Comparison of Fuzzy Diagnosis with K-Nearest Neighbor and Naïve Bayes Classifiers in Disease Diagnosis

Asaad Mahdi
School of Mathematical Sciences, Faculty of Science and Technology, University Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
asaadmahdi@gmail.com

Ahmad Razali
School of Mathematical Sciences, Faculty of Science and Technology, University Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
mahir@ukm.my

Ali AlWakil
School of Mathematical Sciences, Faculty of Science and Technology, University Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia
wakilali@yahoo.com

Abstract

The main objective of this paper is to investigate the performance of fuzzy disease diagnosis by comparing its results with two statistical classification methods used in the diagnosis of diseases namely the K-Nearest Neighbor and the Naïve Bayes classifiers. The comparisons were made using the latest XLMiner® and Medcalc® statistical software's. The first step was using fuzzy relation such as the occurrence relation and confirmability relation on a sample of 149 patients suffering from chicken pox, dengue and flu taken from different general and private hospitals and clinics in Kuala Lumpur to diagnose the three diseases. Fourteen symptoms were used in the diagnoses such as high fever, headache, nausea, vomiting, rash, joint pain, muscle pain, bleeding, loss of appetite, diarrhea, cough, sore throat, abdominal pain and runny nose. The second step was using the K-Nearest Neighbor classification method and the Naïve Bayes classification method on the same sample to diagnose the three diseases. The final step was the comparison between the three methods using performance tests, McNemar and Kappa tests. The result of the comparison between the three methods showed that fuzzy diagnosis outperforms the other two methods in disease diagnosis.

Keywords: Fuzzy set theory, K- Nearest Neighbor Classifier, Naïve Bayes classifier.
McNemar test, Kappa test, performance tests.

1. Introduction

Fuzzy set theory was first presented by Zadeh in 1965 [18], after that Fuzzy logic was developed from fuzzy set theory to reason with uncertain information. The first fuzzy application was created in mid of 1970's (fuzzy control of a steam engine), since then the number of fuzzy application have grown rapidly, especially in Japan. The main reason to develop fuzzy logic from fuzzy set theory was to form a conceptual framework for linguistic information and knowledge. Sets of the conventional set theory are called crisp sets in order to distinguish them from fuzzy sets. Hence fuzzy logic extends conventional crisp sets to handle the concept of the partial truth – the values falling between " totally true" and "totally false" , these values are dealt with using the degree of membership of an element of a set.

The degree of membership can take any real value in the interval [0,1]. This extension of crisp logic to fuzzy logic is made by replacing the functions of crisp logic with fuzzy membership functions. Fuzzy values are assigned to evaluate the truth of propositions, and operations with these values are applied to evaluate composite propositions. Fuzzy set theory and fuzzy logic were used in many disease diagnostic system in recent years such as designing a fuzzy expert system for the determination of coronary heart disease risk [2], fuzzy logic for the diagnosis of diabetics [9]. The diagnosis of urethral obstructions [19], the diagnosis of coronary artery disease [14], expert system for diagnosing the hepatitis B intensity rate [7], human disease diagnosis [4], fuzzy relations were

used for diagnosis. The two well known statistical classifiers K – Nearest neighbor classifier and Naïve Bayes classifier were implemented, trained using the training data sets which were formed using k-fold cross validation method. The data were the same used with fuzzy diagnosis. The number of patients investigated was 149.

K- Nearest neighbor classification method is a non-parametric method sometimes is called instance- based or memory based learning algorithms since what they do is store the training data in a lookup table and interpolate from it. This method was used for disease diagnosis such as classification of ovarian cancer [17], the diagnosis of breast cancer [11].

Naïve Bayes classification method is a practical Bayesian learning method; it is based on the so called Bayesian theorem. Despite its simplicity, Naïve Bayes can often outperform more sophisticated classification methods [5]. This method was also used for the diagnosis of diseases such as medical diagnosis model for infectious diseases [15].

2. Methods and materials

2.1. Fuzzy diagnosis

Adlassing in 1980 found two fuzzy relationships to describe medical knowledge as the relationship between symptoms S_i and diseases D_j [1], namely *Occurrence* which is how often does S_i occur with D_j and *Confirmability* is how strongly does S_i confirm D_j [13]. These functions could be determined by linguistic documentation by medical experts and medical database evaluation by statistical means or a combination of both. We will determine the fuzzy Occurrence and Confirmability relations from expert medical documentation. Since this documentation usually takes the form of statements such as symptom s seldom occurs in disease d or symptom s always indicates disease d, we assign membership grades of 1, 0.75, 0.5, 0.25, 0 in fuzzy matrix Rc for the linguistic terms such as always, frequently, don't know, rarely, and never respectively [10].

In 1973, Zadeh, introduced the combination rule of a *max-min-composition*, that is if we have three sets A, B, and C, to compose fuzzy relations $Q \subseteq L(A \times B)$ and $R \subseteq L(B \times C)$ to get another fuzzy relation $T \subseteq L(A \times C)$, where $L(A \times B)$ and $L(B \times C)$ are the sets of all fuzzy sets in the Cartesian product $A \times B$ and $B \times C$ respectively [16], then $T = Q \circ R$ is defined by the following membership function (1),

$$\mu_T(x,z) = \max_{y \in B} \min \{ \mu_Q(x,y);\ \mu_R(y,z) \},\quad x \in A\ \ y \in B\ \ z \in C \qquad (1)$$

Let $Rs = P \times S$, indicates the degree to which the symptoms s is present in patient p, and it is calculated from the membership functions for each symptom.

Let $Ro = S \times D$ [12], be a matrix that indicates the frequency of occurrence of symptoms s with disease d.

Let $Rc = S \times D$ be the confirmability relation. This matrix indicates the degree which symptom s confirms the presence of disease d. From the relations Rs, Ro, Rc we can calculate three different indication relations

(1) The Occurrence relation R^1

$R^1 = Rs \circ Ro$ defined by (2)

$$\mu_{R^1}(P, D_j) = \max \min(\mu_{Rs}(P, S_i); \mu_{Ro}(S_i, D_j)) \qquad (2)$$

(2) The Non – Occurrence relation

$R^2 = Rs \circ (1 - Ro)$ defined by (3)

$$\mu_{R^2}(P, D_j) = \max \min(\mu_{Rs}(P, S_i); \mu_{Ro}(S_i, D_j)) \qquad (3)$$

(3) The Confirmability indication relation

$R^3 = Rs \circ Rc$ defined by (4)

$$\mu_{R^3}(P, D_j) = \max \min(\mu_{Rs}(P, S_i); \mu_{Rc}(S_i, D_j)) \quad (4)$$

The relations R^1, R^2, R^3 requires multiplication of fuzzy matrix using max- min rules. A computer program was written using Matlab® code to facilitate the calculation of the three relations and the final result was 136 patients with correct disease diagnosis.

2.2. K- Nearest Neighbor Classification (K-NNC)

This method of classification is one of the most fundamental and simple classification methods and should be used for a classification study when there is little or no prior knowledge about the distribution of the data. This method was developed from the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine [8].

The algorithm for this method is

- The k nearest neighbor must be located using the training dataset. The Euclidean distance measure is used to calculate how close each member of the training set is to the target row that is being examined.
- Examine the k- nearest neighbor, which classification or category do most of them belong to? Assign this classification or category to the row being examined.
- Repeat this procedure for the remaining rows in the target set.
- In this software a maximum value for k can be selected, then the software builds models parallel on all values of k up to the maximum specified value and scoring is done on the best of these models.

The First step in using K-Nearest Neighbor classification method in XLMiner® software was determining the training data set, then the input and output variables should be entered. The second step was normalizing the data which will ensure that the distance measure accords equal weight to each variable. The score on best k between 1 and specified value was chosen which builds models parallel on all values of k up to the maximum specified value in which k=9 was chosen and scoring is done on the best of these models. Finally entering the data needed for classification.

2.3. Naïve Bayes classification method

Naïve Bayes classifiers assume that the effect of a variable value on a given class is independent of the values of other variable. This assumption is called class conditional independence. It is made to simplify the computation and in this sense considered to be Naïve. This assumption is often not applicable. However, bias in estimating probabilities often may not make a difference in practice, it is the order of the probabilities not their exact values that determine the classification.

Studies comparing classifications algorithms have found the Naïve Bayesian classifier to be comparable in performance with classification trees and with neural network classifiers. They have also exhibited high accuracy and speed when applied to large databases. This method uses Bayes Theorem. Suppose we let X be the data record (case) whose class label is unknown. Let H be some hypothesis, such as "data record X belongs to a specified class C" for classification, we want to determine P(H|X), the probability that the hypothesis H holds, given the observed data record X. P(H|X) is the posterior probability of H conditioned on X. P(H) is the prior probability of H. similarly, P(X|H) is the posterior probability of X conditioned on H. P(X) is the prior probability of X. Bayes theorem is useful in that it provides a way of calculating the posterior probability, P(H|X), from P(H), P(X), and P(X|H). Bayes theorem is (5):

$$P(H|X) = P(X|H)\ P(H)/P(X) \quad (5)$$

The first step in using the Naïve Bayes classifiers in XLMiner® software is to specify the worksheet under consideration, the data range for the data selected, and the input and output variables. In the second step we must calculate the prior class probability. After choosing the prior class probability, the final step is to specify the training data, and then selecting the new data needed for classification from either database or worksheet.

3. K- Fold cross validation

K- fold cross validation was used to minimize the bias associated with random sampling of training and test data. The data were split in to k subsets with approximately equal sizes classification models were trained and tested k times. In this paper 10 – fold cross validation were used [6]. Each of these 10 folds was used once to test the performance of the classifier while other 9 were used for training. The overall accuracy was calculated by taking the mean of the ten measures for both KNN and NB classifiers [3].

4. Diagnosis and comparison

This section is dedicated to the comparison between the three methods of diagnosis namely fuzzy, KNN and Naïve Bayes classifiers. After conducting the k- fold cross validation process with KNN and NB, ten results were achieved

Table 1. The number of patients with correct diagnosis using fuzzy diagnosis, K- nearest neighbor and Naïve Bayes classifiers.

Fuzzy Diagnosis	K-NNC	Naïve Bayes Classifier
136*	131	108

C = Chicken pox, D = Dengue, F = Flu, KNNC = K- nearest neighbor, *= The best result.

As can be seen in table 1 above fuzzy diagnosis was correct in 136 cases out of 149 which means 91%, KNNC was correct in 131 cases which means 88%, while NB was correct in 108 cases which means 72.4% accuracy. Figure 1 below shows the percentage of patients with correct diagnosis for each of the three methods.

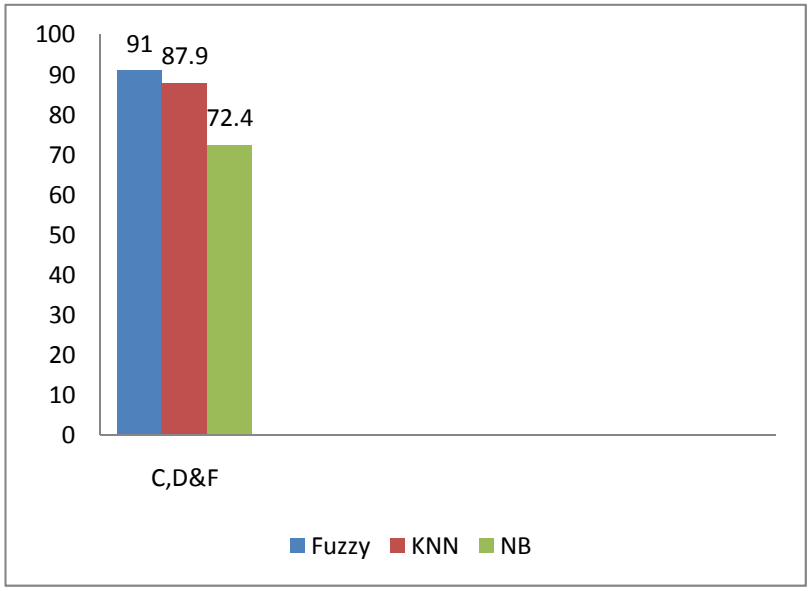


Figure 2 Bar chart represents the percentage of the correct diagnosis using fuzzy diagnosis, K- nearest neighbor and Naïve Bayes classifiers.

5. Measures of performance

5.1. Accuracy, Sensitivity and Specificity

The performance of the three diagnostic methods used in this study were evaluated using the following accuracy, sensitivity, specificity measures and as follows,

Table 2. Classification table

Test result	True condition	
	Present	Absent
Present	TP	FP
Absent	FN	TN

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad (6)$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

where $TP\ TP$, TN, FP and FN denotes true positives, true negatives, false positives and false negatives respectively [3] and [6].

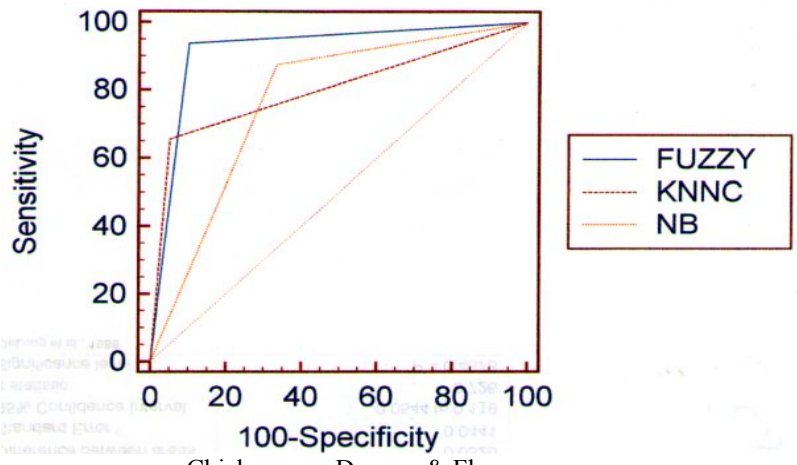
Table 3 The Accuracy, Sensitivity and Specificity for Fuzzy, KNN and NB

Classification Method	Accuracy	Sensitivity	Specificity
Fuzzy	**0.91**	0.94	0.90
KNN	0.88	0.66	0.95
NB	0.72	0.88	0.67

As can be seen from the table 3 above, fuzzy diagnosis outperformed the other two methods. The bold value indicates the best accuracy obtained for that sample using fuzzy diagnosis.

4.2. Area under the ROC Curve (AUC)

The area under the receiver operating characteristic (ROC) curve (AUC) is used to measure the performance of fuzzy diagnosis, KNN and NB. In order to show the difference between AUC for the three methods, a single figure which combines the three AUC for the three methods was used for comparison as shown in figure 1 below.



Chicken pox, Dengue & Flu

Figure 1. The comparison of the Area Under the ROC Curve (AUC)
for fuzzy Diagnosis, KNN and NB

Table 4. The AUC, standard error, confidence interval and the significance level for the comparison of fuzzy diagnosis & KNN, fuzzy diagnosis & NB .

Method	AUC	SE.	95% C.I.	F&KNN Significant	F&NB Significant
Fuzzy	0.918	0.026	0.857 - 0.959		
KNN	0.803	0.044	0.724 - 0.867	P=0.035*	
NB	0.771	0.038	0.689 - 0.840		P=0.001*

F&KNN= comparison between fuzzy diagnosis and KNN
F&NB= comparison between fuzzy diagnosis and Naïve Bayes.

As we can see from the table 4 above, if the P value is < 0.05 then the comparison between any two methods is significant, meaning that there is a significant difference between the two methods. All significant values are denoted by (*).

6. Important statistical tests

6.1. McNemar Test

McNemar's test is a non-parametric method. It is applied to 2 × 2 contingency tables with a dichotomous trait, with matched pairs of subjects, to determine whether the row and column marginal frequencies are equal (marginal homogeneity). It is named after Quinn McNemar, who introduced it in 1947. The test is applied to a 2 × 2 contingency table, which tabulates the outcomes of two tests on a sample of *n* subjects, as follows.

Table 5. The decision matrix table

	Test 2 positive	Test 2 negative	Row total
Test 1 positive	a	b	a+b
Test 1 negative	c	d	c+d
Column total	a+c	b+d	n

The null hypothesis of marginal homogeneity states that the two marginal probabilities for each outcome are the same, that is $p_a + p_b = p_a + p_c$ and $p_c + p_d = p_b + p_d$.
Thus the null hypothesis is $p_b = p_c$.

Here p_a, p_b, p_c, p_d denote the theoretical probability of occurrences in cells with the corresponding label. The McNemar test statistic with Yates' correction for continuity is given by (8):

$$x^2 = \frac{((|b-c|)-1)^2}{b+c} \qquad (8)$$

Under the null hypothesis, with a sufficiently large number of discordants (cells b and c), χ^2 has a chi-squared distribution with 1 degree of freedom.

If either *b* or *c* is small ($b + c < 25$) then χ^2 is not well-approximated by the chi-square distribution.

The binomial distribution can be used to obtain the exact distribution for an equivalent to the uncorrected form of McNemar's test statistic In this formulation, *b* is compared to a binomial distribution with size parameter equal to $b + c$ and probability of success = ½, which is essentially the same as the binomial sign test.

If the χ^2 result is significant, this provides sufficient evidence to reject the null hypothesis, in favor of the alternative hypothesis that $p_b \neq p_c$, which would mean that the marginal proportions are significantly different from each other.

Table 6. The McNemar test results, P values and significant level for the comparison of fuzzy diagnosis and KNNC

Classification Method	McNemar Test		
	KNNC		
	Test value	P value	Sig.
Fuzzy	6.035*	0.014	Yes

Sig.= Significant, (*) Chi-square test with continuity correction

Table 7 The McNemar test results, P values and significant level for the comparison of fuzzy diagnosis and NB

Classification Method	McNemar Test		
	NB		
	Test value	P value	Sig.
Fuzzy	10.256*	0.001	Yes

Sig.= Significant, (*) Chi-square test with continuity correction

As can be seen from table 6, if P is < 0.05 this means that there is a significant difference between the proportions obtained from using fuzzy diagnosis and KNNC, while table 7 shows McNemar test values and significant level for testing the proportions obtained by using fuzzy diagnosis and NB.

6.2. The weighted Kappa test

One of the undesirable properties of Kappa is that all the disagreements are treated equally. So it is preferable to give different weights to disagreement according to each cell's distance from the diagonal. But the weights can only be given to ordinal data, not to nominal data. The reason is that if we change the position of the category in row or in column, the weights will be different. The weighted kappa is obtained by giving weights considering disagreement. It was first proposed by Cohen in 1968. The weights are given to each cell according to its distance from the diagonal.

Suppose that there are k categories, i=1,…,k; j=1,..,k. The weights are denoted by wij they are assigned to each cell and their value range is $0 \le wij \le 1$. The cells in the diagonal ($i = j$) are given the maximum value, wij=1. For the other cells' ($i \ne j$), $wij = wji$ and $0 \le wij < 1$. The observed weighted proportion of agreement are obtained as (9)

$$P_{o(w)} = \sum_{i=1}^{k} \sum_{j-1}^{k} w_{ij} p_{ij} \qquad (9)$$

Where p_{ij} is the proportion of the cell in the ith row and jth column, then it is given weight wij. The observed weighted proportion of agreement is the sum of all the proportion of cells given weights. Similarly, the expected proportion of agreement is (10)

$$P_{e(w)} = \sum_{i=1}^{k} \sum_{j-1}^{k} w_{ij} p_{i.} p_{.j} \qquad (10)$$

It is the sum of all the expected proportion of the cells given weights. And the weighted kappa is then given by

$$\hat{k}_w = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}} \qquad (11)$$

Two kinds of weights are normally used one is suggested by Bartko in 1966, where the formula of the weights is (12)

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2} \qquad (12)$$

The other is suggested by Cicchetti and Allison in 1977 (Fleiss et al. 2003), where the weights are taken as (13)

$$w_{ij} = 1 - \frac{|i-j|}{k-1} \qquad (13)$$

A kappa score over 0.8 indicates very good agreement, 0.6 to 0.8 indicates substantial agreement, 0.4 to 0.6 moderate agreement, 0.2 to 0.4 fair and less than 0.2 is poor (Sprant & Smeeton 2007).

Table 8 The weighted kappa test values for the agreement between fuzzy and KNNC, fuzzy and NB.

Classification Method	Weighted Kappa Test			
	KNNC		NB	
	Test value	Agreement	Test value	Agreement
Fuzzy	0.441	Moderate	0.388	Fair

As can be seen from table 8 above, the values of the weighted kappa are different for both fuzzy diagnosis versus KNNC and fuzzy diagnosis versus NB. The results show the degree of agreement between the three methods which is between poor and moderate.

7. Discussion and conclusion

This study is dedicated to the test of performance, test of proportion and test of agreement between the results achieved from the three methods for diagnosis. The first step was a comparison between fuzzy disease diagnosis which was achieved by using a program written in MATLAB® code for the calculation of fuzzy max-min relations and the diagnosis of diseases using two well known statistical classifiers namely K- NNC and Naïve Bayes classifier using the latest XLMiner® software. The initial results showed that fuzzy was correct in the diagnosis of 136 patients out of 149, which means 91%, while K-NNC was correct in the diagnosis of 131 patients out of 149 which means 87.9% and Naïve Bayes was correct in the diagnosis of 108 patients out of 149 which means 72.4%. The second step was to test the performance of these methods using accuracy, sensitivity, specificity and (AUC). The results showed that fuzzy diagnosis was better and outperformed the other methods and there is a significant difference between the results. The last step was using two tests; firstly the McNemar test was used to test the equality of the proportions obtained by the three methods of classifications. Table 6 shows the significant differences between fuzzy diagnosis and KNNC. Table 7 shows the significant difference between fuzzy diagnosis and NB. Secondly, to test the agreement between the results obtained by the three methods. Table 8 has two parts; the first shows the degree of agreement between fuzzy diagnoses versus KNNC which is moderate agreement. The second part is for fuzzy diagnosis versus NB, the test of agreement shows poor agreement. The comparison results showed that K-Nearest Neighbor and Naïve Bayes Classifier have limited usefulness and fuzzy disease diagnosis is better and more useful in the diagnosis of these diseases than the other two methods. No previous comparison between these three methods has been previously published; the current research is completely new in this field.

8. References

[1] Adlassnig, K-P.(1980). A fuzzy logical model of computer assisted Medical Diagnosis Method. Info Med. 19:141-148

[2] Allahverdi, N., Torun, S. and Saritas, I. (2008). Design of A Fuzzy Expert System for The Determination of Coronary Heart Disease Risk. International Conference on Computer Systems and Technology – CompSysTech07. pp IIIA. (14-1)-(14-8).

[3] Delen, D., Glenn, W., Kadam, A. 2005. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine. Vol. 34, 2 pp 113-127.

[4] Hasan, M., Alam, K. and Chowdhury, A. (2010). Human Disease Diagnosis Using a Fuzzy Expert System. Journal of Computing, Vol.2, issue 6:pp 66-70.

[5] Isam, M., Wu, Q., Ahmadi, M. and Ahmad, M.(2007). Investigating the Performance of Naïve Bayes Classifiers and K-NNC. International Conference on Convergence Information Technology 2007. Pp 1541-1546.

[6] Khan, U., Shin, H., Choi, J., Kim, M. 2009. wFDT – Weighted Fuzzy Decision Trees for Prognosis of Breast Cancer Survivability. The seventh Australian data mining conference. Conference in research and practice in information technology. Vol. 87, pp 141-152.

[7] Neshat, M. & Yaghobi, M. (2009). Designing a Fuzzy Expert System of Diagnosing the Hepatitis B Intensity Rate and Comparing it with Adaptive Neural Network Fuzzy System. Proceeding of the world congress on Engineering and Computer Science, Vol.2,WCECS.

[8] Peterson, L. (2009). K Nearest Neighbor. Scholarpedia 4(2): 1883.

[9] Radha, R.& Rajagopalan, S. (2007). Fuzzy logic Approach for Diagnosis of Diabetics. Information Technology Journal 6(1):pp 96-102.

[10] Raich, V., Sharma, R., Triapathi, R., Pand, D. and Bawa, N. (2010). A Basic Idea to Generate computer program for the study of fuzzy matrix and its application. Advances in fuzzy mathematics, vol. 5, no. 1(2010) , pp. 65-70.

[11] Sarkar, M.& Leong, T. (2000). Application of K-Nearest Neighbor Algorithm on Breast Cancer Diagnosis Problem. American medical informatics association, proceedings for Annual Symposium, pp 759-763.

[12] Schuerz , M., Adlassnig K-P, Lagor, C. Scheider, B. and Grabner, G. (1998). Definition of Fuzzy Sets Representing Medical Concepts and Acquisition of Fuzzy Relationships between them by Semi- Automatic Procedures. European Symposium on Intelligent techniques (ESIT99), Kreta 12568.

[13] Seising, R. (2004). A History of Medical Diagnosis using Fuzzy Relations. (draft paper, Fuzziness in Finland 04).

[14] Setiawan, N., Venkatachalam, P. and Hani, F. (2009). Diagnosis of Coronary Artery Disease Using Artificial Intelligence Based Decision Support System. Proceedings of the International Conference on Man- Machine Systems, Batu Ferringhi, Penang, Malaysia. pp (IC3-1)-(IC3-5)

[15] Wagholikar, K., Deshpande, A. (2008). Fuzzy relation based modeling for medical diagnostic decision support: Case studies. International Journal of Knowledge Based and Intelligent Engineering Systems. Ios press Vol 12, no 5-6 /2008, pp 319-326 .

[16] Vig, R., Handa, N., Bali, H. and Sridhar (2004). Fuzzy Diagnostic System for Coronary Artery Disease. IE(I) Journal, 85: 41-46.

[17] Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor,G., Stone, K., Ward, D., Williams, K.and Zhao, H. (2003). Comparison of Statistical Methods for Classification of Ovarian Using Mass Spectrometry data. Bioinformatics, Vol. 19, No. 13, pp 1636-1643.

[18] Zadeh, LA. (1965). Fuzzy Sets. Information and Control. 1965; 8: 338- 353.

[19] Zuhtuogullari, K., Saritas, I., Arikan, N. (2009). Diagnosis Modelling of Urethral Obstructions Using Fuzzy Epert Systems.International Conference on Computer Systems and Technologies-CompSysTech08: pp IIIA (14-1)-(14-7).