ABSTRACT

The recent advancements in computer innovations and storage technologies have taken an incredible amount of data and information from various sources, for instance, social networks, online databases, and health care data frameworks. Nowadays, various nations around the world are adapting the strategy for implementing health care services to the patients by utilizing the benefits of advancement in computer innovations and communications through e-health.

Breast cancer is a malignant tumor that develops when cells in the breast tissue divide and grow without the normal control on cell death and cell division. Even though a definite cause of breast cancer still remains a mystery but to provide some insight to its possibility of occurrence in women some common risk factors are extracted. These factors include attributes such as age, genetic risk and family history. Breast cancer can recur at any time or not at all, but most recurrences happen initially in the first 5 years after the primary treatment. The recurrence of breast cancer is indicated by the prognostic variables such as age, the involvement of lymph node, large tumors, low levels of estrogen and progesterone receptors, a high synthesis-phase fraction, higher histologic grade and cancer cells with a higher nuclear grade.

The detection and treatment in the initial stages of breast cancer in women are key to the success of patient survival. The challenges of characterizing early breast cancer, a clinical oncologist need to concentrate on detection and its treatment. In the last 2-3 decades, there are remarkable technological innovations with respect to initial stage diagnosis and also detection of recurrence phenomena of breast cancer. There is a need for new research avenues such as knowledge discovery in databases (KDD) which incorporates data mining techniques as a popular research tool. This tool can be used to identify and exploit knowledge & relationships among a large number of expected variables to predict recurrence of breast cancer using the historical cases stored within datasets.

In this thesis, SEER (Surveillance, Epidemiology, and End Results) Public-Use breast cancer dataset 1975 – 2013 program of the National Cancer Institute (NCI) is used to for developing prediction models for the survivability of breast cancer recurrence. For all of the present work, the first step was to select a set of prognostic 17 variables from
more than 120+ variables stored in the SEER dataset. The analysis of the correlation between these variables and final variables reduction are determined by Principal Component Analysis (PCA) to predict breast cancer recurrence for survivability prediction. Subsequently, missing values for continuous variables were substituted utilizing Expectation Maximization (EM) and Multiple Imputation (MI) techniques. After handling missing values, a few variables were dichotomized and were changed to binary or discrete variables.

After data preprocessing machine learning technique like association rule mining technique for successfully detection of breast cancer recurrence in large SEER data set were investigated. It is expected to recognize strong rules from dataset utilizing some measures of interestingness. In order to choose interesting rules from the set of all possible rules, constraints on different measures of noteworthiness and interest are utilized. The best-known constraints are minimum thresholds on support and confidence. Fundamentally, the present algorithm is normally used for the transactional dataset, but in the present investigation, it is used in the tabular form of data yielding a more accurate prediction of breast cancer recurrence.

Further, the traditional data analytic might not have the capacity to handle the enormous amount of data. Due to the rapid growth of information, solutions need to be considered and provided in order to handle useful information or knowledge from these data sets. Moreover, decision makers should have the capacity to increase significant bits of knowledge from such fluctuated and quickly evolving information. Hence, a machine learning approach, big data analytic approach and performance enhancement in Hadoop map-reduce approaches need to be developed for better prediction accuracy and speed of breast cancer recurrence with a minimal execution time. The present work is focused on the design, analysis and testing of SEER data set for predicting breast cancer recurrence for scalable and parallel Map-reduce.

In chapter 1, a brief introduction about machine learning approach, breast cancer & its recurrence, data mining and Hadoop MapReduce approach are discussed. The compressive reviews of the current literature related to data mining and big data analytics approach towards the prediction of breast cancer recurrence are presented.
In chapter 2, it has been clearly brought out that there is a need to design the machine learning approaches for prediction of breast cancer recurrence. The aim of the present work can be briefly stated as follows:

- Association Rule Mining for Predicting Breast Cancer Recurrence in SEER data set.
- Big Data Analytics to Predict Breast Cancer Recurrence on SEER Dataset using MapReduce Approach.
- Parallel Computing to Predict Breast Cancer Recurrence on SEER Dataset using Map-Reduce Approach.
- Data Positioning System to optimize Performance of Big data Processing in Hadoop Map-Reduce for the analysis of Breast Cancer Recurrence of SEER Dataset.

In chapter 3, SEER data set of Program of the National Cancer Institute (NCI) was pre-processed to evacuate inadmissible cases. After using data cleansing and data preparation strategies, the final dataset was constructed for the development of algorithms by the following methods.

METHOD – I: The feasibility of utilizing an Association rule mining to anticipate regardless whether breast cancer will recur for the patient in the perspective of SEER data set are examined.

METHOD – II: To handle the huge volume of data we examine to develop an algorithm in the high-performance platform to efficiently analyze big SEER (Surveillance, Epidemiology, and End Results) breast cancer data set using MapReduce to find the recurrence of breast cancer.

METHOD – III: To optimize the execution, the need for developing a parallel MapReduce algorithm is designed to improve classification accuracy and speedup are presented.

METHOD – IV: To optimize the execution performance of big data analytics, the need for Hadoop data position algorithm are presented.
In Chapter 4, the experimental results for various methods i.e. METHOD – I to METHOD – IV for the prediction of breast cancer recurrence are discussed.

In Chapter 5, Conclusions and scope for future work are presented.