Am J Cancer Res 2017;7(3):610-627
www.ajcr.us /ISSN:2156-6976/ajcr0009215

Original Article

Data mining and medical world: breast cancers' diagnosis, treatment, prognosis and challenges

Rozita Jamili Oskouei[1], Nasroallah Moradi Kor[2,3], Saeid Abbasi Maleki[4]

[1]Department of Computer Science and Information Technology, Mahdishahr Branch, Islamic Azad University, Mahdishahr, Iran; [2]Research Center of Physiology, Faculty of Medicine, Semnan University of Medical Sciences, Semnan, Iran; [3]Student Research Committee, Faculty of Medicine, Semnan University of Medical Sciences, Semnan, Iran; [4]Department of Pharmacology & Toxicology, Urmia Branch, Islamic Azad University, Urmia, Iran

Received April 16, 2015; Accepted June 25, 2016; Epub March 1, 2017; Published March 15, 2017

Abstract: The amount of data in electronic and real world is constantly on the rise. Therefore, extracting useful knowledge from the total available data is very important and time consuming task. Data mining has various techniques for extracting valuable information or knowledge from data. These techniques are applicable for all data that are collected inall fields of science. Several research investigations are published about applications of data mining in various fields of sciences such as defense, banking, insurances, education, telecommunications, medicine and etc. This investigation attempts to provide a comprehensive survey about applications of data mining techniques in breast cancer diagnosis, treatment & prognosis till now. Further, the main challenges in these area is presented in this investigation. Since several research studies currently are going on in this issues, therefore, it is necessary to have a complete survey about all researches which are completed up to now, along with the results of those studies and important challenges which are currently exist in this area for helping young researchers and presenting to them the main problems that are still exist in this area.

Keywords: Data mining, medical data, cancer diagnosis, cancer treatment, cancer prognosis, risk factors

Introduction

Nowadays in all fields of sciences including genetics, education, earth science, agriculture and medicine the amount of data is increasing dramatically. Analyzing these huge amount of data to extract the novel and usable information or knowledge is very complicated and time consuming task. Data mining techniques are useful for this matter.

Generally, in the medical world, there are two phases for making the decisions. These two phases are [1]:

• *Differential Diagnosis (DD):* in this phase, all information of patients including their medical history, symptoms of disease, results of various testing such as blood testing and etc. are perceivedby doctors as the input data. These data are processed by doctors based on their medical knowledge for disease diagnosis. Sometimes several diseases havesome similar symptoms, therefore, medical doctors must be assign arbitrary weights to each one of inputsand make patterns, match these patterns with the patterns of various diseases and finally select the closest match and diagnosis the exact disease.

• *Final or Provisional Diagnosis (FD):* in this phase, the preliminary recommendations and treatments would be start according to the identified disease. In this step, a physician with medical knowledge and his/her logic, continues checkups and records the results of continually perceives or tests, and decides the final treatments and prognosis.

Data mining has various techniques (such as: Classification, Clustering, Regression, Association Rules and etc.,) and algorithms (such as: Decision Trees, Genetic Algorithm, Nearest Neighbor method etc.,) for analyzing the huge amount of raw or multi-dimensional data. In the other words, data mining has capabilities for intelligent data analysis to extract hidden knowledge from large databases of medical or clinical data that are collected from medical centers or hospitals. These knowledge provide

useful information to improve decision support, prevention, diagnosis and treatment in medical world. Further, data mining has ability to identify association rules or establish relationships between various features such as: patient's personal data, disease symptoms and etc. [2].

This investigation attempts to represent the results of several researchworks which are published related to data mining applications in prediction, diagnosis or treatment of breast cancers.

This paper is organized in four sections. Section 2nd includes some basic concepts related to this paper. Section 3rd presents data mining applications or usages in early diagnosis, treatment and prognosis of various cancers. Section 4th concludes this paper and presents our future works.

Basic concepts

This section presents the concepts of data mining and its different techniques.

Data mining

Data mining and knowledge discovery in databases (KDD) are extracting novel, understandable and useful information, knowledge or patterns from huge amount of available data [3]. In the other words, data mining has capabilities for analyzingthe large datasets, finding unexpected or hidden relationships between various attributes and summarizing the extractedinformation more understandable and useful to data users or owners. In the traditional model for transforming data to knowledge, some manual analysis and interpretation are executed. For example, in medical centers, generally doctors or specialists manually analyze current trends, disease and health-care data, then make a report and use this report for decision making or planning for medical diagnosis, treatments and etc. The problem of this type of data analysis is that, this form of manual data analysis is slow, expensive, time consuming, and highly subjective. However, KDD has various data processing steps including [4]:

• *Selection*: selecting target or relevant data based on the goal or data mining task.

• *Pre-processing*: removing missing, incorrect, noisy, and inconsistent or no quality data.

• *Transformation*: includes smoothing, aggregation, generalization or normalization and attribute/feature selection.

• *Data mining:* applying data mining methods or techniques for extracting interesting patterns.

• *Interpretation/Evaluation*: includes statistical validation, qualitative review and etc.

Data mining has two main tasks:

• *Predictive tasks*: with applying various techniques or algorithms, it can make decisions or predict the unknown or future values of other variables. These techniques includes classification, association rule and etc.

• *Descriptive tasks*: describe the data or find human understandable patterns and present the results in tables, diagrams and etc., which can be understand easily by data owners or data users.

The remaining of this section, provides a short review of common data mining techniques.

Classification: Classification is called as supervise learning. It take some of data (named as training set) which has collection of records and each record contain set of attributes and define one attribute named as class. The main goal of classification is producing a model with capability of predicting the value of class attribute in previously unseen records as accurately as possible. A test set is used for predicting the accuracy of the created model [5]. Some applications of classification in medical diagnosis are: classifying tumor cells, analyzing the effectiveness of treatmentand etc.

Several classification algorithms and techniques are proposed such as [6-10]: Decision TreeInduction (ID3 & C4.5, Hunt's Algorithm and etc.,), Rule-Based Methods, Memory-Based Methods (such as: k-Nearest-Neighbor), Genetic Programming [9, 10], Naïve Bayes [11] and Bayesian Classification [12], Artificial Neural Networks [13], Support Vector Machines (SVMs) [14], Ensemble Methods [15] and etc.

Association Rules: Association Rule is one the most important techniques of data mining. It attempts to extract frequent patterns and interesting relationships between different sets of items [16], and etc.

Association rule mining has various applications [17] in biology (exp. for detecting relationships between gens and environment), healthcare settings and Medical Diagnosis [18-20], Critical Medical Applications [21], Maximal-Profit Item Selection (MPIS) [22], Privacy preservation [23], astrophysics [24], crime prevention [25], counterterrorism [26], Business [27, 28], GIS [29] and etc.

There are several association rule mining techniques that are proposed such as: Apriori algorithm and its extension which is called as AprioriTID [30, 31], DIC [32], STEM [33], ICAP (Incremental Constrained APriori [34], CARMA (Continuous Association Rule Mining Algorithm) [35], RARM (Rapid Association Rule Mining) [36], FP-Tree algorithm [37], Goethals FP-Growth , Broglet's FP-Growth, Eclat and SaM algorithm [38], COFI-Tree and CT-PRO [39], Recursive Elimination [40] and etc.

Regression: Regression same as classification attempts to predict the value of an attribute (variable) based on the value/values of other attributes/variables. The main difference between classification and regression is related to the type of target attribute (variable) that must be predict based on the value of other attributes/variables. The target variable in classification is categorical in nature. Whereas in regression, the target variable is numeric or continuous. Further, in classification, classes are created whereas in regression there is no classes, and all data is divided in various split points and for each split point the amount of "error" is equal to square of differences between amount of actual value and predicted value. The amount of split points error across different variables are compared and minimum split point error is selected as the split point/root node. This process recursively continued [41]. In the other words, the main objectives of regression are:

• Dividing the set of data into two continuous variables then describe the associations or relationships between them.

• Find the value of attributes/variables.

• Predict the value of one attribute/variable based on the value of other attribute/variable.

• Control the accuracy of prediction.

Regression has several applications including: estimation of agricultural data [42], Geography [43], marketing [44], business [45], Financial Forecasting [46], medical diagnosis and cancer diagnosis or prognosis [47], Predicting Laptop Retail Price [48] and etc.

Clustering: Clustering is unsupervised learning and divides the data into groups (call as clusters) based on their similar attributes [49]. All objects into one cluster are similar with each other and dissimilar with objects in other clusters. Clustering is widely used in: science and statistics [50, 51], pattern recognition [52, 53], image processing and segmentation [54], Web applications [55-58], DNA analysis in biology [59], GIS [60, 61] and etc.

Different clustering algorithms are available. Some of these algorithms are: Hierarchical Methods [62, 63] (Divisive Algorithms & Agglomerative Algorithms), Partitioning Methods [64] (Relocation Algorithms, K-medoids Methods, K-means Methods, Probabilistic Clustering, Density-Based Algorithms), Grid-Based Methods [65], Constraint-Based Clustering [66], Clustering Algorithms Used in Machine Learning [67], Scalable Clustering Algorithms [68, 69], Algorithms For High Dimensional Data [70] (Subspace Clustering, Co-Clustering Techniques, Projection Techniques) and Methods Based on Co-Occurrence of Categorical Data [71].

Cancers

Cancer is one type of disease. It is happening when cells growth in part of human body becomes out-of-control. In the other words, whenevercells in part of the body divide uncontrollably and damage the other cells, cancer is occurred. Nowadays, more than 100 types of cancers based on the part of body where it's appeared, or cells that are affected, have been classified. Currently, cancer becameone of the main causes of death in allover of the world. Several factors affect the creation or spreading cancers including: gender, age, genetics, marital status, quality of life, living location and etc.

In some cases, cancer makes a masses of tissue in part of the body which is called as tumors. These tumors can grow and effect various organs of body such as nervous, digestive or circulatory systems. However, when a tumor

spreads to other parts of the body, invading or destroying other tissues, it is called as metastasized and this process is called metastasis. When tumor becomes in this stage, it is very difficult to treat that. Therefore, one of the important issues in treatment process of tumors and cancers is related to the time of diagnosis of that cancer or tumor. The early diagnosis of cancers increases the chance of their treatment. For this reason, several researchers attempts to create intelligent systems to assist the doctors for early diagnosis of cancers. For example, P. Ramachandran et al. [72] analyzed the effect of four attributes including age, sex, marital status and educational qualification on cancer creation. Further, these researchers made useful patterns for cancer diagnosis.

Two types of tumors are identified:

• *Benign*: this type of tumors is not dangerous for human body and rarely causes for human death. In this type, tumor grows in one part (spot) of body and has limited growth.

• *Malignant*: this type of tumors is more dangerous andhas two types of effects on human body:

○ A cancerous cell with uncontrolled growth spread with invasion lymph system destroys healthy tissues. In the other words, it metastasis to other tissues and make problem in their general actions or duties.

○ A cancerous cell continually growth and with angiogenesis process makes new blood vessels to feed itself. Therefore, it uses body's blood and can cause Anemia.

Currently, one of the main challenges in the area of cancer treatment is, identifying the most common symptoms which can help for earlier diagnosis of cancers. Several research studies have been conducted to extract the patterns of cancers and create intelligent or fast method for diagnosis of tumors in the early stages and suggest the best treatments. Based on the results of several studies, some of the symptoms that are shown in various cancers have been listed below [73].

• Losing or gaining weight abnormally (in a short time)

• Blood in stool

• Persistent cough

• Problem or Difficulty in swallowing

• Abnormal Hoarseness

• Persistent joint pain or unexplained muscles

• Persistent night sweats or unexpected fever

• Fatigue

• Any changes that are happening in skin color.

• Lump or area of thickness that may be felt in each part of the body or under the skin

• Any changes that may be happen in bladder habits or bowel

Data mining & breast cancers

Breast Cancer is one of the most common type of the cancers in women which is affecting approximately 12.5% of all women in all around of the world. Moreover, developing countries have growing breast cancer epidemic with an increasing number of younger women which are susceptible to the cancer.

There are two types of breast cancers which must be a doctor distinguish in the time of diagnosis and this identification is very important for starting treatment process and Prognosis the time period of recur patients. These two types are:

• *Benign breast lump or non-cancerous*: the size of tumour and its texture is understandable during the roughly examination.

• *Malignant breast lump or cancerous*: clinical diagnosis requires for predicting this type of cancer. Two types of Malignant cancers are exist:

○ *Non-invasive*: The malignant cells have not spread in other tissues.

○ *Invasive*: The malignant cells or cancer have spread into the surrounding tissue.

Since early detection of this cancer can be help for effective treatment, therefore, several efforts are done to achieve early detection of this disease. The main reason of this disease is not known to scientists till now, but some risk factors are recognized that increase the likeli-

hood of breast cancer creation in female, which are broadly classified into non-modifiable and modifiable factors. Some of the non-modifiable factors are [74]:

• Gender (Being a woman)

• Ageing

• A family background of having breast or ovarian cancer

• Starting menopause after age 55

• Having high bone or breast density

• A history of ovarian or breast cancer

Further, some of the modifiable factors are:

• Women with no children

• Number of abortions

• Age at first child birth (if age > 35 risk is high)

• Duration of breast feeding to child

• Having Frequently X-Rays

• Obesity

• Using Estrogen or Estrogen-Plus Progestin hormone

• Alcohol and alcoholic beverages

• Diet and food habits

There are six stages for breast cancer. Among these stages, stage 0 is the most primary stage of this disease and stage IV is the most dangerous and advanced stage [75]. Currently for the aim of diagnosis of breast cancers and identifying the stage of cancer, some common methods are used such as: Mammography, Biopsy, Magnetic Resonance Imaging (MRI) and Positron Emission Tomography (PET) [76]. Further, some data mining techniques are useful for cancers' pattern recognition purpose. In the continue of this sub-section we attempt to cover some of the efforts which are done related to data mining applications in breast cancer diagnosis, treatment or prognosis respectively.

Breast Cancer Diagnosis is recognizing benign from malignant breast cancers and lumps and Breast Cancer Prognosis predicts the high risk people in aspect of breast cancer or predicting recurrence of cancers after treatment or removing their cancers.

The interesting point which we saw in almost all of these research papers was related to the databases which are used. Almost all of them used the following databases in their researches:

• Wisconsin breast cancer dataset (WBCD) has 11 attributes [74, 77, 79, 80, 92, 97 115, 121 and 125]

• Wisconsin Diagnosis Breast Cancer (WDBC) has 32 attributes [92]

• Wisconsin Prognosis Breast Cancer (WPBC) data set from the UCI machine learning repository has 34 attributes [82, 87, 88, 92, 93, 97 and 98]

• SEER dataset [81, 82, 107, 110, 111, 113, 114, 117, 122 and 124]

• The Digital Database for Screening Mammography (DDSM) [82, 107 and 109]

• Breast Cancer Surveillance Consortium (BCSC) database which contains 2.4 million screenings mammograms and associated self-administered questionnaires [113]

• PCE data and PCE colorectal dataset [85]

• National Cancer Institute's Surveillance, Epidemiology, and End Results breast carcinoma data set [85]

This section, covers the results of several research works that have been done related to data mining applications for early diagnosis, treatment and prognosis of breast cancers.

Data mining techniques for breast cancer diagnosis

Currently most of the physicians for identifying type of cancers (benign breast tumours from malignant) prefer to make surgical biopsy. But most of them believed that, biopsy is very critical task and must be prevented as much as possible. Therefore, proposing an intelligent system which can help tophysiciansto identify the type of cancer and avoid unnecessary surgical biopsywould be helpful for both patients and physicians.

In this sub-section we attempt to cover most of the research works that have been done related to diagnosis of breast cancers with applying various techniques of data mining along with their results. For our presentation we will classify all research works based on the interesting points and discuss them.

Based on the main goal of papers, we classified them in the following categories.

• Some of the research works, are compared the accuracy of applying various classification techniques for diagnosis of breast cancers, such as:

○ Vikas Chaurasia et al. [74] applied Simple Logistic, RBF and RepTree for diagnosis of breast cancer. The accuracy of their classification was 74.5%.

Wei-pin Chang et al. [77] made a comparative study for predicting breast cancers by decision tree, neural network, genetic algorithm and logistic regression. They concerned on 10 variable/attribute for creating breast cancer classification model. These variables were included: Clump thickness, Bland chromatin, Uniformity of cell size, Uniformity of cell shape, Bare nuclei, Normal nucleoli, Marginal adhesion, Mitoses, Single epithelial cell size and class variable with two value (benign/malignant). Their experimental results revealed that, decision tree has lowest prediction accuracy and logistic regression model had higher accuracy rate among these applied techniques for predicting breast cancers. Further, genetic algorithm had highest accuracy in the classification of breast cancers and created acceptable classification rules.

○ Based on the results of research which is done by Chaurasi and et al. [74], Simple logistic classifier among the other machine learning algorithms with having accuracy of 74.4% and total time taken for building model in 0.62 seconds, is the best algorithm for diagnosis of breast cancers. Further, in this study, researchers used three tests (including Gain Ratio test, Info Gain test and Chi-square test) for recognizing the variables which are important in diagnosis or treatment of breast cancers such as: Tumour size, patients' Age, Degree of malignancy, Menopause, Breast-quad and etc.

○ Shweta Kharya [78] made a complete survey about applying different classification techniques for diagnosis of breast cancers. She studied different the performance or accuracy rate of various techniques (including Decision Tree, Bayesian Network, Logistic Regression, Support Vector Machines, Naïve Bayes Classifier, Association Rule Mining and ANN) for diagnosis of cancers by analysing factors (genes and etc.) or Digital Mammography images classification, Her study was based on the data which are collected from WBCD and SEER datasets. She claimed that, Decision tree with 93.62% accuracy rate of predicting cancers is the best predictor among the concerned techniques and the Bayesian network is the popular technique which is used in medical world for Brest cancer prognosis and diagnosis.

○ Senturk et al. [79] applied seven algorithms including KNN, Decision Tree, Naïve bayes, logistic regression, multi-layer perceptron, discriminant analysis and Support Vector Machine for diagnosis of breast cancers. Their experimental results declared that, accuracy of classification made by Support Vector Machine was high than others.

○ Ghassem Pour and colleagues [80] made a comparison between a Neural Network classification techniques with Model-based data mining techniques for accuracy of detecting breast cancers. Their experimental results showed that, adding an ensemble oriented approach can improve the results of both techniques. Furthermore, Neural Network approach with ensemble oriented approach had highest accuracy rate of classification in compare with model based data mining techniques.

○ Rajesh et al. [81] for classifying patients into either "Carcinoma in situ" (beginning or pre-cancer stage) or "Malignant potential" group, used C4.5 algorithm. They showed that, C4.5 had accuracy ~93% for diagnosis of breast cancers.

○ Hota [82] several intelligent techniques such as, ANN (Artificial Neural Network, Unsupervised ANN, Statistical and decision tree based techniques used for classifying data related to breast cancer. In this research work, different models are combined and made ensemble model. Experimental results in this study

revealed that, the accuracy rate of ensemble model is better than single individual model.

○ Gupta and et al. [83] made a survey with study the several techniques which are used by many researchers for diagnosis and prognosis of breast cancers. Finally they mentioned that, in both cases, for selecting the best technique or algorithm with high degree of accuracy, can be decided after creating several types of models, trying different techniques or algorithms.

○ Burke HB et al. [84] compared the prediction accuracy of the TNM staging system[1] with that of artificial neural network statistical models. They studied the accuracy of breast cancer prediction based on 5 years and 10 years surveillance data and revealed that, in both case, Artificial Neural Network's prediction was accurate than TNM staging system.

○ Ronak Sumbaly et al. [86] used general (types, risk factors, symptoms and treatment) of breast cancers and applied various data mining techniques for diagnosis of breast cancers in the early step. Their results showed that, decision tree have capability to diagnosis breast cancers in the first stages.

○ Shrivastava et al. [87] made a review of different classification techniques which have been done for diagnosis of breast cancers. Finally they showed that, Neural Network and decision tree are the most popular techniques which are used by various researchers to create decision rules or predictive models from the breast cancer data.

○ Jahanvi Joshi et al. [88] applied various classification and clustering techniques to create pattern of breast cancer patients. For finding the healthy patients, several classifier rules are used. Further, authors claimed that, they used 47 classification algorithms for recognizing healthy people from sick patients. Their experimental results showed that, the results of approximately 13 techniques within those 47 applied techniques were same (24% sick patients and 76% healthy people). These 13 techniques are: Multilayer Perceptron, LMT classifier, Logistic, Classification via Regression, MultiClass Classifier, GD, SMO, J48, Simple Logistic, AdaBoostM1, Bayes Net and Attribute Selected technique.

○ Padmavati et al. [89] for predicting breast cancers used RBF (Radial Basis Function), MLP (Multilayer Perceptron) and Logistic Regression techniques. Their experimental results showed that, RBF has prediction capability of RBF was better that two other techniques. Further, the time taken for prediction by RBF was lesser than other techniques.

○ Aboul [90] applied rough set data and ID3 decision tree classifier algorithm for creating classification rules. Their experimental results showed that, the accuracy of classification rules created by rough set was better than ID3 algorithm. Further, the number of classification rules made by rough set algorithm is reduced in compare with ID3 algorithm. In the other words, rough set algorithm had compact number of produced rules.

○ Gouda I. Salama et al. [91] compared the accuracy and confusion matrix based on 10-fold cross validation method of different classification techniques including Multi-Layer Perception (MLP), decision tree (J48), Instance Based for K-Nearest neighbour (IBK) and Sequential Minimal Optimization (SMO) for diagnosis of cancers in three different databases of breast cancers (WPBC, WDBC, WBC). Their experimental results showed that, the combination of SMO, MLP, IBK and J48 hast the highest accuracy rate in compare with other techniques (in all of three datasets) for diagnosis of benign breast tumours from malignant.

○ One comparison between different neural network techniques (such as MLP, RBF, SOM (Self-Organizing Map) and PNN (Probabilistic Neural Network) is made by Sarvestan et al. [92] for diagnosis of breast cancers and detecting the type of breast tumours. Their experimental results, showed that, the accuracy of identifying breast cancers by PNN technique was high that other techniques. Finally they made a system with applying statistical neural network techniques and predefined accuracy rate for detecting breast cancers.

Challenges: There are several research works related to comparison of various data mining classification techniques, Association rule mining algorithms and etc. for diagnosis of breast cancers in the first stages. But the main challenge has remained and that is, for detecting breast cancers how many attributes are neces-

sary? Which algorithm is applicable for all of databases? How can improve the accuracy of diagnosis and decrease the number of Biopsy and error in detecting malignant cancers? Is it possible we develop a tool which can be automatically without human Interference diagnosis the breast cancer with analysing automatically results of mammography and etc.?

• Several research works have attempted to propose a method or approach to recognize benign from malignant breast tumours.

○ Hassanien and colleagues [93] studied the applications of rough set theory to analysis the medical data and proposed an approach for creating compact classification rules with applying their proposed simplification algorithm. They claimed that proposed approach had classification accuracy of 98% whereas, accuracy of classification made by decision trees was 85.25. Further, the number of classification rules with applying decision trees and their proposed approach was respectively 428 and 30.

○ A simple data mining approach for finding people with high-risk breast cancer is proposed by Orlando [94]. First they made different association rules by default and then made one questionnaire based on that rules and important defined factors which can be related with cancer disease occurrence, and asked from patients to fill that. In that questionnaire several questions were included such as: people habit for drinking alcohol. They made two option for this question including drink along food or no? After analysing the results of questionnaires, they made one decision tree and present important factors that can be help for recognizing high-risk groups of women. They claimed that, their experimental results have been shown that, decision tree has capabilities for finding significant association rules for predicting and diagnosis of breast cancers. However, this methodology needs to be evaluated in a larger set of examples in order to find associations with a higher degree of statistical confidence. Using a larger data set will also enable us to find correlations between a bigger set of genes and SNPs.

○ Einipour [95] combined two methodologies including ACO (Ant Colony Optimization) and Fuzzy System and made an automatically breast cancer diagnosis system named as FUZZY-ACO. The main advantage of the proposed system was high reliability and adequate interpretability in compared with other algorithms. Further the results of comparing the proposed approach with some algorithms such as C4.5, SVM, NN, Naïve Bayes and MLP revealed that, it had accuracy rate higher than other algorithms.

○ Raad and colleagues [96] made an approach for classification of breast cancers based on neural network techniques. Further, they developed a tool for automatic detection of breast cancers based on RBF neural network. They proved that accuracy, reliability and efficiency of RBF in compare with MLP technique was better.

○ Wen-Jia Kuo at el. [97] proposed a new computer aided diagnosis (CAD) system for classification of breast cancers by using decision tree technique. The main goal was reducing the number of unnecessary biopsies and increasing the diagnosis confidence. They used 24 covariance texture features for creating decision tree with ability of identifying benign and malignant breast cancers. Accuracy, Positive Predictive value, Negative Predictive value, Sensitivity and Specificity are concerned as objective indices for estimating performance of proposed system in diagnosis of cancers. Authors claimed that, their system which had been made by decision tree had 96% accuracy rate, 93.33% Positive Predictive Value, 96.69% Negative Predictive Value, 93.33% sensitivity and 96.67% Specificity.

○ Jaimini Majali [98] compared the results of applying FP algorithm for diagnosis of breast cancers with the results of applying various classification techniques (such as: Neural networks, Bayesian classifier and Decision tree algorithm). Finally proposed a system for diagnosis and prognosis of breast cancers. Author used FP (frequent pattern mining) algorithm for recognizing the type of breast cancer (malignant or benign tumour) and Decision Tree algorithm to predict the possibility of breast cancer in context of age.

○ The other approach for early diagnosis of breast cancers with reducing the number of unnecessary biopsies is proposed by Jamarani et al. [99] by applying composition of different

techniques including: ANN (Artificial Neural Network) and multi-wavelet based sub-band image decomposition. They claimed that, their proposed approach can be used as part of Computer Aided Diagnosis (CAD).

○ Sudhir D. Sawarkar et al. [100] have used neural networks and SVM (Support Vector Machine) method for diagnosis of breast cancers and proposed a new algorithm with implementing SVM by using kernel Adatron algorithm. This algorithm has capability for mapping inputs into a high-dimensional space. Further, it can be isolate inputs and separating data into their respective classes. Their experimental results revealed that, the proposed algorithm has high accuracy on diagnosis and detection of breast cancers. Based on their results, the accuracy of cancer diagnosis by surgeons, radiologists or physicians was nearly 85%, whereas, the accuracy of detections made by their proposed algorithm was 97%.

○ Xiao-Hui Wang et al. [101] attempted to combine physical examinations' results, patients' clinical background, histories and features of Mammography images and made a hybrid combination for diagnosis of breast cancers. Their experimental results showed that, applying single classification method (such as logistic regression) for both image and non-image information had higher performance and high accuracy rate in compare with applying hybrid combination which is tested by them.

○ Dina A. Sharaf-elDeen [102] used hybrid case-based approach for proposing a breast cancer diagnosis system. This system extract adaptation rules by integrating case-based and rule-based reasoning. In the proposed system, case based reasoning can be automatically generate the reasoning and the adaptation rules. In this system, both reasoning and adaptation rules are updated automatically with each new cases that are added for solving into system. Therefore, there is no need to create these rules from beginning. Experimental results with mammography images information showed that, the developed approach have reliable accuracy and can assist physicians to make diagnosis decisions with high accuracy rate.

○ Pendharkar PC et al. [103] studied breast cancers and proposed an approach with combining several Association rule mining and classification techniques. Their experimental results revealed that, this approach is capable to predict the occurrence of breast cancers or diagnosis cancers in the first stages.

○ Ta-Cheng Chen [104] proposed an approach based on genetic algorithms (GAs) for extracting breast cancers' pattern, decision rules, threshold values and finally decision-making model with high degree of prediction accuracy. Their experimental results showed that, in the proposed approach, accuracy of prediction was improved. Further, modelling becomes simple. Moreover, this proposed approach had capability for extracting rules and acted on a computer model for prediction or classification purposes in breast cancer data.

Challenges: Regarding breast cancer diagnosis as we have seen in the above, severalalgorithms are proposed for improving the accuracy of breast cancer diagnosis, but the main challenges is: which one is having the highest accuracy rate for all databases? There are several diagnosis ways for doctors to help them for identifying cancers such as: Mammography, Physical test and etc. Now the main question is, is it possible to make a method which can combine the results of all tests or images and can detect the malignant in the first stage? Or propose an algorithm for predicting the possibility of occurring breast cancer for each individual person based on various tests or other personal information?

• Several researchers attempted to apply Regression Data Mining Techniques in breast cancer databases for diagnosis of breast cancers in the first stages.

○ Tayal et al. [105] compared seven regression techniques (Linear Regression, Pace Regression Model, Multiple Linear Regression, Non Linear Regression, Logistic Regression, Regression by Discretization and Isotonic regression) for diagnosis of cancers in first stages. Their experimental results showed that, among these seven regression techniques Logistic Regression with 10.59% of least Relative Absolute and 45.84% of Least Root Relative Squared Error was the best performance for diagnosis of breast cancers.

• Some tools are developed for breast cancer prediction or early diagnosis purpose. Such as:

○ Gauthier and colleagues [106] developed reliable assessment tool for breast cancer prediction and early diagnosis. They collect different profiles from public database and defined four parameters (including: breast density, age, prone to breast biopsy and number of affected first degree relatives) for calculating the risk score. In this research work, authors used k-nearest-neighbor algorithm to compute risk score. Their experimental results showed that, the accuracy of developed tool for identifying high risk people was 63%.

Challenges: Currently there is no tool which can diagnose breast cancer with high accuracy rate and minimum degree of errors or minimum number of required medical tests. Further, there are no tools which can analyze various examination tests (images, physical test's results, MRI and etc.) without need to Biopsy and with high degree of accuracy diagnosis cancers.

Data mining techniques for breast cancer treatment

There are several options for treatment of breast cancers and based on the stage of cancer these approaches are proposed by physicians. Some of these treatments are [107, 108]:

• Breast-conserving surgery (BCS) also called as lumpectomy: if breast cancer is small and in the earlier stage of invasive.

• Mastectomy: When the size of cancer is too big.

• Chemotherapy (chemo): is useful for shrinking of cancer and reducing its size enough and preparing that for BCS. It is recommended usually for big size of cancers.

• Radiation: is done for all women who have Mastectomy or BCS. It is reducing the chance of coming back in the cancer.

• Adjuvant systemic therapy or Adjuvant systemic therapy is recommend after Mastectomy or BCS.

• Neo-adjuvant chemo: for treatment of patients before BCS or Mastectomy.

• Hormone therapy: all women with cancer size of larger than 0.5 cm (or t ¼ inch) will be recommend to continue Hormone therapy for 5 years.

• Targeted therapy: Sometimes for treatment of cancers before/after surgery some drugs are recommended by physicians.

We looked in several papers which are published related to data mining applications on cancers' diagnosis, prognosis and the best way of treatment. But unfortunately there are not enough papers which discuss the application of data mining techniques for selecting the best option among the available options for treatment of each individual based on her/his special characteristics. The paper published by Yadav et al. [93] proposed a procedure for identifying patients whose need urgent Chemotherapy starting without wasting time. It used decision tree and SVM to classify patients into two classes as, Benign and Malignant. The accuracy rate of this two classification techniques in this study was 96% and 98% respectively for decision tree and SVM. Therefore, the classed made by SVM are concerned in the next step. In the second step, a clustering method such as k-means is applied for partitioning the two classes into 3 clusters including intermediate, poor and good for identifying the patients who require chemotherapy as soon as possible. Poor group is crucial group and chemotherapy should start immediately to enhance their survival.

Challenges: Generally, in medical world, all patients based on the stage of cancer lead to follow similar treatments. The main question is, can we apply behavior mining techniques to extract unique follow of treatments for each patient? Can we apply classification or clustering techniques for grouping patients based on their characteristic, stage of cancer and etc. and use these groups' characteristics for identifying high risk people whose treatments follow must be different from other patients? Since we saw whenever similar treatments are started for a group of patients with having same stage of disease sometimes number of patients not become well and their health and size of cancer becomes larger, whereas for other patients size of cancer becomes smaller. Why? There is no clear answer in medical world for this question. Even sometimes a few of patients with stage 2 even stage 3 of breast cancer without any treatment live along time whereas, some other patients whose having same stage cancer and continually treated by physicians, unfortunately no longer live. The

Table 1. Important Attributes in Survivability of Breast Cancers [110]

Nominal Attributes	
Race	Marital Status
Primary Site Code	Histologic Type
Behavioral Code	Grade
Extension	Lymph node involvement
Site Specific Surgery Code	Radiation
Tumor Size	
Numerical Attributes	
Age	Number of Positive Nodes
Number of Nodes	Number of Primaries

question: What is the reason for this event? There is no answer of this question up to now.

Data mining techniques for breast cancer prognosis

Prognosis problem is also called as "analysis of survival or lifetime data". It is predicting the occurrence or recurrence of the breast cancer in each individual person. We divide prognosis in two parts [109]:

• Prognosis of occurrence of breast cancer in a person without cancer background (in the other words, women without any breast cancer symptom but they are identified as a high-risk or at-risk people and the possibilities of occurrence of breast cancer in them is high). This type of prognosis is called as, breast cancer prognosis at treatment stage.

• Prognosis the possibility of recurrence of breast cancer after complete treatment. This type of prognosis is called as, breast cancer prognosis at recurrence stage.

Based on several research results, several attributes are affected on the survivability of breast cancer, some of these attributes are presented in **Table 1**.

There are different types of breast cancer recurrence:

• *Local recurrence:* it means, breast cancer after sometimes (may be 6 months or more) complete treatment, will be back in the same place which it had started before.

• *Regional recurrence:* it means, when the breast cancer happens for the second time, it will appear in the lymph nodes near the place that it happened first time.

• *Distant recurrence:* it means, after treatment, for a second time when breast cancer appears, it will start in some other part of the body not in breast itself, such as: liver, bone, brain or lungs.

Several experimental results of physicians are shown that, approximately most of the women 5 years after diagnosis of breast cancer are alive. However, some people live more than 5 years but generally 5-years survival period is used as a standard rate for discussion about prognosis.

There are several paper which are published for recognizing high risk people those who are prone to cancer. We categorised these papers in two parts:

• For predicting survivability rate of breast cancers, several data mining techniques are compared and performance or accuracy rate of that algorithms are compared.

○ Abdelghani Bellaachia et al. [111] applied three data mining techniques including the back-propagated neural network, Naïve Bayes and the C4.5 decision tree algorithms for predicting the survivability rate of breast cancer patients. Their experimental result showed that, C4.5 algorithm had better performance in comparisonto other techniques for predicting survivability rate of breast cancer patients.

○ Jaree Thongkam et al. [112] applied Modest AdaBoost, k-mean, Real, Gentle, C4.5, Bagging, random forest, C-SVC Algorithms for extracting knowledge from breast cancer survival database in Thailand. Performance of these algorithms is examined by comparing classification accuracy, confusion matrix, sensitivity, specificity and stratified 10-fold cross-validation method. Their experimental results revealed that, the accuracy of prediction by Real, Adaboost and Gentle was better than others.

○ Alshammari et al. [113] applied four classification approaches (including: C4.5, Naïve Bayes and two types of ANN (Artificial Neural Network): Multilayer Perceptron and RBF (Radial Basic Function)) for predicting breast cancer survivability. Their experimental results showed that, among these approaches, the accuracy (0.893), specificity (0.985) and sensitivity (0.891) of predicting breast cancer surviv-

ability by decision tree (C4.5) was higher than others.

○ Kung-Min Wang et al. [114] applied to classification techniques (decision tree and logistic regression model) for predicting 5 year survivability of breast cancer patients. They selected some variables such as: cancer stage, extension of cancer, grade, race, site-specific surgery code for applying their classification. Their experimental results showed that, logistic regression model had better accuracy rate in compare with decision tree model.

○ G. Ravi Kumar [115] compared the accuracy of breast cancer prognosis and prediction by applying six classification techniques including: KNN, Naïve Bayes, Decision Tree, Logistic Regression, MLP and SVM. Their experimental results proved that, SVM is the more suitable for breast cancer prediction since it has the highest accuracy rate among other techniques.

○ YJ Lee et al. [116] used data mining technique (nonlinear smooth support vector machines (SSVMs)) for analysing the effects of chemotherapy on survival time of breast cancer patients. They analysed the effects of several features (feature space, cytological features, pathology features) in tumour size. They classified all breast cancer patients into three classes: Good, Intermediate and Poor. Then showed that, only for patients in good group chemotherapy is not required.

○ D. Delen et al. [117] applied three classification techniques including: decision trees, artificial neural networks and logistic regression for prediction breast cancer survivability. Their experimental results with using 10 fold cross – Validation for each model showed that, decision tree had higher accuracy rate (0.9362) among other techniques.

○ J. Gadewadikar et al. [118] applied Bayesian Belief Network for predicting breast cancer automatically. They developed an interface for radiologists to detect cancers with analysing mammography.

○ A. A. M. Medhat et al. [119] for extracting the mammographic mass features and using these features for predicting survival time of breast cancer patients, applied Tree Boost, SVM and Tree Forest techniques. Their experimental results showed that, SVM technique had highest prediction accuracy in comparison with the other techniques.

○ Gandhi Rajiv K et al. [120] created classification rules by applying Particle Swarm Optimization Algorithm. For selecting feature subset, they used fuzzy and Genetic algorithms. They created smaller fuzzy rule bases system with higher accuracy. These rules are used for classification by applying Particle Swarm Optimization Algorithm and showed the high rate of accuracy.

○ Sudhir D et al. [121] for predicting the type of cancer and avoiding un-necessary biopsy, used SVM and ANN techniques. Their experimental results showed that, both of these techniques had 97% accuracy rate and can be used as assistant for physicians to avoid un-necessary biopsy.

• Several methods are proposed by researchers for prognosis of breast cancers.

○ Saleema et al. [122] proposed a model for identifying prominent response variables (such as: patient age at diagnosis, stage of cancer and patient survival) by applying the standard classifiers. This model had three phases namely: basic level pre-processing, problem specific processing (deals with sampling, feature selection and response variable selection) and classifiers for modelling (such as Decision Tree, KNN and Naïve Bayes) for prediction analysis. Their experimental results showed that, decision tree algorithm had better performance in compare with other classifiers. Further, balanced stratified sampling technique maintained consistency in the performance.

○ Dengju Yao et al. [123] proposed a combination method of multivariate adaptive regression splines (MARS) and random forest algorithms for cancer prediction. They used random forest algorithm for screening variables and give rank for those variables. Then used MARS procedure for creating a model for predicting cancer survivability. The performance of this method by calculating accuracy rate, specificity, sensitivity and 10-fold cross-validation is evaluated and the results showed that, this method had higher accuracy in compare with the models.

○ Saleema et al. [124] proposed a sampling method by examining the effect of sampling techniques (such as stratified sampling, random sampling and balanced stratified sampling) in classifying the prognosis variables. Three classification techniques including K-Nearest Neighbour, Naïve Bayes and Decision Tree are used as a classifiers. Three prognosis factors including stage of cancer, metastasis and survival are used as class labels. The results of experimental results showed that, by applying their proposed method, accuracy of prediction by balanced stratified model is increased continually by increasing the sample size, whereas, it is not true in traditional approaches.

○ Reeti Yadav et al. [93] proposed a procedure for identifying patients whose are require urgent chemotherapy without wasting time. In this procedure, in the first step, they used support vector machines (SVMs) and Decision Tree for classifying all patients into two classes including Malignant and Benign. In the second step, clusters (including: Poor, Intermediate and Good) are made by the help of k-means algorithms for identifying patients whose are need urgently for chemotherapy. In this clusters, whose are belong to Poor cluster, are need for urgent chemotherapy. Based on the experimental results of these researchers, SVM had highest accuracy rate (98%) for classifying the patients.

○ Chul-Heui Lee et al. [125] proposed a classification method based on hierarchical granulation structure (for finding classification rules) and applied the rough set theory. Their experimental results showed that, their proposed method, created minimal classification rules and made analysing of information system easier.

Challenges: The recurrence of cancer is not similar in all of the patients. In one group of patients, never breast cancer recur after treatment, in the other group, it comes back approximately 6 months after treatment and in the third group of patients, within 5 years that will be recur. Therefore, developing a tool or proposing a method for predicting the recurrence of cancer and exact time of recurrence is required. Currently, there is no popular tool with acceptable accuracy rate for this matter.

Conclusions

This paper reviewed several research works which are done for diagnosis, treatment or prognosis breast cancers. Based on the results of this study, most of the research works are concerned on comparing the accuracy rate of data mining various algorithms or techniques. Unfortunately, there is no tool that automatically diagnose or prognoses breast cancer. Further, there is no research work which apply personalized features for proposing the best treatment for patients.

In the future work, we will attempt to develop a tool with the help of intelligent agents and applying data mining tools with the capability of automatically breast cancer diagnosis and proposing the best treatment.

Address correspondence to: Rozita Jamili Oskouei, Department of Computer Science and Information Technology, Mahdishahr Branch, Islamic Azad University, Mahdishahr, Iran. E-mail: rozita2010r@gmail.com

References

[1] Subhagata Chattopadhyay. A Neuro-Fuzzy Approach for the Diagnosis of Depression. Applied Computing and Informatics Journal 2014.

[2] Ivan LS, Ghasemzadeh H, Mortazavi B, Lan M, Alshurafa N, Ong M, Sarrafzadeh M. Remote Patient Monitoring: What Impact Can Data Analytics Have on Cost? Wireless Health'13 Proceedings of the 4th Conference on Wireless Health 2013.

[3] Han J, Kamber M. Data Mining Concepts and Techniques. In: Morgan K, editor. 2nd edition. 2006.

[4] Fayyad UM, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Databases. AI Magazine 1996; 17: 37-54.

[5] Kumar R, Verma R. Classification Algorithms for Data Mining: A Survey. International Journal of Innovations in Engineering and Technology (IJIET) 2012; 1: 7-14.

[6] Kesavaraj G, Sukumaran S. A Study on Classification Techniques in Data Mining. 2012; 1: 4th ICCCNT.

[7] Soundarya M, Balakrishnan R. Survey on Classification Techniques in Data mining. International Journal of Advanced Research in Computer and Communication Engineering 2014; 3: 7550-7552.

[8] Li J, Wong L. Rule-Based Data Mining Methods for Classification Problems in Biomedical Domains. 15th European Conference on Machine Learning (ECML) 2004.

[9] Kumar D, Beniwal S. Genetic Algorithm and Programming Based Classification: A Survey. Journal of Theoretical and Applied Information Technology 2013; 54: 48-58.
[10] Mansuri AM, Verma M, Laxkar P. A Survey of Classifier Designing Using Genetic Programming and Genetic Operators. International Journal of Engineering Research and Reviews (IJERR) 2014; 2: 16-22.
[11] Vidhya KA and Aghila G. A Survey of Naïve Bayes Machine Learning approach in Text Document Classification. (IJCSIS) International Journal of Computer Science and Information Security 2012; 7: 206-211.
[12] Bielza C and Larranaga P. Discrete Bayesian Network Classifiers: A Survey. ACM-Transaction 2014; 47: 1.
[13] Sood A. Artificial Neural Networks- Growth & Learn: A Survey. International Journal of Soft Computing and Engineering (IJSCE) 2013; 2: 103-104.
[14] Pradhan A. SUPPORT VECTOR MACHINE-A Survey. International Journal of Emerging Technology and Advanced Engineering 2012; 2: 82-85.
[15] Bhatt J and Patel NS. A Survey on One Class Classification using Ensembles Method. IJIRST-International Journal for Innovative Research in Science and Technology 2014; 1: 19-23.
[16] Zhao Q, Bhowmick S. Association Rule Mining: A Survey. Technical Report, CAIS, Nanyang Technological University, Singapore 2003; No. 2003116.
[17] Devi MR, sarojini AB. Applications of Association Rule Mining in Different Databases. Journal of Global Research in Computer Science 2012; 3: 30-34.
[18] Serban G, Czibula IG, Campan A. A Programming Interface For Medical Diagnosis Prediction. Informatica 2006; 51: 21-30.
[19] Prati RC, Monard MC and Carvalho A. C.P.L.F. d. Looking for Exceptions on Knowledge Rules Induced from HIV Cleavage Data Set. International Journal Genetics and Molecular Biology 2004; 27: 637-643.
[20] Salleb Turmeaux, Vrain and Nortet. Mining Quantitative Association Rules in an Atherosclerosis Dataset. Proceedings of the Sixth European Conference on Principles and Practice of Knowledge Discovery in Databases 2004; pp. 98-103.
[21] Gamberger D, Lavrac N, Jovanoski V. High Confidence Association Rules for Medical Diagnosis. In Proceedings of IDAMAP99 1999; pp. 42-51.
[22] Wong CW, Fu WC. Association Rule Mining and its Application to MPIS. https://www.cse.ust.hk/~raywong/paper/dataWarehousing05-mpis.pdf.
[23] Eimievski A, Srikant R, Agrawal R, Gehrke J. Privacy Preserving Mining of Association Rules. SIGKDD 02 Edmonton, Alberta, Canada 2002.
[24] Cai JH, Zhao XJ, Sun SW, Zhang JF. Stellar Spectra Association Rule Mining Method Based on Weighted Frequent Pattern Tree. Research in Astron. Astrophys 2013; 13: 334-342.
[25] Zubi ZS, Mahmmud AA. Using Data Mining Techniques to Analyze Crime patterns in the Libyan National Crime Data. Recent Advances in Image, Audio and Signal Processing 2014; 8: 79-85.
[26] DeRosa M. Data Mining and Data Analysis for Counterterrorism. 2004. http://csis.org/files/media/csis/pubs/040301_data_mining_report.pdf.
[27] Xu HS, Wang L. The Application of Association Rule Mining in CRM Based on Formal Concept Analysis. Advances in Intelligent and Soft Computing Volume 2012; 169: 27-32.
[28] Prasad P, Malik L. Using Association Rule Mining for Extracting Product Sales Patterns in Retail Store Transactions. International Journal on Computer Science and Engineering (IJCSE) 2011; 3: 2177-2182.
[29] Lee I. Mining Multivariate Associations within GIS Environments. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.84.8366&rep=rep1&type=pdf.
[30] Hipp J, Guntzer U, Nakhaeizadeh G. Algorithms for Association Rule Mining-A General Survey and Comparison. SIGKDD Explorations 2000; 2: 58-64.
[31] Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules. Proceeding of 20th Intentional Conf. On Very Large Databases VLDB 1994.
[32] Brin S, Motwani R, Ullman JD, Tsur S. Dynamic Itemset Counting and Implication Rules for Market Basket Data. Proceedings of the 1997 ACM SIGMOD international conference on Management of data 1997; pp. 255-264.
[33] Houtsmaa M, Swamib A. Set-oriented Data Mining in Relational Databases. Data and Knowledge Engineering 1995; 17: 245-262.
[34] Thomas S, Chakravarthy S. Incremental Mining of Constrained Associations. Springer-Verlag Berlin Heidelberg 2000; pp. 547-558.
[35] Hidber C. Online Association rule mining. SIGMOD'99 Philadelphia PA ACM 1999; 1-58113-084-8/99/05.
[36] Das A, Ng WK, Woon YK. Rapid Association Rule Mining. ACM 2001; 1-58113-436/01/0011.
[37] Kumar R, Jaiswal A, Rai D. A Survey: FP Tree Algorithm with and Without Trusted Party for Environmentally Distributed Databases. International Journal of Innovative Research in

Computer and Communication Engineering 2013; pp. 2212-2220.
[38] Pramod S, Vyas OP. Survey on Frequent Item set Mining Algorithms. International Journal of Computer Applications 2010; 1: 86-91.
[39] Gupta B, Garg D. FP-Tree Based Algorithms Analysis: FPGrowth, COFI-Tree and CT-PRO. International Journal on Computer Science and Engineering (IJCSE) 2011; 3: 2691-2697.
[40] Girotra M, Nagpal K, Minocha S, Sharma N. Comparative Survey on Association Rule Mining Algorithms. International Journal of Computer Applications 2013; 84: 18-22.
[41] Loh WY. Classification and Regression Tree Methods. Encyclopedia of Statistics in Quality and Reliability, Ruggeri, Kenett & Faltin, Wiley 2008; pp. 315-323.
[42] Bocci C, Petrucci A, Rocco E. An application of Geographically Weighted Regression to Agricultural Data for Small Area Estimates. http://www.unavarra.es/metma3/Papers/PDFS_ORAL/Petrucci.pdf.
[43] Brunsdon C, Fotheringham AS, Charlton ME. Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. Geographical Analysis 1996; 28: 281-298.
[44] Luigi D, Oana S, Mihai T, Simona V. The Use Of Regression Analysis In Marketing Research. Studies in Business and Economics 2012; pp. 94-109.
[45] Pardoe I. Applied Regression Modeling: A Business Approach. 2012; eBook, ISBN: 978-0-470-05265-5.
[46] Kuzhda T. Retail Sales Forecasting With Application the Multiple Regression. Socio-Economic Problems and the State 2012; 6: 91-101.
[47] Li Y, Zhu J. Analysis of array CGH data for cancer studies using fused quantile regression", Bioinformatics 2007; 23: 2470-2476.
[48] Robinson B, Officer J. Data Mining: Predicting Laptop Retail Price Using Regression. http://www.spelman.edu/docs/aspire-research/joibritneypdf.pdf?sfvrsn=2.
[49] Berkhin P. Survey of Clustering Data Mining Techniques. Grouping Multidimensional Data Book, Publisher Springer Berlin Heidelberg 2006; DOI: 10.1007/3-540-28349-8_2, pp. 25-71.
[50] Kaufman L, Pierreux A, Rousseuw P, Derde MP, Detaevernier MR, Massart DL, Platbrood G. Clustering on a microcomputer with an application to the classification of coals. Analytica Chimica Acta 1983; 153: 257-260
[51] Preisach C, Burkhardt H, S-Thieme L, Decker R. Data Analysis, Machine Learning and Applications: Proceedings of the 31st Annual Conference of the Gesellschaft für Klassifikation e.V., Albert-Ludwigs-Universität Freiburg 2007.
[52] Duda, Hart, Stork. Pattern Classification, 2nd Edition", eBook 2000; ISBN: 978-0-471-05669-0.
[53] Bruzzone L, Prieto D. A partially Unsupervised Cascade Classifier for the Analysis of Multitemporal Remote-Sensing Images. Pattern Recognition Letters 2002; 23: 1063-1071.
[54] Basavaprasad B, Ravi M. A Comparative Study on Classification of Image Segmentation Methods with a Focus on Graph Based Techniques 2014; 3: 310-315.
[55] Chavan SB, Meshram BB. Classification of Web Application Vulnerabilities. International Journal of Engineering Science and Innovative Technology (IJESIT) 2013; 2: 226-234.
[56] Web Application Security Consortium: Threat Classification. http://projects.webappsec.org/f/WASC-TC-v1_0.pdf.
[57] Guo Y, Sreedevi Sampath. Web Application Fault Classification-An Exploratory Study. http://userpages.umbc.edu/~sampath/papers/guo.esem08.pdf.
[58] Kung HJ, Tung HL. Chapter XLVIII: Web Application Classification: A Maintenance/Evolution Perspective. IGI Globa 2008; http://www.irma-international.org/viewtitle/21275/.
[59] Manning AM, Brass A, Goble CA, Keane JA. Clustering Techniques in Biological Sequence Analysis. in Proc. 1st European Symp. Principles of Data Mining and Knowledge Discovery 1997; pp. 315-322, 1997.
[60] Wang H, Huang C, Yao L, Qian Y and Jiang X. Application of Gustafson-Kessel clustering algorithm in the pattern recognition for GIS. Przegląd Elektrotechniczny (Electrical Review) 2011; ISSN 0033-2097, pp. 215-219.
[61] Dayal A. Hierarchical GIS clustering using principal components. 2009 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 2009; pp. V68-V71.
[62] Carlsson G, Memoli F. Characterization, Stability and Convergence of Hierarchical Clustering Methods. Journal of Machine Learning Research 2010; 11: 1425-1470.
[63] Murtagh F and Contreras P. Methods of Hierarchical Clustering. 2011; http://arxiv.org/abs/1105.0121v1.
[64] Ayramo S, Karkkainen T. Introduction to Partitioning-Based Clustering Methods with a Robust Example. Reports of the Department of Mathematical Information Technology, Series C. Software and Computational Engineering 2006; http://users.jyu.fi/~samiayr/pdf/introtoclustering_report.pdf.
[65] Caputi P. An introduction to grid-based methods. Personal Construct Methodology, Wiley-

Blackwell, United/kingdom 2012; pp. 149-158.
[66] Anthony, Jiawei, Laks, Raymond Constraint-Based Clustering in Large Databases. Lecture Notes in Computer Science, Springer 2001; 1973: 405-419.
[67] Al-Omary AY, Jamil MS. A New Approach of Clustering Based Machine-Learning Algorithm. Knowledge-Based Systems 2006; 19: 248-258.
[68] Andritsos P. Scalable Clustering of Categorical Data and Applications. PhD thesis 2004; http://dblab.cs.toronto.edu/project/limbo/docs/AndritsosPhDThesis.pdf.
[69] Fahad A, Alshatri N, Tari Z, Alamri A, Khalil I, Zomaya AY, Foufou S, Bouras A. A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. IEEE Transactions On Emerging Topics In Computing 2014; 2: 267-279.
[70] Parsons L, Haque E, Liu H. Subspace Clustering for High Dimensional Data: A Review. Sigkdd Explorations 2010; 6: 90-105.
[71] Gibson D, Kleinberg J, Raghavan P. Clustering Categorical Data: An Approach Based on Dynamical Systems. The VLDB Journal 2000; 8: 222-236.
[72] Ramachandran P, Girija N, Bhuvaneswari T. Healthcare Service Sector: Classifying and Finding Cancer spread pattern in Southern India Using Data Mining Techniques. International Journal on Computer Science and Engineering (IJCSE) 2012; 4: 682-687.
[73] http://www.mayoclinic.org/diseases-conditions/cancer/basics/symptoms/con-20032378.
[74] Chaurasia V, Pal S. Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability. International Journal of Computer Science and Mobile Computing 2014; 3: 10-22.
[75] El-Sebakhy A. Emad, Faisal Abed Kanaan, Helmy T, Azzedin F. and Al-Suhaim F. Evaluation of breast cancer tumour classification with unconstrained functional networks classifier. Computer Systems and Applications, IEEE International Conference 2006; pp. 281-287.
[76] Gupta S, Kumar D, Sharma A. Data Mining Classification Techniques Applied For Breast Cancer Diagnosis and Prognosis. Indian Journal of Computer Science and Engineering (IJCSE) 2011; 2: 188-195.
[77] Chang PW and Liou MD. Comparison of three Data Mining techniques with Genetic Algorithm in analysis of Breast Cancer data. [Online]. Available: http://edoc.ypu.edu.tw:8080/paper/ha/Other/%E5%BC%B5%E5%81%89%E6%96%8C_comparison%20of%20data%20mining%20in%20breast%20cancer.pdf.
[78] Kharya S. Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease. International Journal of Computer Science, Engineering and Information Technology (IJCSEIT) 2012; 2: 55-66.
[79] Senturk ZK, Kara R. Breast Cancer Diagnosis via Data mining: Performance Analysis Of Seven Different Algorithms. Computer Science & Engineering: An International Journal (CSEIJ) 2014; 4: 35-46.
[80] Ghassem Pour S, Mc Leod P, Verma B, Maeder A. Comparing Data Mining with Ensemble Classification of Breast Cancer Masses in Digital Mammograms. 2012; http://ceur-ws.org/Vol-941/aih2012_GhassemPour.pdf.
[81] Rajesh K, Anand S. Analysis of SEER Dataset for Breast Cancer Diagnosis using C4.5 Classification Algorithm. International Journal of Advanced Research in Computer and Communication Engineering 2012; 1: 72-77.
[82] Hota HS. Diagnosis of Breast Cancer Using Intelligent Techniques. International Journal of Emerging Science and Engineering (IJESE) 2013; 1: 45-53.
[83] Gupta S, Kumar D, Sharma A. Data Mining Classification Techniques Applied For Breast Cancer Diagnosis and Prognosis. Indian Journal of Computer Science and Engineering 2011; 2.
[84] Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE Jr, Marks JR, Winchester DP & Bostwick DG. Artificial Neural Networks Improve the Accuracy of Cancer Survival Prediction. Cancer 1997; Vol. 79, pp.857-862. http://www.ncbi.nlm.nih.gov/pubmed/9024725.
[85] http://cancernet.nci.nih.gov&http://www.cancerresearchuk.org/about-cancer/type/breast-cancer/treatment/tnm-breast-cancer-staging.
[86] Sumbaly R, Vishnusri N, Jeyalatha S. Diagnosis of Breast Cancer using Decision Tree Data Mining Technique. International Journal of Computer Applications, Volume 2014; 98: 16-24.
[87] Shrivastava SS, Sant A, Aharwal RP. An Overview on Data Mining Approach on Breast Cancer data. International Journal of Advanced Computer Research 2013; 3: 256-262.
[88] Joshi J, Doshi R and Patel J. Diagnosis and Prognosis Breast Cancer Using Classification Rules. International Journal of Engineering Research and General Science 2014; 2: 315-323.
[89] Padmavati J. A Comparative study on Breast Cancer Prediction Using RBF and MLP. International Journal of Scientific & Engineering Research 2011; 2: 1-5.
[90] Aboul HE and Jafar AH. Rough set approach for generation of classification rules of Breast

cancer data. Journal Informatica, 2004; 15: 23-38.
[91] Salama GI, Abdelhalim MB, Zeid MA. Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers. International Journal of Computer and Information Technology 2012; 1: 36-43.
[92] Sarvestan SA, Safavi AA, Parandeh MN, Salehi M. Predicting Breast Cancer Survivability using data mining techniques. Journal of Software Technology and Engineering (ICSTE) 2010; 2: 227-231.
[93] Yadav R, Khan Z, Saxena H. Chemotherapy Prediction of Cancer Patient by using Data Mining Techniques. International Journal of Computer Applications 2013; 76: 28-31.
[94] Orlando A, Bruno GC, Susana V, Jorge G, Arlindo OL and Jose R. A Data Mining approach for detection of high-risk Breast Cancer groups. Advances in Soft Computing 2010; 74: 43-51.
[95] Einipour A. A Fuzzy-ACO Method for Detect Breast Cancer. Global Journal of Health Science 2011; 3: 195-199.
[96] Raad A, Kalakech A, Ayache M. Breast Cancer Classification Using Neural Network Approach: MLP and RBF. The 13th International Arab conference on Information Technology ACIT'12 2012; pp. 15-19.
[97] Kuo WJ, Chang RF, Chen DR, Lee CC. Data mining with Decision Trees for Diagnosis of Breast Tumor in Medical Ultrasonic Images. Breast Cancer Research and Treatment 2001; 66: 51-57.
[98] Majali J, Niranjan R, Phatak V and Tadakhe O. International Journal of Computer Science and Information Technologies (IJCSIT) 2014; 5: 6487-6490.
[99] Jamarani SM, Behnam H and Rezairad GA. Multi-wavelet Based Neural Network for Breast Cancer Diagnosis. GVIP 05 Conference 2005; pp. 19-21.
[100] Sudhir D, Ghatol Ashok A, Pande Amol P. Neural Network aided Breast Cancer Detection and Diagnosis. 7th WSEAS International Conference on Neural Networks 2006; pp. 158-163.
[101] Wang XH, Zheng B, Good WF, King JL and Chang YH. Computer-Assisted Diagnosis Of Breast Cancer Using A Data-Driven Bayesian Belief Network. International Journal Of Medical Informatics 1999; 54: 115-126.
[102] Sharaf-elDeen DA, Moawad IF and Khalifa ME. A Breast Cancer Diagnosis System using Hybrid Case-based Approach. International Journal of Computer Applications 2013; 72: 14-19.
[103] Pendharkar PC, Rodger JA, Yaverbaum XX, Herman N. Association, statistical mathematical and neural approaches for mining breast cancer patterns. Expert Systems with Applications 1999; 17: 223-232.
[104] Chen TC and Hsu TC. A GAs based approach for mining breast cancer pattern. Expert Systems with Applications 2006; 30: 674-681.
[105] Tayal R, Tickoo A. Performance Analysis of Regression Data Mining Techniques Implemented on Breast Cancer Dataset. International Journal of Latest Trends in Engineering and Technology (IJLTET) 2014; 4: 188-194.
[106] Gauthier E, Brisson L, Lenca P, Ragusa S. Breast cancer risk score: a data mining approach to improve readability. https://www.academia.edu/2874820/Breast_cancer_risk_score_a_data_mining_approach_to_improve_readability.
[107] http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-treating-by-stage.
[108] MD Anderson Cancer Treatment Center. Breast Cancer Treatment by Stage. https://www4.mdanderson.org/pe/index.cfm?pagename=opendoc&docid=50.
[109] American Cancer Society. When Cancer Comes Back: Cancer Recurrence. http://www.cancer.org/acs/groups/cid/documents/webcontent/002947-pdf.pdf.
[110] Alshammari, Sultanah M, Shah, Tawfiq M, Huang, Yan. Data Mining Techniques for Predicting Breast Cancer Survivability Among Women in the United States. UNT Digital Library 2015; http://digital.library.unt.edu/ark:/67531/metadc277303/.
[111] Abdelghani B and Erhan G Predicting Breast cancer survivability using Data Mining Techniques. 2007; http://www.siam.org/meetings/sdm06/workproceed/Scientific%20Datasets/bellaachia.pdf?q=data-mining.
[112] Jaree Thongkam, Xu GD, Zhang YC and Huang FC. AdaBoost Algorithm with Random Forests for Predicting Breast Cance Survivability. International Joint Conference on Neural Networks (IJCNN 2008) 2008; pp. 3062-3069.
[113] Alshammari, Sultanah M, Shah, Tawfiq M, Huang, Yan. Data Mining Techniques for Predicting Breast Cancer Survivability Among Women in the United States. UNT Digital Library 2015; http://digital.library.unt.edu/ark:/67531/metadc277303/.
[114] Wang KM, Makond BJ, Wu WL, Wang KJ and LIN YS. Optimal Data Mining Method for Predicting Breast Cancer Survivability. International Journal of Innovative Management, Information & Production 2012; 3: 28-33.
[115] Ravi Kumar G, Ramachandra GA, Nagamani K. An Efficient Prediction of Breast Cancer Data using Data Mining Techniques. International Journal of Innovations in Engineering and Technology (IJIET) 2013; 3: 139-144.
[116] Lee YJ, Mangasarian OL, Wolberg YW. Survival-Time Classification of Breast Cancer Patients.

Computational Optimization and Applications 2003; 25: 151-166.
[117] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods Artificial Intelligence in Medicine. Journal Artificial Intelligence in Medicine 2005; 34: 113-127.
[118] Gadewadikar J, Kuljaca O, Agyepong K, Sarigul E, Zheng Y and Zhang P. Exploring Bayesian networks for medical decision support in breast cancer detection. African Journal of Mathematics and Computer Science Research 2010; 3: 225-231.
[119] Abdelaal Ahmed Mohamed Medhat and Farouq Wael Muhamed. Using data mining for assessing diagnosis of breast cnacer. In Proc. International multi-conference on computer science and information Technology 2010; pp. 11-17.
[120] Gandhi RK, Karnan M and Kannan S. Classification rule construction using particle swarm optimization algorithm for breast cancer datasets. Signal Acquisition and Processing. ICSAP, International Conference 2010; pp. 233-237.
[121] Sudhir D, Ghatol Ashok A, Pande Amol P. Neural Network aided Breast Cancer Detection and Diagnosis. 7th WSEAS International Conference on Neural Networks 2006.
[122] Saleema JS, Deepa Shenoy P, Venugopal KR, L. Patnaik M. Cancer Prognosis Prediction Model using Data Mining Techniques. Data Mining and Knowledge Engineering 2014; 6: 25-47.
[123] Yao DJ, Yang J, Zhan XJ. A Novel Method for Disease Prediction: Hybrid of Random Forest and Multivariate Adaptive Regression Splines. Journal of Computers 2013; 8: 170-177.
[124] Saleema JS, Bhagawathi N, Monica S, Deepa Shenoy P, Venugopal KR and Patnaik LM. Cancer Prognosis Prediction Using Balanced Stratified Sampling. International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI) 2014; 3: 9-18.
[125] Lee HC, Seo HS and Choi CS. Rule discovery using hierarchial classification structure with rough sets. IFSA World Congress and 20th NAFIPS International Conference 2001; 1: 447-452.