Title of Thesis	**: Data Mining and Medical Decision Making**
Name of Scholar	**: Benaki Lairenjam**
Name of Supervisor	**: Prof. S.K. Wasan**
Department	**: Mathematics**
Faculty	**: Natural Sciences**

ABSTRACT

Data mining is a technique for discovering useful information from large databases. It draws ideas from number of disciplines such as statistics, machine learning and database systems. Decreasing cost of electronic storage media has made it economically feasible for healthcare organizations to maintain large medical databases. Analysis of large medical databases can help in obtaining association rules, indicating relationship between procedures performed on patients and the reported diagnosis.

Modern medicine generate huge amount of patient health care data. These data records are stored electronically by medical community in databases. The information contained in medical data records are interesting and useful for patient care.

Data mining methods are algorithms that are used for building models and for finding patterns in data. In medical data mining building accurate classifier model for predicting serious human diseases is important. Classifier models may assist a physician to accurately diagnose a disease.

Breast cancer is one of the most leading causes of cancer deaths in women. Breast cancer when detected earlier increases the chance of survival of the patient. To detect breast cancer earlier screening mammography is one of the most used methods.

Lot of work has been done to detect breast cancer from mammographic data using different data mining algorithms. Further improvements in both sensitivity and specificity would lead to tremendous benefits both in terms of survival rate of breast cancer patients, and in terms of reduced workload of radiologist. In this thesis we create classifier models using different data mining techniques such as

associative classification algorithm, Naïve Bayes, backpropagation neural network and Bayes theorem on breast cancer mammographic data.

We consider two mammography datasets for finding benign and malignant cases of breast cancer using data mining models. The first dataset is collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006 and name of the dataset is mammographic mass data. The second dataset is collected at University of Wisconsin Hospitals, Madison by Dr. William H. Wolberg and name of the dataset is Wisconsin Breast Cancer Database. The thesis consists of eight chapters. Chapter 1 is the introduction. Chapter 2 is on preliminaries of data mining and the related work. In chapter 3 we present the cancer dataset and the details of breast cancer mammographic data used in our work. We analyze mammographic mass data using classifier algorithms Classification Based on Association rule (CBA), Classification based on Multiple Association Rule (CMAR), Backpropagation Neural Network (BPNN) and Radial Basis Function (RBF) networks to differentiate malignant from benign findings of mammographic mass data. In chapter 4, we propose a new classifier model NNwCMAR based on CMAR and neural network. CMAR is used for creating structure of the network model. It uses bakpropagation algorithm for learning the network and sigmoid activation function. It is tested on mammographic mass data from UCI repository. In chapter 5, we propose an Associative Classifier Algorithm (ACA) and used with NB classifier to develop an ANB (Associative Naïve Bayes) model for classifying breast cancer mammographic data. In our approach we assume that the attributes are not conditionally independent. It is tested on mammographic mass data.

In chapter 6 an improvement is made in the number of average iteration of convergence of NNwCMAR model by introducing initial weights in the hidden layer. Initial weights in the hidden layer are calculated using Bayes theorem. The modified model is called NNC (Neural Network Classifier). The network model is tested on Wisconsin Breast Cancer Database.

In chapter 7, we improve the classification performance and convergence rate of NNC model by introducing additional connection from the input layer to the output layer. The improved model is called HNN (Hybrid Neural Network). The weights connecting the input layer and output layer are calculated using Bayes theorem. It is tested on mammographic mass data.

Finally, we conclude with chapter 8 presenting a brief summary of the work done in the thesis.