MARMARA UNIVERSITY
INSTITUTE FOR GRADUATE STUDIES
IN PURE AND APPLIED SCIENCES

APPLICATION OF DATA MINING ON MEDICAL DIAGNOSIS

MUYESAIER AIHAITI

MASTER THESIS

Department of Computer Engineering

Thesis Supervisor

Assoc. Prof. Dr. Ali Fuat ALKAYA

ISTANBUL,2017

MARMARA UNIVERSITY
INSTITUTE FOR GRADUATE STUDIES
IN PURE AND APPLIED SCIENCES

APPLICATION OF DATA MINING ON MEDICAL DIAGNOSIS

MUYESAIER AIHAITI

(524111921)

MASTER THESIS

Department of Computer Engineering

Thesis Supervisor

Assoc. Prof. Dr. Ali Fuat ALKAYA

ISTANBUL, 2017

MARMARAUNIVERSITY
INSTITUTE FOR GRADUATE STUDIES IN PURE AND APPLIED SCIENCES

Muyesaier AIHAITI, a Master of Science student of Marmara University Institute for Graduate Studies in Pure and Applied Sciences, defended his/her thesis entitled "**Application of Data Mining on Medical Diagnosis**", on July 17, 2017 and has been found to be satisfactory by the jury members.

Jury Members

Assoc. Prof. Dr. Ali Fuat ALKAYA
Marmara University Department of Computer Engineering (SIGN).....

Asst. Prof. Dr. Murat Can GANİZ
Marmara University Department of Computer Engineering (SIGN).....

Asst. Prof. Dr. Kıvanç ONAN
Doğuş University Department of Industrial Engineering (SIGN).....

APPROVAL

Marmara University Institute for Graduate Studies in Pure and Applied Sciences Executive Committee approves that Muyesaier AIHAITI is granted the degree of Master of Science in Department of Computer Engineering, Computer Engineering Program on 18.07.2017 . (Resolution no:). 2017/16-02

Director of the Institute

Prof. Dr.

Prof. Dr. Uğur YAHŞİ
Müdür

ACKNOWLEDGEMENT

It is my great honor and pleasure to gratefully thank to my “Master of Science Thesis” advisor, Assoc. Prof. Dr. AliFuat ALKAYA and Asst. Prof. Dr. Yusuf ŞAHİN for their tremendous support and help at every step of this thesis study; from beginning to the end. Appreciate for their wise guidance and patience which made this dissertation possible. Also provided the key technical insights, critically read through several drafts of this work.

I would like to thank to my family for their endless support during my studies.

I overcame all difficulties with the help of their trust and encouragement.

I wish to express gratitude to all academic and non-academic members of Marmara University of Computer Engineering Department. It is a great honor to study with highly qualified members of this department.

June 2017 **Muyesaier AIHAITI**

CONTENTS

ACKNOWLEDGEMENT ..... **II**

ABSTRACT…… **3**

ÖZET…… **4**

ABBREVIATIONS **5**

FIGURES…… **6**

TABLES…… **7**

1. INTRODUCTION **9**

1.1 Aim 11

1.2. Scope 12

1.3. Methodology 14

1.4. Contributions 14

2. BACKGROUND AND LITERATURE SURVEY **15**

2.1 BREAST CANCER 15

2.2 MAMMOGRAPHY INCLUDING FINDINGS 15

2.3 COMPUTER -AIDED DIAGNOSTICS (CADX) SYSTEMS 18

2.3.1 Data Mining and Data Mining Techniques 18

2.3.2 Meta heuristics 22

2.3.3 Detection Techniques 24

3. MODELING ON CANCER DATA **27**

3.1 Structure of Breast Cancer Data 27

3.2 Data Preparation 29

3.3 Classified with Data Mining Methods 30

3.4 Modeling with Data Mining Methods 33

3.5 Modeling with Meta-Heuristics 34

4. RESULTS AND DISCUSSIONS **38**

5. CONCLUSIONS **55**

6. REFERENCES 56

ABSTRACT

APPLICATION OF DATA MINING ON MEDICAL DIAGNOSIS

Today, breast cancer is the most common non-skin cancer affecting women. Because mammograms do not give clear findings until a certain age for the diagnosis of breast cancer, ultrasounds are applied as an imaging technique. Then, according to the findings in ultrasounds, imaging techniques like mammograms or other detection techniques are applied. Patient will be required mammograms if ultrasound findings look suspicious or the patient is above a certain age. Patients will have biopsy according to the mammogram findings and other risk factors values. Each step in this process will lead the patient anxiety, and expenditures. But the result of not performing these processes will lead one case of cancer and leads the patient more serious problems and expenditures. Therefore, every step of the process is of vital importance to predict the most accurate information available.

CADX models are helpful working on a huge number of variables, and related the risk factors and risk estimation together. Many CADX models provide support for experts by evaluating the mammographic findings and are used by radiologists to increase the detection rate of missed cancer patients, and developing sensitive CADX models for breast cancer detection is an active research topic.

In this study, application of well-known data mining techniques is done on a public breast cancer data set. As a novelty, two meta-heuristics, Simulated Annealing – a well-known optimization meta-heuristic and Migrating Birds Optimization – a fairly new meta heuristic are also applied on the same data set and the results are discussed. It is shown that the two meta-heuristics show comparable performance with the well-known data mining techniques.

JUNE2017 **MUYESAIER AIHAITI**

ÖZET

Veri Madenciliğinin Tıbbi Tanı Konusunda Uygulanması

Günümüzde, Meme kanseri kadınlarda en çok rastlanan deri kanseri olmayan kanseridir. Mamografi, belli bir yaşa Kadar meme kanseri teşhisi için net bulguları vermiyor, çünkü ultrason bir görüntüleme tekniği olarak uygulanır. Ardından, ultrasound bulgularına göre, mamogram veya diğer algılama teknikleri gibi görüntüleme teknikleri uygulanır. Ultrason bulguları şüpheliyse veya hasta belirli bir yaş üzerinde ise hastadan mamografi istenir. Hastanın mamografi bulguları ve diğer risk faktör değerlerine göre hastaya biyopsi uygulanacaktır. Bu süreçteki her adım, hastayı kaygıya ve harcamalara yol açacaktır. Ancak bu süreçten geçmemek eğer kanseri bir durum ise hastayı ciddi sorunlara ve harcamalarına yol açacaktır Bu nedenle, sürecin her adımı en doğru bilgi tahmin etmek için hayati önem taşımaktadır.

Bilgisayar destekli teşhis modelleri (CADX) çok sayıda değişkenlerin işlemi ve risk faktörleri ve risk tahmini arasındaki boşluğa yardım sağlayabilir.

Birçok CADX modelleri mamografik bulguları değerlendirerek uzmanlar için destek sağlamak ve belirisz kanser hastalarının tespit oranını arttırmak için radyologlar tarafından kullanılır ve meme kanseri tespiti için duyarlı CADX modelleri geliştirmek ise aktif bir araştırma konusudur.

Bu çalışmada, iyi bilinen veri madenciliği teknikleri kamu meme kanseri veri seti üzerinde uygulanmıştır Bir yenilik olarak, iki meta-sezgisel tarama, Tavlama Benzetimi - iyi bilinen bir optimizasyon meta-sezgisel ve Göçen Kuşlar Optimizasyonu - oldukça yeni bir metasezgisel yöntemlerle aynı veri üzerinde uygulanır ve sonuçları tartışılmıştır. Iki meta-sezgisel optimize teknik sonuçları ve veri madenciliği teknikleri performansları karşılaştırılarak iyi performans gösterdiği görülmektedir.

HAZİRAN 2017 **MUYESAIER AIHAITI**

ABBREVIATIONS

BC **:** Breast Cancer

ANN **:** Artificial Neural Networks

DM **:** Data Mining

CAD **:** Computer-aided detection

CADX : Computer-aided diagnostic

MBO **:** Migrating Birds Optimization

FN **:** False Negative

FP : False Positive

FPR : False Positive Rate

KDD : Knowledge Discovery in Databases

SVM : Support Vector Machines

TN : True Negative

TP : True Positive

TPR : True Positive Rate

FIGURES

Figure 1. National Mammography Database (NMD) [25] 16

Figure 2. Masses with irregular or micro lobulated margins on magnification view..... 17

Figure 3. Masses with speculated margins 17

Figure 4. Position of DM in KDD Process[33] 20

Figure 5. Data Mining Methods [32]..... 21

TABLES

Table 1. Attributes in data set shows data attribute information 27

Table 2. The variable of the data set and its meaning and explanation 28

Table 3. Encoding Types 29

Table 4. Decision Tree Model and valuable information 30

Table 5. LR Model and valuable information 30

Table 6. Input parameters of ANN (Artificial Neural Network) 31

Table 7. Data information of data set 33

Table 8 . Parameters of the MBO algorithm 34

Table 9. Confusion matrix 38

Table 10. Coincidence Matrix of temperature data set 39

Table 11. Coincidence Matrix of decimal data set 41

Table 12. Summary table of decimal data with the threshold between 2.5-3.6, radius 0.04-0.1 42

Table 13. Summary table of temperature data with the threshold between 8-14, radius 0.01-0.05 45

Table 14. SA High performance threshold range gets the following results for decimal data set 49

Table 15. SA high performance threshold range gets the following results for temperature data set 51

Table 16. Performance Table with different data type set..... 52

1. INTRODUCTION

Today, breast cancer is the most common non-skin cancer affecting women. Breast cancer (BC) can be described as the cells growing and acting abnormally in the texture of the breast that hard to control.

According to the reports of the World Health Organization, every year the number of female will be diagnosed with BC is almost 1.2 million [1].

An estimated 231,840 new cases of invasive breast cancer diagnosed among women in the US during 2015; about 350 new cases men. Excluding cancers of the skin, breast cancer is the most frequently diagnosed cancer in women [2].

Mammography which is one of the imaging test used for diagnosing BC in medical science. Mammogram and ultrasounds are required for proper diagnosis of BC. Within the existing techniques, mammography is the best imaging technique for the diagnosis of BC. Mammographic imaging has been found to reduce mortality rate about 30%~70% among patients having breast cancer [3]. However, Processing and understanding mammograms, especially during the imaging processes are difficult. Moreover, mammograms do not give clear findings until a certain age for the diagnosis of BC, ultrasounds are applied as an alternative imaging technique. Then, according to the findings in ultrasounds, imaging techniques like mammograms or other detection techniques are applied. Patient will be required mammograms if ultrasound findings look suspicious or the patient is above a certain age. Patients will have biopsy according to the mammogram findings and other risk factor values. Each step in this process will lead the patient both emotional and financial costs. But the result of not performing these processes may lead into a case of cancer; thus, into more serious problems. Therefore, every step of the process has vital importance to predict the most accurate information.

Although clinical prediction is important, it is difficult for radiologists who must balance the relative effects of many risk factors. Prediction of medical diagnosis is very important to accurate decision making and can get better patient care. Radiologists used to predict the outcome of a disease rely on the personal training and experience. These experiences also important, but they also lead to systematic errors [4]. Estimation of

breast cancer must put in to account combination of more than 100 risk factors to diagnose a single test, is very task if only rely on experience.

In the last ten years, with increasing attention to cancer research in the technology developers in the sector and the medical science academicians, new and creative way for early detection which helped patients and doctors reduce the cancer death rates, cancer research especially the breast cancer is still reason of concern in the United States [5].

That is why there is developing tools that can calculate an accurate probability of having breast cancer to aid in making decisions. For this reason, the use of computer-based models for diagnostic symptoms, detection, classification and extra deductions is important.

Data mining techniques are wide range used to the early diagnosis of BC besides clinical tests. When literature, data mining techniques have been used for the diagnosis of BC shows good performance.

Computer models can provide assistance and convenient in processing a large number of variables, and bridging the gap between risk factors and risk estimation [6]. As an example, computer models have the potential to help radiologists increase the accuracy of mammography examinations in both detection [7] and diagnosis [8].

Existing computer models in the domain of breast-cancer diagnosis can be classified under 3 broad categories: prognostic, computer-aided detection (CAD) and computer-aided diagnostic (CADX) models. Prognostic models, such as the Gail model [9], use retrospective risk factors and clinical information to predict breast cancer risk during a time interval in the future for treatment or decisions [10]. CAD models [11] are computerized tools that assist radiologists in locating and identifying possible abnormalities in radiologic images, leaving the interpretation of the abnormalities to the radiologists [12]. Diagnostics (diagnostic) models or CADX models [13] helps the radiologists in classifying the findings by producing a suspicion risk score or probability using the information about the findings detected and characterized on a mammogram either by CAD systems or radiologists.

The aim of this thesis is developing a CADX model to get accurate diagnostics for BC. Most of the CADX models in the literature are developed by using ANN techniques [14], logistic regression [15] and decision tree algorithm [16] and also applied in many fields

in medical diagnosis. ANN used to estimate the risk of many cancers, including breast, prostate, liver, ovarian, cervical, bladder, and skin cancer [17]. Logistic regression has been used to estimate disease risk in coronary heart disease [18], breast cancer [18], and prostate cancer [19].

In this study, ANN, logistic regression and decision tree models are built as CADx models and tested. Furthermore, as a novelty, two meta-heuristics: Simulated annealing (SA) [20] and Migrating Birds Optimization (MBO) [21] ,which are not applied to this domain yet are also used to build classifier models. The performance of the meta-heuristics models are compared with the traditional data mining models built in this study and in the literature. It is shown that the meta-heuristics models show comparable even slightly better performance.

1.1 AIM

Computer-aided diagnostics models can provide important support to radiologists for the early diagnosis of breast cancer by evaluating the demographic information and the mammographic findings such as the calcification, mass, bi-directional symmetry distortions and structural defects. Many CADX models developed for the early diagnosis of breast cancer are used by radiologists to increase the detection rate of missed cancer patients, and sensitivity of breast cancer detection.

With the usage of computers with automated tools, availability of data storage, large number of medical data are collected and made them available to the researchers who i interested in making prediction models. At the end, new research way as knowledge discovery in databases (KDD), which includes data mining techniques, become a popular research tool for medical researchers who want to confirm and find patterns and relationships in large number of variables, and able to predict the result of a disease using the historical cases stored in large data sets. Motivation for this research is comes from the reasons like enormous effects of breast cancer, the hopeful results of related research, the potential benefits of the research outcomes and further understand of breast cancer.

In this thesis we developed different kind of models that predict the diagnosed cases for breast cancer. In this thesis we worked on three different types of data mining techniques like artificial neural network (ANN), decision tree, and logistic regression; and two

meta-heuristics: Simulated Annealing and Migrating Birds Optimization to build classifier models which classifies the given patient record as being benign or malignant. Then, the performance of these classifiers is compared with respect to accuracy.

As mentioned above, breast cancer is the most common cancer type and it would cause patients anxiety and expenditures. So, detection and diagnosis of breast cancer is an important research topic and even tiny improvements in this field can bring huge benefits to the patients and also a big lead forward for the medical diagnosis.

As a consequence, this research is focused on computer-aided diagnostics of breast cancer. Beside analysis of the breast cancer diagnostics problem, our aim includes application of not only data mining algorithms and techniques, but also meta-heuristics MBO (Migrating Birds Optimization) and Simulated Annealing (SA) to the problem to build high performance classifier models.

1.2. SCOPE

This thesis exploited the Breast Cancer data set of Wisconsin originated from public Data warehouse. Wisconsin Breast Cancer data set has 683 complete data in 699 samples and 16 instances of them not exist.

In this problem, the aim of the diagnostics system is deciding on the diagnostic as either being benign or malignant.

To build CADX models, we use data mining methods ANN, decision tree algorithms, logistic regression and also two meta-heuristics which minimize the misclassification cost instead of minimizing the error rate.

The aim of most of the applications in the literature is only increasing the classification accuracy. However, the two classification errors – assigning a benign case as malignant and assigning a malignant one as benign – do not result in the same cost. Accordingly, cost-sensitive measures such as precision, recall or specificity should be used to show the real performance of classifier models used in the breast cancer diagnostics model.

As given above, in the BC diagnostic problem, the target of the detection system is assigning a diagnostic as either malignant (cancerous) or benign (not cancerous). Thus, there are four cases that can occur as a result of a class prediction: True Positive (TP), an

actual malignant case is assigned as malignant; True Negative (TN), an actual benign case is assigned as benign; False Positive (FP), an actual benign is assigned as malignant; and False Negative (FN), an actual malignant case is assigned as benign if we assume malignant class value as positive and benign as negative in this binary classification problem.

Assuming that predicting the class label of a diagnostic as its actual class value does not cause any cost, the above given first two assignments (TP and TN) do not cause any cost, while the latter two, assigning an actual malignant instance as benign (FN) and assigning an actual benign instance as malignant (FP) have different costs. The first classification error will bring dangerous results to the patient, even death; but the second one will result in some extra medical processes and will cost some physical paining and economic expenditures.

Nevertheless, because accuracy assumes the effects of these two classification errors as same, we cannot distinguish the effects of different classification errors when we use accuracy as the performance metric. As a consequence, though we may still use accuracy to see the overall performance of a classifier in this problem domain, the real performance can be detected using cost-sensitive performance metrics.

The ultimate goal of this thesis is to propose and implement effective (CADX) models using data mining techniques and meta-heuristics that enhance the efficiency of classification performance of a breast cancer diagnostics. Performance of those classifier models are compared each other using different performance metrics such as accuracy, precision, recall and specificity.

To accomplish this, we identified the following objectives:

- Build training and test sets from the given data set and perform the data preprocessing steps if any required.
- Survey the traditional data mining techniques used in breast cancer that are known to be most successful in detecting the cancer.
- Develop classifier models using the selected well-known traditional data mining techniques.
- Develop classifier models using the meta-heuristics.

- Compare the performance of the classifiers with respect to many performance metrics.

1.3. METHODOLOGY

The methodology taken up in this thesis study is as follows:

Firstly, the data sources that are used as the basis for the training and test sets used during the modeling and testing phases, respectively, of the classification process are identified and analyzed.

In the literature, we can see that for the solution of the problem "How to create a successful computer model for the early detection of breast cancer using by mammogram records," data mining techniques like artificial neural networks, logistic regression, and decision tree algorithms are already used but many meta-heuristics are not applied yet. Therefore, we are going to implement both well-known data mining techniques to use for bench marking and meta-heuristics which are not implemented in the literature to bring an original contribution to the literature.

In accordance with past literature, we randomly sample 80% of the original data set as the training set and the remaining 20% is left as test set using SPSS Clementine 12.0. Afterwards, classifier models are developed using many data mining techniques and two meta-heuristics, Simulated Annealing and MBO.

The performance of the classifier models developed is evaluated with respect to many well-known traditional performance metrics including accuracy, precision, recall and specificity.

1.4. CONTRIBUTIONS

The contribution of this study to the literature is as follows:

Meta-heuristic models are not implemented in previous literature. Accordingly, successful application of meta-heuristics originally defined for optimization to classification problem is a novelty and will bring an original contribution to the literature.

This research will not only make contributions on theoretical aspects, but also will be applicable in real life. That kind of modeling techniques will aid doctor's diagnosis and will raise the accuracy of diagnosis also will help and reduce patients in physical pain and prevent expenditure on financial side.

2. BACKGROUND AND LITERATURE SURVEY

2.1 BREAST CANCER

Breast cancer is development of cancer from breast tissue. [22] Signs of breast cancer may include a lump in the breast, changing in breast shape, dimpling of the skin, fluid coming from the nipple, or a red scaly patch of skin. [23] In those with distant spread of the disease, there may be bone pain, swollen lymph nodes, shortness of breath, or yellow skin. [24]

Until now even scientists still do not know the causes of most breast cancer, but they do know some of the risk factors may increase the probability of breast cancer. These factors contain information like age, genetic risk and family history.

Treatments for breast cancer have two main types, local and systematic. Surgery and radiation are belong to local treatments, chemotherapy and hormone therapy is belong to systematic therapies. To get the best results, we used the two types of treatment together.

2.2 MAMMOGRAPHY INCLUDING FINDINGS

A mammogram search for mark of breast disease when you do not have any breast illness. Mammograms take x-ray pictures of each breast from 2 different angles.

Mammograms analyze a woman's breast if she has a breast sickness or a changing different abnormal seen on a screening mammogram. They could take in to considering the specific information of NMD like demographic information of patient and mammography findings and relationship with BIRADS is shown in Figure 1 which we can diagnose the patient is cancerous or not according to the findings criteria and other screen information. ANN model and the logistic regression are built with the information provided by National Mammography Database (NMD). There are some abbreviations in this chart like Assy. = asymmetric, Br = breast, Ca = cancer, FH = family history, PH = personal history, Trab = trabecular

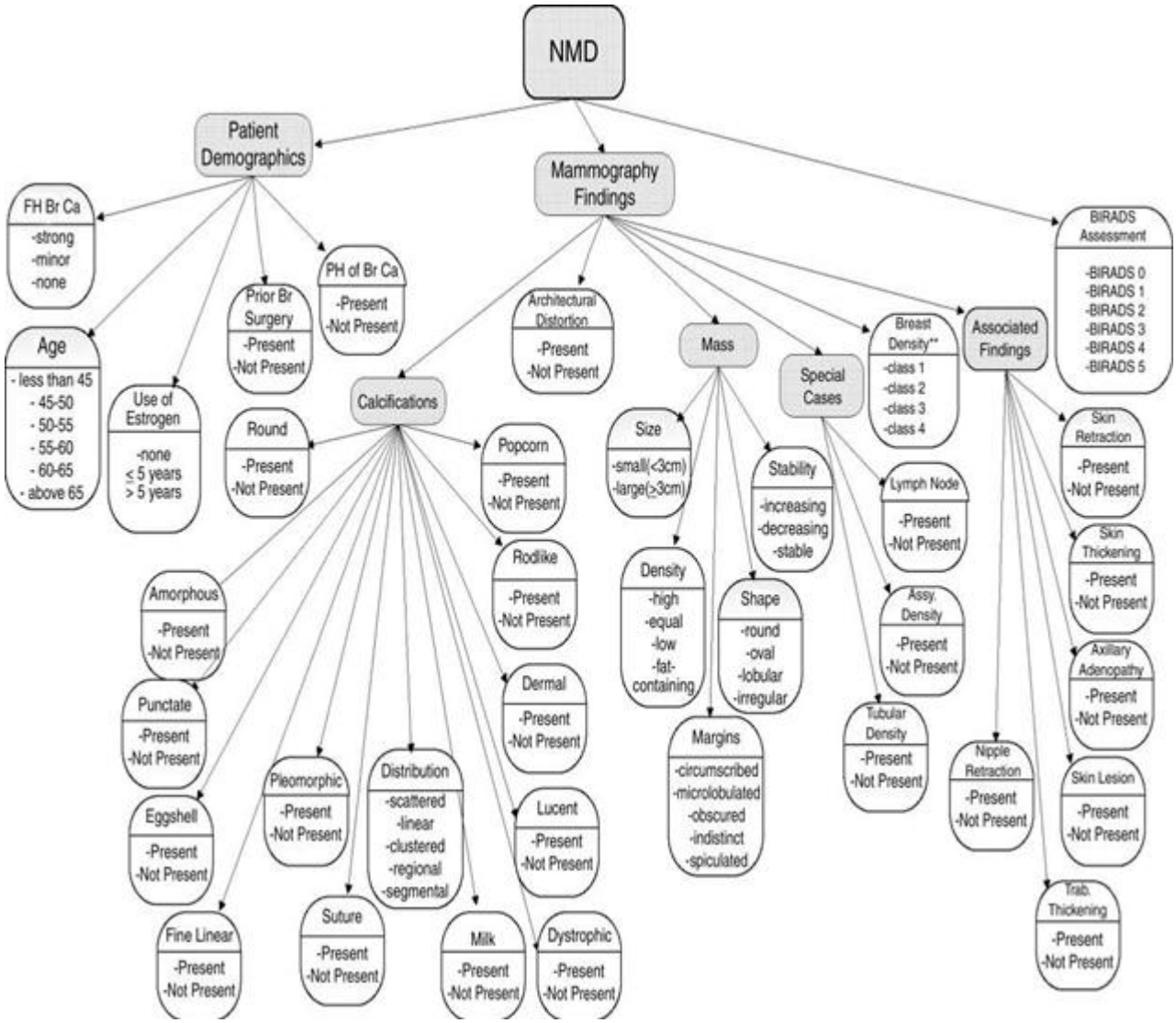


Figure 1. National Mammography Database (NMD) [25]

Analysis only rely on mammograms can't confirm abnormal palace is sign of cancerous, but they can provide further information that will decide more testing is needed or not. Calcification and masses are 2 main types of breast changes on mammogram.

Calcification is kind of very small sediment in the breast organization, which look like small white spots on the mammogram pictures.

A mass is another important changing seen on mammograms. Masses can be many types like cysts (non-cancerous, fluid-filled sacs) and non-cancerous solid tumors, but they could be a cancer. Any mass usually needs to be biopsied if we have no clear information.

Mammograms are used to assess the shape and type of findings of skin.

An irregular shape in Figure2 and Figure 3 pay more attention on its point out unclear or anomaly margins.

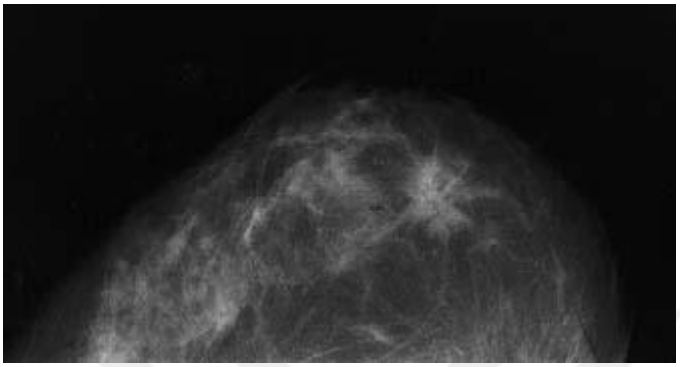


Figure 2. Masses with irregular or micro lobulated margins on magnification view

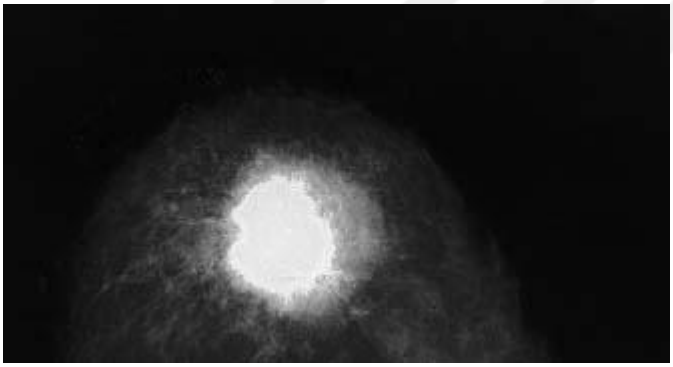


Figure 3. Masses with speculated margins

MAMMOGRAPHIC LESIONS: TYPICALLY BENIGN

Skin calcification

Vascular calcification

Coarse or popcorn like calcification

Small, dense rounded calcification

Milk of calcium.

Eggshell calcification

2.3 COMPUTER -AIDED DIAGNOSTICS (CADX) SYSTEMS

Computer-aided diagnosis (CAD) techniques act like second reader is going to reduce errors. A CAD system could remind the radiologist to review mammograms suspicious place with specialized algorithms.

CAD works following steps:

(1) The radiologist reading the mammogram, then recording any abnormal areas. If exist any abnormality, radiologist will digitally increase the mammographic image to details then confirm abnormal pattern.

(2) CAD controls the mammogram to find any abnormal area.

(3) Radiologist analyzes the CAD system information to confirm whether any suspicious area was left or not.

(4) CAD algorithms may also be used to estimate the probability that a given lesion is malignant or benign, the radiologist gives this diagnoses later.

Investors attracted by the potential benefits of CAD technology and developed some commercial CAD systems, like‘‘Image Checker’’ [26] and ‘‘Second Look’’ [27]. Recently some research has proved that CAD systems efficiently increase a radiologist’s diagnose ability [28]. [29] The authors conclude that the outcome and efficient of a working with CAD is equal to double checking by other way [30] CAD is helpful in screening, where 12,860 types of mammogram were analyzed by CAD system more than 1 year. In this time the number of cancers detected increased by 19.5%, and the scale of early-stage malignancies detected increased from 73% to 78%

The study concludes that CAD can able to improve the detection of early-stage malignancies [31].

2.3.1 Data Mining and Data Mining Techniques

There is a big evolution of information technology. Also the database system industry has witnessed an evolutionary path in the development of the following functionalities: data collection and database creation, data management (including data storage and retrieval, and database transaction processing), and advanced data analysis (involving

data warehousing and data mining). Shows the evolution of database system technology [32].

In the past three decades the progress of computer hardware technology has made powerful computers, data collection equipment, and storage media. This technology provides a great increase to the database and information industry [32]

Nowadays, data can be stored in many different kinds of databases and information repositories and the amount of data kept in computer files and databases is growing with high speed rate. At the same time, the users of these data are expecting more useful information from them.

The position of data mining in the knowledge discovery in databases (KDD) process is given in Figure 4. As seen in Figure 4, DM is the core of KDD process. After the data taken under investigation is prepared through cleaning, integration and selection processes according to the task to be accomplished; DM techniques are employed to find and evaluate the patterns in this data. Due to tremendous amount, high dimensionality and high complexity of data, DM techniques are inevitably used instead of traditional data analysis techniques. Because firms use the knowledge derived by DM techniques in their operational processes, DM plays important roles in customer-oriented working, strategic planning, product placement and fraud detection [33].

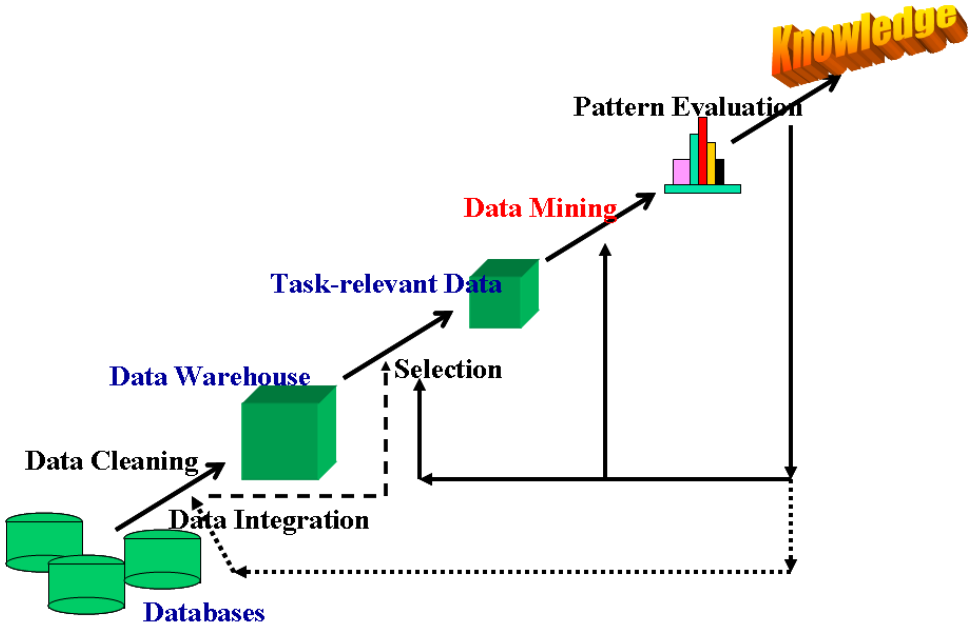


Figure 4. Position of DM in KDD Process[33]

Data used in DM can be derived from any relational public or professional databases or data warehouses; those data can be in any forms of text, multi-media, heterogeneous, spatial, time-series, WWW or transitional data, or data streams. When mining the data, analysis such as characterization, discrimination, evaluation, association, classification, clustering, trend/deviation, outlier analysis can be used. To do these analyses, Online Transaction Processing (OLTP) or Online Analytical Processing (OLAP) techniques, machine learning, statistics and visualization techniques can be employed. Applications based on DM can be used in lots of business sectors such as retail industry, medical science, telecommunications, banking, fraud analysis, bio-mining, and web mining.

The main categories of data mining are classification, regression, time series analysis, clustering and outlier (anomaly) detection, summarization, association rules analysis, sequence and trend discovery.

As seen in Figure 5; classification, regression, time series analysis and prediction are categorized as predictive methods while clustering, summarization, association rules analysis and sequence discovery are categorized as descriptive methods. The objective of predictive tasks is to predict the value of a particular attribute based on the values of other attributes. The attribute to be predicted is commonly known as a target, while the attributes used for making the prediction are known as the explanatory. Predictive

methods use known past features to construct a model which will be used to make predictions on some aspect of new data. Predictive methods are also called as supervised methods because they are supervised by states of the past data. In contrast, descriptive methods are used to show and exposure the characteristics of the examined data when there are no prior sets of past observations. The objective of descriptive tasks is to derive patterns that summarize the underlying relationships in data. Descriptive data mining tasks are often exploratory in nature and frequently require post processing techniques to validate and explain the results .Descriptive methods find the patterns and relationships in the data. Because there is no supervision in this case, descriptive methods are also called as unsupervised methods. While any of these methods can be used individually in one data set, or combination of both supervised and unsupervised ones can be used in together.

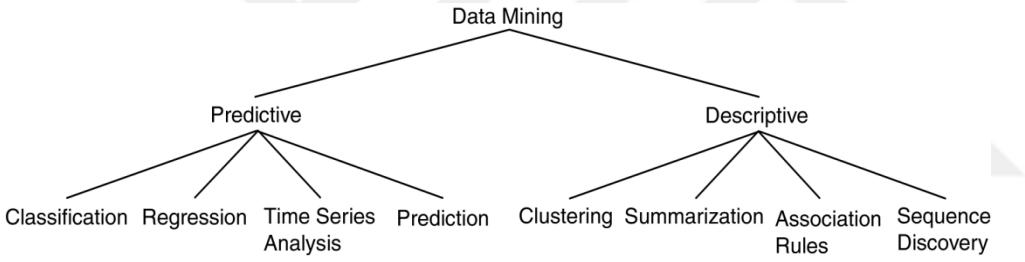


Figure 5. Data Mining Methods [32]

Classification divides the items that get the collection into categories or classes. In the field of data mining, classification is built from historical using a classify model. Predicting the target class for each record in new data accurately is the aim for predictive classification, data that is not in the historical data. A classification task begins with training data. A classification model can also be used to data that was holding aside from the training data to compare the predictions to the known target values; such data is also known as test data. The comparison technique is called testing a model, which measures the model's predictive accuracy. [34] There are some main classification algorithms: Decision Tree Algorithm, Naive Bayes Algorithm, Artificial Neural Network Algorithm, and Support Vector Machine Algorithm, Logistic

Regression. Classification can be used in many fields like classifying diseases (cancer), fraud detection, direct marketing, and other things.

Regression is used to map a data item to a real valued prediction variable. In actuality, regression involves the learning of the function that does this mapping. Regression assumes that the target data fit into some known type of functions and then determines the best function of this type that models the given data.

Regression models have similarity to classification models. The difference between regression and classification technic is that regression work on numerical or continuous target attributes, whereas classification work on discrete or categorical target attributes.

Time Series Analysis, the value of an attribute is examined as it varies over time. The values usually are obtained as evenly spaced time points. A time series plot is used to visualize the time series.

There are two types of clustering algorithms; K- Means Algorithm and O-Cluster Algorithm.

Association Rules refers to the data mining task of discovering interesting relationships hidden in large data sets. The uncovered relationships can be represented in the form of association rules. Association rule is a model that identifies specific types of data association. However association rules can be used to assist retail store management in effective advertising, and marketing.

Sequence Discovery is used to determine sequential patterns in data.

2.3.2 Meta-heuristics

The Meta heuristic methods rely on swarm intelligence are getting more and more popular in optimization, neural network training, scheduling, clustering, classification etc. If there are no satisfactory problem-specific algorithm to solve Metaheuristics are g applied to problems. They are widely used to solve complex problems in industry and services, in areas ranging from finance to production management and engineering [35].

Actually, Meta-heuristic algorithms are the neighborhood search methods and become an important class to improvement algorithms [21]. They search the existing solutions

of neighborhood to get better solutions. The choice of the neighborhood structure, that is, the way in which the neighborhood is defined, is an important matter when designs a neighborhood search algorithm [36].

Many Meta-heuristic algorithms were introduced by many researchers before. Some of the most popular of them, which were introduced firstly, are the genetic algorithm [37], the simulated annealing algorithm [38], the taboo search algorithm [39], the ant colony algorithm [40] and the particle swarm optimization algorithm [41]. Some of the recent meta-heuristic algorithms are the differential evolution algorithm [42], the harmony search algorithm [43], the monkey search algorithm [44], the artificial bee colony algorithm [45], the firefly algorithm [46], the intelligent water drops algorithm [47], the cuckoo search algorithm [48], the bat algorithm [49] and the migrating birds optimization algorithm. Some researchers like to work on the good performances of meta-heuristics [50], like modified and hybrid forms [51] and parallel running methodologies [52] to get better optimization performances.

The MBO algorithm which is a nature inspired meta-heuristic neighborhood search approach was introduced recently by Duman simulates V flight formation of the migrating birds. V shaped flight effectively save energy form for birds .In that V shaped flight the leader bird take the most effort and then other birds get profit from the bird in front position [21].

In the MBO algorithm, the population is initialized firstly. One of the solutions many solutions are chosen to be leader and the other solutions are set on a V formation randomly. Starting from the leader bird claimed as first solution, and keep on the lines from head to the tails, neighbor is want to improve each solution. If the new created neighbor solution giving good performance, then the relevant solution is replaced by that neighbor solution. This neighborhood creation mechanism want to improve all solutions is defined one tour (iteration). These operations proceed many times, and then the position of leader bird is changed. Then changing the position of leader bird to the tail of line then gets the second bird to the leader position.

The notation used in the MBO algorithm is given below = the number of solutions (flock size),

K: the number of neighbor solutions to be considered,

X: the number of neighbor solutions to be shared with the next solution,

M: number of tours to change the leader,

K: maximum iteration number

We are going to use Migrating Birds Optimization (MBO) algorithm and simulated annealing to prepare the performance and accuracy of MBO, and its performance on quadratic assignment problem [21]

2.3.3 Detection Techniques

Many BC diagnostic problems involve huge data sets. So, DM tools used for solving the BC diagnostics problems are many and varied data in different applications can be different in both size and type, but there are common themes. Such tools are essentially based on comparing the observed data with expected values, but expected values can be in various ways depending on the context. They may be single numerical summaries like percentage or simple graphical summaries of some aspect of behavior.

BC diagnostic methods which used in literature can be supervised or unsupervised. In supervised methods, classes of both malignant and benign records are used to construct models.

Data mining techniques used in BC diagnostics are rule-induction techniques, decision trees, neural networks, Naive Bayes, Bayesian learning, logistic regression, k-means clustering and nearest neighbor algorithms. These techniques can be worked alone or used in group together using ensemble techniques to build classifiers.

A. Decision Trees

There are some decision tree algorithms like: ID3, C5.0 and C&RT methods use impurity measures to choose the splitting attribute and the split value/s. ID3 [53] uses information gain while the successor, C5.0 uses gain ratio and C&RT [54] uses Gini index for impurity measurement. Unlikely, CHAID uses chi-square and F statistics to choose the splitting variable [54].

Decision trees such as ID3, C5.0 and C&RT are used for BC diagnostics.

B. Artificial Neural Networks (ANN)

Artificial Neural Network (ANN) is the most commonly used techniques due to their nonlinear nature. ANN is a modeling technique that inspired from the human brains structure. ANN can be used in modeling any complex pattern, so they are well suited for diagnostics of BC problem [55].

The most common neural network is feed-forward back propagation neural networks, also claimed as multilayer perceptron. A set of connected input/output units where each connection has a weight associated with it. Information comes from the input layer through the processing layer(s) to the output layer to make predictions. By adjusting the weights of the connections during training to match predictions to target values for specific records using a proper error function, the network generates better and better predictions through each iteration. However, as the number of iterations reaches a certain level, the network begins to memorize the input data and over fitting occurs. So, the number of iterations must be neither too small so that the network can learn the patterns, nor too large so that over fitting does not occur.

Neural networks are used to build classifiers for the BC diagnostics.

C. Logistic Regression

Logistic regression is for predicting binomial or multinomial results by deciding the contribution of multiple factors to the outcomes. Logistic regression is a member of generalized linear models. Such models include a linear part and a link function where the linear function of the predictor variables is calculated, and the result of this calculation is run through the link function. In logistic regression, the result of the linear function is run through a logistic function. A logistic function starts from 0 at the negative infinity and rises monotonically up to 1.0 at positive infinity. Unlike linear regression where the optimal coefficients are directly calculated; in logistic regression, an iterative fitting procedure is needed in which successive guesses at the right coefficients are incrementally improved like in the case of finding the optimal weights in a perceptron.

There are 2 types of Logistic Regression algorithm like Multinomial and Binomial, first one can make models which have two or more possible values on the target field, second one can make models where the target field is a flag, or binary. Logistic

regression is used in [56] for cıg credit card frauds and in [57] for detecting on insurance frauds.

3. MODELING ON CANCER DATA

This section consists of four parts. In the first part, the information of data structures and the characteristics of the data used will be given. The second part is about the information of data preparation of cancer data in different type. Third part is modeling phases. In the third part, various classifier models are created using well-known traditional data mining methods. In the last part, the details of the newly developed MBO algorithm and the model developed using this algorithm will be given. During the data preparation and modeling with traditional data mining techniques phases, IBM SPSS Clementine 12.0 data mining software is used. It has various data mining algorithms implemented in it. The details of the algorithms used for modeling will be given in the relevant sections.

3.1 STRUCTURE OF BREAST CANCER DATA

Using real world data for BC diagnostic problem is very important. In this thesis we used the same data set to assess the performances of different classic and meta-heuristic classification models. The data set used here was gathered by Dr. William H. Walberg. Diagnosis of Wisconsin breast cancer becomes attractive for many academics these years. The Wisconsin data set consist of 699 samples, 683 of them are complete data and 16 of them not exist. The used to assess nuclear features of fine needle absorbs taken from patients' breasts. Every sample has nine different attributes; they used to measure the changing with nine different variables. The nine variables are got numbered from 1a normal state to 10 the most abnormal state. The class attribute was represented as 2 for benign and 4 for malignant diagnose. All attribute information are given in Table 1 gives the attribute name and data type.

Table 1. Attributes in data set shows data attribute information

#	Attributes	Type	Domain
1	Sample code number	Ordinal	ID number
2	Clump Thickness	Ordinal	1-10

3	Uniformity of cell size	Ordinal	1-10
4	Uniformity of cell shape	Ordinal	1-10
5	Marginal Adhesion	Ordinal	1-10
6	Single Epithelial cell size	Ordinal	1-10
7	Bare Nuclei	Ordinal	1-10
8	Bland Chromatin	Ordinal	1-10
9	Normal Nucleoli	Ordinal	1-10
10	Mitoses	Ordinal	1-10
11	Class	Flag	2=Benign 4=Malignant

To evaluate a tumor in a breast could be malignant (cancerous) or benign (non-cancerous) there are the process used in a pathology report on fine needle absorptions. They are 10 variable of data set which caring important information of patient shown in Table 2. The analysis is probably part of a 3 staged test. First is does cancer cells tend to vary in size and shape. Second the uniformity of cell size/shape will points in a benign direction. Three is bare nuclei; bland chromatin and normal nucleoli are signs of benignity. Biopsy for histological analysis will be necessary if one or both of the other tests points malignancy.

Table 2. The variable of the data set and its meaning and explanation

Clump thickness	Cancer cells are often gathered in multilayers, the opposite benign cells gathered in monolayers.
Uniformity of cell size/shape	Cancer cells want to change in size and shape.
Marginal adhesion	Normal cells want to stay together. Cancer cells lose this ability. So loss of

	adhesion is a symptom of malignancy
Single epithelial cell size	Is related to the uniformity mentioned above. Cells that are notable magnify may be a malignant cell.
Bare nuclei	This is a term used for nuclei that is not surrounded by cytoplasm (the rest of the cell). Those are typically seen in benign tumors
Bland Chromatin	Describes a uniform "texture" of the nucleus seen in benign cells.
Normal nucleoli	Nucleoli are small structures seen in the nucleus. In normal cells the nucleolus is usually very small if visible at all. In cancer cells the nucleoli become more prominent, and sometimes there are more of them.

3.2 DATA PREPARATION

The data which is used for training and test sets are processed using SPSS Clementine. Firstly, we prepared our data in 2 different types which are

1. Decimal encoding
2. Temperature encoding
3. Binary Encoding

Examples which very apparently explain these encoding schemes are as follows:

Table 3. Encoding Types

Encoding scheme	Examples
Temperature	5 -1111100000 8 - 1111111100
Decimal	5 -0000100000 8 -0000000100

Binary	5 – 0101 8 - 1000

There are 444 benign cases and 239 malignant cases in the dataset. In the preparation phase we select the benign and malignant data then split it with the ratio of 80% is for training set and 20% for testing set, according to the literature this ratio gives the best performance among different ratio like (Training70% Testing30%, Training50%Testing50%) [58]. Then emerged the two different classes training and testing parts separately.

3.3 CLASSIFIED WITH DATA MINING METHODS

The chosen methods to build classifier models are C5.0 from decision tree methods; Logistic Regression; and a number of ANN methods defined in SPSS Clementine. Table 3 and Table 4 show parameters of Decision Tree and Logistic Regression model classifier methods.

Table 4. Decision Tree Model and valuable information

Model	Impur.Measure	Pruning Severity	10-Fold Cross -Validate ,boosting	Tree-depth
C5.0	Entropy	75%	False	7

Table 5. LR Model and valuable information

Model	LR
Use partitioned data	false
Calculate variable importance	false

Calculate raw propensity scores	false
Procedure	Multinomial
Model type	Main Effects
Include constant in equation	true
Mode:	Simple
Multinomial Method	Enter

All these methods are used to develop models using the two data sets. These input parameters for the ANN methods used in this study are given in Table 5. Those parameters in 3 Table are the default parameters offered by the relevant methods implemented in SPSS Clementine.

Table 6. Input parameters of ANN (Artificial Neural Network)

Model		**Quick**	**Prune**	**Multiple**		**RFBN**
Model Parameters	Alpha	0.9	0.9	0.9	Alpha	0.9
	Initial Eta	0.3	0.3	0.3	RBF Clusters	20
	High Eta	0.1	0.1	0.1	Eta	0.4
	Low Eta	0.01	0.01	0.01	RBF Overlap	1.0
	Eta Decay	30	30	30		
	Layer 1	20	20	20		
	Layer 2					

	Layer 3					
	Persistent	200	200	200	Persistent	30
	Sample %	50%	50%	50%	Sample%	50%

The code for training C5.0 models is written by Ross Quinlan and licensed from Rule Quest Research Ltd., and the algorithms are proprietary. C5.0 method is implemented in SPSS Clementine and is used as a modeling method in this study. Unlike C&RT method, C5.0 uses gain ratio in the entropy as the impurity measure.

ANN and LR classifier models can make models operating with IBM SPSS Clementine 12.0 in this study. ANN has six different kinds of methods that named named RBFN method, Quick method, Dynamic method, multiple method, Prune method and Exhaustive Prune methods claimed in Clementine 12.0.

Radial basis function network (RBFN) is a special kind of neural network. It contains three layers which named input layer, a hidden layer , and an output layer. The input and output layers have similarity to those multilayer perceptron.

A single neural network is trained, when the quick method is chose to make model. The network has single hidden layer which contains max number of neurons like $\frac{3*(n_i+n_0)}{20}$, where n_i stand for the amount of input neurons and n_o stand for the amount of output neurons. The network is trained it will use back-propagation method.

During the training work the topology of the neural network is changed when the dynamic method is chosen to work, neuron numbers was increased to enhance performance of model when the neural network gets the aimed accuracy. There are two steps of training dynamic method: first is finding the topology the second is training the final network.

When the multiple methods are chose to work, pseudo-parallel fashion way to be used in training. Every specific network is initialized, and all neural networks are trained.

The network gets the best accuracy when stopping criterion is met for all networks is the final model.

The prune methods technic is the opposite of the dynamic method. The prune method begins with enlarge network and prunes useless neurons from the input and hidden layers. Pruning works in two steps: pruning the hidden neurons and pruning the input neurons. The two-step process is repeated until the overall stopping criteria are met.

Table 6 shows classify information of data set, two different types of one data set used in those modeling techniques compare the performances of accuracy which can classify the data close to real classifying information.

Table 7. Data information of data se**t**

Sample	NO of Records	No of Records
Benign (458)	Training Set (516)	Testing Set(167)
Malignant (241)		

3.4 MODELING WITH DATA MINING METHODS

In the modeling phase with the DM methods, the SPSS Clementine 12.0 application is using for make models. For the decision tree models and LR only one model is built for each different method. However, for ANN algorithms, twelve models are built for each method in SPSS Clementine because there is randomness in these methods. Accordingly, modeling for ANN methods takes a long time to complete.

A confusion matrix is a table that is used to describe the performance of a classification model on a set of test data for which the true values are known. The confusion matrix is relatively simple to understand, but the related terminology can be confusing and some terms which can evaluate.

There are formulas from a confusion matrix:

- Accuracy = TP + TN /(TP + TN + FP + FN) (1)
- Sensitivity =TP / (TP + FN) (2)
- Specificity =TN / (FP + TN) (3)
- Precision= TP / (FP+TP) (4)
- Prevalence = FN+TP / (TP+FP+TN+FN) (5)

Where

- TP-True positive, these are cases in which we predicted yes (they have the disease), and they do have the disease.
- FP- False positive, we predicted yes, but they don't have the disease
- TN- True negative, we predicted no, and they don't have the disease.
- FN- False negative, we predicted no, but they do have the disease.

3.5 MODELING WITH META-HEURISTICS

In this part, the MBO algorithm is explained along with the set parameters and their exploited values.

Most Meta heuristic algorithms claimed as neighbor search methods. These are a wide class of improvement algorithms in which an improving solution is found each iteration by searching the "neighborhood" of the current solution. A critical issue in the design of a neighborhood search algorithm is the choice of the neighborhood structure, that is, the manner in which the neighborhood is defined [59]

When migrating birds fly long distances they use V shaped fly forma which is the most famous fly shape. When birds fly they flying in the shape of the letter "V" it is the reason of calling this formation. It is proved to be the efficient way for birds.

In this method we have 13 different parameters shows in Table 7:

Table 8 . Parameters of the MBO algorithm

Parameter Name	Parameter Explanation	Parameter Values

Nob	number of birds	5,11,21,51.101
Non	number of neighbors	3,5,7
Nof	number of flapping	5,10,20
Olf	overlap factor	1,2,3
Fsp	fitness selection factor	10
Threshold	threshold of the value (used for deciding if the patient is malignant or benign	0.5-89.9
Radius	the max distance that new neighbors will be created	0.1,0.2,0.3,0.4,0.5
Interval start/end	the intervals that the weights will be created in	1

There are some parameters which used to scale and test the performance of parameters they are NOx (the running iteration for single searching neighbors process), Average runtime (the average time for single searching process,), Average cost (he average cost of optimization), Min cost (the min cost of optimization), Max cost (the max cost of optimization)

Those methods implemented in the Java language. In its design, MBO is implemented to minimize the misclassification cost which is given in Equation (6).

$$Cost= 1-(TP+TN)/ (TP+FP+TN+FN) \quad (6)$$

Table 9. Parameters name, explanation and value of simulated annealing

Parameter Name	Parameter Explanation	Parameter value
T -initial temperature	The larger this value, the more the inferior exchanges encouraged. In numerical computations	100,1000
R- Number of iterations at each temperature setting	Greater values of R correspond to slower cooling, that is, more exchanges occurring when there is a greater likelihood of inferior exchanges being accepted. In numerical computations	5,20
A - Temperature decrease ratio	After a predetermined number of iterations, T is set to T/a (i.e., T = T/a). When a is large, the temperature decrease is faster and the acceptance of inferior exchanges become less likely at a greater rate. In numerical computations, a is set to either 1.1 or 1.5.	1.1,1.5
b- Increase ratio in iteration number at each setting	After a predetermined number of iterations, R is set to R × b (i.e. R: = R × b). In numerical computations	1.1,1.5
Fsp-fitness selection parameter	fitness selection factor	1
Threshold	threshold of the value (used for deciding if the patient is malignant or benign	0.5-89.5

Radius	the max distance that new neighbors will be created	
Interval-start/end	the intervals that the weights will be created in	1

There are some parameters which used to scale and test the performance of parameters they are NOx (the running iteration for single searching neighbors process), Average runtime (the average time for single searching process,), Average cost (he average cost of optimization), Min cost (the min cost of optimization), Max cost (the max cost of optimization)

4. Results and Discussions

In this section, results of the extensive computational experiments for comparing the selected data mining techniques and implemented algorithms are presented. The selected data mining techniques and meta-heuristic algorithms are ANN (with 14 versions), Decision Tree, LR, MBO and SA. As explained in the previous sections the exploited data is the Wisconsin Breast Cancer data where the objective is to detect breast cancer. The tests are conducted on a workstation having Intel Xeon E5-1620 processor with 4 cores, 32 GB of memory and running Windows 7.

The performance of each model is getting from confusion matrix which in the field of machine learning and specifically the problem of statistical classification. Confusion matrix includes information about actual and predicted classifications done by system. The following table (Table 9) is confusion matrix for a two classifier. Accuracy is the proportion of the total number of predictions that were correct. It is determined using the following equation

$$\text{Accuracy} = (TP+TN)/(TP+FP+TN+FN) \quad (2)$$

Table 9. Confusion matrix

		predicted	
		positive	negative
actual	positive	TP	FP
	negative	FN	TN

In the following, we firstly present the results for the ANN, Decision Tree and LR, then the results of the meta-heuristic algorithms and finally a compact comparison and discussion of all computational experiments.

According to Table 10 they are the coincidence matrix performances of different data sets like decimal data set and temperature data set.

Accuracy of all temperature data is higher than decimal data sets performance including LR has the same accuracy performance and C5.0 models performance is also higher in temperature data than the decimal data s performance except ANN2 and ANN11.

In the sensitivity column 8 of the temperature data sets sensitivity is higher than decimal data sets performance including LR model has the same performance and C5.0 performance is also higher in temperature data set than the decimal data set.

In the specificity column 9 of the temperature data set is higher than decimal data s performance including LR model has the same performance, C5.0 model shows the higher performance in the decimal data set.

In the precision column 8 of them in temperature data set shows the better performance than decimal data set including LR model has the same performance. C5.0 model shows the higher performance in the decimal data set.

Prevalence of both data sets shows the same performance in the all model type.

Table 10. Coincidence Matrix of temperature data set

MFNodel	TP	FP	TN	FN	Accuracy	Sensitivity	Specificity	Precision	Prevalence
LR	47	5	103	12	89.82%	79.66%	95.37**%**	90.38%	35.32%
ANN	57	7	101	2	94.61%	96.61%	93.51%	89.06%	35.32%
ANN 1	59	5	103	0	97.01%	**100%**	95.37%	92.18%	35.32%

ANN 2	55	7	101	4	93.41%	93.22%	93.51%	88.70%	35.32%
ANN 3	56	4	104	3	95.81%	94.91%	96.29%	93.33%	35.32%
ANN 4	58	4	104	1	97.01%	98.30%	96.29%	93.54%	35.32%
ANN 5	56	5	103	3	95.21%	94.91%	95.37%	91.80%	35.32%
ANN 6	59	6	102	0	96.41%	**100%**	94.44%	90.76%	35.32%
ANN 7	55	4	104	4	96.41%	93.22%	**96.29%**	93.22%	35.32%
ANN 8	59	5	103	0	97.01%	**100%**	95.37%	92.18%	35.32%
ANN 9	58	4	104	1	97.01%	98.30%	96.29%	0.974	35.32%
C5.0	57	6	102	2	95.21%	96.61%	94.44%	90.47%	35.32%
ANN 10	58	4	104	1	97.01%	98.30%	96.29%	93.54%	35.32%
ANN 11	55	4	104	4	95.21%	93.22%	96.29%	93.22%	35.32%
ANN 12	59	4	104	0	**97.60%**	**100%**	96.29%	**93.65%**	35.32%
ANN 13	55	4	104	4	95.21%	93.22%	96.29%	93.22%	35.32%
ANN 14	58	4	104	1	95.21%	98.30%	96.29%	93.54%	35.32%

Table 11. Coincidence Matrix of decimal data set

Model	TP	FP	TN	FN	Accuracy	Sensitivity	Specificity	Precision	Prevalence
LR	47	5	103	12	89.82%	79.6%	95.37%	90.38 %	35.32%
ANN	58	4	104	1	**97.01 %**	**98.3%**	96.29%	**93.54 %**	35.32%
ANN 1	57	5	103	2	95.81%	96.61%	95.37%	91.93 %	35.32%
ANN 2	57	5	103	2	95.81%	96.61%	95.37%	91.93 %	35.32%
ANN 3	57	5	103	2	95.81%	96.61%	95.37%	91.93 %	35.32%
ANN 4	57	4	104	2	96.41%	96.61%	**96.29%**	93.44 %	35.32%
ANN 5	55	4	104	4	95.21%	93.22%	96.29%	93.22 %	35.32%
ANN 6	56	4	104	3	95.81%	94.91%	96.29%	93.33 %	35.32%
ANN 7	57	4	104	2	96.41%	96.61%	96.29%	93.44 %	35.32%

ANN 8	58	5	103	1	96.41%	98.3%	95.37%	92.06 %	35.32%
ANN 9	59	5	103	0	97.01%	100%	95.37%	92.18 %	35.32%
C5.0	53	5	103	6	93.41%	89.83	95.37%	91.37 %	35.32%
ANN 10	57	4	104	2	96.41%	96.61%	96.29%	93.44 %	35.32%
ANN 11	58	5	103	1	96.41%	98.3%	95.37%	92.06	35.32%

In this Table12 we use the subtotal to get analysis of the decimal data set, we handle this analysis on the micro soft excel, in the subtotal part we can put the every single parameter particularly threshold and radius separately, then select the function of average and min at the different time, at subtotal selection we choose average cost, min cost, max cost.

Percentage in this table is the misclassification rate of MBO for each parameter, if the percentage is close to zero then it will be better performance and the parameter will has high potential to be best parameters.

Table 12. Summary table of decimal data with the threshold between 2.5-3.6, radius 0.04-0.1

	average			min		
nob	average	min	max	average	min	max

3	5.251%	3.474%	7.340%	3.952%	**1.198%**	5.389%
5	5.19%	3.37%	7.29%	3.95%	1.80%	5.39%
7	5.17%	3.41%	7.26%	**3.89%**	1.20%	**4.79%**
9	5.15%	**3.34%**	**7.21%**	3.95%	1.20%	5.39%
11	**5.14%**	3.38%	7.22%	4.07%	1.20%	5.39%
non	average	min	max	average	min	max
3	5.23%	3.42%	7.35%	3.95%	1.20%	5.39%
5	5.20%	3.40%	7.30%	4.01%	1.20%	5.39%
7	**5.16%**	**3.38%**	**7.21%**	**3.89%**	**1.20%**	**4.79%**
olf	average	min	max	average	min	max
1	5.18%	3.40%	7.28%	3.95%	**1.20%**	**4.79%**
2	5.18%	**3.39%**	7.26%	3.95%	1.20%	5.39%
3	5.18%	3.40%	**7.22%**	**3.89%**	1.20%	4.79%
threshold	average	min	max	average	min	max
2.5	**4.89%**	**3.14%**	**6.91%**	3.95%	1.80%	5.39%
2.6	5.05%	3.31%	7.08%	4.13%	**1.20%**	5.39%

2.7	5.03%	3.32%	7.06%	**3.89%**	1.80%	**4.79%**
2.8	5.04%	3.33%	7.11%	3.95%	1.80%	5.39%
2.9	5.14%	3.35%	7.18%	4.07%	1.20%	5.39%
3	5.17%	3.41%	7.28%	4.19%	1.20%	5.39%
3.1	5.22%	3.38%	7.26%	3.95%	1.80%	4.79%
3.2	5.28%	3.45%	7.45%	4.31%	1.80%	5.39%
3.3	5.29%	3.45%	7.35%	4.31%	1.80%	5.39%
3.4	5.34%	3.52%	7.39%	4.13%	1.20%	5.39%
3.5	5.32%	3.52%	7.47%	4.49%	2.40%	5.39%
3.6	5.41%	3.54%	7.63%	4.25%	1.20%	5.39%
radius	average	min	max	average	min	max
0.04	5.23%	3.40%	7.33%	4.07%	**1.20%**	5.39%
0.05	5.19%	3.43%	7.25%	3.95%	1.80%	5.39%
0.06	5.20%	3.42%	7.31%	3.95%	1.20%	5.39%
0.07	5.18%	3.40%	7.24%	**3.89%**	1.80%	**4.79%**
0.08	5.16%	**3.36%**	7.23%	4.13%	1.20%	5.39%

0.09	5.18%	3.37%	7.28%	4.13%	1.20%	5.39%
0.1	**5.14%**	3.39%	**7.21%**	4.01%	1.80%	5.39%

In this Table 13 we use the subtotal to get analysis of the temperature data set, we handle this analysis on the micro soft excel, in the subtotal part we can put the every single parameter particularly threshold and radius with above mentioned range separately, then select the function of average and min at the different time, at subtotal selection we choose average cost, min cost, max cost.

Table 13. Summary table of temperature data with the threshold between 8-14, radius 0.01-0.05

		average			min		
nob		Average	Min	Max	Average	Min	Max
	5	**2.74%**	**2.39%**	**3.31%**	2.46%	1.80%	2.99%
	11	2.77%	2.39%	3.37%	**2.40%**	1.80%	**2.40%**
	21	2.80%	2.40%	3.44%	2.46%	1.80%	2.99%
	51	2.83%	2.41%	3.53%	2.46%	1.80%	2.99%
	101	2.87%	2.41%	3.62%	2.46%	1.80%	2.99%
non		Average	Min	Max	Average	Min	Max
	3	**2.75%**	**2.39%**	**3.35%**	2.46%	1.80%	2.99%

	5	2.79%	2.40%	3.45%	2.46%	1.80%	2.99%
	7	2.82%	2.40%	3.49%	**2.40%**	1.80%	**2.40%**
nof		Average	Min	Max	Average	Min	Max
	5	2.80%	2.40%	3.46%	**2.40%**	1.80%	**2.40%**
	10	2.80%	2.40%	**3.45%**	2.46%	1.80%	2.99%
	20	2.80%	2.40%	3.45%	2.46%	1.80%	2.99%
olf		Average	Min	Max	Average	Min	Max
	1	2.80%	2.40%	3.46%	2.46%	1.80%	2.99%
	2	2.80%	2.40%	3.46%	**2.40%**	1.80%	**2.40%**
	3	2.80%	2.40%	**3.43%**	2.46%	1.80%	2.99%
Threshold		Average	Min	Max	Average	Min	Max
	8	3.03%	2.44%	3.97%	2.46%	1.80%	2.99%
	8.5	2.95%	2.42%	3.80%	2.46%	1.80%	2.99%
	9	2.88%	2.41%	3.70%	2.46%	1.80%	2.99%
	9.5	2.84%	2.40%	3.58%	2.46%	1.80%	2.99%
	10	2.82%	2.40%	3.53%	2.46%	1.80%	2.99%

	10.5	2.77%	**2.39%**	3.43%	2.46%	1.80%	2.99%
	11	2.74%	2.39%	3.35%	2.46%	1.80%	2.99%
	11.5	2.73%	2.39%	3.29%	2.46%	2.40%	2.99%
	12	2.72%	2.39%	3.26%	2.46%	1.80%	2.99%
	12.5	**2.71%**	2.39%	**3.21%**	**2.40%**	1.80%	2.40%
	13	2.72%	2.39%	3.23%	2.46%	1.80%	2.99%
	13.5	2.74%	2.39%	3.27%	2.46%	1.80%	2.99%
	14	2.76%	2.39%	3.29%	2.46%	1.80%	2.99%
Radius		Average	Min	Max	Average	Min	Max
	0.01	2.97%	2.43%	3.83%	2.46%	**1.80%**	2.99%
	0.02	2.83%	2.40%	3.53%	**2.40%**	2.40%	**2.40%**
	0.03	2.77%	2.39%	3.38%	2.46%	1.80%	2.99%
	0.04	2.73%	2.39%	3.29%	2.46%	1.80%	2.99%
	0.05	**2.71%**	**2.39%**	**3.24%**	2.46%	1.80%	2.99%

Percentage in this table is the misclassification rate of MBO for each parameter, if the percentage is close to zero then it will be better performance and the parameters of this group will have the high potential to be best parameters.

Following is the best performance valuable for temperature data set:

Nob	5
Non	3
Nof	10
Olf	1
Fsp	1
Threshold	12.5
Radius	0.05
Interval-start	0.0
Interval-end	1.0
Average run time	12.946099999999998
Min cost	2.695%
Average cost	2.695%
Max cost	3.593%

For simulated annealing modeling results summary is as following:

Firstly we try to get the threshold range with high performance then get the table 14, table 15 to detect the high performance of threshold range.

For decimal data we can get the following results (only the summary results):

In this work we run the program with the threshold range 0.5-89.9 then look through and ranking the costs under different threshold which can get some high costing threshold data, following make threshold range more narrower and let them more detailed with little distance ,meanwhile look through other parameters performance and make some changing on them to getting high performance.

Table 14. SA High performance threshold range gets the following results for decimal data set

	Average			Min		
T	Average	Min	Max	Average	Min	Max
100	**31.64%**	**4.78%**	**33.61%**	**5.14%**	0.00%	**5.99%**
1000	42.99%	6.67%	46.26%	5.50%	0.00%	7.19%
R						
5	**34.26%**	**5.40%**	**37.29%**	5.48%	0.00%	5.99%
20	40.38%	6.05%	42.59%	**5.14%**	0.00%	5.99%
a						
1.1	53.52%	8.95%	51.56%	10.54%	2.40%	7.36%
1.50	**21.12%**	**2.50%**	**28.31%**	**5.14%**	**0.00%**	**5.99%**
b						
1.1	**26.79%**	**3.46%**	**34.04%**	**5.14%**	**0.00%**	5.99%
1.50	47.84%	7.99%	45.83%	8.17%	1.80%	5.99%
Threshold						
2	63.46%	9.78%	55.38%	8.17%	0.00%	**5.99%**

2.5	48.42%	5.14%	50.28%	8.71%	0.00%	6.59%
3	37.05%	4.21%	42.94%	**5.14%**	0.00%	5.99%
3.5	32.34%	**3.75%**	41.22%	5.50%	0.00%	7.19%
4	28.18%	4.12%	33.20%	5.40%	0.00%	6.59%
4.5	**27.38%**	4.60%	**31.64%**	5.70%	0.00%	6.59%
5	29.03%	6.08%	32.59%	5.80%	0.00%	7.19%
5.5	32.68%	8.13%	32.24%	5.92%	0.00%	6.59%
radius						
0.01	43.15%	7.14%	43.02%	6.28%	0.00%	7.19%
0.02	36.69%	5.72%	39.60%	5.60%	0.00%	5.99%
0.03	**35.15%**	5.33%	38.84%	5.40%	0.00%	5.99%
0.04	35.50%	**5.17%**	**38.18%**	**5.14%**	0.00%	5.99%
0.05	36.10%	5.28%	40.05%	5.50%	0.00%	6.59%

Table 15. SA high performance threshold range gets the following results for temperature data set

		average			min		
		total	min	max	total	min	max
	1000	5.42%	2.11%	3.81%	**4.046%**	1.357%	2.994%
T	100	**5.36%**	**2.10%**	**3.65%**	4.3%	1.357%	2.994%
r	20	5.41%	2.11%	**3.73%**	4.24%	1.36%	2.99%
	5	**5.36%**	**2.10%**	3.74%	**4.05%**	1.36%	2.99%
a	1.1	5.62%	2.34%	3.92%	5.15%	1.80%	2.99%
	1.5	**5.16%**	**1.87%**	**3.54%**	**4.05%**	**1.36%**	2.99%
b	1.5	5.60%	2.33%	3.83%	5.17%	1.80%	2.99%
	1.1	**5.17%**	**1.88%**	**3.63%**	**4.05%**	**1.36%**	2.99%
threshold	13.9	5.39%	2.10%	3.53%	4.40%	1.36%	2.99%
	13.2	**5.32%**	2.10%	**3.52%**	4.34%	1.36%	2.99%
	13.1	5.33%	2.11%	3.55%	4.24%	1.36%	2.99%
	13	5.39%	2.11%	3.70%	4.32%	1.36%	2.99%
	12.9	5.38%	2.11%	3.69%	**4.05%**	1.36%	2.99%

	12.8	5.36%	2.10%	3.63%	4.30%	1.36%	2.99%
	12.7	5.37%	**2.09%**	3.87%	4.30%	1.36%	2.99%
	12	5.56%	2.12%	4.38%	4.32%	1.36%	2.99%

According to the Table16, ANN produced good accuracy results for data sets, but it could not outperform the others. Considering the accuracy, the MBO algorithm shows the best and the LR algorithm shows the worst performances. A support system related to medical decision-making was proposed in this study, using the ANN, LR, DC models and MBO optimization algorithms collectively for the diagnosis of BC. We can get temperature and Binary data set from the decimal original data through the formula that can improve the performance of modeling. In this table we write the mbo and sa results in 100% numbering techniques according our results we should minus our results from 100then get the results in table 16.

Table 16. Performance Table with different data type set

Model name	Temperature Data Set	Decimal Data Set
LR	89.82%	89.82%
ANN	94.61%	97.01%
ANN1	97.01%	95.81%
ANN2	93.41%	95.81%
ANN3	95.81%	95.81%

ANN4	97.01%	96.41
ANN5	95.21%	95.21
ANN6	96.41%	95.81%
ANN7	96.41%	96.41%
ANN8	97.01%	96.41%
ANN9	97.01%	97.01
DC	95.21%	93.41%
ANN10	97.01%	96.41%
ANN11	95.21%	96.41%
MBO	97.30%	98.2%
DC	95.21%	93.41%
ANN10	97.01%	96.41%
ANN11	95.21%	96.41%
MBO	**97.30%**	**98.2%**
SA	94.69%	94.86%

Looking at the results on Table 16, it is clearly seen that the MBO is highly successful in modeling and diagnosis. On the other hand, most ANN versions produced good accuracy results for the data sets, but it could not outperform the MBO. Considering the accuracy, the MBO algorithm shows the best and the LR algorithm shows the worst performances. A support system related to medical decision-making was proposed in this study, using the ANN, LR, DC models and MBO optimization algorithms collectively for the diagnosis of BC.

5. Conclusions

In this thesis, a comparison of decision support systems related to medical decision-making is done. The systems that are evaluated are the NN, DC, LR classification models and two meta-heuristics, namely the MBO and SA.

MBO algorithm is one of the newly introduced metaheuristic algorithms and simulated annealing algorithm is also one of the powerful meta-heuristic algorithms. Both are based on the swarm intelligence, the performance of a recently proposed meta-heuristic on a medical diagnosis problem is presented by comparing with the performance of classical data mining tools.

The compared data mining tools are ANN (with several versions), DC and LR. Comparison study is conducted through an extensive set of computational experiments using benchmark public data. Even though MBO algorithm is one of new introduced meta-heuristic algorithm, its performance outperforms the other aforementioned tools. In the results table, MBO is at least 1.1% and up to 8.3% better than the other tools.

This shows that MBO can also be used in support systems related to medical decision making problems. Therefore, a further study may include application of MBO to other diagnosis problem cases.

Compared the performance of classification and meta-heuristic model to considering the accuracy, the MBO algorithm shows the best and the LM algorithm shows the worst performances. The results show that even though MBO is a recently proposed meta-heuristic it can outperform well known decision support systems in BC diagnosis and it can also be an important player in data mining problems.

6. References

[1] MF. Akay, “Support vector machines combined with feature selection for breast cancer diagnosis”, Expert Systems with Applications, Vol. 36, pp. 3240–3247, 2009.

[2] 2015, American Cancer Society. Cancer Facts & Figures 2015. Atlanta: American Cancer Society; 2015

[3] Rangayyan, Rangaraj M., and Thanh M. Nguyen. "Fractal analysis of contours of breast masses in mammograms." Journal of Digital Imaging 20.3 (2007): 223-237.

[4] Kahneman D, Slovic P, Tversky A. Judgment under uncertainty: heuristics and biases. Cambridge, England: Cambridge University Press, 2001.

[5] Warren J. Cancer death rates falling, but slowly. WebMD medical news; 2003 (http://aolsvc.health.webmd.aol.com/content/Artcile/73/82013.htm)

[6] Ravdin, P. M., De Laurentiis, M., Vendely, T., & Clark, G. M. (1994). Prediction of axillary lymph node status in breast cancer patients by use of prognostic indicators. Journal of the National Cancer Institute, 86(23), 1771-1775.

[7] Tang, Jinshan, et al. "Computer-aided detection and diagnosis of breast cancer with mammography: recent advances." IEEE Transactions on Information Technology in Biomedicine 13.2 (2009): 236-251.

[8] Jalalian, Afsaneh, et al. "Computer-aided detection/diagnosis of breast cancer in mammography and ultrasound: a review." Clinical imaging 37.3 (2013): 420-426.

[9] Gail, Mitchell H., et al. "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually." Journal of the National Cancer Institute 81.24 (1989): 1879-1886.

[10] Ayer, Turgay, et al. "Breast cancer risk estimation with artificial neural networks revisited." Cancer 116.14 (2010): 3310-3321.

[11] Birdwell, Robyn L., Parul Bandodkar, and Debra M. Ikeda. "Computer-aided detection with screening mammography in a university hospital setting 1." Radiology 236.2 (2005): 451-457.

[12] Giger, M. L., Huo, Z., Kupinski, M. A., & Vyborny, C. J. (2000). Computer-aided diagnosis in mammography. Handbook of medical imaging, 2, 915-1004.

[13] Ayer, T., Chen, Q., & Burnside, E. S. (2013). Artificial neural networks in mammography interpretation and diagnostic decision making. Computational and Mathematical methods in Medicine, 2013.

[14] Baker, Jay A., et al. "Breast cancer: prediction with artificial neural network based on BI-RADS standardized lexicon." Radiology 196.3 (1995): 817-822.

[15] Chhatwal, Jagpreet, et al. "A logistic regression model based on the national mammography database format to aid breast cancer diagnosis." American Journal of Roentgenology 192.4 (2009): 1117-1127.

[16] Chao, Cheng-Min, et al. "Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree." Journal of medical systems 38.10 (2014): 106.

[17] Lisboa, Paulo J., and Azzam FG Taktak. "The use of artificial neural networks in decision support in cancer: a systematic review." Neural networks 19.4 (2006): 408-415.

[18] Wilson, Peter WF, et al. "Overweight and obesity as determinants of cardiovascular risk: the Framingham experience." Archives of internal medicine 162.16 (2002): 1867-1872.

[19] Lui, P. D., Terris, M. K., McNeal, J. E., & Stamey, T. A. (1995). Original Articles: Prostate Cancer: Indications for Ultrasound Guided Transition Zone Biopsies in the Detection of Prostate Cancer. The Journal of urology, 153(3), 1000-1003.

[20] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by simulated annealing, Science 220 (4598) (1983) 671¨C680)

[21] E. Duman, M. Uysal, A.F. Alkaya, Migrating Birds Optimization: A new metaheuristic approach and its performance on quadratic assignment problem, Information Sciences 217 (2012) 65, 77

[22] Rangayyan, Rangaraj M., Fabio J. Ayres, and JE Leo Desautels. "A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs." Journal of the Franklin Institute 344.3 (2007): 312-348.

[23] Fan, Lei, et al. "Breast cancer in China." The lancet oncology 15.7 (2014): e279-e289.

[24] Saunders, Christobel; Jassal, Sunil (2009). Breast cancer (1. ed. ed.). Oxford: Oxford University Press. p. Chapter 13. ISBN 978-0-19-955869-8.

[25] American College of Radiology.Breast Imaging Reporting and Data System(BI-RADS).Reston,Va:American College of Radiology,2004.

[26] R2 Technology, Sunnyvale, CA [R2 Technology website, /http://www.r2tech.com/S, accessed on March 6, 2005.]

[27] iCAD, Nashua, NH [iCAD website, /http://www.icadmed.com/S, accessed on March 6, 2005.]

[28]S.M. Astley, F.J. Gilbert, Computer-aided detection in mammography, Clin. Radiol. 59 (2004) 390–399. Ciatto et al

[29]S. Ciatto, M.R.D. Turco, G. Risso, S. Catarzi, R. Bonaldi, V. Viterbo, P. Gnutti, B. Guglielmoni, L. Pinelli, A. Pandiscia, F. Navarra, A. Lauria, R. Palmiero, P.L. Indovina, Comparison of standard reading and computer aided detection (CAD) on a national proficiency test of screening mammography, Eur. J. Radiol. 45 (2003) 135–138.

[30]T.W. Freer, M.J. Ulissey, Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center, Radiology 220 (2001) 781–786.

[31] L.J.W. Burhenne, S.A. Wood, C.J. D'Orsi, S.A. Feig, D.B. Kopans, L.F. O'Shaughnessy, E.A. Sickles, L. Tabar, C.J. Vyborny, R.A. Castellino, Potential contribution of computer-aided detection to the sensitivity of screening mammography, Radiology 215 (2) (2000) 554–562.

[32] Han, J; Kamber, M.: "Data Mining: Concepts and Techniques", 2n Edition, Morgan Kaufman, (2006).

[33] Frawley, William J. Knowledge discovery in databases. Eds. William J. Frawley, and Gregory Piatetsky-Shapiro. Vol. 37. Menlo Park, CA: AAAI Press, 1991.

[34] Taft, Margaret, et al. "Oracle Data Mining Concepts, 10g Release 2 (10.2) B14339-01." (2005).

[35] I. Boussaïd, J. Lepagnot, P. Siarry, A survey on optimization meta-heuristics, Information Sciences 237(2013) 82¨C117.

[36] R.K. Ahuja, O. Ergun, J.B. Orlin, A.P. Punnen, A survey of very large scale neighborhood search techniques, Discrete Applied Mathematics 123 (2002) 75¨C102.

[37] J.H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor,1975.

[38] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by simulated annealing, Science 220 (4598) (1983) 671¨C680.

[39] F. Glover, Future paths for integer programming and links to artificial intelligence, Computers &Operations Research 13 (5) (1986) 533¨C549.

[40] M. Dorigo, Optimization, learning and natural algorithms, Ph.D. Thesis, Politecnico di Milano, Italy,1992.

[41] R.C. Eberhart, J. Kennedy, A new optimizer using particle swarm theory, in: Proceedings of the Sixth International Symposium on Micro machine and Human Science, Nagoya, Japan, 1995, pp. 39¨C43.

[42] R. Storn, K. Price, Differential evolution ¨C a simple and efficient heuristic for global optimization over continuous spaces, Journal of Global Optimization 11 (1997) 341¨C359

[43] Z.W. Geem, J.H. Kim, G.V. Loganathan, A new heuristic optimization algorithm: harmony search,Simulation 76 (2001) 60¨C68.

[44] A. Mucherino, O. Seref, A novel meta-heuristic approach for global optimization, in: Proceedings of the Conference on Data Mining, System Analysis and Optimization in Biomedicine, Gainesville, Florida, 2007, pp. 162¨C173.

[45] D. Karaboga, B. Basturk, A powerful and efficient algorithm for numerical function optimization: Artificial Bee Colony (ABC) Algorithm, Journal of Global Optimization 39 (3) (2007) 459¨C471.

[46] X.S. Yang, Firefly algorithm, in: Nature-Inspired Meta-heuristic Algorithms, Luniver Press, Frome, UK, 2008, pp. 79¨C90.

[47] H. Shah-Hosseini, The intelligent water drops algorithm: a nature-inspired swarm-based optimization algorithm, International Journal of Bio-inspired Computation (IJBIC) Vol. 1, No. 1/2, 2009, 71¨C79.

[48] X.S. Yang and S. Deb, Cuckoo search via L¨¦vy flights, in: Proceeings of World Congress on Nature &Biologically Inspired Computing (NaBIC), India. IEEE Publications, USA, 2009, pp. 210-214. X.S. Yang and S. Deb, Engineering optimization by cuckoo search. Int. J. Math. Modelling & Num. Optimization, Vol. 1, 2010, 330-343.

[49] X.S. Yang, A New Meta heuristic Bat-Inspired Algorithm, in: Nature Inspired Cooperative Strategies for Optimization (NICSO) (Eds. J. R. Gonzalez et al.), Studies in Computational Intelligence, Springer Berlin, 284, Springer, 2010, 65-74. X. S. Yang, Bat algorithm for multi-objective optimization, Int. J. Bio-Inspired Computation, 3 (5), 2011,267-274.

[50] A.E. Eiben, S.K. Smit, Parameter tuning for configuring and analyzing evolutionary algorithms, Swarm and Evolutionary Computation 1 (2011) 19¨C31.

[51] W. Liu, H. Chen, L. Chen, An ant colony optimization based algorithm for identifying gene regulatory elements,Computers in Biology and Medicine 43 (2013) 922¨C932.

[52] T. Rauscha, A. Thomasa, N.J. Campa, L.A. Cannon-Albright, J.C. Facelli, Aparallel genetic algorithm to discover patterns in genetic markers that indicate predisposition to multi factorial disease, Computers in Biology and Medicine 38 (2008) 826¨C836.]

[53] Sahin, Yusuf, Serol Bulkan, and Ekrem Duman. "A cost-sensitive decision tree approach for fraud detection." Expert Systems with Applications 40.15 (2013): 5916-5923.

[54] Wheeler, Richard, and Stuart Aitken. "Multiple algorithms for fraud detection." Knowledge-Based Systems 13.2 (2000): 93-99.

[55] Syeda, Mubeena, Yan-Qing Zhang, and Yi Pan. "Parallel granular neural networks for fast credit card fraud detection." Fuzzy Systems, 2002. FUZZ-IEEE'02. Proceedings of the 2002 IEEE International Conference on. Vol. 1. IEEE, 2002.

[56] Dharwa, Jyotindra N. Data Mining Techniques: Study, Analysis, Prevention & Detection For Financial Cyber Crime And Frauds. Diss. Hemchandracharya North Gujarat University, 2010. [56] Viaene et al., 2007; Phua et al., 2004

[57] A new intelligent classifier for breast cancer diagnosis based on a rough set and extreme learning machine: RS + ELM Yılmaz KAYA Department of Computer Engineering, Faculty of Engineering and Architecture, Siirt University, Siirt, Turkey

[58]R.K. Ahuja, O. Ergun, J.B. Orlin, A.P. Punnen, A survey of very large scale neighborhood search techniques, Discrete Applied Mathematics 123 (2002) 75–102.

[59] Kohavi, Ron, and Foster Provost. "Glossary of terms." Machine Learning 30.2-3 (1998): 271-274.

ÖZGEÇMİŞ

Adı Soyadı **:** Muyesaier AIHAITI

Doğum Yeri ve Tarihi **:**Çin 09.01.1987

Yabancı Dili **:**Çince

E-posta **:** muyassar1.9@gmail.com

Öğrenim Durumu

Derece	**Bölüm/Program**	**Üniversite/Lise**	**Mezuniyet Yılı**
Lise		Şangay QiBao Lisesi	2006
Üniversite	Bilgisayar Bilim ve Teknolojisi	Çin Huazhong Bilim ve Teknik Üniversitesi	2010