Iron-Deficiency Anemia Detection From Hematology Parameters
By Using Decision Trees

Sengul DOGAN¹ and Ibrahim TURKOGLU²
¹Firat University, Department of Bioengineering, 23119 Elazig - TURKIYE
²Firat University, Department of Electronics and Computer Science, 23119 Elazig - TURKIYE
vsenguldogan@hotmail.com
(Received: 14.12.2007; Accepted: 01.03.2008)

Abstract: In this study a decision support system has been designed from the biochemistry blood parameters which will be very helpful for and will make everything easier for the physicians in the diagnosis of Iron-deficiency anemia. Based on Pattern Recognition Process, the system operation is achieved via the decision trees structure which is related as one of the data mining techniques. The basic characteristic of the hematology parameters that is, Serum iron, Serum iron-binding capacity and Ferritin parameters are used in the process of entering the system and finally Anemia (+) and Anemia (-) results have been evaluated at the end of this process. Data of 96 patients are evaluated in the projected system. The results of the decision support system have completely matched with those of the physicians’ decisions.

Keywords: Pattern recognition, Data Mining, Feature extraction, Hematology Parameters, Expert system.

Karar Ağacı Yöntemini Kullanarak Biyokimya Verileri ile Anemi Teşhisi

Özet: Bu çalışmada, kan biyokimya parametreleri ile demir eksikliği anemisi teşhisinde, hekime yardımcı olacak ve kolaylık sağlayabilecek bir karar destek sistemi oluşturulmuştur. Örünüt tanıma süreci esas alınmış olup, sistemin işleyişi veri madenciliği tekniklerinden olan karar ağaçları yapısı ile sağlanmaktadır. Sisteme giriş olarak, biyokimya parametrelerinden demir eksikliği anemisi hastalığı için temel belirleyiciler olan Serum demiri, Serum demir bağlama kapasitesi (SDBK) ve Ferritin enzimleri kullanılarak, çıkış olarak ta Anemi(+) ve Anemi(-) değerlendirmelerinde bulunmuştur. Tasarlanan sistemde 96 hasta verisi değerlendirilmiştir. Karar destek sisteminin sonuçları, doktorun verdiği kararlarla tamamen örtüşmüştür.

Anahtar Kelimeler: Örünüt Tanıma, Veri Madenciliği, Özellikle Çıkarma, Hematolojik Parametreler, Uzman Sistem
1. Introduction

Anemia is defined as the decrease of the hemoglobin amount in the blood under the values accepted as normal according to the age and sex of the person [1]. Although it is physiopathologically defined as the decrease of the total eritrosit mass in the blood circulation, it is functionally defined as the decrease of the oxygen carriage capacity of the blood and so as the tissue hypoxias. The oxygen carriage capacity of blood decreases due to the decrease in the eritrosit numbers and/or the hemoglobin amount. In daily application, hemoglobin concentration (hb), hematotic value (Hct) and eritrosit number are used in the evaluation of the eritrosit mass instead of less common, more expensive and long-lasting method of the direct measurement of the eritrosit mass. For example “these parameters are explained in an order as the number in g/dl, % and mm$^3$ and they reflect the eritrosits in comparison with the plasm, the liquid they are in. That’s why they are affected with the changes in the improving plasm compartment without a change in the eritrosits. So, while evaluating the anemia, it should be noticed whether there is an increase in the plasm volume (hemodulation) or not and the measurements should be done again after the necessary treatment is applied or after a certain time. In every three parameters it shows a certain dispersion in normal population. This dispersion is affected by some important factors such as age, sex, race and the altitude of the living place and the given limits include 95% (arithmathic average +/- two standart difference) of society [2]. It has some kinds such as Iron-Deficiency [3], hemolitic [4], megaloblastic [5], aplastic [6,7] and anemia of chronic illnesses [8]. The Iron-Deficiency Anemia, that we take into consideration among them is defined as decrease of the red blood cells related with low amount of Fe. This is the most common kind of Anemia. Fe is an important part of hemoglobin which is an oxygen carrying pigment in blood. Anemia occurs slowly as a result of the use of the Fe stores in the body and the bone marrow [9]. Observed often especially in the little children and in women in fertility age, The Iron-Deficiency Anemia weakens the immunity system and negatively affects both intellectual and physical development in children and increases the infectious sensitivity [10]. Anemia is seen is the fourt place in an order in women and in the eight in men. It is seen between about the age of 38 and 62 in women and between 33 and 70 in men [11].

As a rule, serum enzyme levels are taken as the base in the diagnosis of the Iron-Deficiency. The Hematology Parameter taken into consideration in this illness are serum, Fe, serum Fe holding capacity and Ferritin [2].

These enzymes form great data stacks when used in each person’s diagnosis. Also, as these kinds of researches in the biology and medical science world increase, this will cause the increase in different kinds of data related with the real-life cases [12]. It became quite difficult for people to be able to comprehend and change hundreds of characteristics and thousands of images into meaningful knowledge in modern medical science [13, 14] DNA [15, 16] and protein synthesis [17], biological state measurements [18], graphs [19] and enzymes [12] are some of the main types of data which are still being used. As the data increased it became quite a boring and difficult work for the medical science experts to prepare a guide that will be used in the processes of reading, simplifying, classifying the findings and making a decision at the end. Also a lot of findings hidden in these data have remained as the data stack in this way. It is necessary to get these data automatic in order to obtain useful information. Getting information from the databases or the data mining is the kind of method mostly used for solving these kinds of problems [20]. Data mining is used for both getting the information related with the text and the information of the image data of an individual patient and then for matching each patient’s entry [21]. In this context data mining has been applied so successfully in many fields of medical science [20-22]. For instance, there are a lot of work that can be seen in the literature related with discovering certain rules in the diagnosis and treatment of acute ailments: applying the determined rules and making complementary data mining for defining the enlargement states of the working procedures of the refined, shared, organized and produced information in the
system [23] and also there are some works related with projecting a Data Acquiring System for the baby deaths due to the Hypoplastic Left Heart Syndrome and making quick and true guesses using the acquired data [24]. In these works, data stores are known as the clinical stores in medical field [12]. Data warehouses in medicine field are called clinical warehouse too. Then, those clinical stores containing biological, clinical and administrating data unit the patients information. Thus, the possibility of the usage of the systems related with the patient is improved [25].

In this study, a decision support system will be helpful for the diagnosis of Iron-deficiency anemia is improved to be used for the aim of classifying the decision tree construction from the data mining techniques. Hematology parameters from biochemistry data were used for application and 96 patient’s data have been evaluated successfully.

2. Theoretical View

In this part, the informations related with the easier conception of the decision support system’s construction in Iron-deficiency anemia diagnosis is introduced in the shape of sub-classifications.

2.1. Pattern recognition

Pattern recognition can be divided into a sequence of stages, starting with feature extraction from the occurring patterns, which is the conversion of patterns to features that are regarded as a condensed representation, ideally containing all the necessary information. In the next stage, the feature selection step, a smaller number of meaningful features that best represent the given pattern without redundancy is identified. Finally, classification is carried out: a specific pattern is assigned to a specific class according to its characteristic features, selected for it. This general abstract model, which is demonstrated in Figure 1, allows a broad variety of different realizations and implementations. Applying this terminology to the medical diagnostic process, the patterns can be identified, for example, as particular, formalized symptoms, recorded signals, or a set of images of a patient. The classes obtained represent the variety of different possible diagnoses or diagnostic statements [26]. The techniques applied to pattern recognition use artificial intelligence approaches [27].

![Figure 1. Pattern Recognition System](image)

2.2. Hematology parameters

**Serum iron and total iron-binding capacity (TIBC):** Iron is necessary for the production of hemoglobin. Iron is contained in several components. Transferrin (also called siderophilin), a transport protein largely synthesized by the liver, regulates iron absorption. High levels of transferrin relate to the ability of the body to deal with infections. Total iron binding capacity (TIBC) correlates with serum transferrin, but the relation is not linear. A serum iron test without a TIBC and transferrin determination has very limited value except in cases of iron poisoning. Transferrin saturation is a better index of iron saturation; it is evaluated as follows: Transferrin saturation(%) = Serum iron x 100 / TIBC

The combined results of transferrin, iron, and TIBC test are helpful in the differential diagnosis of anemia, in assessment of iron-deficiency anemia, and in the evaluation of thalassemia, sideroblastic anemia, and hemochromatosis.

**Ferritin:** Ferritin, a complex of ferric (Fe^{2+}) hydroxide and a protein, apoferitin, originates in the reticuloendothelial system. Ferritin reflects the body iron stores and is the most reliable indicator of total body iron status. A bone marrow examination is the only better test. Bone
marrow aspiration may be necessary in some cases, such as low-normal ferritin and low serum iron in the anemia of chronic disease. The ferritin test is more specific and more sensitive than iron concentration or TIBC for diagnosing iron deficiency. Ferritin decreases before anemia and other changes occur [28].

2.3. Data mining

Data mining is the work of achieving the information that enables us to find the relations among the great numbers of data which will be very helpful for making guesses about the future by using a computer programme. That is, the aimed information is obtained by processing the data. Data mining is used in many fields such as biomedicine, gene functions, data analysis of DNA arrangement pattern, diagnosis of illnesses, retail data, telecommunication industry, guesses about selling, financial analysis and astronomy [29]. For instance, some standart softwares used in a medical diagnosis system were problematic because of the unsystematic data, the absence of control, the usage of too many various kinds, being unable to make a consistent and systematic analysis on databases, comparing the examples and determining the critical differences. So, the physicians were in the opinion that these softwares were short of many characteristics. In this context, data mining which makes it possible to make a basic analysis on the graphs related with both the digital parameters and morphemes by the help of an expert information system and which stores the processes, administers the necessary information and also which includes a very useful construction such as a reference for decision, took an active position. Data mining and techniques are also accepted in the medical science world because they make everything easier for the experts and provides a necessary and important help for the practitioners [30]. In fact data mining is evaluated as a part of the information discovery process both in medical science and in the other fields [31]. Data mining stages are presented in Figure 2 [32].
Data mining has an interaction with the user and the database. Interesting data patterns are showed to the user. And also, they can be saved in the database if wanted. According to this, data mining goes on till the hidden data patterns are found [33].

Firstly the necessary data are taken, classified and then processed while obtaining meaningful information from the databases. It is an important problem to be able to save the state of a patient and to predict the characteristic of the data such as all the laboratory test results, findings and signals of all the patients. These points are also problems in machine learning and data mining which works in many fields such as classification and problem detection [14]. Classification in data mining is used for an automatic definition of the interesting object in great data and for the information discovery in the applications including the classification of the trend in the market [34]. There are a lot of methods used for classifying these data. In data mining the decision trees among these techniques, which have the characteristics of making guesses and defining, have the most common usage in the classification models because of the following reasons: [35]

- Inexpensiveness of the organization,
- Easiness in interpretation,
- Their possibility of being easily integrated with the database systems,
- More reliability.

2.4. Decision trees

After the data are created in decision trees, the rules can be written down on the tree from root to the leaf (If-Then rules). Getting decision in this way provides the confirmation of the result of the work of data mining. These rules can be checked in terms of application by showing them to an expert whether they are meaningful or not [36]. The decision trees are used in the analysis and applications of classifying various cases into low-mid or high risk groups; creating rules in order to predict the future events; definition of the relations of the certain subgroups; the union of the categories and getting the most effective decisions by the help of the medical observation [37,27]. Structure of rule extraction is given in table 1 [38].

Table 1. Rule Extraction

<table>
<thead>
<tr>
<th>Rule 1</th>
<th>Decision 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 2</td>
<td>Decision 2</td>
</tr>
<tr>
<td>Rule N</td>
<td>Decision N</td>
</tr>
</tbody>
</table>

Here each one of decision class's parts that will be taken in rules are represented with \( \Omega \) and the result that will be arrived from that class parts are represented with \( \Psi \). The preparation of such a decision list constitutes an important step in a medical application and a decision tree can be created easily according to these rules [20].

A decision tree algorithm that is created according to these rules constitutes a decision tree in the context of information obtainment [39].

Figure 3. Decision Tree Structure

3. Developed Method

The most common and main tests that help the physicians decide on a diagnosis are biochemistry tests which are mostly successful in diagnosing. An Iron-deficiency anemia diagnosis can be obtained by controlling the hematology parameters which were taken into consideration for the Iron-deficiency anemia diagnosis in biochemistry test results. The construction of the improved decision support system is showed in Figure 4 based upon the pattern recognition given in part 2.1 [26]. The functions of system are constituted of these following steps:
Collecting the data is the step in which the ones those are proper for the goal are chosen. The necessary data were obtained by collecting the biochemistry test results of the patients applied for the internal medicine and by analyzing the ones related with the hematology parameters with the direction of the physicians. The database was created by choosing the test results of the Anemia suspected-ill from the data collected.

**Step – 2: Feature Extraction and Classification**

This is the most important step of decision tree based decision support system which was developed for processing the data in the collected database. The data collected are being processed in this stage. For this, the rules underlined in table 2 and determined according to the physician’s knowledge and then the decision tree in figure 5 was obtained in relation with this knowledge.

<table>
<thead>
<tr>
<th>Decision Tree Index</th>
<th>Rules</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>if (sd &lt;= 37) and (sdbk &gt;= 340) and (ferritin &lt;= 13)</td>
<td>Anemia(+)</td>
</tr>
<tr>
<td>b</td>
<td>else if (sd &lt;= 37) and (sdbk &gt;= 340) and (ferritin &gt; 13)</td>
<td>Anemia(+)</td>
</tr>
<tr>
<td>c</td>
<td>else if (sd &lt;= 37) and (sdbk &lt; 340)</td>
<td>Anemia(+)</td>
</tr>
<tr>
<td>d</td>
<td>and (ferritin &lt;= 13)</td>
<td>Anemia(-)</td>
</tr>
<tr>
<td>e</td>
<td>else if (sd &lt;= 37) and (sdbk &lt; 340) and (ferritin &gt;= 340)</td>
<td>Anemia(-)</td>
</tr>
<tr>
<td>f</td>
<td>and (ferritin &gt; 13)</td>
<td>Anemia(-)</td>
</tr>
<tr>
<td>g</td>
<td>else if (sd &gt; 37) and (sdbk &lt; 340) and (ferritin &lt;= 13)</td>
<td>Anemia(-)</td>
</tr>
<tr>
<td>h</td>
<td>and (ferritin &gt; 13)</td>
<td>Anemia(-)</td>
</tr>
</tbody>
</table>

Among the Hematology parameters given in the biochemistry test results the normal values of hematology parameters are given in table 3.

<table>
<thead>
<tr>
<th>Hematology Parameters</th>
<th>Normal values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serum Iron</td>
<td>37-157 µg/dl</td>
</tr>
<tr>
<td>Serum iron-binding capacity</td>
<td>210-340 µg/dl</td>
</tr>
<tr>
<td>Ferritin</td>
<td>13-400 ng/ml</td>
</tr>
</tbody>
</table>
Step – 3: Evaluation

In this step, the data obtained after the feature extraction and classification process are presented after the evaluation. The results reached to \( a, b, c, d \) leaves of the decision tree structure, (figure 5) projected for this purpose, are evaluated with the Iron-deficiency anemia (+) diagnosis and the results reached to the \( e, f, g, h, j \) leaves of the tree are evaluated directly as Iron-deficiency anemia (-).

4. Results

The developed decision support system is experienced on the biochemistry test results of the 96 patients. In the system Iron-deficiency anemia (+) diagnosis was certain in 76 of the 96 patients and the other 20 were healthy. And we saw that these results have successfully matched with the physicians’. This success rate has clearly and undoubtedly showed the reliability and the effectiveness of the projected decision support system because of the perfect match between the physicians’ decisions in the Iron-deficiency anemia diagnosis and the results of the developed system.

Table 4. Performance of The Decision Support System

<table>
<thead>
<tr>
<th>Anemia</th>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of samples</td>
<td>76</td>
<td>20</td>
</tr>
<tr>
<td>Correct classification</td>
<td>76</td>
<td>20</td>
</tr>
<tr>
<td>Incorrect classification</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

5. Discussion and Conclusion

The increase in the knowledge obtained under the light of biomedical searches, the processes of reading, simplifying, classifying these findings and making a decision are being so complex day by day. These processes are automated after the data mining took a part in this field. It has considerably helped the medical experts and made it easier to prepare a guide. Also, a lot of findings hidden between the data mining and the data stacks obtained in these kinds of fields, has turned to be useful information in this way.

The advised system is based on pattern recognition on which so many clever diagnosis systems are constituted. There, the decision trees, which take an important part in data mining for the feature extracting and classification stage in the pattern recognition process, have been used. Thus, a decision support system is projected based on pattern recognition and the function of the system is based upon the data mining methods.

The developed decision support system will considerably be helpful for the expert physicians and practitioners for the interpretation of the illnesses. This system construction can also be used in the diagnosis of every illnesses in which the criteria having certain parameters can be controlled.

Acknowledgments - We want to thank, Department of Biochemistry, Ufuk Medical Center, Elazığ, TURKIYE for providing the Biochemistry tests to us Dr. Bilal USTUNDAG and Dr. Osman Metin IPCIOGLU.

References

10. Toksöz, P., Ceylan, A. (2000). Hemotokrit measurement of school-age children in Diyarbakir district and the determinanition of
anemia frequency, *Dicle Medical Magazine (Journal of Medical School)* C:27 S:1.


34. Agrawal, R., Mehta, M. (1996). The Quest Data Mining System, IBM Almaden Research Center San Jose 6s California, U.S.A.


