

## A proposed Mining System for Recognition of the Peripheral Arterial Disease and Analyzing Risk Factors

نظام تنقيبي مقترح لتمييز مرض قصور الشرايين المحيطية وتحليل العوامل المسببة

Assist Prof Ghazi Johnny  
Computer Science/Technical College of  
Management – Baghdad/dept. of IT  
Middle Technical University

Lecturer. Sarah Saadoon Jasim  
Computer Science/Technical College of  
Management – Baghdad/dept. of IT  
Middle Technical University

### **Abstract:**

Association rule techniques is considered as one of the most well-known techniques which has been employed to find the most dominant associations and correlations of the items within large data set. Thus, it can be used as useful method to analyze and fix many data mining issues. In this research, the association rule has been adopted as a proposal method for the cases that are related to peripheral arterial disease (PAD) risk factor.

PAD is one of the most common problem with high morbidity and mortality, which has many potential risks factors that might affect its severity. Such factors can be modified, and beneficial to good assessment and diseases' progress control and the effect of those treatments and its modalities, by the understanding of some relations between those factors. The gender factor among other factors was not discussed because the ratio of male to female is almost the same. Our data collected out from 50 patient persons with PAD which was analyzed and recognized. By the use of the proposed method that implemented within two stages: firstly we implemented stage a back propagation-NN was trained on records of patients with PAD, and secondly the association rules encoding was performed, here the relations between pre-disposing factors was extracted.

Our proposed mining system offers a faster and better data analysis and recognition; also it reduces time, efforts, and also explores the relations between factors. Therefore, we got the relations those been mentioned above, in which they are important in the process of PAD management in the future.

There was a strong relations among alcoholism, smoking ,diabetes and

hypertension with the presence of dyslipidaemia, and many other relations among these factors. That result goes with the results obtained by other studies that use traditional statistical methods.

**Keywords:-** Association Rule, PAD, Hypertension, Diabetes, Dyslipidaemia, Smoking , Alcoholism ,neural network, genetic.

## 1. Introduction:

### 1.1 Knowledge Discovery And Data Mining:

The vast growing and rapid development in data usage and internet technologies, has come out with the necessity of developing robust tools in order to manipulate such data sets and repositories. These tools are being used to analyze and interpret to extract the most significant and meaningful knowledge that contribute strongly in decision-making processes.

The term data mining which is also called Knowledge-Discovery in Data-bases (KDD), denotes to the techniques of insignificant extraction of implicitly unknown and potentially beneficial information out of databases' contents. Although data mining and KDD are popularly considered as equivalents, data mining is an aspect of the wider KDD process. However, the KDD process involves some steps that arrange from data collecting operations to the new knowledge formations. This recursive process includes the steps in below:

- Data cleansing or cleaning: in this step noisy and meaningless is illuminated the data set.
- Data integration: here various sources of data (frequently heterogeneous data), are gathered in a joint form.
- Data selection: here at this level, we decide and retrieve the data that are relevant to the required analysis of data collection.
- Data transformation or data consolidation: in this stage, the data that have been selected from the last step is transformed to other forms which are appropriate to the mining process.

- Data mining: this step is considered critical; however in this step we apply smart methods for extracting potentially-useful patterns.
- Evaluation of patterns: at this stage, the rigorously useful patterns that represent knowledge can be identified basing on some specific measurements.
- Knowledge representation: in this step the knowledge that has been discovered is represented to the user in visualized form. This step helps users to easily understand the results of mining the data [1,2].

### 1.2 Data Mining Techniques:

Under this context, there is a variety of methods that can be employed can be used in data mining: neural networks (NN), rough sets techniques, genetic algorithms (GA), Cluster Analysis, support vector machines (SVM), induction, association-rules mining, and data visualization [3].

### 1.3 The Neural Networks:

The neural networks (NN) are methods of computations involve the development of mathematical forms that have learning capabilities from data. Those methods represent the results of scientific investigations for nervous-system learning modeling [4,5].

Neural networks are seriously able to conclude the meaning out of imprecise/ complicated data, also they could be used for pattern extraction and detection in which are difficult to be recognized by humans and other simple techniques [5].

#### 1.3.1 Multi Layer Perceptrons (MLP):

Perceptron method is considered as the simplest method of neural networks, as in a single-layered perceptron NN operates as one neuron, and the multi-layered perceptron NN is simply an amplification of the basic perceptron idea which is usable for solving more complex problems [5]. A perceptron NN composed of single input layer, at least one hidden layers, and one output layer. Hidden layers is the power of the network that is it allows the network in extraction features of the input data. Back Propagation algorithm is one of the popular approaches used for training the MLP. BP passes twice each layer of the NN (first pass is forward and

the second pass is backward). This algorithm is trained in three stages [5,6,7]: (i) the training pattern is input forwardly, (ii) then the error of the associated data is calculated in BP, and (iii) finally the re-calculation of weights.

#### 1.4 Association Rules Mining:

Association rules are refer to the process of finding a particular relationships that link multiple objects among large data set. This process helps to explore a number of association rules over multiple layers from particular data set. For instance, a number of diseases may come with similar symptoms that appears frequently together. Since, association rules can come out with interesting associations in data bases and big data sets, they also can be employed to discover a number of meaning full patterns that serve different fields such as marketing, financial medical diagnosis. The thing that makes association rules very attractive field of study.

In this research, we present the conventional definition of A-priori algorithm. Which is used to implement association rules. The *support* of an item set  $I$  [ $\text{sup}(I)$ ], is known as the number of processes in the database containing  $I$ . The *minimum support* ( $\text{min\_sup}$ ), is a user predefined threshold. An item set is *frequent* if its support is not less than the  $\text{min\_sup}$ . An item set with  $k$  items is called a  $k$ -item set.

Let  $D$  be a set of transactions and  $I = \{i_1, i_2, \dots, i_m\}$ , an item set is a subset of  $I$ . Given  $X$  and  $Y$  are item sets, an *association rule* is of the form  $X \Rightarrow Y$ , where  $X \subset I$ ,  $Y \subset I$ , and  $X \cap Y = \Phi$ , where the  $\text{sup}(X \cup Y) \geq \text{min\_sup}$ , and the *confidence* of  $X \Rightarrow Y$  is not less than a predefined threshold,  $\text{min\_conf}$ , the *confidence* of  $X \Rightarrow Y$

is  $\frac{\text{Sup}(X \cup Y)}{\text{Sup}(X)}$ .

After discovering all frequent item sets, the algorithm of generating association rules uses the subsets of a frequent item set as antecedents to generate the rules [9,10,11].

The form  $X \Rightarrow Y$  with confidence 60% where  $X$  = computer and  $Y$  = software for example means that (60% of the customers who purchased a computer also bought the software).

And the form of  $X \Rightarrow Y$  is not equal to the form  $Y \Rightarrow X$  because

$$\frac{\text{Sup}(X U Y)}{\text{Sup}(X)} \text{ NOT EQUAL } \frac{\text{Sup}(Y U X)}{\text{Sup}(Y)} \text{ unless } X=Y \text{ which is called perfect rule.}$$

This must be taken into consideration when analyzing the results of association rules.

Also we used the Mean of the items confidence (MOC) which can be defined as follows:

$$\text{MOC}(A) = \{\text{conf}(A \rightarrow B) + \text{conf}(A \rightarrow C) + \text{conf}(A \rightarrow D)\} / 3.$$

The formula gives the mean of the occurrence of item A among the items B, C and D which means the influence of item A on the others [12].

### 1.5 Peripheral Arterial Disease (PAD):

PAD is one of the most common diseases that appears when fatty deposits hinder the blood flow toward legs. It is also known as peripheral vascular disease (PVD).

Several patients of PAD have no clear symptom. Nevertheless some of them have terrible pain in their legs while walking, this pain usually gone after a while. The medical term for this is "intermittent claudication".

PAD commonly captured through physical examination through measuring the blood pressure in arms and ankles.

PAD commonly classified under cardiovascular disease (CVD), since it influenced by blood pressure.

The main reason of having PAD is due to the fatty deposits that found inside the walls of the leg arteries which known as atheroma which mainly caused by cholesterol.

Atheroma causer narrowing the blood arteries and eventually hinder the blood flow toward the legs. This called atherosclerosis.

The chance of getting PAD increases when someone get older. The possibility of that is each one person of five is prom to get PAD especially when the age is over to years. The studies refer that the men are tend to get PAD more than women. Men tend to develop the condition more often than women.

There are some reasons of getting affected by PAD which are:

- Which considered as the most affecter reason,
- Diabetes,
- Hypertension,
- Cholesterol,

By inspecting the reasons of developing PAD, we may be able to reduce the chance of widespare such diseases.

Having PAD refers to the possibility of having more terrible and bad diseases with serious danger of CVD. Including heart attack.

If PAD gets worse that meant the legs are in the way of death (gangrene). [Figure 1] [13, 14,15, 16].



Figure (1) Gangrene of the limb

## 2. Methods:

### 2.1 Reaserch:

A prospective observational study was conducted from July the 1<sup>st</sup> 2014, to August the 1<sup>st</sup> 2015.

### 2.2 Study sample:

In this research, we collect a sample from two public hospitals which are located in Baghdad. All cases that have been records were belong to patients who were diagnosed to have PAD. The collected data was in the form of name, age, gender, medical history (especially blood pursue issues), smoking, alcohol, family history of atherosclerosis, blood sugar levels (HbA1c), recent serum lipid profile, dopplex scan results, and angiography results if performed.

Six risk factors (age, diabetes mellitus, hypertension, dyslipidaemia, smoking, alcohol intake) was selected to be examied using association rule. The patient sample was composed of 250. Using association rule and confidence of means (82) records analyzed [10]. After that we added 168 not recognized new records for the test of proposed method and make a comparison between old and new results we get.

### 2.3 The suggested method:

The system proposed includes two phases as shown in the diagram below:

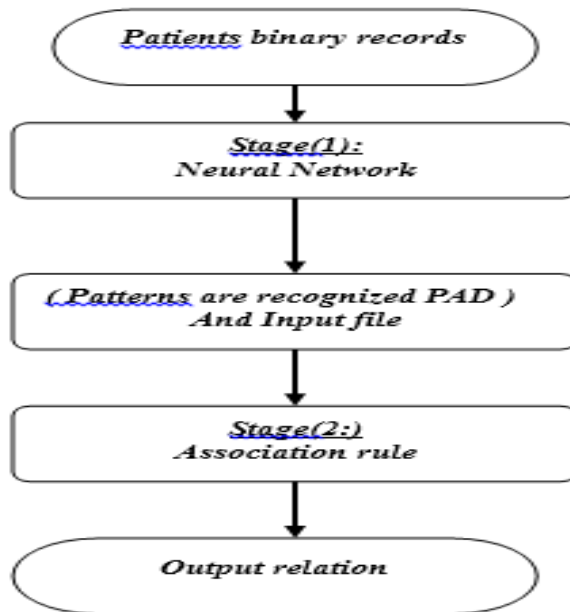


Figure (2) the suggested system diagram

#### 2.3.1. Neural network's stage:

The structure of back-propagation neural network is: 9 neurons in input layer. In hidden layer there are 7 neurons. And there are 3 neurons in the output layer.

In any layer, the number of neurons can affect the connections to the next layer. As much as the connections matrix is big, it means that the time and cost consuming is high. In other word a connection needs a mathematical operation (( $w_{ij} \times \text{input } i$ ), the input layer multiplied by the weight matrix). And the same mathematical operations are needed when we increase the hidden layer. This can be implemented with the rest of layers as shown is figure (3).



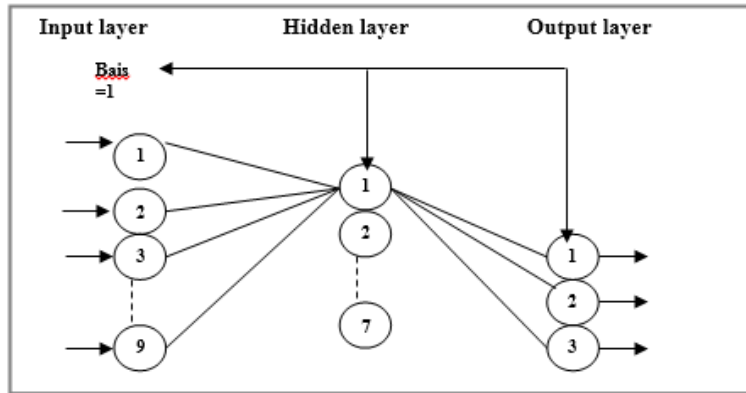


Figure (3) MLP neurons of the proposed method

### 2.3.2. Association rules' stage:

The Interactive KDD system is implemented in this system to speedup association rules process and confidence's mean. All the techniques above combined in the proposed method.

### 2.3.3. Implementation:

Some of doctors chose specific patterns that have been applied to train the neurons of (ANN) as shown in the Figure (4).

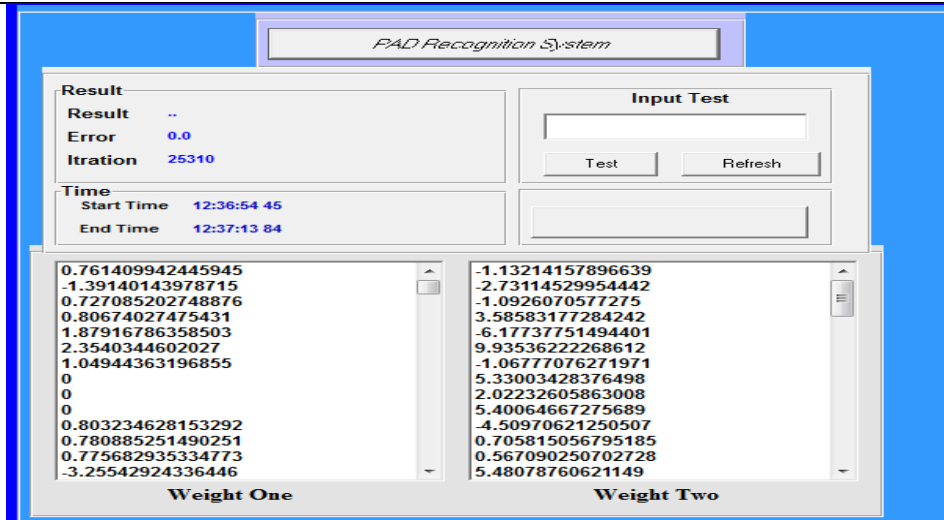


Figure (٤) The training phase of neural-network

Every passed pattern from out the first level combined into the file of input data for the second layer (AR) and the analysis are shown in figures (5,6).

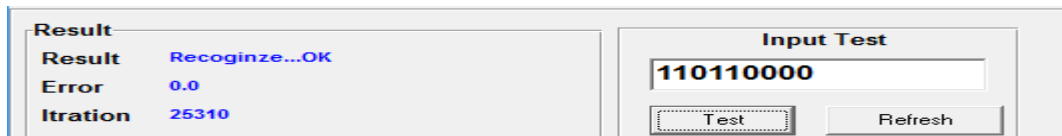


Figure (٥) displays the recognizing process of patterns

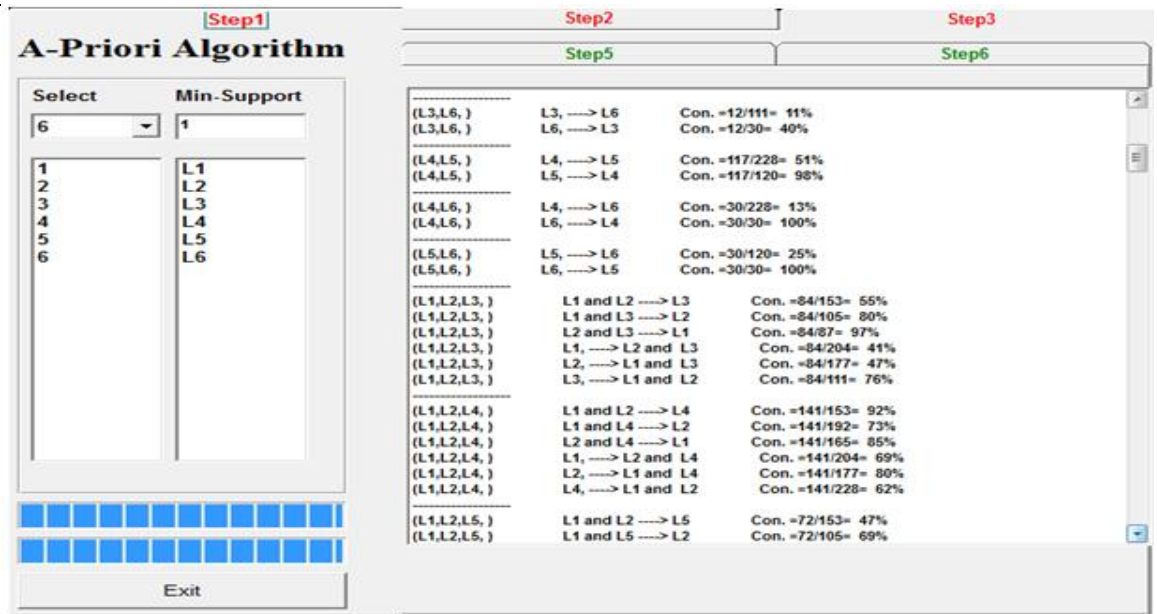


Figure (٦) shows a sample of extracted relations could be analyzed

### 3. Experiments and Discussion:

3.1 Analyzing the outcome relations of the proposed system in consideration to the risk factors of PAD, we have found many relations amongst those factors. In specific, the significant relations that might effect seriousness, pathogenesis, and course's disease, were extracted.

#### 3.1.2 Association-Rules results:

We worked on 6 parameters which are L1 =Age >=50 years, L2 =Diabetes, L3 =Hypertension, L4 =Dyslipidaemia, L5 =Smoking, and L6 =Alcohol consumption.

Regarding the age, 61% of patients were >=50 years, 71% were diabetics, 44% were hypertensive, 91% have dyslipidaemia, 48% were smokers and only 12 % were alcoholics[figure 7].

According to the previously extracted relations [10], and what we extracted in

this research, doctors preferred to depend on the rules that represent the most significant relations which come with high confidence values figure (6).

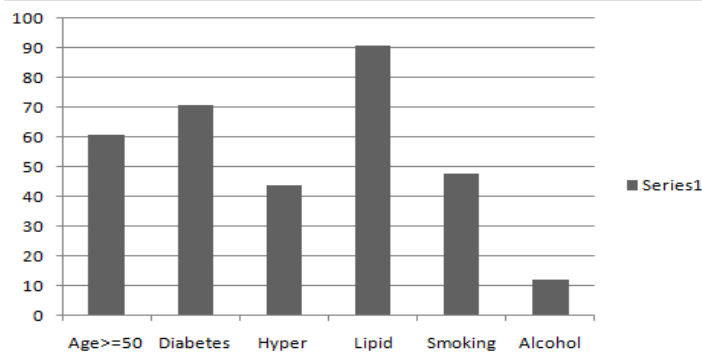


Figure (7) factors percentage from the given data

Among all the relation that have been extracted from the data, we select only the most significant relations that reflect high confidence.

1. All alcoholic patients (100%) were dyslipidaemics , and all were smokers ,the proposed system shows the same.
2. All smoker patients, who are  $\geq 50$  years of age, were dyslipidaemics, the proposed system shows the same.
3. All diabetics who smoke were dyslipidaemics, the proposed system shows the same.
4. All diabetics who are alcoholics were dyslipidaemics (and were smokers), the proposed system shows the same.
5. All hypertensive patients aged  $\geq 50$  years, who smokes were dyslipidaemics, the proposed system shows the same.
6. All hypertensive who are alcoholics, were aged  $\geq 50$  years, and all were diabetics, dyslipidaemics, and all are smokers, the proposed system shows the same.
7. All patients, with hypertension who smokes and have dyslipidaemia, were  $\geq 50$

years, the proposed system shows the same.

8. 98% of smoker patients have dyslipidaemia, the proposed system shows the same.

9. 95% of hypertensive were dyslipidaemics, the proposed system shows the same.

10. 94% patients aged  $\geq 50$  years have dyslipidaemia, the proposed system shows the same.

11. 93% of diabetics have dyslipidaemia, the proposed system shows the same.

12. 87% of diabetics aged  $\geq 50$  years, the proposed system shows the same.

13. 84% of dyslipidaemics, aged  $\geq 50$  years, the proposed system shows the same.

14. 79% of hypertensive patients were diabetics, the proposed system shows the same.

15. 78% of hypertensive patients with dyslipidaemia, are diabetics, the proposed system shows the same.

### 3.1.3 The proposed system:

In this section we present the results of mean of confidence values, and factors' effects.

#### 1. Occurrence of age ( $\geq 50$ years) with the other factors, [Figure8].

Dibetes 75%, hypertension 51%, dyslipidaemia 94%, smoking 52%, and alcoholism 12%.

Where the previous research [10 ] shows smoking 51%.

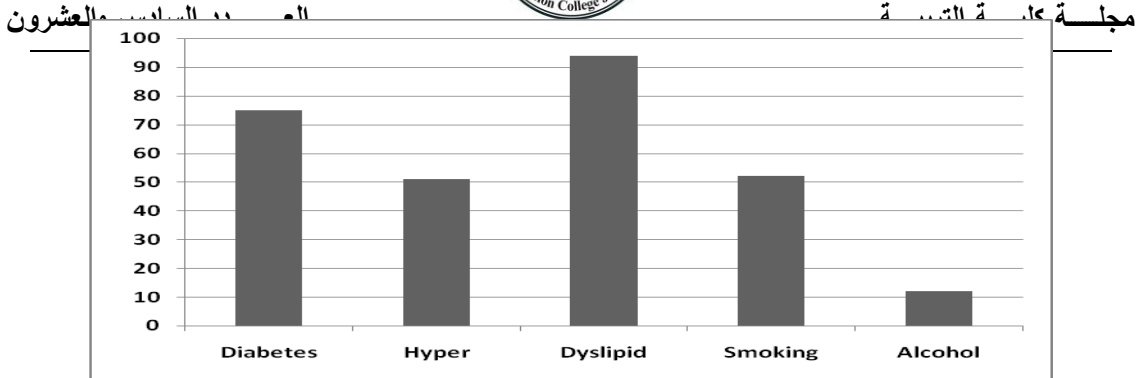


Figure (8) shows the Occurrence of age with the other factors

## 2. Occurrence of diabetes with the other factors, [Figure 9].

Age 87%, hypertension 49%, dyslipidaemia 93% , smoking 48%, and alcoholism 13%.

Where the previous research [10 ] shows smoking 47%, Age 86%, alcoholism 14%.

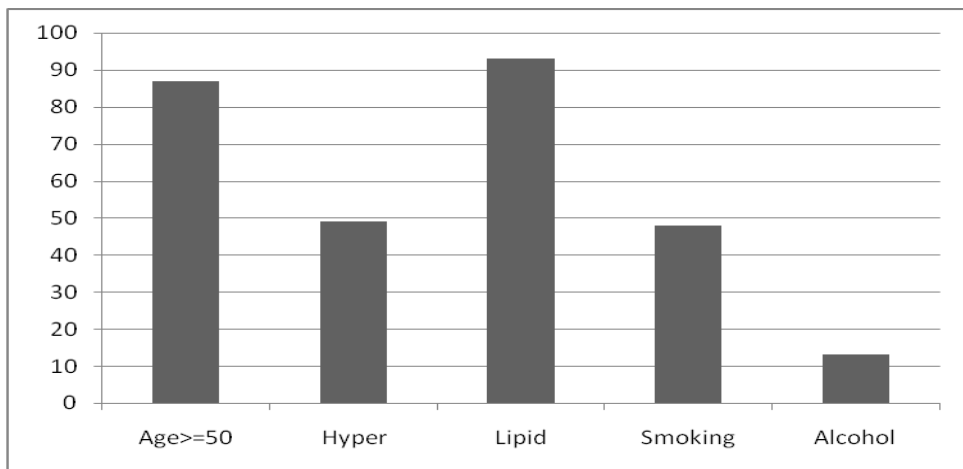


Figure (9) shows the Occurrence of diabetes with the other factors

## 3. Occurrence of hypertension with the other factors, [Figure 10].

Age 95%, diabetes 79%, dyslipidaemia 95% , smoking 49%, and alcoholism 11%.

Where the previous research [10] shows diabetes 78%.

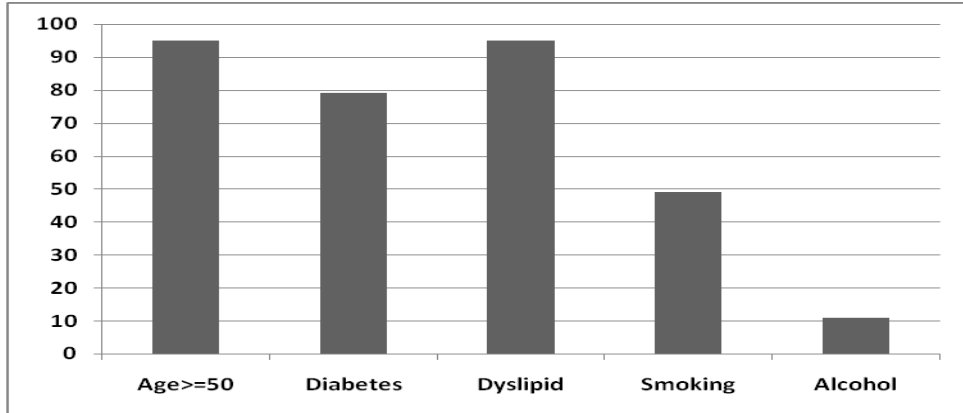


Figure (10) shows the Occurrence of hypertension with the other factors

4. Occurrence of dyslipidaemia with the other factors, [Figure 11 ].

Age 84%, diabetes 73%, hypertension 46 % , smoking 52%, and alcoholism 13%.

Where the previous research [10 ] shows diabetes 72%, smoking 51%,

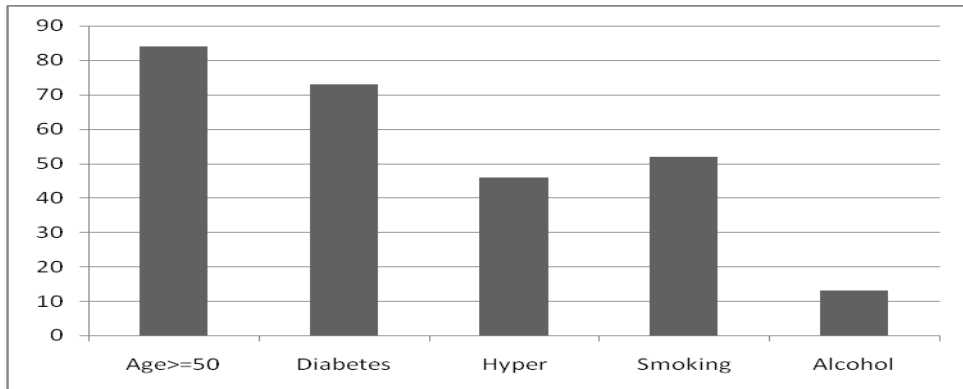


Figure (11) shows the Occurrence of dyslipidaemia with the other factors

5. Occurrence of smoking with the other factors, [Figure 12 ].

Age 88%, diabetes 71%, hypertension 45%, dyslipidaemia 98 % , and alcoholism 24%.

Where the previous research [12] shows diabetes 70%, and alcoholism 25%.

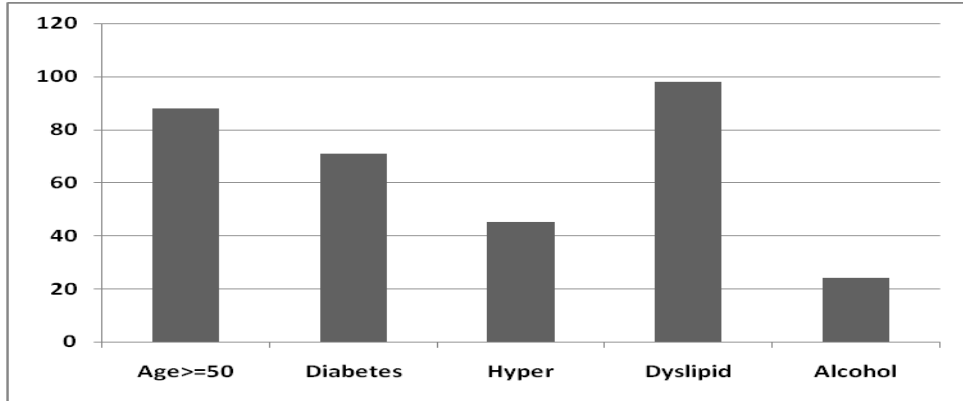


Figure (12) shows the Occurrence of smoking with the other factors

6. Occurrence of alcoholism with the other factors, [Figure 13 ].

Age 80%, diabetes 80%, hypertension 40%, dyslipidaemia 100%,and smoking 100%.

Where the previous research [10 ] shows the same.

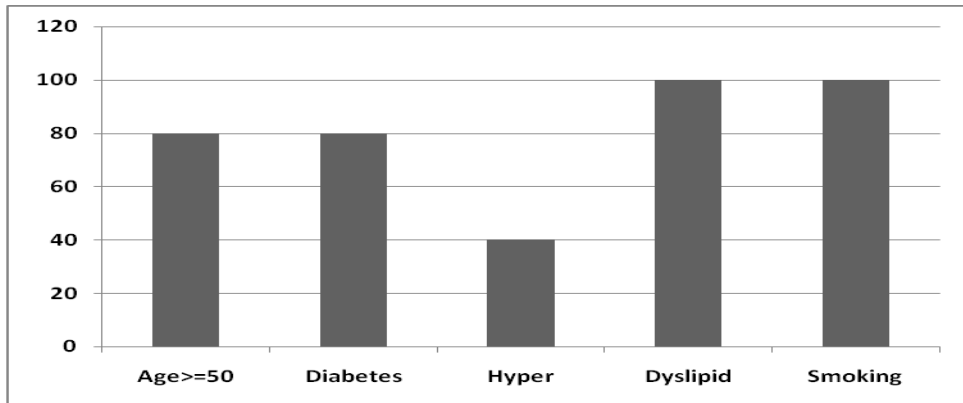


Figure (13) shows the Occurrence of alcoholism with the other factors

7. Mean of confidence (occurrence of the other factors with each single factor), [Figure 14].



Age 57%, diabetes 58%, hypertension 66%, dyslipidaemia 54% , smoking 65%, and alcoholism 80%.

Where the previous research [10 ] shows dyslipidaemia 53%.

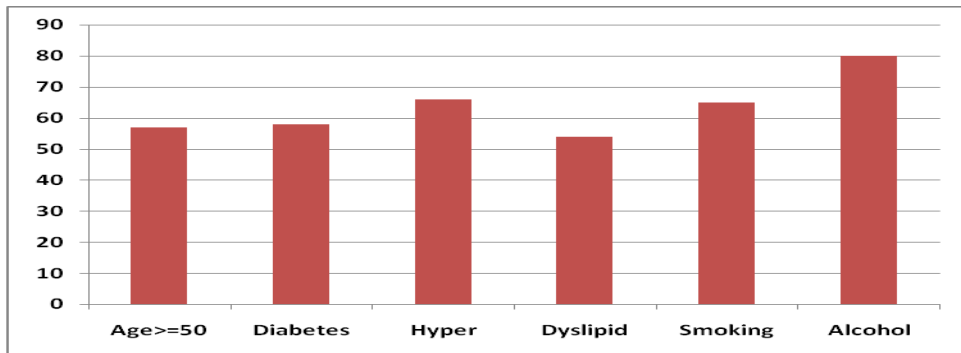


Figure (14) shows the Mean of confidence

### 3.2 Discussion:

PAD is a very common problem with high mortality. However, there are many risk factors that might affect its course, response to a treatment and seriousness. Generality of those risk factors can be modified. In addition, an understanding of their relationships can give more beneficial information for the purpose of better assessment, disease control progression, and our treatment model's performance modalities.

In this study, the ratio of male to female was almost the same, so that risk factor has been neglected.

Our proposed mining-system for recognition provides an interactive-powerful system for both faster and better analysis of input data. In this system, we accomplished the time and time saving processes, and discovered relationship amongst the risk factors. So, we use this method for analysis in this study, and we get the above mentioned relations, which are important for the future management of PAD.

#### 4. Conclusions:

The proposed system provides results that are too close to the previous research's results [10]. By adding 168 of recognized records (i.e. the system works well with the PAD standards researches) the following conclusions can be staffed:

which mean that the system goes with the standards of the PAD researches. The following conclusions can be staffed:

1. Alcoholism causes hyperlipidaemia and all PAD patients who are alcoholics, should be investigated for serum lipid levels, and treated according the results of serum lipid profile, and cessation of alcohol will decrease serum lipids, and will positively affect the course of the disease.
2. Smoking alone, has a very strong effect(98%) in causing dyslipidaemia, and this effect increases even more with diabetes(100%), and with age  $\geq 50$  years(100%). So any smoker with PAD should be investigated for serum lipid levels, and treated according the results of serum lipid profile, and cessation of smoking will decrease serum lipids, and will positively affect the course of the disease.
3. All diabetic patient who were smokers have dyslipidaemia, so diabetic patients should give up smoking for better control of disease, and better response to treatment.
4. Hypertension alone, has a very strong effect (95%) in causing dyslipidaemia, and this effect increases even more with age  $\geq 50$  years and smoking(100%), so cessation of smoking will eliminate two risk factors.
5. Hypertensive patients  $\geq 50$  years of age ,who are alcoholics, have dyslipidaemia , so give up drinking will eliminate two risk factors.
6. Any patient with PAD, who is  $\geq 50$  years should be investigated for dyslipidaemia because 94% of them will have increased serum lipids.
7. Any patient with PAD, who is diabetic, should be investigated for dyslipidaemia because 93% of them will have increased serum lipids, whatever their age is.

8. The results and percentages we get during this study, were very close to those obtained by other studies, that uses standards statistical methods.

The first seven conclusions show the traditional analysis of the association rule mining. By using the Mean of confidence, the following conclusions can be staffed:-

9. Regarding age, 94% of those  $\geq 50$  years, were dyslipidaemic, and 75%, were diabetic, so any patient  $\geq 50$  years with PAD, should be investigated for dyslipidaemia and diabetes, and control them in the affected patients, and prevent them in non affected patient, [figure 8].

10. 93% of diabetic patients were dyslipidaemic. A well known fact about the effect of diabetes on increased serum lipid levels, and this emphasizes the necessity of better diabetic control, to reduce two risk factors,[figure 9].

11. Among hypertensive patients, 95% of patients were  $\geq 50$  years, and this might be due to the fact that both hypertension and PAD are due to atherosclerosis, which increases with age. 95% were dyslipidaemics, and this might be the cause of hypertension due to atherosclerosis. 79% were diabetic, this might reflect the indirect effect of diabetes through atherosclerosis in causing hypertension,[figure10].

12. Among dyslipidaemics, 84%  $\geq 50$  years , this is also a known fact that dyslipidaemia is an aging process. 73% were diabetics, and this means that 28% of dyslipidaemics, are non diabetics, and this means that we should investigate and control serum lipids even if PAD patients, are not diabetics,[figure11].

13. The high occurrence of dyslipidaemia with smoking (98%), indicates, the obvious effect of smoking on elevating serum lipids, and PAD pathogenesis,[figure12].

14. 100% of alcoholics with PAD were dyslipidaemics, which indicates the strong effect of alcoholism on serum lipid. So giving up alcohol in those patients, will eliminate two risk factors,[figure 13].

15. [Figure 14] summarizes all the previous findings [Figure 8-13].

16. It is recommended to include the genetic factor along with other factors in the



---

system as it does not recognize patient record's (i.e. full zero patterns).

17. Finally, we recommend employing association rule mining and mean of confidence in other medical researches for better and faster analysis.

## REFERENCES :

- [1] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, eds.,  
Advances in Knowledge Discovery and Data Mining. Menlo Park, CA:AAAI/MIT  
Press, 1996.
- [2] K. J. Cios, W. Pedrycz, and R. Swiniarski, Data Mining Methods for  
Knowledge Discovery. Dordrecht: Kluwer, 1998.
- [3] U. Fayyad, G. P. Shapiro, and P. Smyth, "The KDD process for extracting useful  
knowledge from volumes of data," Communications of the ACM, vol. 39, pp. 27-  
34, 1996.
- [4] H Lu, R Setiono, H Liu. Effective Data Mining Using Neural Network. IEEE  
Transactions on Knowledge and Data Engineering, 1996, 8(6):957-961.
- [5] A. Al-taei, Automated classification of game players among the participant  
profiles in massive open online courses, Cankaya University, 2015.
- [6] L. Fasett, "Fundamentals of neural networks" Prentice Hall International Inc.,  
1994.
- [7] M. Yilmaz, A. Al-taei, Rory V O'Connor, "A Machine-Based Personality  
Oriented Team Recommender for Software Development Organizations". In  
Systems, Software and Services Process Improvement, Sep 30, pp. 75-86, Springer  
International Publishing, 2015.
- [8] C. gershenson "Artificial Neural networks for beginners",  
Gershenson@sussex.ac.uk.  
1993.
- [9] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", In  
Proc. of Int. Conf. on Very Large Data Bases (VLDB'94), Santiago, Chile,  
September 1994, pp. 487-499.
- [10] Johnny G, Developing A-priori Algorithm for Fast Mining Association Rules ,

Al-Taqani, Refereed Scientific Journal, Foundation of Technical Education, Vol. 22 , No.5, 2009, Baghdad-Iraq.

[11] Johnny G, Interactive KDD system for fast mining association rules, Al-Taqani, Refereed Scientific Journal, Foundation of Technical Education, Vol. 22 , No.5, 2009, Baghdad-Iraq.

[12] G. Johnny and Dr M. ALassel “Association Rules Mining Analysis of the Assessment of the peripheral Arterial Disease Risk Factors” 2nd conference, Technical College of

Management – Baghdad 28-29 /11/2012.

[13] Schwartz's Principles of surgery f.Charles Brunicardi ,8<sup>th</sup> Edition,chap 22,2004.

[14] Rutherford's Vascular surgery 7<sup>th</sup> edition vol1 Jack.l.Cronenwett and Johnson,K.wayne Johnston. Saunder Elsevier,Section3,chap 25,chap 26,chap 27,chap 28,chap 29, 2010.

[15] Sabiston DC:Disorder of the arterial system.In Textbook of surgery edited by David C. Sabiston .14 ed ,vol.1 philadelphia W,.B.Saunders Company,1991 1:1618-1722.

[16] Norgren L, Hiatt WR, Dormandy JA, Nehler MR, Harris KA, Fowkes FG. Inter-Society Consensus for the Management of Peripheral Arterial Disease (TASC II). Eur J Vasc Endovasc Surg 2007;33:S1–S75.