Breast Cancer Detection Using Extreme Learning Analysis Using Big Data

Sangeetha Vimalraj[1], Megala B[2], Nithyashree R[3]
Assistant professor[1], Students[2, 3]
[1, 2, 3] Dept of Computer Science and Engineering
[1, 2, 3] R.M.K. College of Engineering and Technology

Abstract- *Breast cancer is the second cause of dead among women. Early detection followed by appropriate cancer treatment can reduce the deadly risk. It is a hereditary disease and does not result from a single cause. The diagnosis of cancer starts with a biopsy. A computer-aided diagnosis (CAD) system based on mammograms enables early breast cancer detection, diagnosis, and treatment. However, the accuracy of existing CAD systems remains unsatisfactory. Various methods are used to detect and recognize cancer cells, from microscopic images and mammography to ultrasonography and magnetic resonance images (MRI). In the present study, Extreme Learning Machine (ELM) classification was performed for 9 features based on image segmentation in the Breast Cancer Wisconsin (Diagnostic) data set in the UC Irvine Machine Learning Repository database. Big Data technology is used to analyze these datasets in a database for accurate exploration and detection over benign and malignant breast masses. Extensive experiments demonstrate the accuracy and efficiency of our proposed mass detection and breast cancer classification method.*

I. INTRODUCTION

Cancer begins when healthy cells in the breast change and grow out of control, forming a mass or sheet of cells called a tumor. A tumor can be cancerous or benign. A cancerous tumor is malignant, meaning it can grow and spread to other parts of the body. A benign tumor means the tumor can grow but will not spread. Breast cancer spreads when the cancer grows into other parts of the body or when breast cancer cells move to other parts of the body through the blood vessels and/or lymph vessels. This is called a metastasis.

This guide covers early-stage and locally advanced breast cancer, which includes stages I, II, and III. The stage of breast cancer describes where the cancer is located, how much the cancer has grown, and if or where it has spread.Although breast cancer most commonly spreads to nearby lymph nodes, it can also spread further through the body to areas such as the bones, lungs, liver, and brain. This is called metastatic or stage IV breast cancer.

If breast cancer comes back after initial treatment, it can recur locally, meaning in the breast and/or regional lymph nodes. The regional lymph nodes are those nearby the breast, such as the lymph nodes under the arm. It can also recur elsewhere in the body, called a distant recurrence or metastatic recurrence.

Types of breast cancer

Breast cancer can be invasive or noninvasive. Invasive breast cancer is cancer that spreads into surrounding tissues. Noninvasive breast cancer does not go beyond the milk ducts or lobules in the breast. Most breast cancers start in the ducts or lobes and are called ductal carcinoma or lobular carcinoma:

Ductal carcinoma. These cancers start in the cells lining the milk ducts and make up the majority of breast cancers.Ductal carcinoma in situ (DCIS). This is cancer that is located only in the duct. Invasive or infiltrating ductal carcinoma. This is cancer that has spread outside of the duct. Invasive lobular carcinoma. This is cancer that starts in the lobules.

Less common types of breast cancer include:

Medullary
Mucinous
Tubular
Metaplastic
Papillary

Inflammatory breast cancer is a faster-growing type of cancer that accounts for about 1% to 5% of all breast cancers.

Paget's disease is a type of cancer that begins in the ducts of the nipple. Although it is usually in situ, it can also be an invasive cancer.

Breast cancer subtypes

There are 3 main subtypes of breast cancer that are determined by doing specific tests on a sample of the tumor. These tests will help your doctor learn more about your cancer and recommend the most effective treatment plan.

Testing the tumor sample can find out if the cancer is:

Hormone receptor positive. Breast cancers expressing estrogen receptors (ER) and/or progesterone receptors (PR) are called "hormone receptor positive." These receptors are proteins found in cells. Tumors that have estrogen receptors are called "ER positive." Tumors that have progesterone receptors are called "PR positive." Only 1 of these receptors needs to be positive for a cancer to be called hormone receptor positive. This type of cancer may depend on the hormones estrogen and/or progesterone to grow. Hormone receptor-positive cancers can occur at any age, but are more common in women who have gone through menopause. About 60% to 75% of breast cancers have estrogen and/or progesterone receptors. Cancers without these receptors are called "hormone receptor negative."

HER2 positive. About 10% to 20% of breast cancers depend on the gene called human epidermal growth factor receptor 2 (HER2) to grow. These cancers are called "HER2 positive" and have many copies of the HER2 gene or high levels of the HER2 protein. These proteins are also called "receptors." The HER2 gene makes the HER2 protein, which is found on the cancer cells and is important for tumor cell growth. HER2-positive breast cancers grow more quickly. They can also be either hormone receptor positive or hormone receptor negative. Cancers that have no or low levels of the HER2 protein and/or few copies of the HER2 gene are called "HER2 negative."

Triple negative. If a tumor does not express ER, PR, or HER2, the tumor is called "triple negative." Triple-negative breast cancer makes up about 15% to 20% of invasive breast cancers. Triple-negative breast cancer seems to be more common among younger women, particularly younger black and Hispanic women. Triple-negative cancer is also more common in women with a mutation in the BRCA1 or BRCA2 genes. Experts recommend that all people with triple-negative breast cancer younger than 60 be tested for BRCA gene mutations.

II. EXISTING SYSTEM

Existing concept deals with providing backend by using MySQL which contains lot of drawbacks i.e. data limitation is that processing time is high when the data is huge and once data is lost we cannot recover so thus we proposing concept by using Hadoop framework.

III. MATERIALS AND METHOD

The methodology of the Cochrane Handbook for Systematic Reviews [17] was followed through a search of several electronic databases including MEDLINE, Cochrane Database of Systematic Reviews, Cumulative Index to Nursing and Allied Health (CINAHL) and SCOPUS. Searches were restricted to research published in the English language peer-reviewed journals, as well as grey literature, till December 2014. From these publications, the bibliographic lists were also hand-searched for additional papers.

Index terms (MeSH terms) used for the search were, ((breast neoplasms OR breast cancer OR breast health) AND (awareness OR knowledge OR attitude OR education* OR programme) AND (women OR female OR health worker* OR health professional*) AND (risk factor* OR risk assessment) AND India). We also searched for qualitative studies on breast cancer using the above mentioned search terms. No qualitative study was found on breast cancer awareness among women in India. The initial search yielded 120 studies on the basis of terms in the titles and abstracts (where available) identified from the search strategy. Studies that focused on awareness of screening, or treatment modalities alone for breast cancer were excluded as we focused exclusively on the literacy levels of risk factors and causes of breast cancer. After applying the inclusion criteria, full-text articles were retrieved for 20 studies, of which 13 fulfilled the eligibility.

We considered risk factors (Table 1) summarised in a systematic review conducted by an expert panel committee of the International Agency for Research on Cancer (IARC), the World Cancer Research Fund (WCRF) and the American Institute of Cancer Research (AICR). They classified breast cancer risk factors on the basis of the strength of existing evidence such as sufficient/convincing evidence; insufficient/weak evidence and no conclusive evidence [15,16].

IV. SYSTEM TECHNIQUES

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence

of the name MapReduce implies, the reduce task is always performed after the map job.

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.

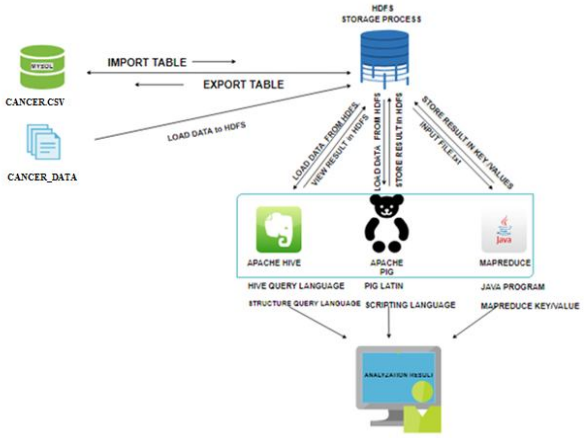
The Algorithm

Generally MapReduce paradigm is based on sending the computer to where the data resides!MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.

Map stage: The map or mapper's job is to process the input data. Generally the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.

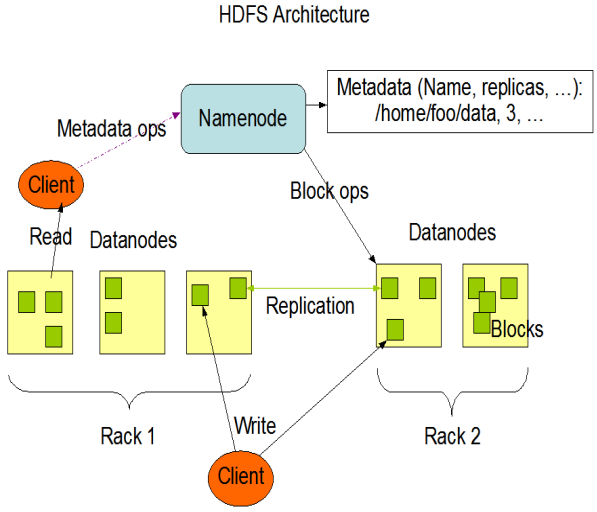
Reduce stage: This stage is the combination of the **Shuffle** stage and the **Reduce** stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.

V. SYSTEM ARCHITECTURE



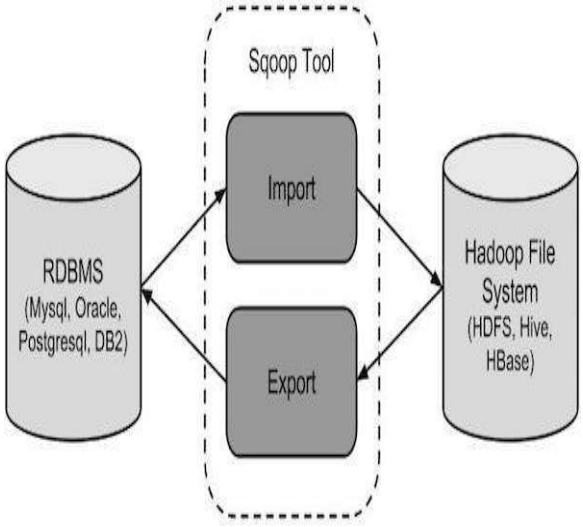
HDFS Architecture

Given below is the architecture of a Hadoop File System. HDFS follows the master-slave architecture and it has the following elements.



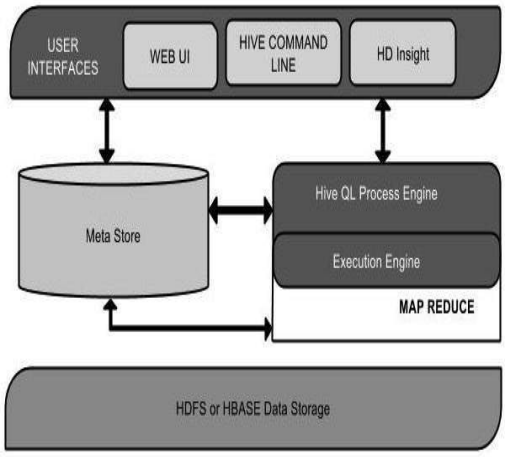
Sqoop Working

The following image describes



Architecture of Hive

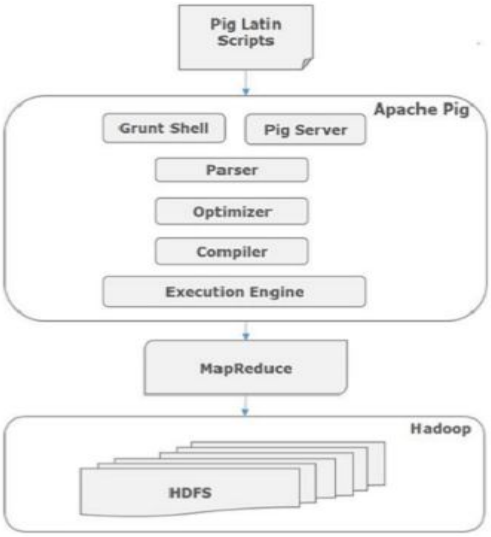
The following component diagram depicts the architecture:



Pig architecture

The language used to analyze data in Hadoop using Pig is known as **Pig Latin**. It is a high-level data processing language which provides a rich set of data types and operators to perform various operations on the data.

To perform a particular task Programmers using Pig, programmers need to write a Pig script using the Pig Latin language, and execute them using any of the execution mechanisms Grunt Shell, UDFs, and Embedded. After execution, these scripts will go through a series of transformations applied by the Pig Framework, to produce the desired output.Internally, Apache Pig converts these scripts into a series of MapReduce jobs, and thus, it makes the programmer's job easy. The architecture of Apache Pig is shown below.



VI. CONCLUSION

In this paper, we presented a study on Breast Cancer data and prediction regarding research paper using Extreme learning method. To analyze the Breast Cancer data in Hadoop ecosystem and to improve the accurate analysis of breast cancer. Hadoop ecosystem is using hive, pig, map reduce tools for processing so that output will take less time to process and result will be very fast. Hence in this project, Breast Cancer data which is traditionally going to store in RDBMS going to less performance hence by using Hadoop tool it will be faster and efficiently processing the data.

REFERENCES

[1] S. V. Liu, L. Melstrom, K. Yao, C. A. Russell, and S.F. Sener, "Neoadjuvant therapy for breast cancer," JSurg Oncol, vol. 101, no. 4, pp. 283-91, Mar 2010.
[2] J. P. O'Connor et al., "Imaging biomarker roadmap forcancer studies," Nat Rev Clin Oncol, vol. 14, no. 3, pp.169-186, 03 2017.
[3] Y. C. Chang et al., "Delineation of Tumor Habitatsbased on Dynamic Contrast Enhanced MRI," Sci Rep,vol. 7, no. 1, p. 9746, Aug 2017.
[4] D. D. Lee and H. S. Seung, "Learning the parts ofobjects by non-negative matrix factorization," Nature,vol. 401, no. 6755, pp. 788-91, Oct 1999.
[5] P. S. Tofts, "Modeling tracer kinetics in dynamic GdDTPA MR imaging," J Magn Reson Imaging, vol. 7,no. 1, pp. 91-101, 1997 Jan-Feb 1997.