Journal of Advances in Computer Research
Quarterly ISSN: 2008-6148
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 4, No. 3, August 2013), Pages: 119-134
www.jacr.iausari.ac.ir

An Efficient Artificial Intelligence Based Technique in Diseases Staging and Forecasting

Negar Ahmadi * **, Alfredo Milani**

Department of Mathematics and Computer Science, University of Perugia, Perugia, Italy

negar.ahmadi@dmi.unipg.it; milani@unipg.ac.it

Received: 2013/06/04; Accepted: 2013/8/27

Abstract

Artificial Intelligence (AI) techniques offer powerful objective algorithms for analysis of multimodal and high-dimensional data. Recently, these techniques have become a reliable tool in the medical domain. This paper describes an efficient technique for building an application that is capable of forecasting and classifying healthcare information using machine learning as a subfield of AI methods. The algorithm predicts a label for each sample. The sample is a single set of feature data and the label is what category the sample falls into. The algorithm takes many of these samples as the training set, builds an internal model and finally predicts the labels of other samples, called the testing set. We apply this methodology to the breast cancer staging and also to forecast the myocardial infarction and examine the risk assessment using fuzzy clustering and Framingham heart study. The results show that the proposed technique obtains credible outputs that could be integrated in an application to be used in the health care field.

Keywords: *Disease Forecasting, Artificial Intelligence, Computational Biology, Breast Cancer, Myocardial Infarction, Framingham Study*

1. Introduction

Nowadays, people care about their health extremely and want to be in control of their health and healthcare more than ever. Disease forecasting provides warning that a certain amount of disease or probability of risk may occur at a particular time in the future. The prediction of disease happening ensures that control measures are used more efficiently. So, clinicians and patients need reliable information about an individual's risk of disease. In other words, they would have accurate information completely and would be able to use a perfect model to estimate risk. A perfect model would even be able to predict the timing of the disease's onset. In any disease, many of the known important risk factors such as blood pressure, low density lipoprotein (LDL) and cholesterol level cannot be measured with adequate exactness to support risk assessment with the adequate degree of certainty. Furthermore, our knowledge of the disease aetiology is not complete, in terms of both which risk factors are independently important and how they should each be weighted. Thus, no such perfect model exists.

Recently, artificial neural networks, machine learning and decision tree algorithms are being used in a wide range of applications ranging from detecting and classifying various diseases such as cancers [1-3] and CRT images [4] to the classification of malignancies from proteomic and microarray assays [5].

Lundin *et. al* [6] applied artificial neural network to predict 5, 10 and 15 years breast cancer specific survival. They used the value of ROC curve as a measure of accuracy of the prediction model. Delen *et. al* [7] implemented artificial neural network, decision tree and the logistic regression techniques to predict the breast cancer and showed that the decision tree and artificial neural network with 93.6% and 91.2% accuracy are more superior to logistic regression method with 89.2% accuracy. Bellaachia and Gauven [8] proposed three classification techniques in data mining, namely the Naïve Bayes, the back-propagated neural network and *C4.5* rule to predict the survivability rate of breast cancer patients. Their analysis did not include records with missing data. Their study showed that the preliminary results are promising for the application of the data mining methods into the survivability prediction problem in medical databases. Note that, *C4.5* rule is a specific rule which is used to learn rules from the new training data set. Researches on various diseases show that C4.5 rule could generate rules with good comprehensibility which profits from rule induction and strong generalization ability, which profits from artificial neural network [8, 9].

Leung *et al.* [10] presented a classification method based on the data mining framework for the real world hepatitis B data sets. The results showed that the classifier has high predictive accuracy. Uhmn *et al.* [11] used the machine learning techniques, i.e. decision rule, decision tree and support vector machine (SVM), to predict the susceptibility of the different diseases such as chronic hepatitis and liver disease from the single nucleotide polymorphism (SNP) data. The experimental results showed that decision rule is able to distinguish chronic hepatitis from normal with the maximum accuracy of 73.20%, whereas decision tree is with 72.68% and SVM is with 67.53%. Ozyilmaz and Yildirim [11] presented three neural network algorithms namely radial basis function (RBF), multi-layer perceptron (MLP) and CSFNN for diagnosis of hepatitis diseases and compared the results with some statistical methods. They showed that the RBF algorithm gives promising results. However, CSFNN has the best classification accuracy for hepatitis diagnosis. Furthermore, they demonstrated that using a hybrid network CSFNN that combines MLP and RBF is more reliable for the diagnosis.

Ordonez [12] represented the weighted association rule that uses search constraints to reduce the number of rules, searches for association rules on a training set, and finally validates them on an independent test set. The author exhibited important rules with high confidence that remain valid on the test set of several runs. Anbarasi *et. al* [13] used a genetic algorithm to determine the attributes which indirectly reduces the number of needed tests which are to be taken by the patient. They implemented three classifiers to predict the diagnosis of patients. In their research, the intensity of the disease based on the results was unpredictable. Carlos [14] implemented efficient search for diagnosis of heart disease comparing association rules with decision trees.

Harleen *et. al* [15] examined the application of the decision tree, the rule induction and the artificial neural network as classification techniques for diagnosis of diabetic patients. Also, in [16] the authors used data mining algorithm for testing the accuracy in predicting diabetic status. The authors used fuzzy systems for solving a wide range of problems in different application domain. In [17] the authors studied a new approach, called the Homogeneity-Based to optimally control the over-fitting and overgeneralization behaviors of classification on the dataset. The approach was used in conjunction with classification approaches such as artificial neural networks and decision trees to enhance the classification accuracy. By comparing with the

experimental results, they indicated that the proposed approach significantly outperforms the other current approaches. However a disease prognosis can only come after a medical diagnosis and a prognostic prediction must take into account more than just a simple diagnosis [18].

The present paper focused on the implementation of a ML method to support medical decisions in the staging and the prognosis of breast cancer and myocardial infarction. The aim is to stage and predict previously identified diseases, through interdisciplinary work that included the collection and processing of data (or risk factors) from hospitals and clinics, the implementation of *J48* decision tree algorithm with specific application to medical prognosis [9], and the classification of the risky groups using the fuzzy clustering algorithm to determine the 10-year percentage of the risk for the MI patients.

2. The Algorithm Used

2.1 Machine Learning Algorithms:

Machine Learning (ML) is as a subfield of AI that employs a variety of statistical, probabilistic and optimization algorithms to learn from past data and to then use that prior training to improve performance over time, classify new data, find patterns in data or predict novel trends [19]. Machine learning techniques can employ Boolean logic (True, False), absolute conditionality (AND, OR, NOT) and conditional probabilities (IF, THEN, ELSE) to model data or classify patterns.

There are two major types of ML algorithms namely supervised and unsupervised methods [16]. Supervised methods are trained with labeled set of training data; that is, cases that have known outcomes. In fact, the labeled data are the training set that the system tries to learn about or to learn how to transfer the input data to the desired output. Unsupervised methods are trained with unlabeled data and group data based on similarity. These methods include such methods as self-organizing feature maps [20], *k*-means and hierarchical clustering algorithms which create clusters from unclassified and unlabeled data. Almost, all the ML algorithms used in diseases prediction and prognosis employ supervised learning methods belong to a specific category of classifiers on the basis of conditional classification algorithms.

Usually, the objective of the classification is to assign a class to find previously unseen records or data as accurately as possible. Consider that there is a training set or collection of records (e.g. patient 1, patient 2 and etc.) and each record contains a set of attributes (such as: sex, age, LDL value and etc.). The aim is to use a model to classify the attributes. For this purpose, the data set is divided into the training set (which is used to build the model) and the test set which is used to validate the training set and determine the accuracy of the model. The artificial neural networks [21], the genetic algorithms [22], the linear discernment analysis [23], the *k*-nearest neighbor algorithms [24] and the decision trees [25] are the major types of the conditional classification algorithms.

2.2 Data Mining:

Data mining is the computational process of extracting hidden knowledge from large volumes of data set. The overall goal of the data mining is to extract previously unknown, implicit and potentially useful information from a data set and transform it into an understandable structure for further use [26]. Medical data mining involves the

conceptualization, extraction, analysis and interpretation of available clinical data for practice clinical decision making. In other words, medical data mining describes a practice-based research strategy for systematically collecting and analyzing available medical data and has great potential for discovering the hidden patterns in the data sets of the medical domain. In fact, prediction and description are two primary goals of the medical data mining. Prediction involves some variation in the data set to predict unknown or future values of other variables of interest and description focuses on finding patterns describing the data that can be interpreted by humans [27]. Medical data can be obtained from various sources like medical transcript files which usually are voluminous, widely distributed and heterogeneous in nature. Thus, the data should be collected in the structured forms to provide a user oriented approach to novel and hidden patterns in the data.

As mentioned before, the language of the training set is based on the attribute-value pairs. In this research, the following steps are used to remove the errors due to the missing attributes or the missing values of the training data set.

Step 1(Missing attributes): In this case, the missing attribute is added to the patient record and its value is obtained based on step 2.

Step 2 (Missing values): the abundant value in the related class is selected for symbolic missing value (e.g. true/false) and the average of the values in the related class is calculated and set to the missing numeric values.

2.3 Decision Tree:

A decision-tree system recursively partitions the dataset into smaller subsets, at each level of the recursion choosing one attribute to *branch* on (creating two or more subsets); each branch is labeled with an attribute and value (or set of values). When the recursion stops, the final subsets are called the *leaves* of the decision tree that has been formed by the recursive process; each leaf is labeled with a class. Each level of branching represents an attribute-value pair, so the results from a decision tree run may also be written as a rule set. The evaluation measure, the splitting criterion and the stopping criterion are the basic components of a decision tree. The evaluation measure assigns a value to the quality of the partition obtained when the current subset is branched on a specific attribute, using the specified splitting criterion. The specified splitting criterion determines how a particular subset should be partitioned, using the specific attributes. The stopping criterion indicates that the recursion should end at this level.

One of the best known decision tree systems is *J48* which implements Quinlan's algorithm [28] for generating pruned or un-pruned *C4.5* decision tree. Note that, *C4.5* algorithm is an extension of the Quinlan's *ID3* algorithm. *J48* builds decision trees from a set of labeled training data set and can be used for classification purposes. The algorithm examines the information gain that results from choosing an attribute for splitting the data. To make the decision, the attribute with the highest normalized information gain is used. Then *J48* algorithm recurs on the smaller subsets and the splitting procedure stops if all instances in a subset belong to the same class and finally a leaf node is created telling to choose that class. *J48* provides an option for pruning trees after creation and can handle continuous and discrete attributes, attributes with differing costs and training data with missing attribute values [29]. In this article *J48* algorithm is used to make a decision tree.

2.4 Fuzzy C-Means Classifier:

In this research, the fuzzy c-means classifier is used to classify the risky myocardial infarction class to four sub-classes (i.e. low, mean, high and extremely high risk) based on the Framingham heart studies [30]. For brevity, in the sequel we abbreviate fuzzy c-means as FCM. The FCM algorithm was first reported by Dunn [31] which was subsequently modified by Bezdek [32]. FCM is a clustering method which allows one piece of data to belong to two or more clusters. In other word, FCM algorithm works by allotting membership to each data point corresponding to each cluster center on the basis of Euclidean distance between the data point and the center of the cluster or centroids [33]. More the data is near to the center of the cluster more is its membership towards the particular centroid. Note that, the membership summation of each data point should be equal to one. The algorithm is composed of the following steps:

Step 1: Let $\boldsymbol{X} = \{x_1, x_2, x_3 ..., x_n\}$ be the data points set and $\boldsymbol{V} = \{v_1, v_2, v_3 ..., v_n\}$ be the centroids set and select the number of centroids, *c*, and initialize the fuzzy membership matrix as following:

$$u_{ij} = 1 \Big/ \sum_{k=1}^{c} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}} \tag{1}$$

where *n* is the number of data points, *m* is the fuzziness index $(1 \le m \le \infty)$, c_j is j^{th} centroid and $\|x_i - c_j\|$ represents the Euclidean distance between i^{th} data and j^{th} centroid. Note that, the membership matrix should obey the following rules [34, 35]:

$$u_{ij} \in [0,1] \quad \forall i, j, \qquad \sum_{j=1}^{c} u_{ij} = 1 \quad \forall i \ , \qquad 0 < \sum_{i=1}^{n} u_{ij} < n \quad \forall n \tag{2}$$

Step 2: update the centroids set, c_j, using:

$$c_j = \sum_{i=1}^{n} u_{ij}^m \times x_i \Big/ \sum_{i=1}^{n} u_{ij}^m \quad \forall \ j \in 1,2,...,c \tag{3}$$

Step 3: calculate the new fuzzy membership matrix and objective function using equation (1) and (3) respectively:

$$J_m = \sum_{i=1}^{N} J_i = \sum_{i=1}^{n} \left(\sum_{j=1}^{c} u_{ij}^m \|x_i - c_j\|^2 \right) \tag{4}$$

Step 4: The main objective of FCM algorithm is to minimize the objective function. Thus, Repeat step 2 and 3 until the minimum objective function value is achieved as following:

$$\max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} < e \tag{5}$$

where *k* is iteration step and e is the termination criterion between [0, 1].

3. Results and Discussion

In this section, the results of the proposed algorithm for the breast cancer and the myocardial infarction as two prevalent diseases are presented. At first, the application of algorithm for breast cancer staging is presented and then we apply the methodology to forecast the myocardial infarction. At the end of this section, the risk assessment of the myocardial infarction is examined. All of the required training data set has been collected from the hospitals of Tehran city in a period of ten years and also the information on the people who had visited the clinics for a health checkup.

3.1 Breast Cancer Staging

Breast cancer can begin in different areas of the breast, usually the ducts (tubes that carry milk to the nipple), the lobules (glands that make milk) or in some cases, the tissue in between. It occurs in both men and women, although male breast cancer is rare. In fact, the female sex and older age are the primary risk factors for breast cancer. Lack of childbearing or breastfeeding, higher hormone levels, diet and obesity are the other potential risk factors of breast cancer [36]. A staging system is a standardized process to summarize information about how far a cancer has spread within the breast or to other parts of the body. The stage of a breast cancer can be based either on the results of physical exam, Sentinel lymph node biopsy, Chest X-ray, CT scan and bone scan or on the results of these tests plus the results of surgery. It is crucial to know the stage in order to plan therapy.

The most common system used to describe the stages of breast cancer is TNM (Tumor, Node, Metastasis) system. This system has been proposed by the American Joint Committee on Cancer (AJCC) and classifies cancers based on their T, N, and M stages [37]. Table (1) shows the details of the TNM staging system. Once the T, N, and M characteristics have been determined, this information is combined and the pathologist can use them to assign a stage to the cancer. Cancers with similar stages are often treated in a similar way. Stages are expressed from zero to IV. Generally, stages 0 and I (includes IA and IB) represent the earliest detection of breast cancer development, i.e. the cancer cells are confined to a very limited area. The cancer has begun to grow or spread in stage II (includes IIA and IIB), but it is still in the earliest stages and contained to the breast area. Stage III (includes IIIA, IIB and IIIC) is considered advanced cancer with evidence of cancer invading surrounding tissues near the breast and stage IV indicates that cancer has spread beyond the breast to other areas of the body.

A training set that consists of more than 900 samples is used to train the machine and generate *J48* pruned decision tree. The generated decision tree is shown in figure (1). The attributes of each sample are based on the TNM staging system and the generated tree has 18 leaves. The confusion matrix of the used decision tree is presented in figure (2). A confusion matrix contains information about actual and predicted classifications done by a classification system. This matrix shows that the machine classifies the samples with a very good accuracy. The descriptions of the stages are summarized in Table (2).

Table 1: Details of TNM staging system.

T: describes the size of tumor and spread to the skin or to the chest wall under the breast.	
T0	no evidence of primary tumor.
Tis	carcinoma in situ (e.g. Paget disease of the nipple with no associated tumor mass)
T1	size of tumor ≤ 2cm
T2	2cm < size of tumor ≤ 5cm across.
T3	Size of tumor > 5cm across.
T4	tumor of any size growing into the chest wall or skin. This includes inflammatory breast cancer.
N: indicates whether the cancer has spread to lymph nodes and, if so, how many lymph nodes are affected.	
N0	nearby lymph nodes do not contain cancer.
N1	cancer has spread to 1 to 3 axillary lymph node(s), and/or tiny amounts of cancer are found in internal mammary lymph nodes.
N1mi	the areas of cancer spread in the lymph nodes under the arm are 2 mm or less across.
N2	cancer has spread to 4 to 9 lymph nodes under the arm, or cancer has enlarged the internal mammary lymph nodes.
N3	The cancer has spread to 10 or more lymph nodes under the arm or to the infraclavicular lymph nodes OR the cancer has spread to the internal mammary nodes with axillary node involvement OR to the supraclavicular lymph nodes.
M: indicates whether the cancer has spread to distant organs e.g. the lungs or bones.	
M0	No clinical or radiographic evidence of distant metastases.
M1	distant metastasis is present.

M = M0
|**N = N0**
| |T = Tis: Stage 0
| |T = T1: Stage IA
| |T = T2: Stage IIA
| |T = T3: Stage IIB
| |T = T4: Stage IIIB
|**N = N1mi**: Stage IB
|**N = N1**
| |T = T0: Stage IIA
| |T = T1: Stage IIA
| |T = T2: Stage IIB
| |T = T3: Stage IIIA
| |T = T4: Stage IIIB
|**N = N2**
| |T = T0: Stage IIIA
| |T = T1: Stage IIIA
| |T = T2: Stage IIIA
| |T = T3: Stage IIIA
| |T = T4: Stage IIIB
|**N = N3:** Stage IIIC
M = M1: Stage IV

Figure 1: J48 decision tree for breast cancer staging.

=== Confusion Matrix ===

a	***b***	***c***	***d***	***e***	***f***	***g***	***h***	***i***	<-- classified as
24	0	0	0	0	0	0	0	0	***a*** = Stage 0
0	24	0	0	0	0	0	0	0	***b*** = Stage IA
0	0	48	0	0	0	0	0	0	***c*** = Stage IB
0	0	0	72	0	0	0	0	0	***d*** = Stage IIA
0	0	0	0	48	0	0	0	0	***e*** = Stage IIB
0	0	0	0	0	105	0	0	0	***f*** = Stage IIIA
0	0	0	0	0	0	63	0	0	***g*** = Stage IIIB
0	0	0	0	0	0	0	105	0	***h*** = Stage IIIC
0	0	0	0	0	0	0	0	420	***i*** = Stage IV

Figure 2: confusion matrix of J48 decision tree for breast cancer staging.

Table 2: description of the breast cancer stages.

Stage	Sub-stage	TNM Description
0	---	**Tis, N0, M0:** it is called noninvasive cancer and describes that cancer is only in the ducts and lobules of the breast tissue and has not spread to the surrounding tissues.
I	***IA***	**T1, N0, M0:** The tumor measures up to 2cm or less and has not spread outside the breast; no lymph nodes are involved.
	IB	**T0 or T1, N1mi, M0:** The size of tumor is 2 cm or less across (or is not found) with micro-metastases in 1 to 3 lymph nodes and the cancer has not spread to distant sites.
II	***IIA***	**T0, N1, M0:** There is no evidence of a tumor in the breast, but the cancer has spread to the lymph nodes but not too distant parts of the body. **T1, N1, M0:** The size of tumor is 2cm or less and has spread to the lymph nodes. **T2, N0, M0:** The size tumor is more than 2cm but not larger than 5cm and has not spread to the lymph nodes.
	IIB	**T2, N1, M0:** The size of tumor is more than 2cm but not greater than 5cm and has spread to one to three lymph nodes. **T3, N0, M0:** The size of tumor is more than 5cm but has not spread to the lymph nodes.
III	***IIIA***	**T0 to T2, N1, M0:** The size of tumor is not more than 5cm and has spread to four to nine lymph nodes or it has enlarged the internal mammary lymph nodes but it hasn't spread to distant sites. **T3, N1 or N2, M0:** The size of tumor is more than 5cm but does not grow into the chest wall or skin. It has spread to 1 to 9 lymph nodes or to internal mammary nodes but it hasn't spread to distant sites.
	IIIB	**T4, N0 to N2, M0:** The tumor has grown into the chest wall or skin and caused swelling or an ulcer AND may have spread to up to nine lymph nodes OR may have spread to lymph nodes near the breastbone. In this stage, The cancer hasn't spread to distant sites.
	IIIC	**Any T, N3, M0:** A tumor of any size that has not spread to distant parts of the body but has spread to 10 or more lymph nodes or the lymph nodes OR it has spread to lymph nodes around the collarbone OR it has spread to lymph nodes or to lymph nodes near the breastbone.
IV	---	**Any T, any N, M1:** The cancer can be any size and may or may not have spread to nearby lymph nodes. It has spread to distant organs or to lymph nodes far from the breast (bones, liver, brain, etc.).

3.2 Myocardial Infarction Forecasting

Myocardial infarction (MI) is the technical name for a heart attack and occurs when blood flow to a part of your heart is blocked for a long enough time that part of the heart muscle is damaged or dies (called an infarct). This part of the heart muscle is at risk of dying unless the blockage is quickly removed. MI is common and mostly occurs in

people aged over forty five and it becomes more common with increasing age [38, 39]. Briefly, the risk factors that can increase the chance of MI include:

- *Age:* The risk of MI increase in men aged over 45 and women over 55.
- *Sex:* Women have a significantly higher risk for MI, than their male counterparts [40].
- *Family History:* the risk is increased if there is a family history of MI that occurred in father/brother aged below 55, or in mother/sister aged below 65.
- *Diabetes:* People with diabetes mellitus (DM) has a higher risk of having the MI.
- *LDL:* it is also called *bad* cholesterol. The normal bound is 130 and levels above 130 raise the risk.
- *HDL:* The normal bound is from 30 to 70. The higher levels decrease the risk of MI, more.
- *Cholesterol:* Cholesterol is a fat-like substance that is made in the body. High blood cholesterol increases the chance of having the MI.
- *Triglyceride:* Triglycerides (TG) are one of the particles that transport fat around the body. The contribution of TG to the development of heart disease has been less clear (compared to LDL and HDL). The normal TG level is less than 150.
- *High Blood Pressure:* High blood pressure or hypertension is a common condition in adults that can lead to major health problems like a MI.
- *Smoking:* Relative risk of MI increased with tobacco consumption in people.
- *Chest Pain:* Chest pain may be a symptom of a number of serious conditions such as MI.

In this section we will try to forecast the MI based on the training set data. The machine training set consists of 124 samples and each sample has 11 attributes or risk factors as listed above. Figure (3) shows the generated decision tree which has six leaves. Finally, the samples are classified two classes namely risky and non-risky. A part of the training set is shown in table (3). The last row of the table (3) presents the result of classification for each sample. Note that, the most important risk factors in MI are consecutively: chest pain, high cholesterol level, high blood pressure and DM. Thus, the machine generates the decision tree based on these major factors.

Table 3: A part of training data set and its classification results for MI prognosis

	Patient1	Patient2	Patient3	Patient4	Patient5	Patient6	Patient7	Patient8	Patient9
Age	72	73	43	39	42	40	24	32	38
Sex	M	M	M	M	F	F	F	F	M
Familiar History	True	True	True	True	True	True	False	False	False
DM	False	True	True	True	True	True	False	False	False
LDL	140	166	130	148	150	132	115	112	121
HDL	43	36	42	40	50	43	50	44	51
Cholesterol	220	249	158	130	160	139	165	124	115
TG	180	200	121	112	131	122	90	102	114
Blood Pressure	True	True	True	True	False	True	False	False	False
Smoking	False	True	False	False	False	False	False	False	False
Chest Pain	True	True	False	False	True	False	False	False	False
CLASS	**Risky**	**Risky**	**Risky**	**Risky**	**Risky**	**Risky**	**Non-Risky**	**Non-Risky**	**Non-Risky**

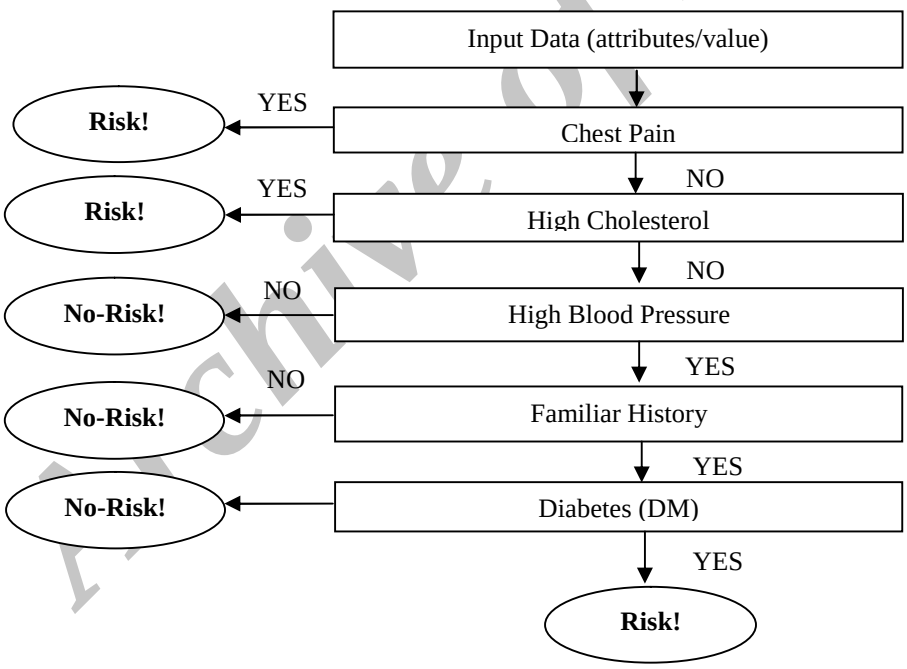

Figure 3: J48 decision tree for MI forecasting.

Figure (4) shows the summary of the simulation. It is clear that 123 samples out of 124 samples are correctly classified. Thus, the training set is classified with 99.1935 % of accuracy. By the way the constructed tree can classify the test set with the extremely high accuracy 98.72%.

Correctly Classified Instances	123	**99.1935 %**
Incorrectly Classified Instances	1	0.8065 %
Mean absolute error	0.0138	
Root mean squared error	0.0831	
Relative absolute error	2.7763 %	
Root relative squared error	16.6629 %	
Total Number of Instances	124	

Figure 4: Summery of the generated decision tree for MI forecasting.

The interpretation of the results is as follows: 1) the risk of MI is certain for a person who suffers from chest pain continuously. Note that, more than 90% of MI patients suffer from chest pain before hospitalization. 2) high blood pressure is the second major risk factor after chest pain. So, if the patient doesn't suffer from chest pain, the value of cholesterol is checked by the machine. Since high cholesterol level is another major risk factor for MI, the machine forecasts the risk for the patient. 3) the blood pressure is checked by the machine if the patient doesn't suffer from chest pain and high cholesterol level. The class of patient is non-risky if he/she doesn't suffer from the high blood pressure; otherwise the familiar history is checked by the machine. 4) if the patient has no familiar history, he/she is classified in the non-risky class. 5) the DM is checked by the machine if the patient has a MI familiar history. If the patient suffers from DM, he/she is classified in the risky class.

3.3 Classifying the MI Risky Class

In this section, we will try to classify the risky samples and estimate the percentage risk of MI in the period of ten years based on the Framingham heart study [30]. In the Framingham study, the risky samples are divided into two main groups (i.e. men and women) and then five most important risk factors are examined (i.e. age, cholesterol, high blood pressure, HDL and smoking), and each sample receives a score according to these major factors and its 10-year risk in percentage is specified. Then, we break down the 10-year risk to four classes (i.e. low risk, mean risk, high risk and extremely high) and the fuzzy classification algorithm is employed to classify the risky samples to these four classes. In the following the procedure is described.

The Framingham risk score was first developed to estimate the ten year risk of developing heart disease and is one of a number of scoring systems used to determine an individual's chance of developing cardiovascular disease [41].http://en.wikipedia.org/wiki/Framingham_Risk_Score - cite_note-1 This scoring system gives an estimate of the probability that a person will develop cardiovascular disease within 10 years. Because the Framingham risk scores give an indication of the likely benefits of prevention, they are useful in helping decide whether lifestyle modification and preventive medical treatment and for patient education for future cardiovascular events. The Framingham risk score for women and men are tabulated in Table (4) to (8). According to the tabulated scores, the 10-year risk in percentage for any patient is calculated (see table 9).

Table 4: Framingham risk score for different ages of men and women

Age	20-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79
Men	-9	-4	0	+3	+6	+8	+10	+11	+12	+13
Women	-7	-3	0	+3	+6	+8	+10	+12	+14	+16

Table 5: Framingham risk score for people with different cholesterol levels.

	Cholesterol level					
	Age	Under 160	160-199	200-239	240-279	280 or higher
Men	20-39	0	+4	+7	+9	+11
	40-49	0	+3	+5	+6	+8
	50-59	0	+2	+3	+4	+5
	60-69	0	+1	+1	+2	+3
	70-79	0	0	0	+1	+1
Women	20-39	0	+4	+8	+11	+13
	40-49	0	+3	+6	+8	+10
	50-59	0	+2	+4	+5	+7
	60-69	0	+1	+2	+3	+4
	70-79	0	+1	+1	+2	+2

Table 6: Framingham risk score for people with different blood pressure levels.

		Blood pressure level				
		Under 120	120-129	130-139	140-159	160 or higher
Men	Untreated	0	0	+1	+1	+2
	Treated	0	+1	+2	+2	+3
Women	Untreated	0	+1	+2	+3	+4
	Treated	0	+3	+4	+5	+6

Table 7: Framingham risk score for people with different HDL levels.

	HDL level			
	60 or higher	50-59	40-49	Under 40
Men	-1	0	+1	+2
Women	-1	0	+1	+2

Table 8: Framingham risk score for smokers.

Smoker's age	Non-smoker	20-39	40-49	50-59	60-69	70-79
Men	0	+8	+5	+3	+1	+1
Women	0	+9	+7	+4	+2	+1

Table 9: The 10-year risk in percentage for men and women based on the Framingham study.

	Score (Men)														
	0	1-4	5	6	7	8	9	10	11	12	13	14	15	16	17 or more
%risk	<1%	1%	2%	2%	3%	4%	5%	6%	8%	10%	12%	16%	20%	25%	> 30%

	Score (Women)													
	under 9	9-12	13-14	15	16	17	18	19	20	21	22	23	24	25
%risk	<1%	1%	2%	3%	4%	5%	6%	8%	11%	14%	17%	22%	27%	>30%

Here, we break down the 10-year risk (table 9) to four classes as: low risk (less than 5%), mean risk (6% to 16%), high risk (17% to 30%) and extremely high (above 30%) for both men and women's groups. The initial centroids of the classes are displayed as a 4*1 matrix in which the rows show the clusters and the columns shows the centroids. According to Framingham heart study, the initial centroids are set to 5, 12, 15, and 16.5 for the men's group and 12, 19, 22.5 and 25 for the women's group. These values are used to calculate the initial value of the membership matrix. Note that, these values are averages of the scores in each selected class. In the FCM algorithm, the termination criterion is set to $\boldsymbol{\varepsilon}$ =0.03 and the fuzzy exponent is set equal to 2, which leads to extremely good results. After, 500 iterations the final membership matrix will be achieved and the membership matrix includes the best and optimum values for the centroids. Now, we can calculate the membership matrix for any new risky sample using the optimum centroids.

For example, consider a risky sample with the major risk factors as: Male, age: 73, cholesterol: 220, HDL: 43, high blood pressure and non-smoker. The score of this sample is equal to 14 based on the Framingham study. Thus, the score inputs to the fuzzy algorithm and finally the percentage of belonging to each risky class is specified. The result of clustering for this sample is presented in Table (10). It is clear that, the selected sample belongs to the high risk class mostly.

Table 10: percentage risk of MI for a selected risky sample.

Risky Groups	% of belonging to the group
Low Risk	0.8%
Mean Risk	16.5%
High Risk	70.2%
Extremely High Risk	12.5%

4. Conclusion

Medical diagnosis has become highly attributed with the development of the computer science. The artificial intelligence algorithms have improved the medical forecast to a greater extent. Application of data mining in analyzing the medical data is a good method for considering the existing relationships between variables. In our work we have tried to stage the breast cancer patients using the machine learning algorithm and decision tree. Also, we used *J48* decision tree to predict the MI in the patient. Furthermore, the risk assessment of the MI is examined and percentage of belonging to the selected risky classes for the patients is specified using the Framingham heart study and a fuzzy clustering algorithm. The classification accuracy, sensitivity, and specificity have been found to be high thus making it a good option for the prediction and classification of other diseases.

5. References

[1] D.V. Cicchetti, " Neural networks and diagnosis in the clinical laboratory: state of the art", *Clin. Chem*., 38: 9–10, 1998.

[2] J.X. Wang, B. Zhang B, and J.K. Yu, "Application of serum protein fingerprinting coupled with artificial neural network model in diagnosis of hepatocellular carcinoma", *Chin. Med. J*., 118: 1278–1284, 2005.
[3] J. A. Cruz, and D.S. Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis", *Cancer Inform*., 2: 59–77, 2006.
[4] E.H. Bollschweiler, S.P. Monig, K. Hensler, and et al., "Artificial neural network for prediction of lymph node metastases in gastric cancer: a phase II diagnostic study", *Ann. Surg. Oncol*., 11:506-511, 2004.
[5] M. Dettling, "Bag Boosting for tumor classification with gene expression data", Bioinformatics, 20: 3583-3593, 2004.
[6] M. Lundin, J. Lundin, B.H. Burke, S. Toikkanen, L. Pylkkänen, and H. Joensuu, "Artificial Neural Networks Applied to Survival Prediction in Breast Cancer", *Oncol. Int. J. Cancer Res. Treat*., 57: 281-286, 1999.
[7] D. Dursun , W. Glenn, and K. Amit, "Predicting breast cancer survivability: a comparison of three data mining methods", *Artificial Intelligence in Medicine*, 34: 113-127, 2005.
[8] B. Abdelghani, and E. Guven, "Predicting Breast Cancer Survivability using Data Mining Techniques", *6th SIAM International Conference on Data Mining*, Maryland, USA, 2006.
[9] Z.H. Zhou, and Y. Jiang, "Medical Diagnosis with C4.5 Rule Proceeded by Artificial Neural Network Ensemble", *IEEE Trans. Inform. Technol. Biomedicine*, 7: 37– 42, 2003.
[10] S. Uhmn, D.H. Kim, J. Kim, S. W. Cho, and J. Y. Cheong, "Chronic hepatitis classification using SNP data and data mining techniques", *Frontiers in the Convergence of Bioscience and Information Technologies*, Korea, 2007.
[11] L. Ozyilmaz, and T. Yildirim, "Artificial neural networks for diagnosis of hepatitis disease", *International Joint Conference on Neural Networks*, Portland, USA, 2003.
[12] C. Ordonez, "Association rule discovery with train and test approach for heart disease prediction", *IEEE Trans. Inform. Technol. Biomedicine*, 10: 334-343, 2006.
[13] M. Anbarasi, E. Anupriya, and N. Iyengar, "Enhanced prediction of heart disease with feature subset selection using genetic algorithm", *Int. J. Eng. Sci. Technol*., 2: 5370- 5376, 2010.
[14] C. Ordonez, "Comparing association rules and decision trees for disease prediction", *ACM, HIKM '06*, Virginia, USA, 2006.
[15] H. Kaur, and S. K. Wasan, "Empirical study on applications of data mining techniques in healthcare", *J. Comput. Sci*., 2: 194-200, 2006.
[16] M. S. Sapna, and A. Tamilarasi, "Data mining – fuzzy neural genetic algorithm in predicting diabetes", *Res. J. Comput. Eng*., 46-50, 2008.
[17] H. N. Pham, and E. Triantaphyllou, "Prediction of diabetes by employing a new data mining approach which balances fitting and generalization", *Comput. Inform. Sci*., 131: 11-26, 2008.
[18] R.G. Hagerty, P.N. Butow, P.M. Ellis, and et al., "Communicating prognosis in cancer care: a systematic review of the literature", *Ann Oncol*, 16: 1005-1053, 2005.
[19] M.H. Dunham, *Data Mining: Introductory and Advanced Topics*, Prentice Hall, 2003.
[20] T. Kohonen, "Self-organized formation of topologically correct feature maps", *Biol. Cybernetics*, 43:59-69, 1982.
[21] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning representations by back-propagating errors", *Nature*, 323: 533-536, 1986.
[22] J.H. Holland, "*Adaptation in Natural and Artificial Systems*", University of Michigan Press, Ann Arbor, 1975.
[23] N. Belhumeur, J.P Hespanha, and D.J. Kriegman, "Eigenfaces vs Fisherfaces: recognition using class specific linear projection", *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19: 711-720, 1997.
[24] D. Coomans, and D.L. Massart, "Alternative k-nearest neighbour rules in supervised pattern recognition : Part 1. k-Nearest neighbour classification by using alternative voting rules". *Analytica Chim. Acta*, 136: 15–27, 1982.
[25] J.R. Quinlan, "Induction of decision trees", *Machine Learning*, 1: 81-106, 1986.
[26] J. Han, and M. Kamber, "*Data Mining Concepts and Techniques*", Morgan Kaufmann Publishers, 2006
[27] S.Vijiyarani, and S.Sudha, "Disease prediction in data mining technique – a survey", *Int. J. Comput. Appl. Inform. Technol*., 2: 17-21, 2013.
[28] J.R. Quinland, "Induction of decision trees", *Machine Learning*, 1: 81-106, 1986.

[29] R. Quinlan, "*C4.5: Programs for Machine Learning*", Morgan Kaufmann Publishers, San Mateo, 1993.
[30] T.R. Dawber, G.F. Meadors, and F.E. Moore, "Epidemiological approaches to heart disease: the Framingham Study", *Am J Public Health Nations Health*, 41: 279–281, 1951.
[31] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact, well-separated clusters", *J. Cybern*., 3: 32-57, 1973.
[32] J. C. Bezdek, C. Coray, R. Gunderson, and J. Watson, "Detection and characterization of cluster substructure", *SIAM J. Appl. Math*., 40: 339-372, 1981.
[33] N. Ahmadi, "Using fuzzy clustering and TTSAS algorithm for modulation classification based on constellation diagram", *Eng. Appl. Artif. Intel*., 23: 357-370, 2010.
[34] K. Chintalapudi, and M. Kam, "A noise-resistant fuzzy c-means algorithm for clustering", *IEEE international Conference on Fuzzy systems*, 2: 1458-1463, 1998.
[35] N. Shinozaki, T. Yuasa, and S. Takata, "Cigarette smoking augments sympathetic nerve activity in patients with coronary heart disease", *Int. Heart J*., 49: 261-272, 2008.
[36] J.D. Yager, "Estrogen carcinogenesis in breast cancer", *New Engl. J. Med*., 354: 270–82, 2006.
[37] S.B. Edge, D.R. Byrd, C.C. Compton, and et al., "*AJCC Cancer Staging Manual*", 7th ed., Springer, 2010.
[38] T. Mallinson, "Myocardial Infraction", *Focus on First Aid*, 15: 1-25, 2010.
[39] M. Kosuge, K. Kimura, T. Ishikawa, and et al., "Difference between men and women in terms of clinical features of ST-segment elevation acute myocardial infarction", *Circulation J.*, 70: 222–226, 2006.
[40] J.W. Nance, S.C. McConnells, J. Schoepf, and et al., "Gender differences in the predictive value of the presence, extent, and composition of coronary atherosclerotic plaque as measured by cardiac CT angiography", *Radiological Society of North America 2911 Scientific Assembly and Annual Meeting*; Chicago, USA, 2011.
[1] [41] P.W. Wilson, R.B. D'Agostino, D. Levy, A.M. Belanger, H. Silbershatz, and W.B. Kannel, "Prediction of coronary heart diseases using risk factor categories", *Circulation*, 97: 1837–1847, 2013.