SULEYMAN DEMIREL UNIVERSITY
DEPARTMENT OF POST-GRADUATE EDUCATION

UDK 004.855.5

MERALIYEV MERARYSLAN MERRUSLANOVICH

Detection of diseases using machine learning algorithms

Specialty: "6M070400 – Computing systems and software"

Academic Degree: Master of Computer Sciences

"Admitted to defense":
Director of the department of
Post-graduate education
PhD. Mussabekov M.N.
«01» 06 2017 г.

Head of Department PhD. Alimanova M.O.

Scientific advisor PhD. Zhaparov M.K.

Kaskelen, 2017

ТҮЙІН

Компьютерлік технологияларды және сақтау мүмкіндіктері соңғы жетістіктер деректер мен ақпарат, мысалы, әлеуметтік желілер сияқты көптеген көздерден, онлайн деректер базасын, және денсаулық сақтау ақпараттық жүйелердің керемет сомасын өндірді. Қазіргі таңда, әлемнің көптеген елдері электронды денсаулық сақтау арқылы компьютерлік технологиялар мен коммуникация саласындағы жетістіктерді артықшылықтарын пайдалану арқылы науқастар мен адамдарға медициналық көмек жүзеге асыру жолын өзгертеді.

Бұл үлкен сома есептеу және машина оқыту құралдарын пайдалану арқылы деректерді білу және одан да көп пайдалы түрге баптауға болады. Ол деректер осы сома инженерлік сияқты сүт безі обырының ретінде қауіп төндіретін аурулардан кейбір жету мерзімін диагностикалау және болжау дәрігерлер көмектесе алады шешім қолдау үшін сараптамалық жүйелерді дамыту көмектесе алады деп саналады. шешім қолдау үшін Эксперттік жүйелер көп зерттеу үшін құнын, күту уақыты және еркін адам сарапшыларды (дәрігерлер) азайту, сондай-ақ, шаршағандар байланысты адам жасаған болуы мүмкін қателер мен қателіктер азайтуға болады. Алайда байланысты әкелуі мүмкін ерекшеліктері (атрибуттары) үлкен саны, және мүмкіндіктерін анықтау үшін іс-шаралар барысында, үшін мөлшері, денсаулық деректерді пайдалану процесі тиімді, мұндай жоғалған ерекшеліктері құндылықтарды проблема ретінде көптеген міндеттерді қамтиды, қарғыс дәлірек қорытындысы (дәлірек диагноз) үшін. Тиімді машина оқыту құралдары, мысалы, сүт безі қатерлі ісігінің сияқты ауруларды ерте анықтау бойынша көмек алады, және бұл диссертацияда ағымдағы жұмыс машиналық оқыту құралдар негізінде сүт безі қатерлі ісігін анықтаудың роман тәсілдерді тергеу назар аударады, және салу, жаңа әдістерін әзірлеуді және процесс хабарсыз кеткен ерекшеліктері құндылықтар, түрлі ерекшелігі іріктеу әдістері тергеу, және қалай диагностика процесіне оларды жұмысқа орналастыру.

Кілт сөздер: *Машина оқыту, сүт безі обыры, деректер тау-кен, алгоритм, статистика, дамыту, эволюциясы.*

АННОТАЦИЯ

Недавние достижения в области компьютерных технологий и средств хранения позволили получить невероятный объем данных и информации из многих источников, таких как социальные сети, онлайн-базы данных и информационные системы здравоохранения. В настоящее время многие страны мира меняют способ оказания медицинской помощи пациентам и людям, используя преимущества продвижения компьютерных технологий и коммуникаций посредством электронного здравоохранения.

Этот огромный объем данных может быть настроен на знание и более полезную форму данных с использованием вычислительных и машинных средств обучения. Считается, что разработка этого объема данных может помочь в разработке экспертных систем для поддержки принятия решений, которые могут помочь врачам в диагностике и прогнозировании некоторых изнурительных опасных для жизни заболеваний, таких как рак молочной железы. Экспертные системы поддержки принятия решений могут сократить затраты, время ожидания и свободных экспертов (врачей) для большего количества исследований, а также уменьшить ошибки и ошибки, которые могут быть сделаны людьми из-за усталости и усталости. Однако процесс эффективного использования данных о здоровье включает в себя множество проблем, таких как проблема недостающих значений признаков, проклятие размерности из-за большого количества функций (атрибутов) и ход действий для определения функций, которые могут привести к К более точным результатам (более точный диагноз). Эффективные инструменты машинного обучения могут помочь в раннем выявлении таких заболеваний, как рак молочной железы, и текущая работа в этом тезисе фокусируется на исследовании новых подходов к диагностике рака молочной железы на основе средств машинного обучения и включает разработку новых методов для создания и обработки недостающих функций Значения, исследовать различные методы выбора объектов и способы их использования в процессе диагностики.

Ключевые слова: *машинное обучение, рак грудной клетки, сбор и обработка данных, алгоритмы, статистика, развитие, эволюция.*

ANNOTATION

The recent advancements in computer technologies and storage capabilities have produced an incredible amount of data and information from many sources such as social networks, online databases, and health information systems. Nowadays, many countries around the world are changing the way of implementing health care to the patients and the people by utilising the benefits of advancements in computer technologies and communications through electronic health.

This huge amount of data can be tuned into knowledge and more useful form of data by using computing and machine learning tools. It is believed that engineering this amount of data can aid in developing expert systems for decision support that can assist physicians in diagnosing and predicting some debilitating life threatening diseases such as breast cancer. Expert systems for decision support can reduce cost, the waiting time and free human experts (physicians) for more research, as well as reduce the errors and mistakes that can be made by humans due to fatigue and tiredness. However, the process of utilising health data effectively, involves many challenges such as the problem of missing features values, the curse of dimensionality due to a large number of features (attributes), and the course of actions to determine the features, that can lead to more accurate results (more accurate diagnosis). Effective machine learning tools can assist in early detection of diseases such as breast cancer, and the current work in this thesis focuses on investigating novel approaches to diagnose breast cancer based on machine learning tools, and involves development of new techniques to construct and process missing features values, investigate different feature selection methods, and how to employ them into diagnosis process.

Keywords: *machine learning, breast cancer, data mining, algorithm, statistics, development, evolution.*

CONTENTS

INTRODUCTION

1.1 Overview

The advancement of information technology, software development, and system integration techniques have produced a new generation of complex computer systems. These systems have presented challenges to information technology researchers. Challenges include the compatibility between heterogeneous systems, security and privacy issues, systems management, sharing of data, and re using and benefiting from the existing resources and data. An example of complex systems is the healthcare system. Recently, there has been an increased interest to utilise the advancement of communication and data mining technologies in healthcare systems. Also, many countries are changing the way of conducting healthcare systems towards a global healthcare system across the country by setting healthcare standardization in communication and building the electronic healthcare records.

The aim of the current work is to investigate the aspects of utilising health data for the benefit of humans by using novel machine learning and data mining techniques. The idea is to propose an automated method for diagnosing diseases based on previous data and information. However, there are several problems associated with effectively utilising this previously acquired patient data, which can make any electronic healthcare system less efficient. These problems are: the problem of missing values and how to process them, the issue of huge features and attributes and how to select the most beneficial features and the problem of extracting accurate diagnostic markers that can predict the early onset of the disease, and the monitoring of different stages of the disease. This thesis will try to investigate these issues and propose methods for breast cancer disease as an example, based on the power of automated technologies and the previous evidence or data. The scope of the thesis is limited to the problems outlined above, and does not include other equally important issues like privacy and security. In this research, UCI Machine Learning Repository will be used as data sources for developing automatic data mining and machine learning techniques, so as to produce useful patterns and decision support logic for automatic computer aided diagnosis. For pursuing investigations for this project, the study used well-known datasets available publicly for research purposes. It is envisaged, that the novel algorithms and techniques developed and validated on this dataset can be extended to real clinical environments by integrating them into clinical computer aided diagnosis and decision support systems. This database is the trial dataset before integrating the proposed methods into the real clinical environments.

1.2 Research Motivation

In Kazakhstan and all over the world, people are suffering from limited medical resources and long waiting times to receive medical services. The increasing population of Kazakhstan, the ageing population, the modern lifestyle, the climate change, and the new diseases that come into view have presented challenges for the Australian health organisations and state governments to set procedures and plans to manage and cope with the available medical resources, infrastructure, and to deliver a decent healthcare services for residents despite the shortages in medical personnel and equipment. In addition, medical services are essential for all individuals and it is the nation's responsibility to develop and sustain the medical infrastructures and services for all residents and citizens. In addition to the shortages in medical personnel and technology, incidents of prescription errors have been increasingly causing minor to major problems for patients. For example, serious health problems may occur because of Adverse Drug Effects (ADE). ADE caused by mistaken prescription, errors in dosage, miscommunication between physicians and pharmacy, dispensing and administering of drugs, and inappropriate number of drug intake [1]. For example, a study [2] shows that ADE may rank as the sixth leading cause of death in the United States after heart diseases, cancer, stroke, pulmonary diseases, and roads accidents. Those problems may be avoided by a systematic information transfer between different healthcare providers (hospitals, medical centres, pharmacies, pathologies, etc.).

Breast cancer has become a common disease around the world. Yearly, millions of women suffer from this debilitating life threatening disease, making it the second common non-skin cancer after lung cancer, and the fifth cause of death among cancer diseases in the world [3]. Discovering the disease in its early stages may reduce the breast cancer tragedy. Computing technologies and machine learning tools can be used to assist physicians in diagnosing and predicting the disease so they can provide the necessary treatment and prevent the impact, including the possibility of death. More specifically, breast cancer cause about 22.9% of all cancers in women excluding skin cancers [4]. For example, breast cancer caused 458,503 deaths worldwide in 2008 [4]. Breast cancer is targeting women 100 times more than men, although men tend to have poorer outcomes due to delays in diagnosis [5]. Survival rates for breast cancer vary greatly depending on the cancer type, stage, treatment, and geographical location of the patient. For instance, survival rates in the Western world are high. However, in developing countries survival rates are much poorer. Therefore, this work provides a hope, that this research and the related future work makes some contributions that can help in a better diagnosis of breast cancer for men and women worldwide, especially for countries with poor health services.

1.3 Research Objectives

Computing and machine learning tools can significantly help in solving the health care problems by developing expert systems that can assist physicians in diagnosing and predicting diseases in early stages. These systems can reduce the cost, the waiting time, and free human experts for more research as well as reduce the errors and mistakes made by medical personnel [6]. Computer Aided Diagnosis (CAD) and medical expert systems and tools have become one of the foremost research areas in the field of medical diagnoses. The aim of CAD is to design an expert system that combines the human expertise and the technology intelligence to achieve more accurate diagnosis effectively [6]. CAD can be used to assist physicians in diagnosing and predicting diseases. Accordingly, physicians can provide a necessary treatment promptly to prevent loss, including the possibility of death.

The aims of this research work are:

- To utilize patient's histories, health information, and databases for discovering and diagnosing diseases, and provide decision support to medical professionals. The research is expected to establish some models that can assist physicians in diagnosing diseases and grouping patients in useful patterns based on different risk factors, and how machine learning techniques can identify such patterns. This can help in detecting early onset of the disease, identification of disease stages and treatment plans.
- To address an important issue related to missing values that can play an important role in determining the performance improvements achieved by data mining and machine learning algorithms.
- To work with large number of features and attributes in the dataset, and identify the significance of some features over others. Large number of features can lead to curse of dimensionality, and can render a machine learning algorithm or technique limited in terms of accuracy, precision and specificity

This work envisages that the outcomes of this research in terms of an integrated computer aided decision support and diagnosis system with a principled workflow of different algorithmic techniques for using machine learning based classification can enhance the accuracy with which benign and malignant forms of the disease can be identified.

1.4 Research Methodology

Knowledge discovery from the databases or data mining refers to extracting useful relationships and patterns from large databases. Due to the amount of data and to obtain useful outcomes, a systematic method must be

applied. It has become a fact that quality data will imply more accurate outcomes than dirty data. Dirty data is a common term in data mining that describe some unwanted data characteristics such as incompleteness, noisy, and inconsistency. In this research, the method proposed involves different data mining processes starting by appropriate data collection (in our case, online datasets), data selection, data pre-processing, applying learning based classifier methods, evaluation, and finally visualisation and evaluation of results in terms of tables and diagrams. Details of each stage are described in Chapter 3.

1.5 Thesis Overview

The thesis is organised into six chapters:

- Chapter 1: Introduction
- Chapter 2: Background Study
- Chapter 3: Research Methodology
- Chapter 4: Breast Cancer Diagnosis based on Machine Learning Algorithms.
- Chapter 5: Discussion and Future Work

This introductory chapter presents the problem description, motivation and objectives of this work, the contribution to the scientific knowledge, the methodology, and the road map of thesis. Chapter two provides a review of canonical machine learning and data mining techniques, features selection methods, and data mining algorithms in the field of healthcare. This work has combined them into one chapter because they are strongly related as a background study. Chapter three describes data mining methodology used in this work. Chapter four shows a new method for diagnosing breast cancer based on optimization of parameters of learning algorithms for achieving best accuracy rates. The idea is to obtain a hybrid approach that is based on selecting best parameters to achieve best performing learning algorithms. Finally, Chapter five presents some of the conclusions drawn from this work and scope for future work.

1.6 Chapter Summary

This Chapter presented an overview of the thesis, the motivation and the objectives of the proposed research, and described the need of current research to assist in solving the shortages in automated computer aided decision support technologies, the contribution of current research, and the thesis road map. The next chapter will present a background study about machine learning, data mining in healthcare, feature selection techniques, and missing features values.

BACKGROUND STUDY AND LITERATURE REVIEW

2.1 Overview

Machine learning (ML) can be interpreted as a group of topics that emphasizes on making and testing algorithms that can assist the process of classification, prediction, and pattern recognition, using computer models obtained from existing data (previous data). Machine learning can produce classifiers to be used on the available resources. In addition, machine learning does not involve much human interaction. The objective behind limited human involvement is that the use of automatic preprogrammed methods can reduce human biases. The process of proposing the algorithm and its functionality to classify objects or predict new cases are to be built on solid and reliable data [7].

2.2 Background Study

In general, machine learning can be defined as a scientific domain that aims to design and develop algorithms that allow computers to learn and behave to solve a real time problem based on previous data or under a certain instructions and rules. There are many presentations of machine learning; data mining is the most used application of machine learning [8]. Data mining is a science to discover knowledge from databases. The database contains a collection of instances (records or case). Each instance used by machine learning and data mining algorithms is formatted using same set of fields (features, attributes, inputs, or variables). When the instances contain the correct output (class label) then the learning process is called the supervised learning. On the other hand, the process of machine learning without knowing the class label of instances is called unsupervised learning. Clustering is a common unsupervised learning method (some clustering models are for both). The goal of clustering is to describe data. However, classification and regression are predictive methods. In the current research, my focus is on supervised machine learning [8].

2.3 Classification

Classification and regression are common models in supervised learning. The current research will concentrate on classification. However, it is useful to distinguish between them. Regression algorithms attempt to map input to domain values (can be real values). For Example, a regressor can forecast a certain goods sales by considering goods features. At the same time, classifiers can map the input space into pre-defined classes. For example, a classifier can predict a new case of patient whether benign (healthy) or malignant (suffer from a certain disease) [9]. Classification is the process of learning the target function

that maps between a set of features (inputs) and a predefined class labels (output). The input data for the classification is a set of instances. Each instance is a record of data in the form of (x, y) where x is the features set and y is the target variable (class label). Classification model is a tool that used to describe data (Descriptive Model) or a tool to predict the target variable for new instance (Predictive Model). Examples of classification models are decision tree, artificial neural network, Naïve Bayes, and nearest neighbour's classifier [10]. The general approach for solving classification problem is shown in Figure 1. The training data consists of instances whose class labels are known. The classification model can be built based on the training data. The model then can be evaluated and tested by using the testing data which contains records with missing class labels. The evaluation of model performance is based on the number of testing instances that are correctly forecasted [10]. The result of performing the model on the testing data produces the confusion matrix.

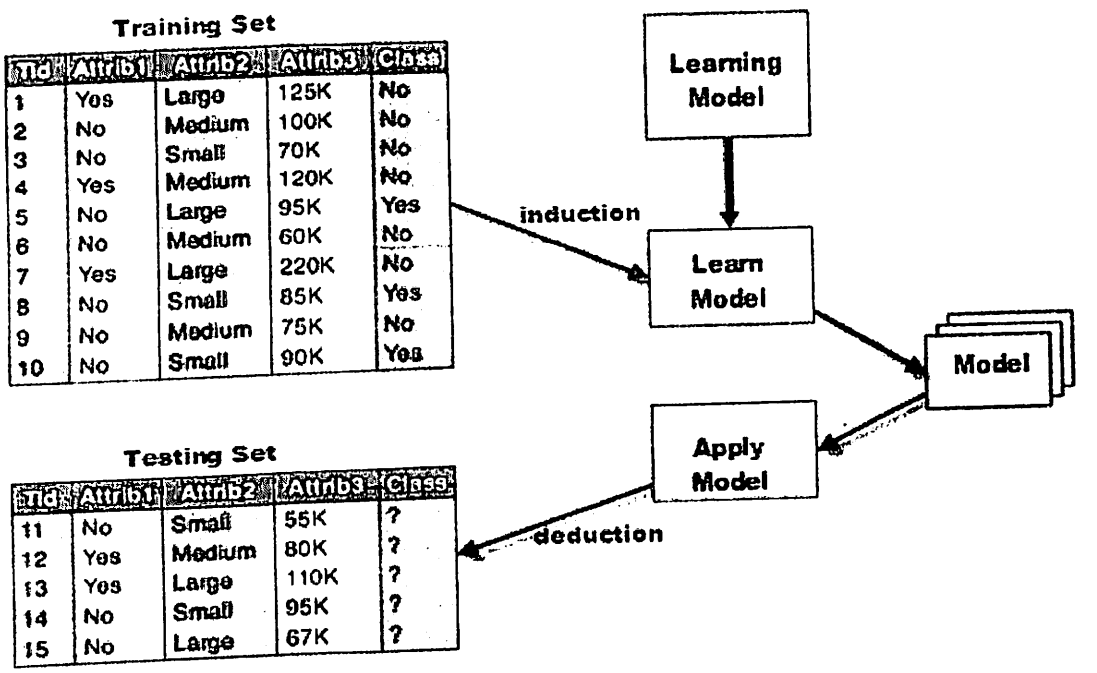

Figure 1 General approach for building a classification model

Suppose the goal is to classify some objects $i = 1, \ldots, n$ into k predefined classes, where k represent the number of classes. For example, if the aim of classification is to diagnose a patient whether or not suffering from breast cancer then the value of k will be 2 corresponding to either benign or malignant. Database (available data) can be organised as $n \times p$ matrix X, where xij represent the feature value j in the record i. Every row in the matrix X is represented by a vector xi with p features and a class label yi . The classifier can be denoted as $c(x)$. One method to evaluate the classifier is by calculating the error estimation based on the confusion matrix. To explain the error estimation, let us consider an example. Suppose the aim of a certain classifier

$c(x)$ is to train and test input vectors x into two possible classes benign and malignant. Suppose the result of classification of the classifier $c(x)$ on vectors x is as shown in the confusion matrix in Table 1.

Table 1 The confusion matrix for classifier c(x) on matrix X that contains 160 records.

		Predicted
true	*benign*	*malignant*
benign	*60*	*15*
malignant	*5*	*80*

The error rate (Er) of algorithm is the total number of incorrect classified samples divided by the total number of records in the matrix X. In the example above, Er = (15 + 5) / 160 = 0.125. On the other hand, the classification accuracy of the model can be calculated as Acc = 1 − Er = 0.875.

2.3.1 k-Nearest Neighbors algorithm

The k Nearest Neighbour algorithm (k-NN) is an instance based machine learning algorithm. k-NN is very simple to understand but works amazingly well [11]. The idea behind k-NN method for classifying objects is based on the closest training cases in the feature space. The k-NN finds the k closest instances to a predefined instance and decides its class label by identifying the most frequent class label among the training data that have the minimum distance between the query instance and training instances. The distance is determined by the distance metric. Preferably, the distance metric minimise the distance between similar instances and maximise the distance between different instances. The following pseudo-code shows an illustration for k-NN implementation [12]. Examples of approaches to define the distance are the Euclidean and Manhattan methods. Figure 2 shows an example of kNN.

procedure K-NN-Learner(TestingDataSet)
for each testing instance
{ find the k most nearest instances of
the training set according to a distance metric (Euclidean
distance or Manhattan distance)
Resulting Class = most frequent class
label of the k nearest instances}

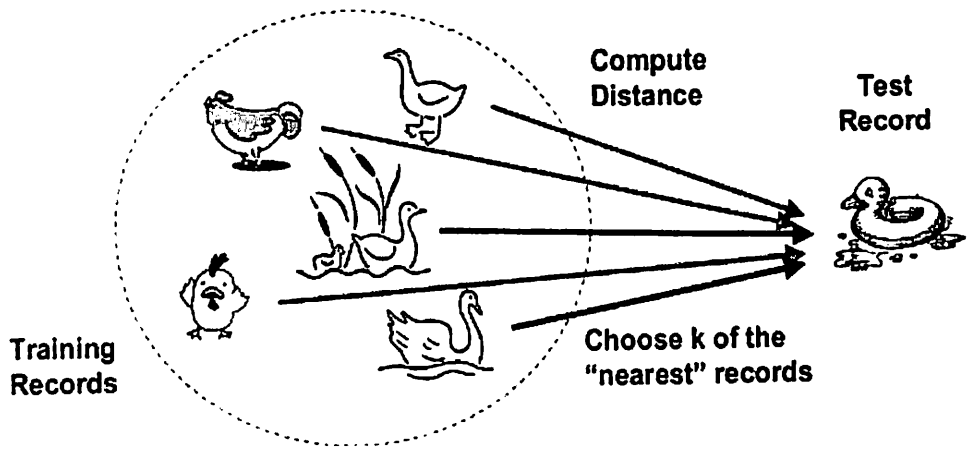

Figure 2 Example of k-NN [10]

Distance Functions

1. Euclidean Distance

The most regularly used metric to compute the distance between data points is the Euclidean distance. Euclidean distance is the square root of the sum between two points. For n-dimensional data, the distance is giving by the formula number 1, where d denote to distance, x and y are two different cases in the dataset, n is the total number of cases in the dataset [13].

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (1)$$

2. Manhattan distance

Manhattan distance is one of well-known measuring distance. Manhattan distance is calculated by summing the absolute value of the difference of data points. Manhattan distance is less costly to calculate in compare to Euclidean distance. The formula for Manhattan distance is giving by the formula number 2, where d denote to distance, x and y are two different cases in the dataset, n is the total number of cases in the dataset [14].

$$d(x, y) = \sum_{i=1}^{n} |xi - yi| \quad (2)$$

3. Minkowski distance

Minkowski function is a geometric distance between two points and uses a scaling factor, r. The main use is to find the similarity between objects. When r = 2 then it become the Euclidean distance. When r = 1 then it become the Manhattan distance. The distance is giving by the formula number 3, where d denote to distance, x and y are two different cases in the dataset, n is the total number of cases in the dataset [14].

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^r\right)^{\frac{1}{r}} \quad (3)$$

4. Chebyshev distance

Chebyshev distance function calculates the absolute differences between the coordinates of two points. Example of common application for using Chebyshev distance is Fuzzy C-means Clustering [15], where d denote to distance, x and y are two different cases in the dataset.

5. Canberra distance

Canberra distance is the sum of absolute values of the differences between ranks divided by their sum, thus it is a weighted version of the Manhattan distance function, where d denote to distance, x and y are two different cases in the dataset, n is the total number of cases in the dataset [16].

$$d(x,y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|} \quad (4)$$

Before using the k-NN, it is a good approach to list the advantages and disadvantages of k-NN to ensure that the k-NN is appropriate for the dataset and the learning process.

k-NN advantages:

- Is a very efficient pattern recognition method and can be easily carried out [17].
- k-NN simplicity of use [18].
- Strong against noisy data[18].
- Can be used for large and small datasets [18].
- Suitable for linear and nonlinear functions[18].

- The ability to add additional instances with no need to train the data set [19]
- Weight is used to measure features significance [19].
- missing values can be easily imputed [20].
- flexibility (nonparametric model except the value of k) [21].

k-NN disadvantages

- The need calculate the distance between the query instance and all other instances[20].
- The need for huge memory[20].
- Not useful for multidimensional dataset because of high error rate [20].
- The option of using many distance function which may lead to different accuracy level [20].

Figure 3 shows the features of learning for k-Nearest Neighbour algorithm [22]:

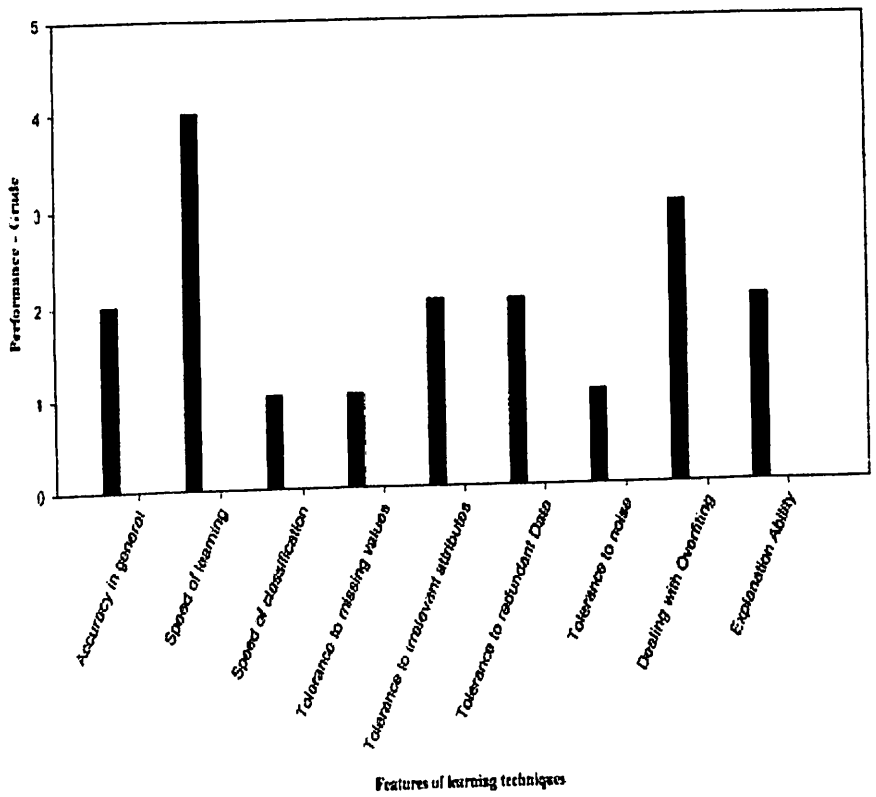

Figure 3 k-NN characteristics in regards to some learning features.

2.3.2 Artificial Neural Network

Artificial Neural Networks date back to nineteenth century when William James and Alexander Bain comprehended the ability of constructing a manmade system based on neural models [23]. In the middle of twentieth century, McCulloch and Pitt found the capability of learning a group of neurons, and Donald Hebb had developed tuning method that shows how neurons use enforcement to strengthen the connections from important input. In 1950s, based on Hebb methodology, Farley and Clark established the first artificial neural networks where neurons were randomly connected followed by development of perceptron for pattern classification by Frank Rosenblatt. Regrettably, the system was not able to perform complex classification and the research was stopped in 1960s [23]. During that era, the Adaptive linear Element (ADALINE) was developed by Widrow and Hoff that ultimately used to eliminate the echoes in telephone systems based on adaptive signal processing [24]. Despite the limited research on neural networks during 1970s, some researchers had developed self-organising neural model based on physiological studies on nerves systems called adaptive resonance theory (ART) [25]. In 1974, Paul Werbos had developed a learning rule based on error minimization approach in which the error is propagated in reverse by adjusting the weights using the Gradient descent model. Paul's technique is the back propagation error algorithm which is the most used artificial neural networks model that spread widely in mid 1980s by a group of researchers[23]. During 1980s and 1990s, computers have extended in speed about hundred times quicker since the beginning of the research, academic programs appeared, new courses were introduced, and funding becomes available. All the mentioned factors encouraged researchers to concentrate on neural networks application, development, and new approaches for prediction, forecasting, and diagnoses. For example, many studies[26, 27] demonstrates the potential applications of ANN for clinical decision making. Now a day, Major evolution in neural networks that attracts funding for further research in many fields such as the hybrid neural networks and how to combine it with other technologies. The artificial neuron is a computer simulated model stimulated from the natural neurons. Natural neurons receive signals from synapses located on the surface of the neuron. The neuron is starting to work and send a signal through the axon once the signal extent to a certain threshold. This signal then transfers through to other neurones and may get to the control unit (the brain) for a proper action. Figure 4 shows how the human neuron looks like [28].

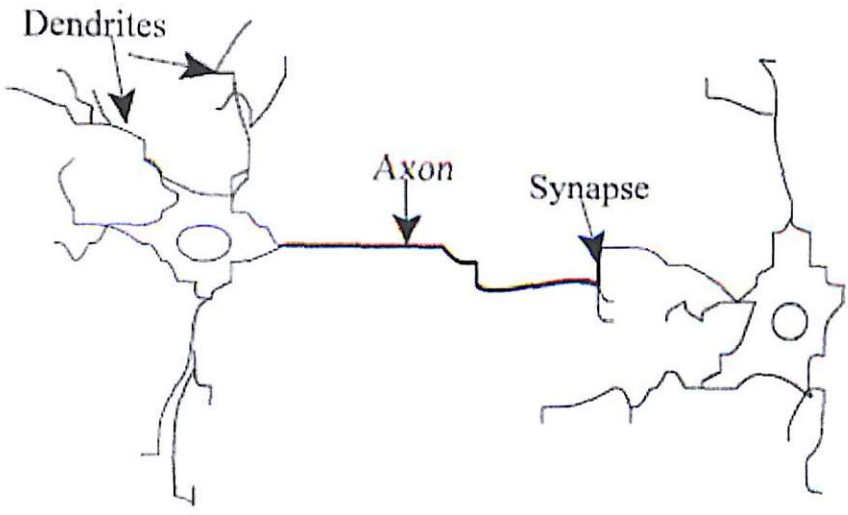

Figure 4 Human neuron [28]

The Artificial Neuron (AN) simulates the functionality of real neuron. AN have a set of inputs associated with weights. Inputs and weights are calculated by a mathematical equation to control when the AN activated. ANN is a combination of artificial neurons that process information [28]. Figure 5 shows a simple artificial neuron.

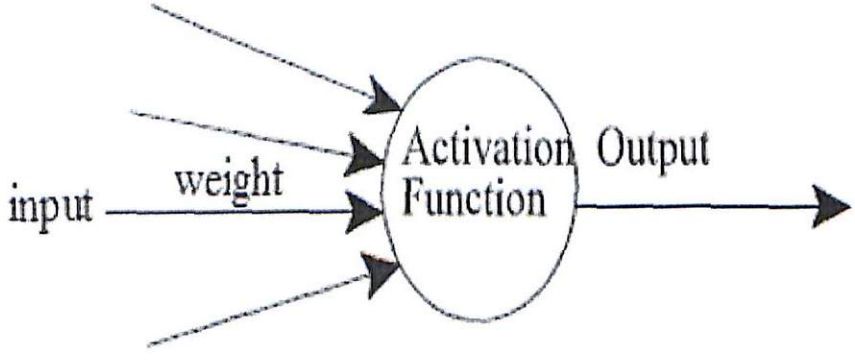

Figure 5 Artificial Neuron

In general, the artificial neuron operation is modelled by the data flow diagram as in Figure 6

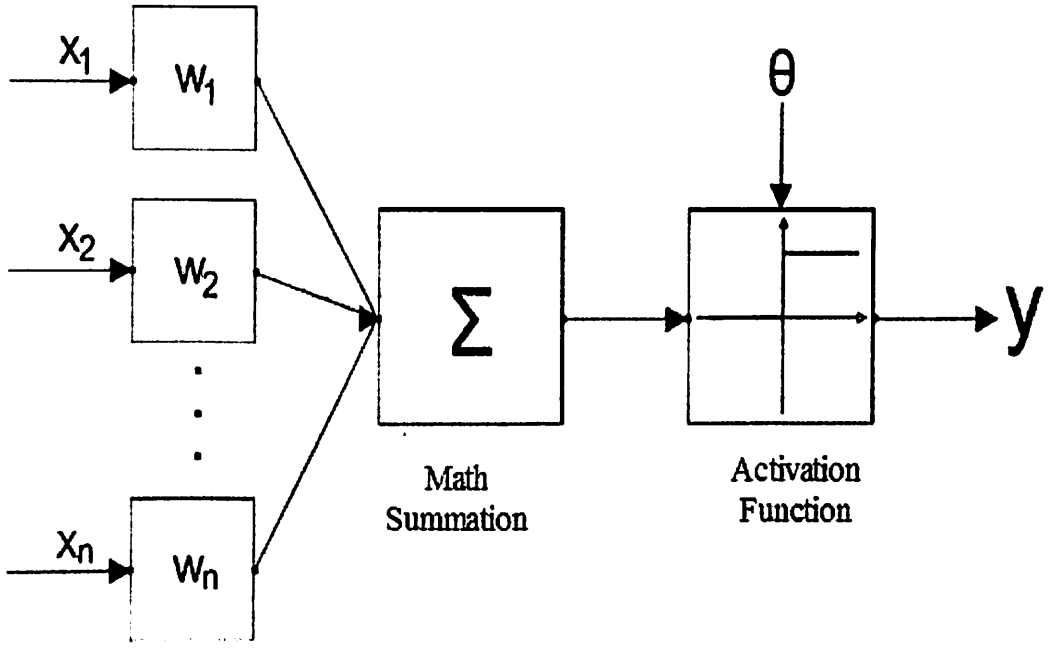

Figure 6 Simplified neuron operation

After briefly describing the artificial neuron, the ANN is described next. ANN is a set of connected artificial neurons. The most used ANN model is the Feed Forward Networks. Figure 7 shows a three layer topology of Feed Forward Networks. The outcome of ANN is subject to input and the value of weight [29].

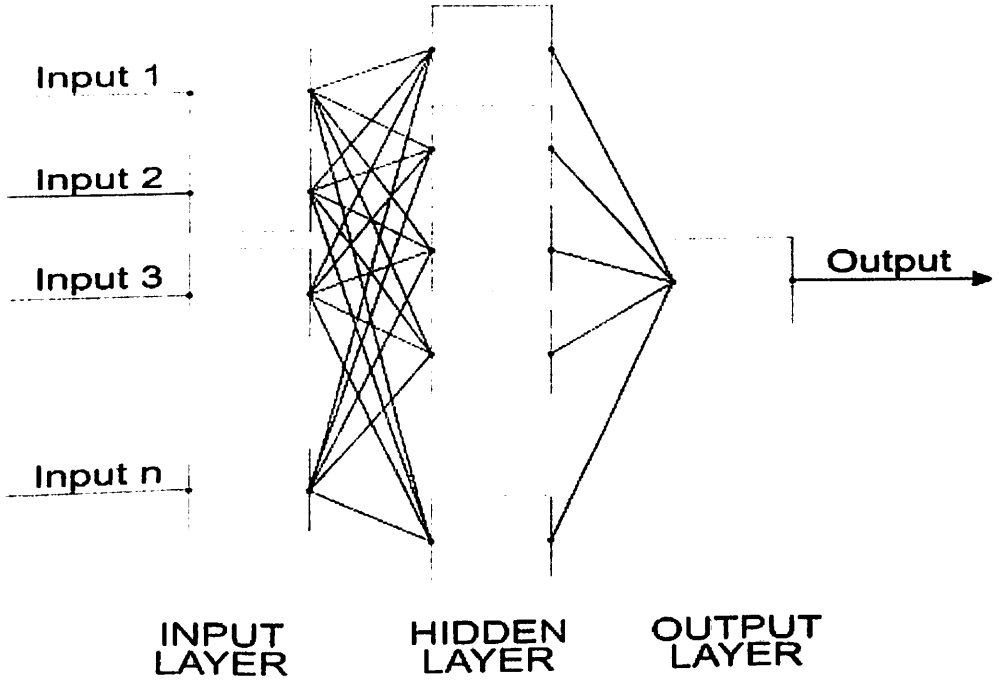

Figure 7 ANN architecture

Figure 8 shows the features of learning for Artificial Neural Network[22]:

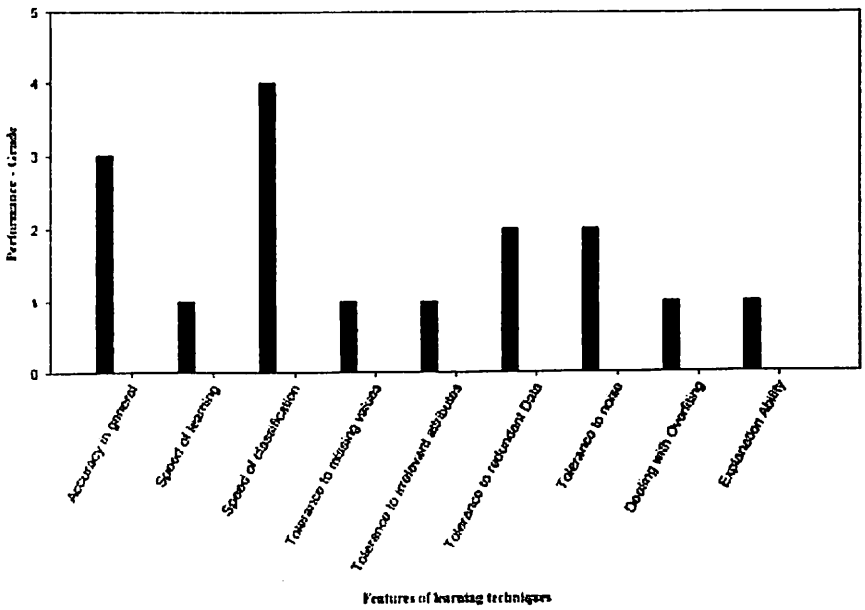

Figure 8 ANN characteristics in regards to some learning features

2.3.3 Decision Tree

Decision tree is a classification method which contains nodes, branches, and leafs. The first node on the tree or the top node is called the root node. Each node in the tree is connected with one or more nodes using branches, the last node in the tree that contains no outgoing branches is called leaf node. The leaf node indicate to termination or the outcome value [10] [30]. Figure 9 shows an example of a simple decision tree. Figure 9 shows how to solve a real time problem based on making questions and answers about attributes in the testing records. The terminology of such classification method is to keep asking question until conclusion is reached. The set of questions and answers could form a decision tree with set of nodes. The tree could contain three types of nodes [10]:

- Root node that has zero or more outgoing nodes and no incoming nodes, as well as, it contains the testing condition that separate records.
- Normal nodes, those nodes are internal nodes and each has one and only one incoming node and two or more outgoing edges. It also

contains the testing condition that separate records.

- Leaf nodes, those nodes hold the class labels, have no outgoing edges, and only one incoming edge.

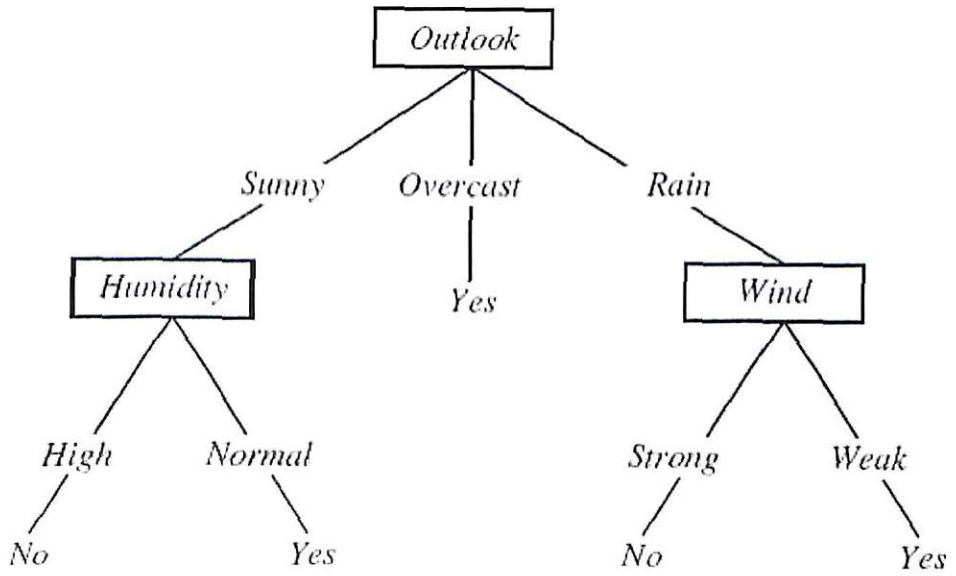

Figure 9 Simple Decision Tree

2.3.3.1 Building Decision Tree

Building an optimal decision tree is a difficult task because there are many decision trees that can be built from a set of attributes. In addition, constructing an optimal decision tree is computationally costly [31]. Generally speaking, the methods of constructing decision trees can be grouped into two types: top-down and bottom-up method with favourite to the first group according to the literature [31]. There are many types of top-down decision tree for example CART, C4.5, and ID3.

2.3.3.2 ID3

The ID3 is a top-down decision tree. The algorithm proposed by Quinlan in 1986. ID3 features the simplicity among other classifiers; it uses information gain to split instances and building the tree. ID3 is simple to perform. However, it doesn't handle missing values and no pruning procedures [9].

2.3.3.3 C4.5 decision tree

C4.5 is a better version of ID3 found by the founder of ID3 in 1993. The purpose of C4.5 is to overtake the disadvantages of the early version (ID3). The

process of splitting instances is done by gain ratio or the information gain. The algorithm runs as long as the number of instances to be split is more than a predefined threshold. Unlike ID3, the new version is capable to treat missing values and can handle numeric attributes [9].

2.3.3.4 CART

CART proposed by Breiman in 1984, it is shortcut for Classification and Regression Tree. CART has become a common method for constructing decision tree model due to the capability to deal with different data types, handling missing values, and the ability to produce rules which are understandable by human. CART may name binary tree because the tree is constructed by splitting a node into two child nodes with exactly two outgoing edges from the internal nodes. The splits are selected using the towing criteria (represent the quality of the connection between a parent and child decision nodes) [9] .

Figure 10 shows the features of learning for Artificial Neural Network.

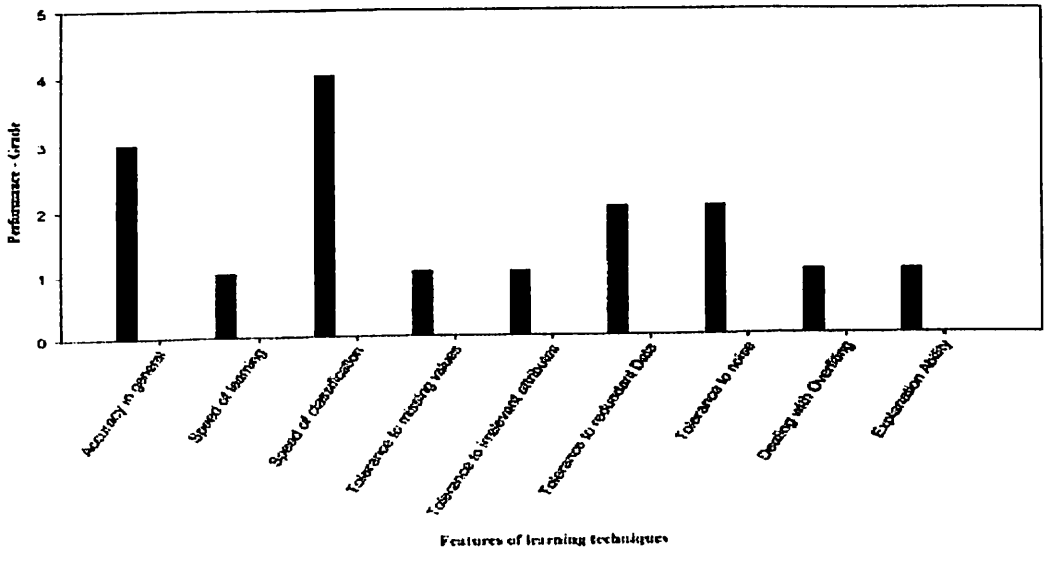

Figure 10 Decision Tree characteristics in regards to some learning features

2.3.4 Naïve Bayes Classifier

Naïve Bayes classifier in data mining is a mathematical classifier based on independency and probability (Bayes theorem). The Naïve Bayes classifier adopts the idea that the existence of a certain feature of an object is unrelated to the existence of any other feature, given the class variable. For example, an

animal may be considered to be a cat if it is hunt, play with kids, has four legs, has a head, and weight about 3 kilograms. Naïve Bayes algorithm treat all features independently and how they make a prediction of this animal is a cat, with no feature depends on others features values [38]. Naïve Bayes algorithm is significant classifiers; it is easy to construct, does not requires parameter estimation, easy to interpret. Therefore, Naïve Bayes can be performed by expert and inexpert data mining developers. Finally, Naïve Bayes generally performs well in comparison with other data mining methods [39]. The literature shows two types of Naïve Bayes, Multinomial model and Multivariate Bernoulli model. In these models, the classification is performed by the following Naïve rule [40, 41]:

$$P(c_j|x_i) = \frac{P(c_j).P(x_i|c_j)}{P(x_i)} \tag{5}$$

Where cj is the instance class label, xi is the test attribute, $P\ (cj \mid xi)$ is the posterior probability of the class label cj given the attribute xi , $P\ (cj)$ is the prior probability of class label cj , $P\ (xi \mid cj)$ is the likelihood which is the probability of attribute xigiven the class label ci . Assume that each attribute xi is conditional independent of every other attribute xj then the conditional distribution over the class variable c is:

$$P(c|x_i) = P(c)\prod\prod_{i=1}^{n} P(x_i|c) \tag{6}$$

The advantage of Bayesian classifier over other classification methods is the opportunity of considering the prior information about a given problem. The main disadvantages of Bayesian classifier are (1) the numerical attributes require discretization in most cases; (2) it is not suitable for large data sets which contain many attributes (time and space issues) [22]. Figure 11 shows the features of learning for Bayesian classifier.

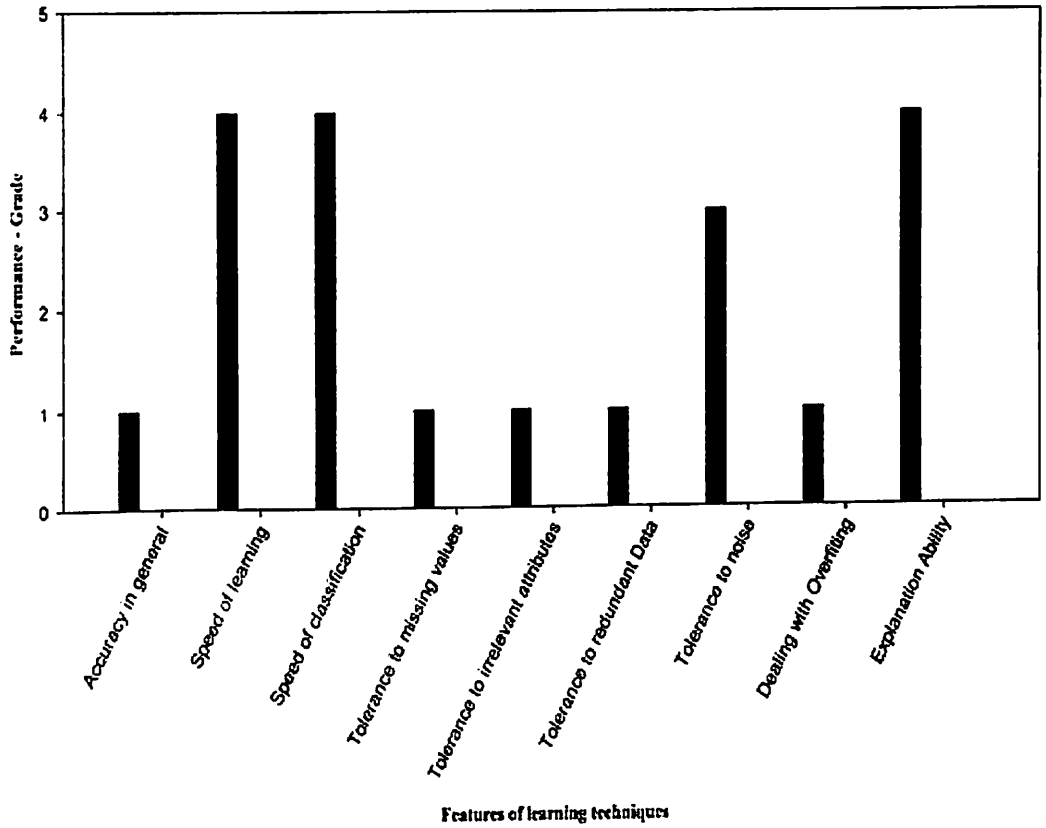

Figure 11 Bayesian classifier characteristics in regards to some learning features

2.4 Data Mining in Healthcare

Data mining methods have been commonly used in healthcare and diagnostic applications since of their ability to predict new cases. The main feature of data mining algorithms is the capability to learn from previous cases and it is ability to produce a prediction model. The resulting model is used to predict the new arrival cases, produce conclusion (knowledge) form a large amount of data, or to classify the data into useful patterns. There are enormous benefits for data mining applications in healthcare sector. In the main, data mining benefits in healthcare sector can be categorized in the following: treatment effectiveness, healthcare management, customer relationship management, fraud and abuse detection, and medical diagnosis (computer aided diagnosis) [36]. The focus of current research will be on computer aided diagnosis. However, other data mining application will be described briefly.

2.4.1 Treatment Effectiveness

Data mining applications can be used to assess the level of medical treatments. The aim is to develop a model that compares between the

symptoms, causes, and the treatment procedures. At the end, data mining model may produce an analysis for all treatment procedures to produce knowledge about the best practice procedure. For example, data mining model can find the best performing drug for a certain disease, grouping the side effects for a drug and how to reduce its risk on patients. Finally, data mining may play a role to determine the successful treatment and make it a standard approach among healthcare providers [36].

2.4.2 Healthcare Management

In regards to healthcare management, a good design of data mining can benefit to better classify and track some dangerous diseases and some patients who may infect others, design appropriate intervention, and reduce the number of hospital admissions and claims. For example, investigating readmission cases in a certain hospital and comparing its data with current scientific literature can make an efficient utilization of medical resources. As an another data mining example in healthcare management, data mining can be used to decrease patient length of stay (compare a certain case with previous cases, the length of stay should not exceed the average stay of previous cases, of course the final decision is to be taken by the doctor. However, data mining models can provide an approximate length of stay), provide information to physicians (data mining model may use as a second opinion for physicians or as a consultant), and to develop best practices [36, 37].

2.4.3 Customer Relationship Management

As in other industries, data mining applications can be used to improve customer satisfaction [38]. A study performed by Milley [39] stated that mining patient survey data can help to determine the waiting times expected for a patient before seen by a physician, how to improve customer service for patients, and assist healthcare providers with knowledge about patient expectations. Also, Hallick [40]suggested that Customer Relation Management in healthcare can help promote disease education, prevention, and wellness services.

2.4.4 Fraud and Abuse

Data mining applications can work to help governments (Medicare for instance) and healthcare insurance companies to control and reduce the fraud made by some healthcare providers. The idea is to establish a model that recognises strange claims made by customers (patients), physicians, pharmacies, labs, or other healthcare providers [36]. Once the model predicted fraud case, the fraud control department may investigate the case further before action is

taken against violating healthcare provider.

2.4.5 Computer Aided Diagnosis

Computer Aided Diagnosis (CAD) is an assistance method for diagnosing diseases and estimating the level for illness. CAD can be categorised as an expert system which utilizes human knowledge and experience to solve problems automatically or with a little support from human experts. The use of CAD systems is not replacing the role of medical personnel. However, CAD systems work as a second opinion, or for assisting in decision support in the diagnosis process. The final decision is to be made by the physicians [6].

Types of CAD Systems

There are two types of CAD systems: CADe for Computer Aided detection and CADx for computer aided diagnosis. CADe involve the use of computer analyses to indicate whether or not a certain case suffer from a certain disease (the target disease such breast cancer). CADx on the other hand, is to evaluate the process of detection. In both types, the final diagnosis and patient management is performed by the physician or the medical personnel [41].

CAD Systems Characteristics

Computer Aided Diagnosis (CAD) and medical expert systems and tools have become one of main areas of research in the field of medical diagnoses. The aim of CAD is to design an expert system that combines the human expertise and the technology intelligence to achieve more accurate diagnosis effectively. In addition, it can speed up the diagnoses; reduce the errors and mistakes made by human being, and free human expertise for further research. CAD can be used to assist physicians in diagnosing and predicting diseases so physicians can provide a necessary treatment promptly and prevent loss, including the possibility of death. In general, Computer Aid Diagnosis and expert systems have a number of attractive advantages [6, 42]:

- Fast response and reduce the cost. The cost of providing expert system per user is lower than providing human expert. In addition, some cases require immediate response, especially in emergencies. In such cases, diagnosis can be obtained by expert systems to produce an approximate and reliable result for the situation.
- Increase availability and reliability. Experts are available 24/7 on a suitable computer with internet connectivity for e-experts systems. Medical expert systems also increase the confidence about decisions made by physicians and may help to break arguments between human experts in case of different opinions.

- Steady, unemotional, and complete response at all the time.
- Human experts are not permanent. However, the expert systems last indefinitely.
- The knowledge in expert systems can be examined and corrected since it is explicitly known instead of being implicit in the mind of human experts mind.
- Justification and warranting. Medical expert systems can explain in more details the steps and reasons for taking the decision than the human being. Moreover, expert systems confirm that the knowledge has been correctly used.

2.4.6 Ethical, Legal, and Social Issues

About three-quarter billions of people from North America, Europe, Asia have their medical information collected in electronic form. Therefore, there must be a form of protection for human data. This work will discuss some ethical, legal, and social limitations on data collection and distribution that limit researchers and industries when utilizing human data to prevent the abuse of patients and the use of their data for commercial purposes. The main points of the ethical, legal, and social issues in mining medical data may be organized into the following categories[43]:

Data Ownership

Cios and Moore [43] discussed the term of data ownership and raised some questions that require legal professionals assistance to answer. In legal theory, ownership is determined by who is entitled to sell a particular item of property[44]. The problem of identifying the ownership of data is typically unresolved because human data and tissues are not supposed to be sold for any reason. Therefore, cannot apply the legal theory in regards to human data. At the same time, human medical data are available for data mining without prior consent in some cases. The question of patients' information ownership is still unsettled and further investigation by legal specialists and researchers is needed to solve this issue. Example of questions that need to be answered: (1) who own the data; the patients, the government, or healthcare providers? (2) Does the medical doctor own the data? (3) Do insurance providers own the data? (4) If insurance providers do not own their customers data, can they refuse to pay for the collection and storage of the data? (5) Who organize and mine the data? Hence, data ownership brought a debate about who own the data and who is legally allowed to give authorities for data usage in scientific field such as data mining.

Fear of Lawsuits

An important aspect of medical data mining is the fear of lawsuits against medical practitioners and health care providers. Some medical practitioners and health care providers are reluctant to hand over data to data miners. However, providing individuals data to data mining may lead to lawsuits. Therefore, health care providers and governments should work together to give patients the right to decide whether their data to be involved for research purposes or not. In addition, patients should choose the field of research involved and the location of their data storage. In my opinion, this could facilitate the process of mining data and may avoid the lawsuits. Moreover, this could reduce the efforts and time during mining process [43] [44].

Privacy and Security of Human Data

Medical data which obtained by healthcare providers and medical practitioners from individuals may contain private and confidential data. Individuals' data have to be handled with enormous care to protect people privacy and confidentiality. To meet these requirements, there are four forms of patient data identification [43]:

- *Anonymous data*: Individual identification is deleted during data collection. Therefore, no way to recover the patient identity in the future.
- *Anonymized data*: Individual identification is recorded initially during data collection and then removed. In this type of identification, there is a chance to reidentify the patient because patient information has been recorded at some stage.
- *De-identified data*: Individual identification is recorded initially during data collection, which is subsequently encoded or encrypted. There still some chances to identify the person using computer technology (The country privacy law and guidelines). This method could help researchers to remove duplicate records that related to same patient. However, there is still a technique to identify the patient in the future by decrypting the identification field.
- *Identified data*: Individual identification is recorded initially during data collection. This method requires receiving a written consent from patient to be identified. This method should adhere to the country privacy law and guidelines.

Expected Benefits

The World Wide Web and the internet are convenient ways to share and store data, accessible almost from everywhere, and can help researchers who may have legitimate reasons for access private information. For example, researchers who hold reasonable claims to mine data because the data is rare and they could not mount the financial and administrative resources to collect and mine private data. The use of individuals' data must be justified to the authorities. In addition, researchers who want to apply methods on data must show some expected benefits for the science or the society [43] [44].

Administrative Issues

Researchers and data miners are not the only people dealing with private data. Therefore, some countries including the United State did specify administrative guidelines for patient privacy. The guidelines include [43] [44]:

1. The establishment of security measures and policy to ensure privacy and security is in place in research centers and all institutions and organizations that hold and have access to people information.
2. There must be legal agreement between the healthcare providers and researchers (or institutions) that use patients' medical information. The agreement should force researchers and institutions to protect patients' data.
3. There must be up to date plans to protect patients' information against natural disasters including disasters plans and data backup.
4. There must an authorization and identification scheme for employees to limit access to authorized personnel only.
5. There must be an ongoing internal review of authorization and privileges procedures to ensure that the right person have an access to patients data.
6. There must be training sessions for employees in security and privacy issues. The training should be regular to cope with the technology advancements in regards to security and privacy.
7. There must be a daily update to security infrastructure including anti-virus and internet security software.

2.4.7 Challenges of Data Mining in Healthcare

Data mining and the advancement in computer technology can help the healthcare industry in many applications as mentioned earlier in this chapter. However, utilizing data mining in healthcare have some limitations. The first limitation is the type of data in healthcare databases. The types of data in healthcare database are heterogeneous. Some patients' examinations results are in numeric form, text form, and images. The process of mining such a mixed data types bring a challenge to developers. The source of data is different, such as laboratories, medical centers, physicians and more. Therefore, data collection and integration is time consuming. To overcome this problem, some authors recommended that a data warehouse to be built before data mining process. However, this can be time consuming and may not reliable for previous data[36]. Secondly, the nature of data is to be unorganized (not processed data), this include missing features values, corrupted files, inconsistent with patient history or family history. The problem of missing features values can be solved by constructing or estimating the missing features values. However, the mining process will be more efficient with complete data. Thirdly, mining data that contains large number of cases and attributes may led to patterns that are random and not real [45]. For this reason, not all significant patterns are necessary to be useful. Fourthly, mining healthcare data requires expertise's that combine the knowledge in data mining and knowledge discovery as well as knowledge in medical science. Since it is pretty uncommon to find expert people who have knowledge in the domain area (data mining and medical science), mining healthcare data may requires collaboration between expertise in data mining and expertise in medical science[36]. Finally, resources for developing data mining application should be allocated by healthcare organizations including budget, time, and efforts, and expertise. Data mining developments can produce a negative outcome for some causes, such as lack of management, limited support, and lack of cooperation between mining and medical expertise[36].

2.5 Related Work on Breast Cancer Diagnosis

In this section some of the related prior work on data mining methods for breast cancer diagnosis is discussed. Song et al. [46] presented a new approach for automatic breast cancer diagnosis based on artificial intelligence technology. They focused to obtain a hybrid system for diagnosing new breast cancer cases in collaboration between Genetic Algorithm (GA) and Fuzzy Neural Network. They also showed that inputs reduction (features selections) can be used for many other problems which have high complexity and strong non-linearity with huge data to be analysed. Arulampalam and Bouzerdoum [47] proposed a method for diagnosing breast cancer and called Shunting Inhibitory Artificial

Neural Networks (SIANNs). SIANN is a neural network stimulated by human biological networks in which the neurons interact among each other's via a nonlinear mechanism called shunting inhibition. The feed forward SIANNs have been applied to several medical diagnosis problems and the results were more favourable than those obtained using Multilayer Perceptions (MLPs). In addition, a reduction in the number of inputs was investigated. Setiono [48]proposed a method to extract classification rules from trained neural networks and discussed its application to breast cancer diagnosis. He also explained how the pre-processing of data set can improve the accuracy of the neural network and the accuracy of the rules because some rules may be extracted from human experience, and may be erroneous. The data pre-processing involves the selection of significant attributes and the elimination of records with missed attribute values from Wisconsin Breast Cancer Diagnosis dataset. The rules generated by Setiono's method were more brief and accurate than those generated by other methods mentioned in the literature. Meesad and Yen [49] proposed a hybrid Intelligent System (HIS) which integrates the Incremental Learning Fuzzy Network (ILFN) with the linguistic knowledge representations. The linguistic rules have been determined based on knowledge embedded in the trained ILFN or been extracted from real experts. In addition, the method also utilized Genetic Algorithm (GA) to reduce the number of the linguistic rules that sustain high accuracy and consistency. After the system being completely constructed, it can incrementally learn new information in both numerical and linguistic forms. The proposed method has been evaluated using Wisconsin Breast Cancer Dataset (WBC) data set. The results have shown that the proposed HIS perform better than some well-known methods.

2.6 Feature Selection Techniques

Nowadays, the capability of collecting and generating data is more than before. Contributing factors include the steady progress of computer hardware technology for storing data and the computerization of business, scientific, and government transactions. In addition, the use of the internet as a wide information system has flooded us with incredible amount of data and information.

Data mining has attracted a big attention to information system researcher in the recent years due to the wide availability of big amount of data and the need for tuning such data into knowledge and useful patterns. The gained knowledge and patterns can be used in many fields such as marketing, business analysis, and health information systems [59]. The quality of data, the large amount of data, the existence of low quality data, unreliable, redundant and noisy artefacts and outliers; all mentioned factors do affect the process of

extracting knowledge and useful patterns, and then knowledge discovery during training phase is more difficult. Experts in machine learning and data mining stated that the classification performance (such as accuracy) decrease when the dataset contains many features that are not relevant to the process of prediction. For example, performing decision tree C4.5 on Monk1 problem produced error rate of 24.3% because of three irrelevant features. However, the error rate dropped to 11.1% by ignoring the irrelevant features [60]. The k nearest neighbour algorithm degrade to irrelevant attributes and training set size for a certain accuracy level grows exponentially with the number of irrelevant attributes [61]. Therefore, researchers have felt the necessity for producing more reliable data from large amount of records such as using feature selection methods. Feature selection or attribute subset combination is the process of identifying and utilizing the most relevant attributes and removing as many redundant and irrelevant attributes as possible [62, 63].

Variables, features, inputs, or attributes selection have become the focus of much research in many areas where the number of cases and attributes are huge. The purpose of feature selection is to obtain less number of features than the original number of features in a certain dataset to: (1)enhance the prediction accuracy,(2)obtaining a quicker classifier, (3) ignore the less important or irrelevant features, (4) improve data quality, (5) avoid over fitting (6) and help solve the problem of incredible amount of available data and how to utilise it effectively [64]. The literature shows that feature selection techniques can be divided upon the induction algorithm and how it works with feature selection search. According to that, feature selection techniques can be divided into three types: filter methods, wrapper methods, and embedded methods [65].

2.6.1 Wrapper Feature Selection Technique

The wrapper approach was proposed by Kohavi and Paeger in 1994 in Stanford university AI lab [66]. In wrapper methods, the feature selection algorithm located as a wrapper around the learning algorithm. The process starts with a search for relevant subset of attributes by using the learning algorithm. The learning algorithm itself is used to evaluate the feature subset which obtained by the search. Figure 12 illustrates how the wrapper approach performs on the training set and the evaluation process. The learning algorithm is treated as a black box with no modification to the learning algorithm itself. The learning algorithm assesses the subsets of features obtained by the search method. The learning algorithm obtains a hypothesis about the quality and the relevance of a certain feature subset. Features subset with the highest estimated value is chosen as the final set on which to run the learning algorithm. The final step is to evaluate the model on new dataset (not used by the search) to ensure the independency between the training process and the testing process. The

result is an estimated accuracy by using the highly relevant features subset on the desired learning algorithm [67].

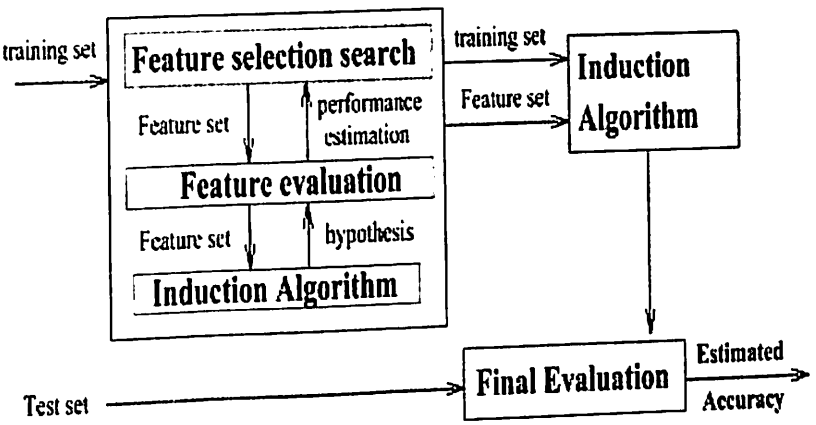

Figure 12 The Wrapper approach for features subset selection [67]

Table 2 shows the main advantages and disadvantages of using wrapper as a feature selection method, as well as, examples of existence methods that utilize the wrapper approach [65].

Table 2 Examples, advantages, and disadvantages of wrapper feature selection [65]

Advantages	disadvantages	Examples
• Simple to use and easy to implement • Interact with learning classifier • Models feature dependencies	• The risk of over fitting • Computationally intensive	• Sequential Forward Selection • Sequential backward elimination

2.6.2 Filters Feature Selection Techniques

Filter techniques examine the significance of features by investigating the real characteristics of the data. In most cases feature rank is calculated, and low ranking features are ignored during the learning process. Afterwards, the high ranking subset of features is used as training set to the classification algorithm [65]. The main difference of filter in compare with wrapper is that filter ignores the learning algorithm during features subset search. Figure 13 shows the filter approach; it shows that features subset extraction is totally independent from the learning classifier.

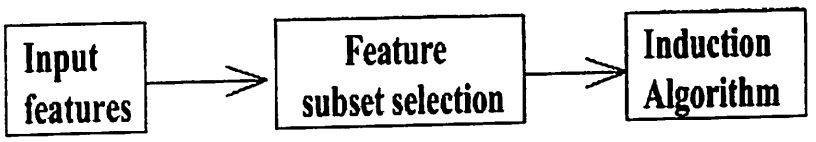

Figure 13 The filter approach [49]

Some advantages of filter techniques include that they are able to be performed on large databases that contain large number of attributes and cases, simple computation, fast in comparison to wrapper and embedded methods, and they are independent of the classification algorithm. The aim behind the independency between filters and learning classifier is that feature selection needs to be performed only once and then different classifiers can be used to evaluate the subset. On the other hand, the independency between filter methods and learning algorithms may cause low level of classification accuracy [65].

Table 3 summarise the main advantages and challenges of filter methods and some examples of popular filter methods.

Table 3 Examples, advantages, and disadvantages of filter feature selection [65]

Advantages	disadvantages	Examples
• Relatively fast • Scalable • Independent of classifier	• Ignores feature dependencies • Ignores interaction with the classifier	• Correlation-based feature selection (CFS) • Relief

2.6.3 Embedded Feature Selection Techniques

Embedded Methods (EM) vary from other feature selection methods in how classification methods and feature selection cooperate. In filter methods, there is no corporation between learning classifiers and feature selection. In wrapper methods, the learning classifier are used to measure the quality of subsets of features without intervenes in the structure of the classification. In contrast to filter and wrapper approaches, the embedded feature selection methods and learning process cannot be taken apart [68]. The process of finding the optimal subset of features is combined into the classifier construction. EM computation cost is less than wrapper methods and the fact that there is interaction between the classifier and EM. [65].

Table 4 shows some advantages and disadvantages of using such a method along of examples.

Table 4 Examples, advantages, and disadvantages of embedded feature selection [65]

Advantages	disadvantages	Examples
• Interacts with the classifier • Better computational complexity than wrapper	• Classifier dependent selection	• Decision tree • Weighted naive Bayes

2.4.6 Feature Selection Techniques Used

Information Gain

The information gain method was proposed to approximate quality of each attribute using the entropy by estimating the difference between the prior entropy and the post entropy [69]. This is one of the simplest attribute ranking methods and is often used in text categorization. If x is an attribute and c is the class, the following equation gives the entropy of the class before observing the attribute:

$$H(x) = -\sum_{x} P(x) \log_2 P(x) \tag{7}$$

Where $P(c)$ is the probability function of variable c. The conditional entropy of c given x (post entropy) is given by:

$$H(c|x) = -\sum_{x} P(x) \sum_{c} P(c|x) \log_2 P(c|x) \tag{8}$$

The information gain (the difference between prior entropy and postal entropy) is given by the following equations:

$$H(c,x) = H(c) - H(c|x) \tag{9}$$

$$H(c,x) = -\sum_{c} P(c) \log_2 P(c) - \sum_{x} \left(-P(x) \sum_{c} P(c|x) \log_2 P(c|x) \right) \tag{10}$$

Correlation based selection (CFS)

CFS is among the simplest feature selections methods. CFS ranks features and discover the merit of features or subset of features according to a correlation between features. The main aim of CFS is to find a subset of features space that highly correlated with the class label. CFS usually removes

uncorrelated features and redundancy. CFS's feature subset evaluation function is shown as follows [70].

$$Merit_s = \frac{kr_{cf}}{\sqrt{K + (K+1)r_{ff}}} \tag{11}$$

Where$Merits$ is the worth of feature subset s that contain k features, rcf is the average feature correlation to the class, and rff is the average feature to feature correlation. In order to apply this equation to calculate approximately the correlation between features, CFS uses a modified information gain method called symmetrical uncertainty to compensate the information gain bias for attributes with more values as follows [71].

$$SU = \frac{H(x_i) + H(x_j) - H(x_i, x_j)}{H(x_i) + H(x_j)} \tag{12}$$

Relief

Relief is ranked among the well-known feature selection techniques and [72]. Its aim is to rank the features quality giving their ability to predict instances of different classes. Relief uses instance based learning (lazy learning such as k-Nearest Neighbour) to assign a grade to each feature. Features are ranked by weight and those that exceed a threshold -determined by the user- are selected to form the promising subset. For each instance, the closest neighbour instance of the same class and the closest instance of a different class are selected. The following equation compute the average of distance between the nearest match and nearest miss [72].

$$W_x = W_x - \frac{diff\,(x,r,h)^2}{m} + \frac{diff\,(x,r,h')^2}{m} \tag{13}$$

Where Wx is the grade for the attribute x, r is a random sample instance, h is the nearest hit, m′ is the nearest miss, and m is the number of samples.

Principle Components Analysis (PCA)

The purpose of Principal Components Analysis (PCA) is to decrease the dataset dimension that contains a large number of correlated attributes by transforming the original attributes space to a new space in which attributes are uncorrelated. The algorithm then ranks the variation between the original dataset and the new one. Transformed attributes with most variations are kept; meanwhile discard the rest of attributes. It is also important to mention that

PCA is valid for unsupervised data sets because it does not take into account the class label [73].

Consistency based Subset Evaluation (CSE)

Consistency based Subset Evaluation (CSE) adopts the class consistency rate as the evaluation measure. The idea is to obtain a set of attributes that divide the original dataset into subsets that contain one class majority [62]. One of well-known consistency based feature selection is consistency metric [74] proposed by Liu and Setiono:

$$\text{Consistency}_s = 1 - \frac{\sum_{j=0}^{k} |D_j| - |M_j|}{N} \tag{14}$$

where s is feature subset, k is the number of features in s, Dj is the number of occurrences of the jth attribute value combination, Mj is the cardinality of the majority class for the jth attribute value, and N is the number of features in the original data set. For continuous values, Chi2 can be used [75]. Chi2 automatically discretize the continuous features values and removes irrelevant continuous attributes.

2.6.5 Related Work on Feature Selection Techniques

Numerous feature selection methods have been broadly used in different domains. Xie and others [76] have proposed a hybrid features selections algorithms to build an efficient diagnostic models based on a new accuracy criterion, generalized F-score (GF) and SVM. The hybrid algorithms adopt Sequential Forward Search (SFS), Sequential Forward Floating Search (SFFS), and Sequential Backward Floating Search (SBFS), respectively, with SVM to accomplish hybrid features selections with the new accuracy criterion to guide the procedure. They were called as modified GFSFS, GFSFFS and GFSBFS, respectively. The mentioned hybrid methods combine the advantages of filters and wrappers to select the optimal feature subset from the original dataset to build the efficient classifiers. Experimental results showed that the proposed hybrid methods construct efficient diagnosis classifiers with high average accuracy when compared with traditional algorithms.

Liao and others [77] proposed a hybrid features selections method along with k-NN and SVM. They aimed to identify the most significant genes that demonstrate the highest capabilities of discrimination between the classes of samples. They first used filter method to rank the genes in terms of their expression difference, and then a clustering method based on k-NN principles for clustering gene expression data. A support vector machine is applied to

validate the classification performance of candidate genes. Their experimental results demonstrated the effectiveness of their method in addressing the problem.

Vijayasankari and Ramar [78]also proposed a novel hybrid features selections method to select relevant features and cast away irrelevant and redundant features from the original dataset using C4.5 and Naïve Bayes classifier. The efficiency and effectiveness of the proposed method was demonstrated through extensive comparisons with other methods using real world data of high dimensionality. Experimental results on datasets revealed that the proposed algorithm increases the classifier accuracy with less error rate.

Hall and Holmes [62] presented a benchmark comparison of several attribute selection methods for supervised classification. Attributes selections is achieved by cross-validating the attribute rankings with respect to a classification learner C4.5 and Naïve Bayes. The results conclude that features selections methods can enhance the performance of some learning algorithms. The findings also include that Correlation based feature selection method has produced the best result among six different feature selections methods.

Saeys, et al [65] reviewed the importance of feature selection approach in a set of well-known bioinformatics applications. They focused into two main issues: the large input dimensionality, and the small sample sizes. The authors found that features selections methods could help researchers solve the mentioned issues. They also believed that features selections application will become fundamental in dealing with the high dimensional applications.

The literature also showed two categories of feature selection- that are wrapper and filter. The wrapper evaluates and select attributes based on accuracy estimated by the target learning algorithm. Using a certain learning algorithm, wrapper basically searches the feature space by omitting some features and testing the impact of feature omission on the prediction metrics. The feature that make significant difference in learning process implies it does matter and should be considered as a high quality feature. On the other hand, filter uses the general characteristics of data itself and work separately from the learning algorithm. Precisely, filter uses the statistical correlation between a set of features and the target feature. The amount of correlation between features and the target variable determine the importance of target variable [79]. A further category is by sorting attributes using algorithms that rank a features or set of features in which attributes are ranked in regards to their improvement to a subset of attributes [59].

2.7 Missing Features Values

Missing features values are common in many medical databases for different reasons such as some features values are not specified because they are not available. For example, diagnosing patients without blood test result. Another reason for missing attributes values is that the attribute values might be forgotten or mistakenly erased or not filled. Moreover, some interviewers decline to respond for some private information such as the income or the age [80]

2.7.1 Types of Missing Values

Donald Rubin classified the missing features values from the literature into three types: missing completely at random, missing at random, and missing not at random [81].

Missing Completely At Random

Missing Completely At Random (MCAR) is a term that describes how the missingness occurred, in which the probability that a feature value is missing is unrelated to the feature value or to the value of any other features in the dataset. For example, the data may be missing because equipment malfunctioned, the weather was terrible and could not record the observation for a certain day, people got sick, or the data were not entered correctly [82].

Notice that, the main concern is the value of feature not the missingness itself. For instance, person who refused to mention the personal income is also likely to refuse to mention the family income, the data obtained is still to be considered MCAR as long as the reasons have no relation to the income value itself [82].

Missing At Random

Missing At Random(MAR) is the case when the existence of missing feature value does not depend on the feature value itself and may depend on other features values in the dataset. For example, depressed person is more likely not to report income which may lead to find that not reporting income may due to depression [82].

Missing Not At Random

Missing Not At Random (MNAR) is the case when the missing feature value is not missing at random or completely at random. For example, if a person suffers depression and a person who suffer depression is more likely not report his mental status, then the data are not missing at random. Respectively,

if a person refuses to tell the age, then the missing data are not random [82]. In data mining and machine learning applications, missing features values that matters but still missing creates a challenge for researchers. Handling unknown attributes values with the most appropriate values is a common concern in data mining and knowledge discovery. The process of constructing missing values is a vital process in most supervised and unsupervised data mining researches because it may affect the quality of learning and the performance of classification algorithms [83]. In general, the performance of classification accuracy is particularly affected by the presence of missing feature values because most of learning classifiers such as neural networks do not consider the probability of having missing features values and cannot not deal with it automatically [83].

2.7.2 Handling missing data

As mentioned earlier, the process of handling missing feature values is vital process in most supervised and unsupervised data mining researches because it may affect the quality data itself, which may affect the classifier performance. The literature showed many attempts to treat missing features values. The most popular methods for dealing with missingness are omitting instances, imputation, and expectation maximisation. All of these methods can be applied in conjunction with any classifier that operates on complete data [83].

Omitting Instance

In omitting instances method, any record of data that contains missing features values is deleted from the data set. After omitting instances that contain missing features values, the classification process run on the remaining instances. The main disadvantage of this method is discarding important information in some cases. Therefore, it is not a common method. However, it could be used if there is a small amount of missing data [83].

Features Imputation

Features imputation is a well-known method for constructing missing features values in the datasets for learning purposes. The imputation method can be divided into two major types: single imputation and multiple imputation [83].

In single imputation, the missing features values are substituted by the correspondence features values according to certain rules such as the features values means, mode, median, or learning algorithm. For example, the mean imputation calculate the mean of feature f in the dataset that contain values. The mean value for feature f is then used to fill the features f that has missing values.

Another example is the regression imputation. Regression imputation is a method for dealing with missing features values by building regression models that construct missing features values based on observed features (features that contain values). The regression models are used to predict the values of missing attributes [83].

The scenario for constructing missing features values in multiple imputations is similar to the scenario for single imputation. However, the multiple imputation use more than one value to fill missing features values in the dataset, such as mean of observed feature values, the mode of observed feature values, and regression method.

The multiple imputations approach drawbacks include the computational cost is higher than in single imputation. However, the classification performance (accuracy) is higher than single imputation [83].

Expectation Maximization

The two most important methods to deal with missing features values in datasets in the recent literature are expectation maximization and multiple imputation [781]. Expectation maximization is one of the most effective methods for handling missing data [84]. To demonstrate expectation maximization, consider the data as shown in table 5. Missing features values are depression, age, and height [85].

Table 5 Extract of data to demonstrate Expectation Maximization [85]

ID	depression	age	height	wage
1	5	32		32, 010
2		17	173	31, 600
3	7		169	48, 020
4	5	24	186	17, 400
.	.	.	.	.
.	.	.	.	.
100	4	45	201	7, 800

To perform expectation maximization, firstly, the mean, variance, and covariance are estimated from instances whose data is complete such as row number 4 in table 5. In particular, expectation maximization will calculate the following values as shown in table 6:

- The mean of depression, age, height, and weight is 4.71, 37.50, 183.21, and 45504.43 respectively.
- The variance of depression, age, height, and weight is 3.55, 9.43, 194.43, and 14403.12 respectively and appears in the diagonals

Other cells are the values of covariance between each pair of variables

Table 6 The calculations of mean, variance, and covariance for the features depression, age, height, and weight.

	depression	age	height	wage
Depression	3.55			
Age	7.42	9.43		
Height	184.42	1643.32	194.43	
Wage	43042.345	143254.43	14425.54	14403.12
Mean	4.71	37.50	183.21	45504.43

Secondly, Expectation maximization uses maximum likelihood procedures to estimate regression equations calculate the relationships between variables. For example, maximum likelihood algorithm may produce the following equations:

- Depression = -15.3 + .01 x age + .004 x height + .0005 x wage
- Age = 7.3 + .34 x depression + .002 x height + .0003 x wage
- Height = 19.2 + .53 x depression + .021 x age + .0004 x wage
- Wage = 7.3 + .44 x depression + .031 x age + .0021 x height

The purpose of maximum likelihood is to ensure these equations predict the means, variances, and covariance more accurately than any other equations [81, 84] .Thirdly, these equations can be used to estimate the missing values. The procedure of estimating the missing feature values is shown below:

- Consider the equation Depression = -15.3 + .01 x age + .004 x height + .0005 x wage.
- This equation can then be used to estimate the depression for individuals who did not provide his/her information.
- For the second case, 17, 173, and 31600 would be substituted into this equation.
- Depression for this person would be 1.362.

For other missing features values, the same procedure is used after considering the right equation. The constructed missing features values are shown in bold in table 7.

Table 7 The final data set after performing Expectation Maximization method.

ID	depression	age	height	Wage
1	5	32	181.43	32, 010
2	1.362	17	173	31, 600
3	7	19.53	169	48, 020
4	5	24	186	17, 400
.	.	.	.	.
.	.	.	.	.
100	4	45	201	7, 800

2.8 Chapter Summary

This chapter presented a background study of main machine learning and data mining technologies used in the present research. It also presented data mining in the field of healthcare. Some related prior work on different data mining techniques.

The next chapter will present the research methodology used in current research and the details of the datasets used.

RESEARCH METHODOLOGY

3.1 Introduction

Two major research paradigms have been identified in the Western tradition of science, Positivist (called scientific) and Interpretive (known as anti-positivist) [50]. However, Dash [51] stated three major types of research paradigms: positivism, antipositivism, and critical thinking. The positivism paradigm is based on observation and reasoning as a tool of understanding a certain problem or behavior. This paradigm usually involves manipulation to variables and predictions on the basis of previous observation or previous history. Positivist researchers are concerned about what has caused a particular relationship and what the effects of this relationship are, they also prefer quantitative data which can be transformed into numbers and statistics.

Anti-positivism or qualitative research approach concentrate on subjectivist approach to studying social phenomena which have importance to a range of research techniques. Anti-positivism researchers criticize positivists because they believe that statistics and numbers are not useful about human behavior .Similarly, critical theory research approach describes critique and action research as research methods to investigate a certain problem [51].

Despite the fact that each research tradition has its own approaches and research methods, the researcher may adopt research methods cutting across research traditions to solve the problem or to answer research questions [51].

This research can be placed as a positivism research utilizing the principles of prediction upon previous history and data manipulations (the term manipulating in this regard does not involve change in data structure or values. However, manipulating data is the process of filling missing features values, treating noisy data, data normalization and more).

3.2 Data Mining Methodology

Knowledge discovery from the databases or data mining refers to extracting useful relationships and patterns from large databases. Due to the amount of data and to obtain useful outcomes, a systemically method must applied. It is became a fact that quality data will imply more accurate outcomes than dirty data. Dirty data is a common term in data mining that describe some unwanted data characteristics such as incompleteness, noisy, and inconsistency. In this research, our method involves different data mining processes as shown in figure 14:

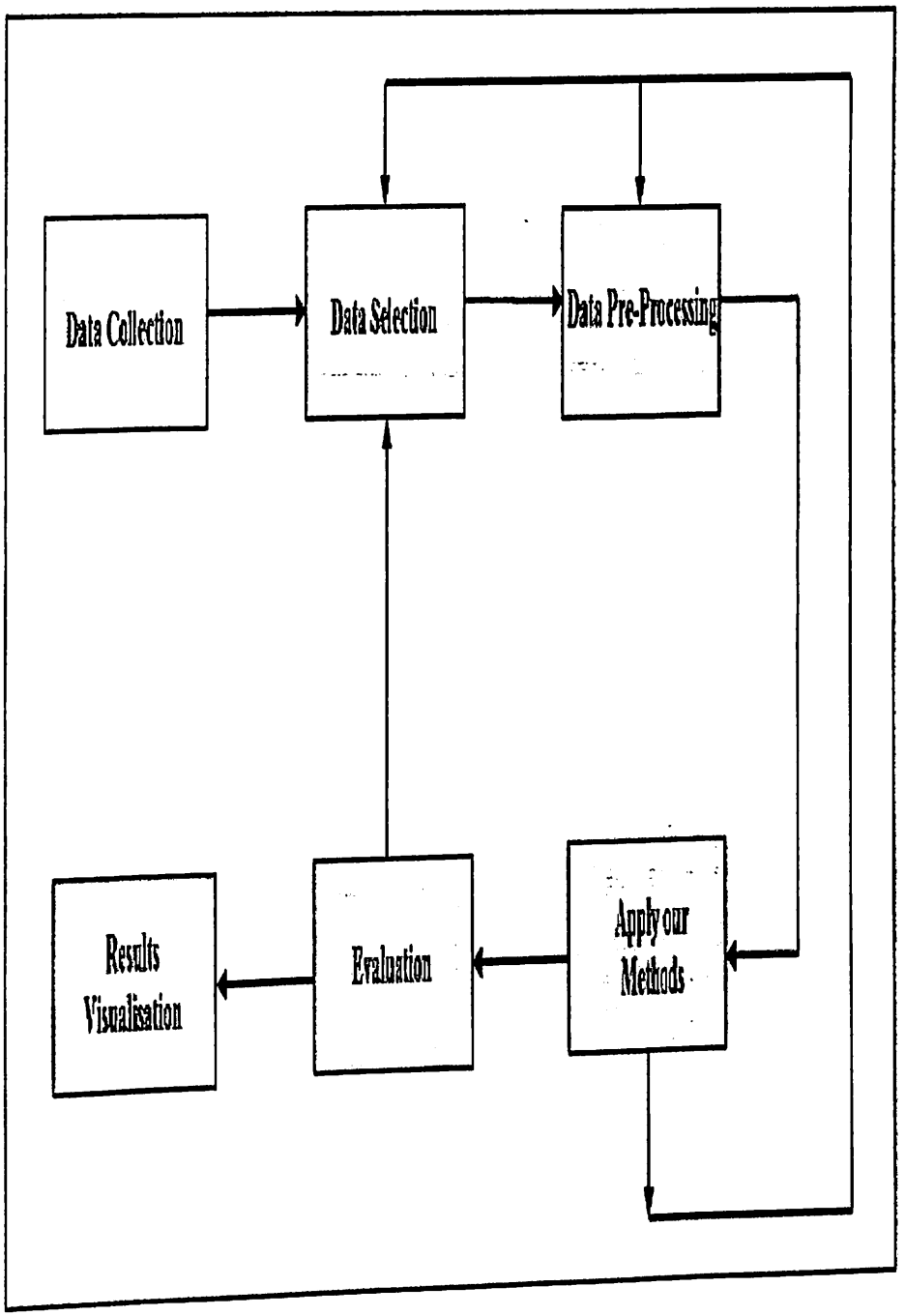

Figure 14 Research Method Overview

3.2.1 Data Collection

It is very important to acquire high quality data, which is highly reliant on the quality of the data collection process. The research data were proposed to be collected from Canberra hospital and some healthcare providers in the Kazakhstan. However, access to patient data could not gain due to privacy police in healthcare providers in Kazakhstan. The second option was to collect data from overseas. This option failed due to the cost involved. Therefore, this research relied on third option which utilizes online databases. Online databases

are available publicly and are collected from clinical environment, and have undergone proper organisational ethics approval processes and available freely for research proposes. The advantage of using online databases is the ability to make comparison between our methods and the existing methods by using the same databases. One of the most popular machine learning repositories is UCI machine learning repository. UCI is a collection of databases, domain theories, and data generators that are used by machine learning researchers to train and test machine learning algorithms. The repository was created in 1987 by David Aha and fellow graduate students at UC Irvine. Since that time, it has been widely used by students, educators, and researchers all over the world as a primary source of machine learning databases [52]. This research has used the dataset Wisconsin Breast Cancer Dataset (WBC) from UCI repository. WBC contains 699 records, each record consists of 9 features plus the class attribute. Table 8 shows a sample of WBC dataset. In addition to WBC, I recently found two other versions of breast cancer diagnosis, Wisconsin Diagnosis Breast Cancer (WDBC), and Wisconsin Prognosis Breast Cancer (WPBC). WDBC contains of 569 instances, 32 attributes, and 2 class labels where WPBC have 198 instances, 34 attributes, and 2 class labels.

3.2.2 Data Selection

Data selection or feature selection have been an active research area in pattern recognition, statistics, and data mining. The aim behind feature selection is to select a subset of records variables by ignoring features with little or less important information. For example, physician can make a decision based on some features whether a dangerous surgery is necessary for treatment or not. In the current research, feature selection methods have been used to minimise the number of features in the dataset before commencing the mining process.

3.2.3 Data Preprocessing

Data collection phase may produce dataset that contains incomplete, inaccurate, and inconsistent data. Inaccurate data is having incorrect attribute values; this may due to data entry errors, faulty in data collection tools, errors in data transmission, and users may submit incorrect values just to fill mandatory fields during surveying [53]. Incomplete data can occur for many reasons. For example, some attributes values were not important during data entry and some attributes values were not always available. Inconsistency occurs when there is a record that is in conflict with other records on the dataset [53]. Completeness, accuracy, and consistent data are the elements that define data quality. Data pre-processing is an important step in data mining process to satisfy data quality elements. Therefore, the current research is to utilize data pre-processing tasks to ensure the dataset is ready for mining process in order to produce accurate

results as possible. The study has proposed a new approach for constructing missing feature values to satisfy the completeness element; also a comparison has been made between feature selection methods to find the best method that suites datasets, and performed some techniques to eliminate noise and outliers. At the end of the current phase, data should be ready for mining process.

Table 8 Sample of Wisconsin Breast Cancer Diagnosis dataset

Uniformity of Cell Size	Uniformity of Cell Shape	Normal Nucleoli	Bare Nuclei	Single Epithelial Cell Size	Clump Thickness	Marginal Adhesion	Bland Chromatin	Mitoses	Class
5	1	1	1	2	1	3	1	1	2
5	4	4	5	7	10	3	2	1	2
3	1	1	1	2	2	3	1	1	2
6	8	8	1	3	4	3	7	1	2
4	1	1	3	2	1	3	1	1	2
8	10	10	8	7	10	9	7	1	4
1	1	1	1	2	10	3	1	1	2
2	1	2	1	2	1	3	1	1	2
2	1	1	1	2	1	1	1	5	2
4	2	1	1	2	1	2	1	1	2
1	1	1	1	1	1	3	1	1	2
2	1	1	1	2	1	2	1	1	2
5	3	3	3	2	3	4	4	1	4
1	1	1	1	2	3	3	1	1	2
8	7	5	10	7	9	5	5	4	4
7	4	6	4	6	1	4	3	1	4
4	1	1	1	2	1	2	1	1	2
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
8	4	5	1	2	?	7	3	1	4
1	1	1	1	2	1	3	1	1	2
5	2	3	4	2	7	3	6	1	4
3	2	1	1	1	1	2	1	1	2
5	1	1	1	2	1	2	1	1	2
2	1	1	1	2	1	2	1	1	2
1	1	3	1	2	1	1	1	1	2
3	1	1	1	1	1	2	1	1	2
2	1	1	1	2	1	3	1	1	2
10	7	7	3	8	5	7	4	3	4
2	1	1	2	2	1	3	1	1	2
3	1	2	1	2	1	2	1	1	2

3.2.4 Applying Data Mining Methods

At this stage, the data are ready for the mining process with no or little data preprocessing. The processed data will be used to evaluate the proposed methods. This work proposed a method called Greedy Search and 5 algorithms: k-nearest neighbours algorithm, support vector machines, artificial neural networks, logistic regression, naïve bayes.

3.2.5 Evaluation

Evaluation phase is an important part of data mining process. In this phase, the aim of the data mining experts is to test and assess the proposed model. If the model does not satisfy the expectations, then data mining experts usually rebuild the model by changing its parameters until the desired results are achieved.

In this study, the evaluation of proposed methods is performed by comparing the model results with the real data values (class features). According to that, the classification accuracy and error rate are calculated. The error rate (Err) of the classifier is defined as the average number of misclassified samples divided by the total number of records in the dataset. On the other hand, the classification accuracy of the model can be calculated as one minus the error rate. If the classification accuracy is less than a certain threshold let say 80% for example, then some changes has to be perform to the method, the feature selection, or the pre-processing phase until obtaining satisfying outcomes. Another approach for evaluating the results is by making a comparison between the results obtained by the proposed methods and previous methods in the literature. In most cases, the dataset used in the proposed method should be the same dataset used by other methods on the literature to ensure that a competitive method has been obtained.

In order to evaluate the performance of pattern classification systems (developed using supervised machine learning techniques such as ANNs and SVMs), the binary classification performance has to be measured. The performance of a binary classifier cannot be described by a single value and is usually quantified by its accuracy during the test phase, i.e., the fraction of misclassified points on the test set. The performance of a binary classifier can be 81 best described in terms of its sensitivity and specificity, quantifying its performance to false positive (FP) and false negative (FN) instances[86].

In a Receiver Operator Characteristics (ROC) curve, the sensitivity, which in this research is the portion of malignant tumors that are correctly classified by the learning machine, is plotted against 1-specificity, the share of benign tumors that are falsely classified by the learning machine, for different

cut-off values. The ROC analysis generally is used to determine an optimum cut-off value referred to as criterion for use in medical diagnostic tests. It is possible to achieve an optimal balance between sensitivity and specificity that is needed for a certain purposes. This can be achieved by changing the cut-off value of the system. Also, if the cost of not detecting a particular disease becomes high to society, the cut-off value can be changed to achieve a very high sensitivity, but lower specificity[86].

Table 9 Relation between, TP, TN, FP and FN — Confusion matrix

Confusion Matrix	Positive (p^a)	Negative (n^a)
Positive (p^P)	True Positive (TP)	False Positive (FP)
Negative (n^P)	False Negative (FN)	True Negative (TN)

Table 10 Binary classification performance measures

Performance Measure	Definition
True Positive (TP)	Tumor marked as malignant by a biopsy, which is also classified as malignant by the learning machine.
False Positive (FP)	Tumor marked as malignant by a biopsy, which is classified as benign by the learning machine.
True Negative (TN)	Tumor marked as benign by a biopsy, which is also classified as benign by the learning machine.
False Negative (FN)	Tumor marked as benign by a biopsy, which is classified as malignant by the learning machine.

The confusion matrix in Table 9 indicates the relationship between different performance indices for binary classification. For the computerized classification 82 of malignant and benign abnormalities in digital mammograms, the four performance indices (TP, TN, FP and FN) in Table 10 are calculated by comparing the predicted output from the learning machine with the real labels determined by a biopsy. Using these four performance measures, relative measurements for binary classification can be calculated. Sensitivity is defined as the ratio of tumors which are marked and classified as tumor, to all marked tumors, given by:

$$Sensitivity = \frac{\text{Positives correctly classified}}{\text{Total positives}} = \frac{TP}{TP+FN} \quad (15)$$

Specificity is defined as the ratio of tumors which are not marked and also not classified as tumor, to all unmarked tumors, given by:

$$Specificity = \frac{\text{Negatives correctly classified}}{\text{Total negatives}} = \frac{TN}{TN+FP} \quad (16)$$

The overall accuracy is the ratio between the total number of correctly classified instances and the test set size, given by:

$$Accuracy = \frac{\text{Instances correctly classified}}{\text{Total instances}} = \frac{TP+TN}{TP+TN+FP+FN} \quad (17)$$

In order to visualize ROC curves of the binary classification performance, the performance metrics True Positive Fraction (TPF) and False Positive Fraction (FPF) can be computed using:

$$TPF = (Sensitivity) = \frac{\text{Positivites correctly classified}}{\text{Total positives}} \quad (18)$$

$$FPF = (1\text{-}Specificity) = \frac{\text{Negatives correctly classified}}{\text{Total Negatives}} \quad (19)$$

In a medical diagnosis test, sensitivity gives the percentage of correctly classified diseased individuals and specificity indicates the percentage of correctly classified individuals without the disease. So, ROC curves are two-dimensional representations the relative tradeoff between the sensitivity (TPF) and the 1- specificity (FPF) of medical diagnostic test.[86]

3.2.6 Machine Learning Software Development Tools

The current work has used scikit library. Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means, DBSCAN and machine learning methods for data pre-processing, classification, regression, clustering, association rules, visualization, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy. [54]

3.2.7 Results Visualization

At the end of evaluation phase, data mining experts decide how to present the data mining results. The aim of data visualization is to let the end user view and utilize the obtained results. Since data mining usually involves extracting not existing information from a database, the end users may raise some questions about information source and how to utilize it. However in databases, the end users expect the information is already reside on the database. This research hasn't investigated data visualization in depth. The reason is because the current study is for research purposes and not business oriented. However, tables, scatter chart, bar chart, and figures have been used to demonstrate the obtained results.

3.3 Chapter Summary

This chapter presented the research methodology used in current research and the source of dataset used. It also describes the main methodology of data mining process. Further, the next chapter will describe a new approach for diagnosing breast cancer based on the combination of a new a new information gain and adaptive neuro fuzzy inference technique.

BREAST CANCER DIAGNOSIS BASED ON MACHINE LEARNING ALGORITHMS

4.1 k-nearest neighbours

In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression.[55] In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

During the creation of best model of KNN algorithm we used Greedy Search to fit algorithm with best parameters. And for KNN we tried to test number of neighbours from 1 to 41. The best results of our work on KNN algorithm are shown in Table 11.

Table 11 Results of KNN model

	K-Fold 2	K-Fold 3	K-Fold 4	K-Fold 5	K-Fold 6	K-Fold 7	K-Fold 8	K-Fold 9	K-Fold 10	K-Fold 11
ACCURACY	0.94	0.95	0.96	0.97	0.95	0.99	0.97	0.98	0.98	0.98
SENSITIVITY	0.93	0.88	0.91	0.93	0.92	0.96	0.94	0.95	0.96	1
SPECIFICITY	0.94	0.99	0.97	1	0.98	1	1	1	1	0.97
PPV	0.83	0.98	0.9	1	0.98	1	1	1	1	0.92
NPV	0.98	0.93	0.97	0.96	0.92	0.98	0.95	0.98	0.97	1
F-SCORES	0.87	0.93	0.9	0.96	0.95	0.98	0.97	0.98	0.98	0.96
G-SCORES	0.87	0.93	0.9	0.96	0.95	0.98	0.97	0.98	0.98	0.96

4.2 Support Vector Machines

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

When data are not labeled, supervised learning is not possible, and an unsupervised learning approach is required, which attempts to find natural clustering of the data to groups, and then map new data to these formed groups. The clustering algorithm which provides an improvement to the support vector machines is called support vector clustering.[56]

We tried to find best parameters for SVM model. And tried to change kernel and C parameters. For kernel we used linear and rbf types of kernel and for C parameter we tried values from 1 to 10. Results of best SVM model shown in Table 12.

Table 12 Results of SVM model

	K-Fold 2	K-Fold 3	K-Fold 4	K-Fold 5	K-Fold 6	K-Fold 7	K-Fold 8	K-Fold 9	K-Fold 10	K-Fold 11
ACCURACY	0.94	0.96	0.96	0.97	0.98	0.99	0.98	1	0.98	1
SENSITIVITY	0.97	0.93	0.97	0.95	0.96	0.96	0.95	1	0.94	1
SPECIFICITY	0.93	0.98	0.96	0.97	0.99	1	0.98	1	1	1
PPV	0.81	0.97	0.89	0.95	0.96	1	0.95	1	1	1
NPV	0.99	0.96	0.99	0.97	0.99	0.98	0.98	1	0.98	1
F-SCORES	0.88	0.95	0.93	0.95	0.96	0.98	0.95	1	0.97	1
G-SCORES	0.89	0.95	0.93	0.95	0.96	0.98	0.95	1	0.97	1

4.3 Logistic Regression

In statistics, logistic regression, or logit regression, or logit model is a regression model where the dependent variable (DV) is categorical. This article covers the case of binary dependent variables—that is, where it can take only two values, such as pass/fail, win/lose, alive/dead or healthy/sick. Cases with more than two categories are referred to as multinomial logistic regression, or, if the multiple categories are ordered, as ordinal logistic regression.[57]

In Logistic Regression algorithm we tried to change solver parameter. And as values for this parameter we used 'newton-cg ', 'lbfgs', 'liblinear' and 'sag'. Results of the best model are shown in Table 13.

4.4 Decision Tree Classification

Decision tree learning is a method commonly used in data mining.[58] The goal is to create a model that predicts the value of a target variable based on several input variables. An example is shown in the diagram at right. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.

In Decision Tree Classification algorithm we used min_samples_split and

min_samples_leaf, both of this parameters we tried to change from 1 to 5. And the results of our best test are shown in Table 14.

Table 13 Results of Logistic Regression model

	K-Fold 2	K-Fold 3	K-Fold 4	K-Fold 5	K-Fold 6	K-Fold 7	K-Fold 8	K-Fold 9	K-Fold 10	K-Fold 11
ACCURACY	0.95	0.97	0.95	0.97	0.97	0.98	0.97	0.98	0.98	0.98
SENSITIVITY	0.97	0.94	0.92	0.95	0.94	0.96	0.95	1	0.96	1
SPECIFICITY	0.94	0.98	0.97	0.97	0.98	0.98	0.98	0.98	1	0.97
PPV	0.83	0.97	0.97	0.95	0.97	0.96	0.95	0.95	1	0.93
NPV	0.99	0.97	0.94	0.97	0.97	0.95	0.98	1	0.97	1
F-SCORES	0.9	0.96	0.94	0.95	0.96	0.96	0.95	0.98	0.98	0.96
G-SCORES	0.9	0.96	0.94	0.95	0.96	0.96	0.95	0.98	0.98	0.96

Table 14 Results of Decision Tree Classification model

	K-Fold 2	K-Fold 3	K-Fold 4	K-Fold 5	K-Fold 6	K-Fold 7	K-Fold 8	K-Fold 9	K-Fold 10	K-Fold 11
ACCURACY	0.92	0.95	0.96	0.96	0.96	0.98	0.97	0.97	0.98	0.98
SENSITIVITY	0.9	0.94	0.88	0.93	0.96	0.95	1	1	1	1
SPECIFICITY	0.95	0.96	0.98	0.98	0.96	1	0.96	0.96	0.98	0.97
PPV	0.95	0.93	0.93	0.93	0.96	1	0.89	0.88	0.93	0.92
NPV	0.9	0.97	0.96	0.98	0.96	0.96	1	1	1	1
F-SCORES	0.92	0.94	0.9	0.93	0.96	0.97	0.94	0.94	0.96	0.96
G-SCORES	0.92	0.94	0.9	0.93	0.96	0.97	0.94	0.94	0.96	0.96

4.5 Artificial Neural Network

Neural network models in artificial intelligence are usually referred to as artificial neural networks (ANNs); these are essentially simple mathematical models defining a function $f : X \rightarrow Y$ or a distribution over X or both X and Y, but sometimes models are also intimately associated with a particular learning algorithm or learning rule. A common use of the phrase "ANN model" is really the definition of a class of such functions (where members of the class are obtained by varying parameters, connection weights, or specifics of the architecture such as the number of neurons or their connectivity).

For ANN algorithm we tried to work with activation and learning_rate parameters. For activation we used 'identity','logistic', 'tanh', 'relu' values and for

learning_rate we tried 'constant', 'invscaling' and 'adaptive' values. The result of our best model shown in Table 15.

Table 15 Results of ANN model

	K-Fold 2	K-Fold 3	K-Fold 4	K-Fold 5	K-Fold 6	K-Fold 7	K-Fold 8	K-Fold 9	K-Fold 10	K-Fold 11
ACCURACY	0.9	0.94	0.94	0.96	0.95	0.98	0.94	0.97	0.95	0.98
SENSITIVITY	0.91	0.86	0.78	0.9	0.88	0.96	0.9	0.85	0.89	0.92
SPECIFICITY	0.9	0.95	0.98	1	0.98	0.98	0.97	1	1	1
PPV	0.73	0.84	0.93	1	0.97	0.96	0.97	1	1	1
NPV	0.97	0.96	0.94	0.95	0.94	0.98	0.93	0.96	0.91	1
F-SCORES	0.81	0.85	0.85	0.95	0.92	0.96	0.94	0.92	0.94	0.96
G-SCORES	0.81	0.85	0.85	0.95	0.92	0.96	0.94	0.92	0.94	0.96

DISCUSSION AND FUTURE WORK

The main purpose of current research is to participate in the efforts of enhancing the quality of healthcare services, proposing technology as one of solutions for the problem of medical shortages in regards to staff shortage and technology lack. This thesis presented the challenges that face many countries in the field of healthcare services; the increase in population, the new culture, the climate change, and other factors have derived more demand on healthcare services. However, the process of utilising technology in healthcare services is a comprehensive process and involves many stages and steps. It is very important to discuss all related issues to conclude with new system that derive the expected services.

The current research focused on common issues that related to huge databases; missing features values and feature selections methods and how can be used to diagnosis and predict the examination for new cases.

Missing features values are a concern when dealing with databases, especially, large databases. Therefore, an approach for constructing missing features values based on iterative k-nearest neighbours and the distance functions has been proposed. The approach is an iterative approach until finding the most suitable features values that satisfy classification accuracy. The proposed approach showed improvement of 0.005 of classification accuracy on the constructed dataset than the original dataset on both Euclidean and Minkowski distance functions. The study found that Manhattan, Chebychev, and Canberra distance metrics produced lower classification accuracy on the new dataset than the original dataset. The study also noticed that classification accuracy depend greatly on the number of k-Folds. The experiment showed that less neighbours may lead to more accuracy. The reason for that, in my opinion, is the amount of noise produced from conflict neighbours. Finally, the maximum classification accuracy was on k-Fold=7 which was 0.99.

During this research, we also used Greedy Search algorithms to test different parameters for our models, to fit and find the best model for our dataset.

According to Table 16, the obtained results of modeling show that the algorithms SVM and KNN are the best ones for breast cancer prediction.

Table 16 Best results of modeling for each algorithm

	ANN	DTC	KNN	LOGIT	SVM
ACCURACY	0.98	0.98	0.99	0.98	1
SENSITIVITY	0.96	0.95	0.96	1	1
SPECIFICITY	0.98	1	1	0.98	1
PPV	0.96	1	1	0.95	1
NPV	0.98	0.96	0.98	1	1
F-SCORES	0.96	0.97	0.98	0.98	1
G-SCORES	0.96	0.97	0.98	0.98	1

Future work can be described as follows. The current research resided mainly on classification accuracy as the main criteria for measuring the performance of proposed approaches. However, future work will focus in other criteria such as classification speed and computational cost. Future work can also broaden disease options and which has been started with some attempts on thyroid and hepatic. In addition, breast cancer dataset used in this study has binary outcomes (class label). Clinical practice; however, is often more complex and outcomes maybe in different format. It is envisaged that the future work can contribute to the knowledge and improve the accuracy and reliability of established system by broaden the databases and expanding the criteria for measuring the performance of established systems.

We also aim to contact the Kazakhstan Ministry of Healthcare data to obtain real data about breast cancer patients. And we will optimize our models for this data. Then we are going to create web based interface for clinics and patients to provide support for the medical community in clinical decision making by the web system given results.

REFERENCES

1. Sorwar, G. and S. Murugesan, Electronic medical prescription: an overview of current status and issues, in Biomedical knowledge management: infrastructures and processes for e-health systems, M. Cooper and M. Gururajan, (Editors). 2010, IGI Global Hershey, PA. p. 61-81.
2. Lazarou, J., B. Pomeranz, and P. Corey, Incidence of adverse drug reactions in hospitalized patients: A meta-analysis of prospective studies. Journal of the American Medical Association, 1998. 279(15).
3. Mammography Screening Can Reduce Deaths from Breast Cancer. 2002 [sighted 2011 20/05/ 2011]; Available from: http://www.iarc.fr/en/mediacentre/pr/2002/pr139.html.
4. Most Frequent Cancers in Men and Women. 2008 [sighted 2012 20/01/2012]; Available from: http://globocan.iarc.fr/factsheets/populations/factsheet.asp?uno=900.
5. General Information About Male Breast Cancer. 2012 [sighted 2012 30/12/2012]; Available from: http://www.cancer.gov/cancertopics/pdq/treatment/malebreast/Patient/page1
6. Giarratano, J. and G. Riley, Expert Systems Principles and Programming. 2 ed. Vol. 1. 1994, Boston: PWS Publishing Company
7. Tarca, A.L., et al., Machine Learning and Its Applications to Biology. PLoS Comput Biol, 2007. 3(6).
8. Rokach, L., Data mining with decision trees: theory and applications. Vol. 69. 2007: World scientific.
9. Rokach, L. and O. Maimon, eds. Data Mining and Knowledge Discovery Handbook. Second ed. 2010, Springer Science and Business.
10. Tan, P.-N., M. Steinbach, and V. Kumar, Introduction to Data Mining. 2006: Addison-Wesley.
11. Thirumuruganathan, S. A Detailed Introduction to k-Nearest Neighbor (kNN) Algorithm. 2010.
12. Pevsner, J., Bioinformatics and Functional Genomics. 2 ed. 2009, New York: Wiley-Blackwell.
13. Weisstein, E.W. Euclidean Metric. [sighted 2011 19/08/2011]; Available from: http://mathworld.wolfram.com/EuclideanMetric.html.
14. Young, M., et al. Distance Metrics Overview. 2004 [sighted 2011 03/08/2011]; Available from: http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Clustering_Parameters/Distance_Metrics_Overview.htm
15. Grabusts, P., The Choice of Metrics for Clustering Algorithms, in Proceedings of the 8th International Scientific and Practical Conference. Volume I1. 2011. p. 70-76.
16. Jurman, G., Riccadonna, S., Visintainer, R., & Furlanello, C., Canberra distance on ranked lists. In Proceedings, Advances in Ranking–NIPS 09 Workshop , 2009, p. 22-27.
17. Mei, Z., Q. Shen, and B. Ye, Hybridized k-NN and SVM for gene expression data classification. Life Science Journal, 2009. 6(1).
18. Parvin, H., H. Alizadeh, and B. B, MKNN: Modified k-Nearest Neighbor, in Proceeding of the World Congress on Engineering and Computer Science. 2008: USA.
19. Cedeno, W. and D. Agrafiotis, Using particle swarms for the development of QSAR

models based on k-nearest neighbor and kernel regression. Journal of Computer-Aided Molecular Design, 2003. 17(2-4).
20. Crookston, N. and A. Finley, Impute: An R Package for k-NN Imputation. Journal of Statistical Software, 2007. 23(10).
21. Wolberg, W. and L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proceedings of the National Academy of Sciences, 1990. 87: p. 9193 - 9196.
22. Kotsiantis, S., Supervised Machine Learning: a Review of Classification Technigues. Informatica, 2007. 31: p. 249-268.
23. Priddy, K. and P. Keller, Artificial neural networks: introduction. 2005, Washington: SPIE.
24. Widrow, B. and M. Hoff, Adaptive Switching Circuits, in WESCON Conference Record. 1989. p. 709-717.
25. Grossberg, S., Adaptive Resonance Theory, in Encyclopedia of Cognitive Science. 2006, John Wiley & Sons, Ltd.
26. Javed, K., et al., Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, 2001. 7: p. 673 - 679.
27. Baxt, W.G., Use of an artificial neural network for data analysis in clinical decision-making: the diagnosis of acute coronary occlusion. Neural Comput., 1990. 2(4): p. 480-489.
28. Gershenson, C., Artificial neural networks for beginners. arXiv preprint cs/0308031, 2003.
29. Deligiorgi, D. and K. Philippopoulos, Spatial Interpolation Methodologies in Urban Air Pollution Modeling: Application for the Greater Area of Metropolitan Athens, Greece. Advanced Air Pollution. 2011.
30. Larose, D., Discovering Knowledge in Data: An Introduction to Data Mining. 2005, New Jersey: John Wiley & Sons, Inc.
31. Rokach, L. and O. Maimon, eds. Data Mining With Decision Trees. 2008, World Scientific Publishing.
32. Mitchell, T.M., Machine Learning. 2005: McGraw Hill
33. Arzucan, O., Supervised and Unsupervised Machine Learning Technique for Text Document Categorization, in Graduate Program in Computer Engineering. 2004, Bogazici University.
34. Caruana, R. and A. Niculescu-Mizil, An empirical comparison of supervised learning algorithms, in Proceedings of the 23rd international conference on Machine learning. 2006, ACM: Pittsburgh, Pennsylvania. p. 161-168.
35. Wu, X., et al., Top 10 algorithms in data mining. Knowl. Inf. Syst., 2007. 14(1): p. 1-37.
36. Hian, K. and T. Gerald, Data Mining Applications in Healthcare. Journal of Healthcare Information Management, 2005. 19(2): p. 64-72.
37. Dakins, D., Center takes data tracking to heart. Health Data Management, 2001. 9(1): p. 32-36.
38. Biafore, S., Predictive solutions bring more power to decision makers. Health Management Technology, 1999. 20(10): p. 12-14.
39. Milley, A., Healthcare and data mining. Health Management Technology, 2000. 21(8): p. 44-47.
40. Hallick, J., Analytics and the data warehouse. Health Management Technology, 2001.

22(6): p. 24-25.
41. Feng, D.D., Biomedical information technology. 2008, Amsterdam; Boston: Elsevier/Academic Press.
42. Ennis, R.L., et al., A continuous real-time expert system for computer operations. IBM J. Res. Dev., 1986. 30(1): p. 14-28
43. Cios, K.J. and G.W. Moore, Uniqueness of medical data mining. Artif. Intell. Med., 2002. 26(1-2): p. 1-24.
44. Moore, G.W., et al., A prototype Internet autopsy database. 1625 consecutive fetal and neonatal autopsy facesheets spanning 20 years. Arch Pathol, 1996. 120(8): p. 782-785.
45. Hand, D.J., Data mining: statistics and more. The American Statistician, 1999. 52(2): p. 112-118.
46. Song, H., et al., New methodology of computer aided diagnostic system on breast cancer, in Proceedings of the Second international conference on Advances in Neural Networks - Volume Part III. 2005, Springer-Verlag: Chongqing, China. p. 780-789.
47. Arulampalam, G. and A. Bouzerdoum. Application of shunting inhibitory artificial neural networks to medical diagnosis. in Intelligent Information Systems Conference, The Seventh Australian and New Zealand 2001. 2001.
48. Setiono, R., Generating Concise and Accurate Classification Rules for Breast Cancer Diagnosis. Artificial Intelligence in Medicine, 2000. 18(3): p. 205- 219.
49. Meesad, P. and G. Yen, Combined numerical and linguistic knowledge representation and its application to medical diagnosis. Component and Systems Diagnostics, Prognostics, and Health Management II, 2003. 4733: p. 98-109.
50. Galliers, R., Choosing Information Systems Research Approaches, in Information Systems Research: Issues, Methods and Practical Guidelines, R. Galliers, (Editor). 1992, Alfred Waller: Henley-on-Thames. p. 144-162.
51. Dash, N.K. Module: Selection of the Research Paradigm and Methodology. 2005
52. UCI Machine Learning Repository [sighted 2010; Available from: http://archive.ics.uci.edu/ml/about.html.
53. Ian, W. and F. Eibe, Data Mining Practical Machine Learning Tools and Techniques with Java Implementations. 2000: Morgan Kaufmann.
54. Ian, W. and F. Eibe, Data Mining Practical Machine Learning Tools and Techniques with Java Implementations. 2000: Morgan Kaufmann.
55. Altman, N. S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". The American Statistician. 46 (3): 175–185.
56. Ben-Hur, Asa, Horn, David, Siegelmann, Hava, and Vapnik, Vladimir; "Support vector clustering" (2001) Journal of Machine Learning Research, 2: 125–137.
57. Walker, SH; Duncan, DB (1967). "Estimation of the probability of an event as a function of several independent variables". Biometrika. 54: 167–178.
58. Rokach, Lior; Maimon, O. (2008). Data mining with decision trees: theory and applications. World Scientific Pub Co Inc.
59. Han, J. and K. M, Data Mining Concepts and Techniques. Vol. 3. 2011, San Franscisco: Morgan Kaufmann.
60. Thrun, S.B., et al., The Monk's Problems-A Performance Comparison of Different Learning Algorithms. 1991, Carnegie Mellon University: Pittsburgh, PA.
61. Langley, P. and S. Sage, Induction of Selective Bayesian Classifiers. ;In UAI(1994). Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence, 1994: p. 399-406.

62. Hall, M.A. and G. Holmes, Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. IEEE Transactions on Knowledge And Data Engineering, 2003. 15(3).
63. Ashraf, M., et al., A New Approach for Constructing Missing Features Values. International Journal of Intelligent Information Processing, 2012. 3(1): p. 110-118.
64. Guyon, I., et al., An introduction to variable and feature selection. J. Mach. Learn. Res., 2003. 3: p. 1157-1182.
65. Saeys, Y., I. Inza, and P. Larrañaga, A review of feature selection techniques in bioinformatics. Bioinformatics, 2007. 23(19): p. 2507-2517.
66. Kohavi, R., D. Sommerfield, and J. Dougherty, Data Mining using MLC++ -- A Machine Learning Library in C++. IEEE, 1996.
67. Kohavi, R. and G.H. John, Wrappers for feature subset selection. Artif. Intell., 1997. 97(1-2): p. 273-324.
68. Lal, T., et al., Embedded Methods, in Feature Extraction, I. Guyon, et al., (Editors). 2006, Springer Berlin Heidelberg. p. 137-165.
69. Kononenko, I., Estimating attributes: Analysis and extensions of RELIEF, in Machine Learning ECML-94. 1994.
70. Hall, M.A., Correlation-based Feature Selection for Machine Learning, in Department of Computer Science. 1999, The University of Waikato: Hamilton
71. Rutkowski, L., et al., eds. Artificial Intelligence and Soft Computing, Part I. ed. L.N.i.C.S. 6113. Vol. 1. 2010, Springer: Poland. 487-498.
72. Guyon, I. and A. Elisseeff, An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 2003. 3: p. 1157-1182.
73. Jolliffe, I.T., Principal Component Analysis. 2002, Springer: NY
74. Liu, H. and R. Setiono. A probabilistic approach to feature selection: A fiter solution. in Proceedings of the 13th International Conference on Machine Learning. 1996. Morgan Kaufmann.
75. Liu, H. and R. Setiono. Chi2:Feature selection and discretization of numeric attributes. in Proceedings of the 7thIEEE International Conference on Tools with Articial Intelligence. 1995.
76. Xie, J., et al., Novel Hybrid Feature Selection Algorithms for Diagnosing Erythemato-Squamous Diseases, J. He, et al., (Editors). 2012, Springer Berlin Heidelberg. p. 173-185.
77. Liao, B., et al., A Novel Hybrid Method for Gene Selection of Microarray Data. Journal of Computational and Theoretical Nanoscience, 2012. 9(1): p. 5-9.
78. Vijayasankari, S. and K. Ramar, Enhancing Classifier Performance Via Hybrid FeatureSelection and Numeric Class Handling- A ComparativeStudy. International Journal of Computer Applications, 2012. 41(17): p. 30-36
79. Leach, M., Parallelising Feature Selection Algorithms. 2012, University of Manchester: Manchester.
80. Grzymala-Busse, J.W. and W.J. Grzymala-Busse, Handling Missing Attribute Values Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, (Editors). 2010, Springer US. p. 33-51.
81. Rubin, D.B., Inference and missing data. Biometrika, 1976. 63(3): p. 581- 592.
82. Howell, D. Treatment of Missing Data. 2009.
83. Marlin, B., Missing Data Problems in Machine Learning, in Department of Computer Science. 2008, University of Toronto: Canada

84. Dempster, A., N. Laird, and D. Rdin, Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal Of The Royal Statistical Society, 1977. 39(1): p. 1-39.
85. Moss, S. Expectation maximization--to manage missing data. 2009
86. Veropoulos, K. 2001, Machine Learning Approaches to Medical Decision Making. PhD thesis, University of Bristol, United Kingdom.