.

- Additionally to validate our model we divided the labeled dataset into training and testing and used only the training data for generating the model. In the end, we cross-referenced the accuracy of the model by using the test data.

Data Sources

1) **Name**: Breast Cancer Histopathological Image Classification (BreakHis)

   Description: It is composed of 9,109 microscopic images of breast tumor tissue collected from 82 patients using different magnifying factors (40X, 100X, 200X, and 400X). To date, it contains 2,480 benign and 5,429 malignant samples.

   Size of data: 4GB

2) **Name**: Breast Cancer Wisconsin (Diagnostic) Data Set

   Description: This dataset contains tumor features acquired from a digitized image of a fine needle aspiration biopsy of the tumor. This is a labeled dataset with 32 tumor attributes of 569 subjects, where the patients are labeled as benign or malignant of breast cancer (357 benign and 212 malignant).

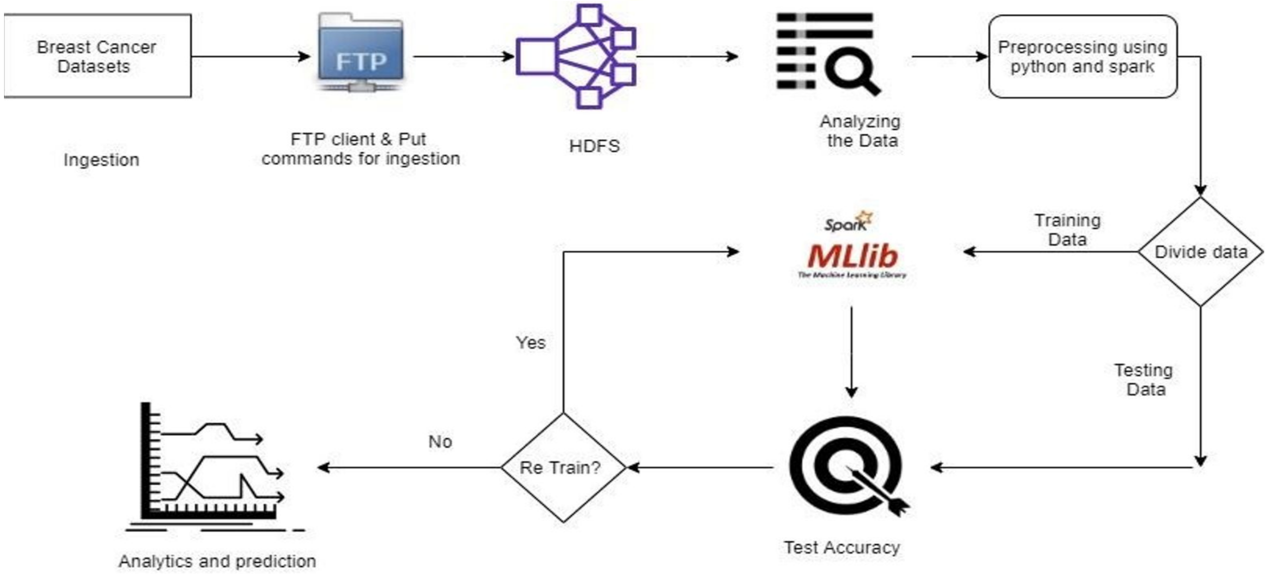
   Size of data: 1-2GB

3) **Name**: BreCaHAD: A dataset for breast cancer histopathological annotation and diagnosis

   Description: A dataset of 162 breast cancer histopathology images, namely the breast cancer histopathological annotation and diagnosis dataset (BreCaHAD) which allows researchers to optimize and evaluate the usefulness of their proposed methods. The dataset includes various malignant cases.

   Size of data: 3.47GB

Design Diagram



Platform(s) on which the analytic ran:

- NYU HPC Dumbo Cluster
- Local machine running Ubuntu - 18.04.2 LTS

Early Detection of Breast Cancer Using Spark MLlib and its Analysis

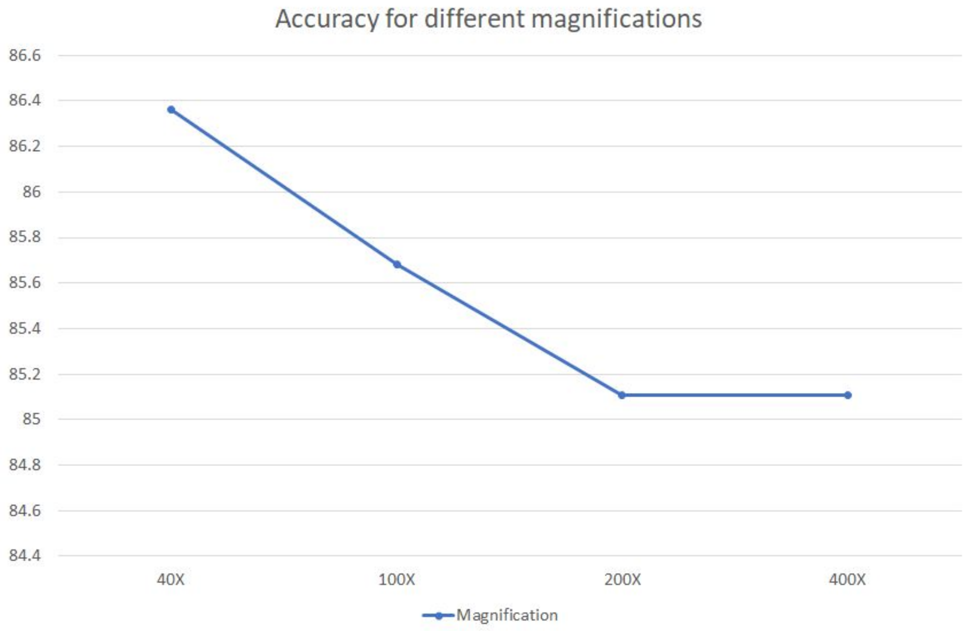
Results

- Highest accuracy for the 40X magnified image after running it against multiple models

- Helpful to reduce the total training time by 75%

- Better accuracy than a model running on a single machine

- Big advantage of reducing the time for code execution using distributed processing on a cluster of machines.

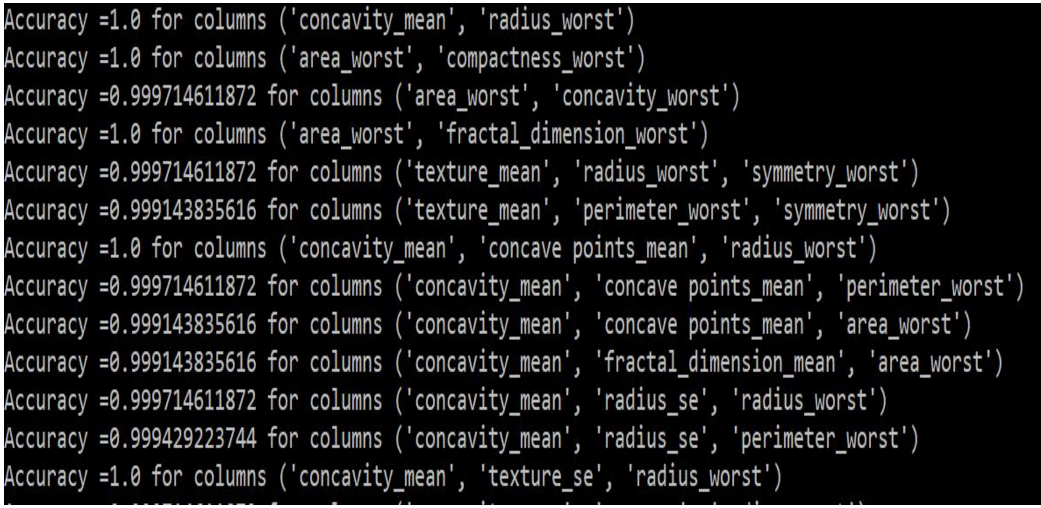


Results

- Total number of permutations for 30 features = 1073741823. Est. time to run on Dumbo is 30 years and on a single machine 130 years

- Goal to find the minimum relevant features for the best accuracy

- Running a model against the possible combinations of features, we found that the following were enough to accurately distinguish between tumors: concavity_mean, radius_worst, area_worst, compactness_worst, texture_mean, symmetry_se



```
Accuracy =1.0 for columns ('concavity_mean', 'radius_worst')
Accuracy =1.0 for columns ('area_worst', 'compactness_worst')
Accuracy =0.999714611872 for columns ('area_worst', 'concavity_worst')
Accuracy =1.0 for columns ('area_worst', 'fractal_dimension_worst')
Accuracy =0.999714611872 for columns ('texture_mean', 'radius_worst', 'symmetry_worst')
Accuracy =0.999143835616 for columns ('texture_mean', 'perimeter_worst', 'symmetry_worst')
Accuracy =1.0 for columns ('concavity_mean', 'concave points_mean', 'radius_worst')
Accuracy =0.999714611872 for columns ('concavity_mean', 'concave points_mean', 'perimeter_worst')
Accuracy =0.999143835616 for columns ('concavity_mean', 'concave points_mean', 'area_worst')
Accuracy =0.999143835616 for columns ('concavity_mean', 'fractal_dimension_mean', 'area_worst')
Accuracy =0.999714611872 for columns ('concavity_mean', 'radius_se', 'radius_worst')
Accuracy =0.999429223744 for columns ('concavity_mean', 'radius_se', 'perimeter_worst')
Accuracy =1.0 for columns ('concavity_mean', 'texture_se', 'radius_worst')
```

Obstacles

- Very few online resources / forums for our problem statement as not much research was done on this approach using distributed computing. And the ones that were available were outdated.

- Spark is highly version dependent when combining with other libraries like image processing and analytics. Difficulty in changing the versions for different libraries on Dumbo without disrupting the workflow of others on the cluster

Summary

- For image analysis: our methods gave a higher accuracy of 85.4% when compared against other Breast Cancer Prediction approaches Using Spark MLlib and ML Packages where the accuracy achieved was only 83%.
- Images of 40X magnification give us the best accuracy.
- For textual analysis: we found the minimum features that could accurately distinguish between the tumors.
- Image analysis combined with the textual analysis allows for the effortless analysis in the absence of clear images that are clinically rendered unusable for research.

Acknowledgements

- We would to like to thank the NYU HPC team for their continued support and Cloudera for providing CDH through the Cloudera Academic Partnership
- Additionally, we would also like to extend our gratitude to, Ms. Nikita Meghani (MS in Biology, New York University), our domain expert, for providing her expertise in helping us build our machine learning model and validating the entire process from a clinical perspective.

References:

[1] Umesh D. R, B. Ramachandra: Big Data Analytics to Predict Breast Cancer Recurrence on SEER Dataset using MapReduce Approach

[2] David A. Omondiagbe, Shanmugam Veeramani Machine Learning Classification Techniques for Breast Cancer Diagnosis

[3] Heyam H. Al-Baity, Sara Alghunaim: On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context

[4] Madhuri Gupta, Bharat Gupta: Survey of Breast Cancer Detection Using Machine Learning Techniques in Big Data

[5] Mitko Veta*, Josien P. W. Pluim, Paul J. van Diest, and Max A. Viergever: Breast Cancer Histopathology Image Analysis: A Review

[6] S.Sujithra, Dr.L.M.Nithya, Dr.J.Shanthini: Predicting Breast Cancer using Apache Spark Machine Learning Logistic Regression.

[7] Duy Hung, Tran Duc Hanh, Vu Thu Diep, Breast Cancer Prediction Using Spark MLlib and ML Packages. I

[8] P.J. Sudharshan, Caroline Petitjean, Fabio Spanhol, Luiz Eduardo Oliveira, Laurent Heutte, Paul Honeine, Multiple instance learning for histopathological breast cancer image classification, Expert Systems with Applications Volume 117, 1 March 2019, Pages 103-111

[9] Xinpeng Xie, Yuexiang Li, Linlin Shen, Active Learning for Breast Cancer Identification, arXiv:1804.06670, 18 Apr 2018

[10] David A. Omondiagbe, Shanmugam Veeramani, Amandeep S. Sidhu, Machine Learning Classification Techniques for Breast Cancer Diagnosis, et al 2019 IOP Conf. Ser.: Mater. Sci. Eng. 495 012033

[11] K. ShailajaM. A. Jabbar, Prediction of Breast Cancer Using BigData Analytics, September 2018 DOI: 10.14419/ijet.v7i4.6.20480

[12] Anusha Bharat,Pooja N, R Anishka Reddy: Using Machine Learning algorithms for breast cancer risk prediction and diagnosis, 2018, IEEE Third International Conference on Circuits, Control, Communication and Computing

Early Detection of Breast Cancer Using Spark MLlib and its Analysis

Thank You