Early Detection of Breast Cancer Using Spark MLlib and its Analysis

Ishan Khanka
CIMS, New York University
New York, United States of America
ik1304@nyu.edu

Shashank Rajiv Lochan
CIMS, New York University
New York, United States of America
shashank.lochan@nyu.edu

Anisha Salunkhe
CIMS, New York University
New York, United States of America
anisha.salunkhe@nyu.edu

***Abstract*—**

Cancer is a disease in which cells in the body grow and divide beyond the control. Breast cancer is the second most common disease after lung cancer in women. Incredible advances in health sciences and biotechnology have prompted a huge amount of gene expression and clinical data. Machine learning techniques are improving the prior detection of breast cancer from this data. The research work carried out focuses on the application of machine learning methods, data analytic techniques, tools, and frameworks in the field of breast cancer research with respect to cancer survivability, breast cancer prediction, and detection using text and image analysis. Apache Spark data processing engine is found to be compatible with most of the machine learning frameworks.

Keywords—analytics, machine learning, spark, histology, Breast cancer, computer-aided diagnosis, digital pathology, histopathology, image analysis, Wisconsin breast cancer data, digital mammogram

I. Introduction

Breast cancer is causing widespread deaths and is one of the most common causes of cancers encountered around the world. It is close to one-fourth of all cancers and cancer-related deaths in women across many countries. These deaths can be reduced by early detection of cancerous cells. It is thus a clarion call for the early detection of the cancerous tumor that gives us the opportunity to take preventive action in a timely manner during its initiation to work towards its treatment and potentially eradicating the risk of proliferation. Given the availability of open and public data about patients along with crucial data like breast scans and tumor cell characteristics, it provides us with the foundations to build upon our solution to detect the early onset of breast cancer in a time-efficient manner with high accuracy that can prove to be a boon for all the medical practitioners in making informed decisions.

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. Machine learning algorithms are used to predict the type of cancerous cells and efficiently and accurately[17].

The analytic deals with the diagnosis of breast cancer by analyzing survey images of breast cancer cells and raw text data consisting of the characteristics of the cell nuclei present in the image which play a significant role in the early detection and predictions, based on symptoms that would help in providing customized treatments to cancer patients. However, until now we have been relying on pathologists for the stage - based diagnosis, which has significantly appeared to be a time-consuming process. In this approach, we will use the distributed computation framework of Hadoop through Apache Spark for faster training along with the timely and early prediction of breast cancer. The objective of these predictions is to assign patients to either as ‘benign’ group that is noncancerous or a ‘malignant’ group that is cancerous.

II. Motivation

The traditional data analytic might not have the capacity to handle an enormous amount of data with accuracy. Due to the rapid growth of information, solutions need to be contemplated and provided in order to handle and extract value and knowledge from these data sets. In this paper, we address the problem of breast cancer prediction in the big data context. The research work carried out focuses on the application of machine learning methods, data analytic techniques, tools, and frameworks in the field of breast cancer research with respect to cancer prediction.

It can approximately take 10 minutes up to 2 days for a pathologist on an average to examine a stained image to come up to a rough approximation of whether the image could indicate the cells to be malignant or benign. Although there have been certain implementations of machine learning models for automated detection of breast cancer, the training stage of this analytic is the biggest bottleneck in the detection mechanism system. The focus of this research is to integrate

these machine learning techniques with feature selection/feature extraction methods on a high performance computing cluster using distributed processing frameworks and compare their performances to identify the most suitable approach to perform the analytics.

III. Related Work

Numerous techniques have been implemented for breast cancer prediction in the past few decades. The most recent publications of breast cancer classification are presented in this section.

The paper "Predicting Breast Cancer using Apache Spark Machine Learning Logistic Regression" [9] provides a lucid explanation of how Breast Cancer Diagnosis poses a great challenge to the researchers and how many scientific technologies have rich information in making medical decisions but that might not be accurate and properly used to its potential. The paper also discusses how the use of machine learning and data mining techniques has revolutionized the whole process of breast cancer Diagnosis. The authors in their system have summarized different types of data mining algorithms to obtain a good mortality rate. They made us of strong and sophisticated algorithms like Bagging Logistic Regression, Support Vector Machine, k-Nearest Neighbors algorithm, Decision tree, and Artificial Neural Networks and concluded that there was not a single best algorithm depending on the features of the large dataset, but it was also a challenge for single-node tools with limited memory and computing power. In their research, the main goal is to identify the sample observation as malignant or not by their exact attributes through logistic regression analysis which improves performance through intelligent optimizations and also to achieve single node analysis through spark framework. The key difference in our implementation is that we will be harnessing the computational power of distributed processing of enormous medical data that can be efficiently handled by Apache Spark for large scale data analysis over a diversity of workloads and that makes it one of the best open-source tools to harness the power of big data and create solutions for social good.

This paper serves as a reference point for us to ponder upon the optimum approach to go about breast cancer detection solutions and provides us an insight into one of the attractive options available in the open-source world ie. Apache Spark and the analysis of different algorithms to solve the problem. Consider this paper as a starting point or a milestone along the journey of researching and developing the optimal approach towards a breast cancer detection system.

The paper Breast Cancer Prediction Using Spark MLlib and ML Packages [10] discusses the advances of machine learning and how it can be applied to various applications, especially in healthcare. The dataset used here is the routine blood analysis for detecting breast cancer. They argue that various factors such as Leptin, Adiponectin, MCP, BMI, etc. that can be utilized for testing breast cancer. The columns in the dataset are Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP1. Furthermore, they have used the Decision tree algorithm for training the model. In the conclusion of the paper, the author discusses that various factors influence and cause breast cancer and hence we can get better results if we don't limit the input to the to just 10 factors in blood testing for generating feature vectors. They propose that other sensors with the advents of IoT can be used in healthcare to gather more data and in turn lead to the inception of models with better accuracy. This paper is helpful for the research that we are performing as it proves that Hadoop packages such as MLLib can be used for machine learning and provides a baseline for accuracy. But there would be key differences in our approach. Firstly, we will use microscopic images of breast tumor tissue as the dataset rather than blood analysis samples. Additionally, we will use image classification techniques for image analysis to give us better accuracy rather than the decision tree model used here.

The paper "Survey of Breast Cancer Detection Using Machine Learning Techniques in Big Data" [4], states that as reported by the World Health Organization, breast cancer is the most prominent problem in the area of medical diagnosis, which is in increasing numbers every year. Machine learning techniques are improving the prior detection of breast cancer from available data such as from magnetic resonance imagery (MRI), superresolution digital microscopy, mass spectrometry, etc. The research work carried out on this paper focuses on the application of machine learning methods, data analytic techniques, tools, and frameworks in the field of breast cancer research with respect to cancer survivability, cancer recurrence, cancer prediction, and detection. This paper primarily explains the basic categorization of Breast Cancers by describing their stages and their different types of malignity.

It further proceeds with the introduction of Machine learning as to how the algorithms work to train the data. This background information was provided to link their usages derived by their purpose in order to understand and analyze the different types of applications that can be used to implement the analytics in terms of 2 phases. The Survey of Machine Learning Application in Breast Cancer explains the Metrics used for measuring the performance of different techniques, which are accuracy, error rate, survival time, specificity, sensitivity, etc. The paper further explains the different frameworks which could be used for performing machine learning. This reference is used for understanding why spark can be used for the most appropriate machine learning techniques and the solutions that can be achieved along with the accuracy gains. The research essentially does a compare and contrast of exploring different technologies and frameworks and gives you an analytic on how different use

cases of breast cancers can best fit with different types of frameworks thereby deriving the right set of comparative analytics This research work can be further extended by incorporating feature reduction techniques to improve performance metrics.

IV. Design and Implementation

A. Design Details

The design diagram fundamentally describes the overall workflow of the analytic that we have performed over the Breast cancer datasets. The workflow begins with migrating the datasets which we have selected for training the model into the HDFS via the FTP client from an external source to the machine hosting Hadoop. These datasets are related to the domain of Breast Cancer and their early prognosis. The HDFS further analyses the data in a distributed manner which helps in finding out meaningful patterns for us to derive towards a certain direction of conclusions. After undergoing certain preprocessing where we link and filter out certain data we reach a stage where we can segregate the usage of it for training purposes. We furtherly divide the data into Testing or Training for Machine Learning and increasing the result accuracy respectively using the libraries. We are using MLLib for training the model over the datasets. It undergoes a loop of retraining until a certain standard is achieved to then lead to the final stage. This is the result that would further lead to drawing visualized analytics and predictions for enabling a faster approach to have early detection of malignant cells along with further classification or the degree of malignity observed in a patient's body.

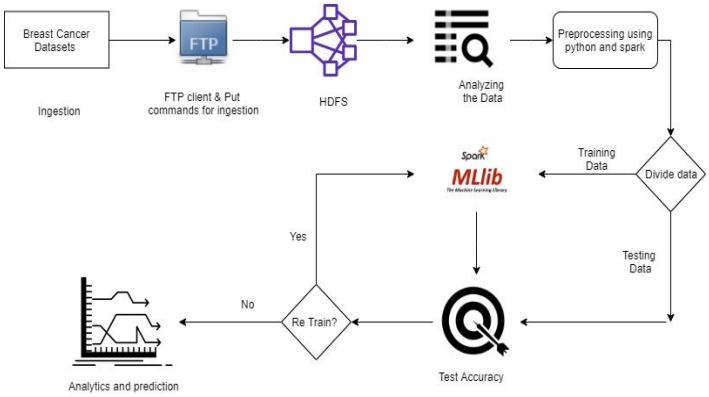

Figure 1: Data Flow Diagram

V. Datasets

A. Breast Cancer Histopathological Image Classification (BreakHis)

It is composed of 9,109 microscopic images of breast tumor tissue collected from 82 patients using different magnifying factors (40X, 100X, 200X, 400X). It has 2,480 benign and 5,429 malignant samples of resolution 700X460 pixels.

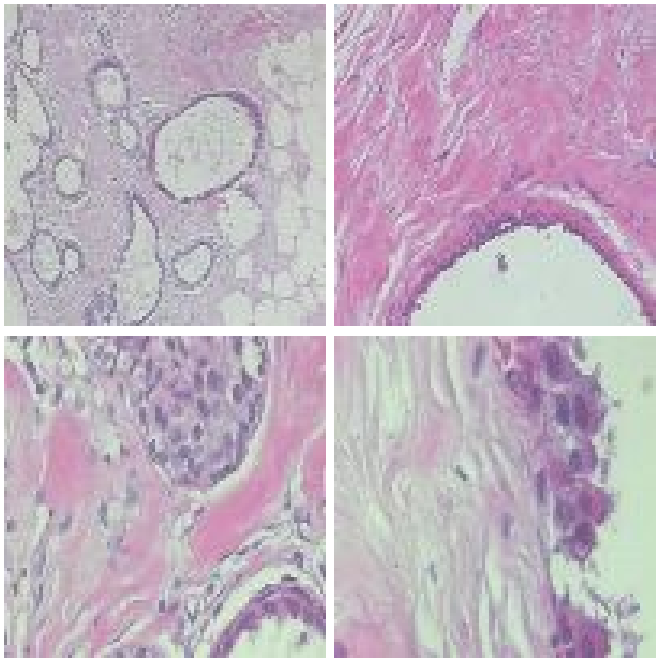

Figure 2: From left to right sample and top to bottom, sample images of Benign Cells of magnifications 40, 100, 200, 400 respectively

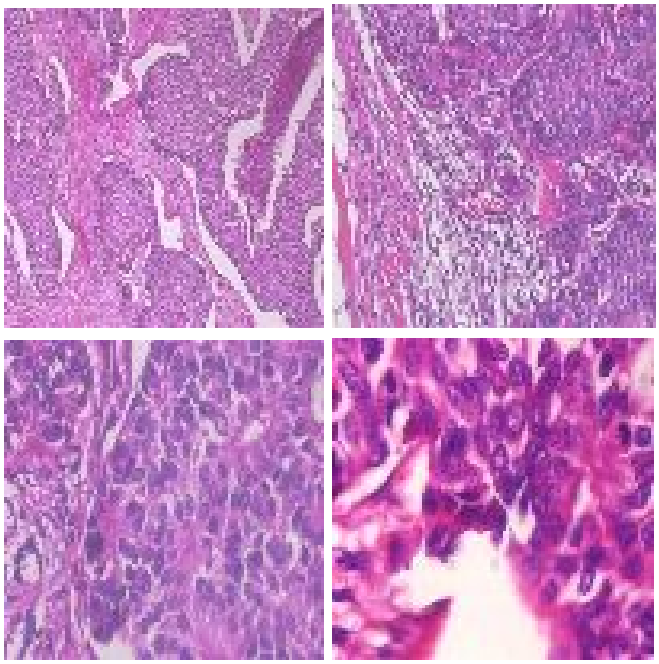

Figure 3: From left to right sample and top to bottom, sample images of Malignant Cells of magnifications 40, 100, 200, 400 respectively

Link:https://web.inf.ufpr.br/vri/databases/breast-cancer-histopathological-database-breakhis/

B. UCI Machine Learning Laboratory: Breast Cancer Wisconsin (Diagnostic) Data Set

For research purposes, the Wisconsin Diagnostic data from Breast Cancer Data Set of UCI machine learning repository is used for making predictions. This dataset is selected mainly for its reliability and is a publicly available real-world breast cancer data set. 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). For every patient, 10 attributes of cell nuclei are gathered: radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, concave points and fractal dimension. Then various statistical measures like mean, standard error and "worst" of these 10 attributes are calculated which outcomes of 30 features.

Table 1: Features Used

Radius	Mean of distances from the center to points on the perimeter
Texture	The standard deviation of grey-scale values
Perimeter	The total distance between the snake points constitutes the nuclear perimeter.
Area	Number of the pixel on the interior of the snake and adding one-half of the pixel in the perimeter
Smoothness	Local variation in radius length, quantified by measuring the difference between the length of a radial line and the mean length of lines surrounding it.
Compactness	Perimeter ^2 / area
Concavity	The severity of concave portions of the contour
Concave points	Number of concave portions of the contour
Symmetry	The length difference between lines perpendicular to the major axis to the cell boundary in both directions.
Fractal dimension	Coastline approximation. A higher value corresponds to a less regular contour and thus to a higher probability of malignancy.

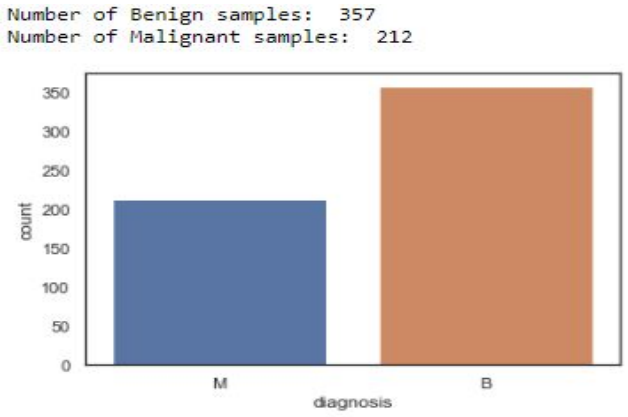

Figure 4: Count of patients diagnosed with benign and malignant cancer

Link:https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

C. BreCaHAD: A Dataset for Breast Cancer Histopathological Annotation and Diagnosis

This dataset consists of 1 .xlsx file, 2 .png files, 1 .json file and 1 .zip file. The distribution of annotations in the previously mentioned six classes (mitosis, apoptosis, tumor nuclei, non-tumor nuclei, tubule, and non-tubule) is presented in an Excel spreadsheet. There are two sets of images, annotated and original. In the annotated images, The circles represent different types of nuclei and their malignity. The annotations for the BreCaHAD dataset are provided in JSON (JavaScript Object Notation) format. An archive file which contains the dataset includes 3 sets of images (original images), ground truth (JSON files), and groundTruth_display (ground truth applied on original images)

Link:https://figshare.com/articles/BreCaHAD_A_Dataset_for_Breast_Cancer_Histopathological_Annotation_and_Diagnosis/7379186

VI. RESULTS

Table 2: Test environment configuration for image analysis:

Spark	2.3
Python	3.6
Tensorflow	1.4.0
JDK version	1.8.0_222
Image resolution	100x100
Min Memory	16 GB
OS	Ubuntu
Numpy	18.04.2 LTS
Keras	2.1.5
OS Version	1.17.4
Pillow	6.2.1

Table 3: Test environment configuration for textual analysis:

Spark	1.6
Python	2.7.13
Databricks package	2.10
Hive	1.1.0
Operating System	Linux (Centos 6.9)
Master Nodes	2x12-core Intel "Haswell" CPUs 256GB memory 8TB RAID1 disk

Though the experiment we have uncovered the potential use case for using SparkML for analytics on a big dataset of Histopathological images of breast cancer tissue and the Wisconsin breast cancer text dataset. Spark gives us a similar and sometimes better accuracy than a model running on a single machine. But running the ML model on Spark provides a big advantage of reducing the time for code execution using distributed processing on cluster of machines.

All the images were downscaled to 100x100 for faster training and prediction. Figure 5 shows the remodeled dataset used for the experiments.

```
AVG HxW =  100 100 for folder  filtered_dataset/b40  No of images: 625
AVG HxW =  100 100 for folder  filtered_dataset/b100  No of images: 644
AVG HxW =  100 100 for folder  filtered_dataset/b200  No of images: 623
AVG HxW =  100 100 for folder  filtered_dataset/b400  No of images: 588
AVG HxW =  100 100 for folder  filtered_dataset/m40  No of images: 1370
AVG HxW =  100 100 for folder  filtered_dataset/m100  No of images: 1437
AVG HxW =  100 100 for folder  filtered_dataset/m200  No of images: 1390
AVG HxW =  100 100 for folder  filtered_dataset/m400  No of images: 1232
```

Figure 5: Average resolution of the images in the dataset and the images for each folder. For the folder name -Letter B stands for Benign and M for malignant followed by a number representing image magnification

Additionally, we found out that for the images of different magnifications (40X, 100X, 200X, 400X), the highest accuracy was obtained for the least magnified image(40X). This can be helpful for model training as it will reduce the total training time by 75%. As shown in Figure 6 below.

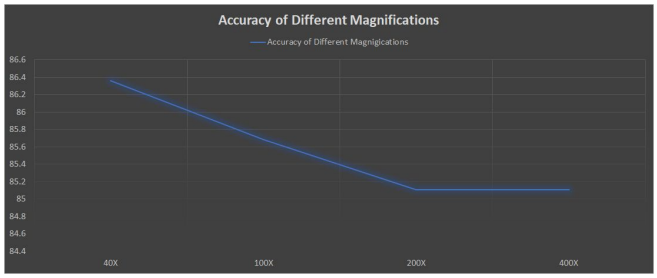

Figure 6: Accuracy of a model using different magnifications

For textual analytics, we used a linear regression model and calculated new statistical features using the original dataset, namely- radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concavepoints_mean, symmetry_mean and fractal_dimension_mean. By using exhaustive combinations of different features and running a linear regression model against them, we found that any new statistical feature calculated above was enough to accurately distinguish between benign and malignant tumors as shown in figure 7.

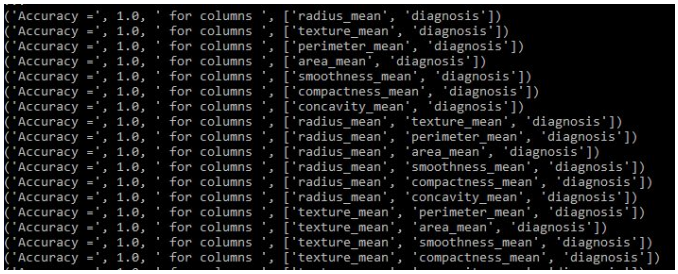

Figure 7: Accuracy for some of the combinations of newly calculated statistical features.

VII. Future Work

In the consideration of future research and implementations, we can enhance the detection by sub-categorizing the malignancy of cells and detect the stage at which cancer resides. It can help therefore to bring out customized treatment solutions for cancer-prone patients and recommendations for therapy based treatments and medications.

Furthermore, the work can be extended to the appropriateness of machine learning techniques such as an artificial neural network (ANN) and Deep learning on Hadoop and Spark framework. Furthermore, this can also be implemented on a cloud platform for ease of usage.

Our results show that using any of the newly calculated statistical features or its combination is enough to accurately distinguish between the tumors namely benign and malignant. Hence it becomes imperative to have additional data points to concretely conclude the astounding accuracy of our linear regression model.

VIII. Conclusion

In conclusion, our methods gave a higher accuracy of 85.4% when compared against Breast Cancer Prediction Using Spark MLlib and ML Packages[10] where the accuracy was only 83%. In addition, this approach reduces the time to train and predict by using distributed computing in contrast to using a single machine for the same[17]. 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.

Also, we found out that among all the tested magnifications - 40X, 100X, 200X, 400X, images of 40X give us the best accuracy. So, this could further reduce the training time by 75% if we concentrate only on 40X magnified images. Furthermore, there would also be a significant reduction of manual effort if we focused on collecting and sampling only the images of said magnification rather than all the varying magnifications.

In textual analysis, we ran an exhaustive combination of newly calculated statistical features against a linear regression model. We found that any of the above-mentioned features or its combination was enough for the model to accurately distinguish between benign and malignant tumors.

Acknowledgment

We would like to sincerely thank the people of P&D Laboratory, BreCAD And the Research Laboratory of the University of Wisconsin for providing us with the useful datasets and information required by us to proceed with the analytic successfully. We would sincerely like to thank NYU HPC for the continuous and rigorous support for the installation of libraries and guiding us through technical roadblocks. Additionally, we would like to thank our Professor, Dr. Suzanne McIntosh for constantly mentoring us and guiding us to derive this analytic successfully. We would also like to extend our gratitude to our domain expert, Ms. Nikita Meghani (MS in Biology, New York University) for providing her expertise in helping us build our machine learning model and validating the entire process from a clinical perspective. In the end, we would like to acknowledge our sincere gratitude for our friends and family for their cooperation and support along with New York University for giving us this opportunity.

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] T. White. Hadoop: The Definitive Guide. O'Reilly Media Inc., Sebastopol, CA, May 2012.

[7] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. In proceedings of 6th Symposium on Operating Systems Design and Implementation, 2004.

[8] S. Ghemawat, H. Gobioff, S. T. Leung. The Google File System. In Proceedings of the nineteenth ACM Symposium on Operating Systems Principles – SOSP '03, 2003.

[9] S.Sujithra, Dr.L.M.Nithya, Dr.J.Shanthini: Predicting Breast Cancer using Apache Spark Machine Learning Logistic Regression. In Proceedings of the International Journal for Scientific Research & Development| Vol. 4, Issue 11, 2017 ISSN (online): 2321-0613 organized by SNS College of Technology, Coimbatore, India

[10] Duy Hung, Tran Duc Hanh, Vu Thu Diep, Breast Cancer Prediction Using Spark MLlib and ML Packages. In ICBRA '18 Proceedings of the 2018 5th International Conference on Bioinformatics Research and Applications

[11] B.M.Gayathri, C.P.Sumathi, T.Santhanam, BREAST CANCER DIAGNOSIS USING MACHINE LEARNING ALGORITHMS –A SURVEY, International Journal of Distributed and Parallel Systems (IJDPS) Vol.4, No.3, May 2013

[12] 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

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

[14]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

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

[16]SARA ALGHUNAIM, HEYAM H. AL-BAITY, On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context, IEEE Digital Object Identifier 10.1109/ACCESS.2019.2927080

[17]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