Chapter – 5: Conclusion and Scope for Future Work

5.1 Conclusion

The main focus of the present work is the design and implementation of a knowledge-based approach for the prediction of breast cancer recurrence on SEER dataset. Data mining can reveal new knowledge from retrospective data. The findings of the breast cancer recurrence could be acquired from the dataset without the requirement of domain experts. Historical data that are stored in the health records provide new possibilities for researchers to use these data sources for a wide range of studies. However, it should be kept in mind that the complexity of these data and the necessity for pre-processing before data mining are important aspects of such analysis. Further, the quality of data is critical to the success of knowledge discovery. Analyzing medical data with data mining along with big data analytics is a challenging task and should be done by following the KDD methodology in order to get satisfactory results.

The results of medical knowledge discovery can be used to support decision-making for a clinical oncologist in the prediction of breast cancer recurrence. One of the challenges for researchers in the field of biomedical informatics is to make this possible. New methodologies for revealing knowledge from retrospective medical data has been adopted in accord with circumstances in medicine. Both the biomedical informatics and medical disciplines will be enriched by developing methods and applications in this field.

The following conclusions have shown the potential in the present work.

- The arbitrary sample of breast cancer data set has been extracted from the huge volume of SEER data for analysis.
- It is observed that dimensionality reduction on the SEER breast cancer dataset of 17 attributes from 120+ attributes was sufficient to predict accuracy for this research work.
- It is observed that handling of missing values on a reduced attributes has been effectively handled through Multiple Imputation and Expectation Maximization algorithm on a SEER dataset.
• METHOD – 1: Association rule mining used in SEER breast cancer dataset found to be the optimized data mining technique for the prediction of breast cancer recurrence on limited samples of SEER dataset.

• METHOD – 1 algorithm results show that error rate is more accurate compared to other data mining techniques in the prediction of breast cancer recurrence.

• METHOD – 2: Big Data Analytics MapReduce approach shows the system behavior on a single node for two concurrent maps and reduce tasks.

• METHOD – 2 shows the task timeline for the breast cancer recurrence system workload corresponding to map, shuffle, merge and sort.

• METHOD – 2 also shows the CPU utilization and CPU iowait for Hadoop MapReduce contrast to Hadoop MapReduce online.

• The new concept developed on the MapReduce framework in METHOD – 2 offers an improvement of prediction of breast cancer recurrence over METHOD – 1.

• The new concept developed on the MapReduce framework in METHOD – 3 shows the parallel structure in the deterministic execution of breast cancer recurrence on a SEER dataset.

• METHOD – 3 algorithm shows the utilization of the same Big Data Analytics MapReduce algorithm in all computing hubs for computation.

• The METHOD – 3 systems on MapReduce framework results show the same or more prediction accuracy and gain significant speedup in a parallel domain over METHOD – 2.

• The new concept developed on data positioning system in METHOD – 4 provides the significant improvement on data locality and positioning system in Hadoop Distributed File System over Hadoop MapReduce framework.

• METHOD – 4 proposed system provides an improvement in execution time over the default data position systems of standard HDFS and HDFS balancer.

• METHOD – 4 shows the modification of the Hadoop MapReduce is prominently noticeable in the write-read benchmark when compared with the TeraSort benchmark.
The present work has achieved an accuracy of 91.4% for the prediction of breast cancer recurrence in large SEER dataset.

It is observed that the improvisation of breast cancer recurrence learning model also depends on the number of attributes and records of SEER dataset.

5.2 Scope for Future Work

- New data preprocessing techniques need to be designed to aim towards some more relevant attributes and handling missing values that help with prediction of breast cancer recurrence accuracy in a SEER dataset.

- Further investigations are required for METHOD – 3 i.e. Parallel computing MapReduce system, to have distinctive algorithms on various computing hubs to build the accuracy.

- Further investigations are required for METHOD – 4 trails, to have distinctive data partitioning algorithms that can enhance the execution performance.

- The present breast cancer recurrence system algorithms may also be used for various other applications such as pharmaceutical, agricultural, food industries, online shopping, and so forth.

- The possible extension of the present breast cancer recurrence system can be adopted for investigating other types of cancer dataset.