Automated Diagnosis of Iron Deficiency Anemia and Thalassemia by Data Mining Techniques

Maysam Hasani a,b, Ali Hanani a,c*
a Department of Computer Engineering, College of Technical and Engineering, Kermanshah Branch, Islamic Azad University, Kermanshah, Iran.
b Department of Computer Engineering, College of Technical and Engineering, Kermanshah Science and Research Branch, Islamic Azad University, Kermanshah, Iran.
c Department of Computer Engineering, Songhor and Koliaei Branch, Islamic Azad University, Songhor and Koliaie, Iran

Corresponding Author. Tel.: +98 912 309 9599
E-mail addresses: Ali.Hanani@iauksh.ac.ir (A. Hanani).

Abstract
In the present paper, three types of anemia including iron deficiency anemia (IDA), β-thalassemia trait and α-thalassemia trait (cis and trans) have been investigated. Detection of these three types of anemia is difficult, as their blood characteristics are similar to each other. Also, specialists use some tests to diagnose these disorders that those tests are very time consuming and costly. Thus, providing a model for accurate diagnosing of these kinds of anemia is extremely important. The present study mainly focused on a simple complete blood count (CBC) test instead of using some tests to detect and differentiate between these kinds of anemia in Weka software. In order to suggest an algorithm with the highest accuracy and the lowest mean absolute error, five classification algorithms and a vote algorithm were used and the performance of the vote algorithm was compared with the performance of those five algorithms. The results of this study indicated that combining J48, IBK and Naive Bayes algorithms using voting algorithm with all the features and reduced features had the highest performance with the accuracy of 96.343 and 96.2169, respectively. Using hybrid algorithm (vote) demonstrated that hybrid algorithm increases diagnosis accuracy and decreases error rate in comparison with the single classifiers.

Keywords:
α-thalassemia trait, β-thalassemia trait, Iron deficiency anemia (IDA), Vote algorithm, Complete blood count (CBC), Data mining techniques

1. Introduction

Data mining analyzes and reviews visible data, finds unknown relationships among the data and also summarizes them in an understandable and valuable way. Diagnosis of the diseases is one application of data mining techniques in medicine [1]. Anemia is called as the most common nutritional deficiency in the world and the most common blood disorder in infancy and childhood. Iron deficiency anemia (IDA) is very common among children and women throughout the world, especially in developing countries [2]. Another common anemia is thalassemia. The most common hereditary hemoglobinopathies is thalassemia in the world [3] that is one of the problems of many countries and also the most common inherited disorder in Iran [4]. Thalassemia includes alpha and beta, types of alpha are silent carrier, α-Thalassemia trait, Hb H disease and Hb Bart's hydrops fetalis syndrome [5] and types of beta are β-thalassemia trait, thalassemia intermedia and thalassemia major [6].

It is a challenging task to diagnose accurately the blood diseases just using Complete Blood Count (CBC) test, because these blood disorders (iron deficiency anemia, β-thalassemia trait and α-thalassemia trait) have many similar characteristics [5]. If specialists do not examine these three disorders on CBC test simultaneously, they will probably obtain a wrong conclusion. Regarding their blood similarities and the number of the patients, data mining can be used to recognize optimal disease pattern whereas specialists have a lot of difficulties to detect them using just CBC and they need other tests to diagnose accurately.

Some studies about blood disorders have been done all over the world. For instance, in an article, Nurul Amin and Ahsan Habib aimed to diagnose some types of blood disorders using classification techniques. The CBC test, age and gender were used with J48, MLP and Naive Bayes algorithms [7].

Similarly, Saichanma, et al. used J48 algorithm and data mining techniques to predict the abnormality of peripheral blood smear. RBC was the main focus of this study [8].

In a study, Setsirichok, et al. classified blood characteristics by C4.5, naive Bayes and multilayer perceptron to screen thalassemia. Two characteristics of CBC test and six known types of haemoglobin via high performance liquid chromatography were used [9].

In another study, Wongseree, et al. made a distinction between people with symptoms of thalassemia and thalassemia patients to screen thalassemia by neural networks and genetic programming based on decision tree

Manuscript received April 5, 2017
Manuscript revised April 20, 2017

and some tests (20 features) including red blood cells (RBC), reticulocytes, platelets and their subsets [10].

In the previous studies, IDA and different types of thalassemia have been investigated using some tests and data mining techniques or just two types of Thalassemia (alpha carriers and β-thalassemia trait) have been examined by a test and data mining techniques [11]. However, the present study examined IDA, β-thalassemia trait and α-thalassemia trait includes cis (deletion of two α-genes on the same chromosome) and trans (deletion of two α-genes on the different chromosome). Regarding that the first test for the diagnosis of anemia is CBC test, this study aimed to help specialists to detect the type of anemia accurately, avoid doing irrelevant tests, reduce time and the price of detection only using the CBC test. This study also offered a hybrid model to improve the efficiency and accuracy of diagnosis and decrease the error. Here, the blood characteristics were reduced to understand whether it can improve the accuracy or not.

So blood characteristics of CBC test were collected, discretized and after that classified using 10 fold cross validation. Since J48 (C4.5) [7], Naive Bayes [9], Random Forest [12], k-nearest neighbor (KNN) or IBK [13], and Multi-Layer Perceptron (MLP) algorithms [10, 11] had high accuracy in previous works to detect diseases, especially anemia, they have been used in the data analysis of this study as well. Vote algorithm is the proposed algorithm in this study, which is the combination of three algorithms including KNN, naive Bayes, and J48. These three classification algorithms were combined for the following reasons: 1) they are from different classifier families and produce different models that classify entries differently. The combination of algorithms will be useful when some classifiers give different ideas about the inputs [14], 2) these algorithms showed high accuracy in previous studies and the current study. By examining different combinations of these five algorithms, the three above mentioned algorithms were selected as the best combination which combined based on majority voting.

The organization of this article is as follows: In Section 2 materials and methods are fully explained, in section 3 results and discussion are given, finally conclusion is drawn.

2. Materials and Methods

Since the main purpose of this paper was the diagnosis and discrimination of IDA, α-thalassemia trait and β-thalassemia trait, to detect these subjects some tests including CBC, hemoglobin electrophoresis, ferritin and PCR in Hafeziyeh, Farhangian and Reference (centeral) laboratories in Kermanshah, Iran were investigated. The data set for this study consisted of 793 samples that 184 individuals of them were with IDA, including 71 males and 113 females, 200 of them were normal subjects including 100 males and 100 females, 203 β-thalassemia trait including 105 males and 98 females, and also 206 α-thalassemia trait including 104 males and 102 females. Also in this study, data such as age and gender of all of these subjects were checked, all the women and men were over the age of 12 years. Then, these people have been divided in different groups according to diagnostic criteria.

2.1. Different Types of Diagnostic Tests

Typically, four tests use to diagnose the anemia disorders which are as follows:

2.1.1. CBC Test

CBC is a laboratory test which is one of the most frequent blood tests to measure overall health and determine a wide range of diseases [15]. The CBC test features include white blood cells (WBC), red blood cells (RBC), hemoglobin (Hb), hematocrit (HCT), mean corpuscular volume (MCV), mean corpuscular hemoglobin (MCH) and mean corpuscular hemoglobin concentration (MCHC), red blood cell distribution with (RDW) and platelets (PLT).

2.1.2. Hemoglobin Electrophoresis

Hemoglobin electrophoresis test is a kind of blood test which is used to detect subsets of hemoglobin [16].

2.1.3. Ferritin

Ferritin is a combination of Ferric Hydroxide, Protein and Apoferritin that is made from reticuloendothelial system. Ferritin is the iron stored in the body and a test to detect IDA and it is more sensitive than other iron tests [17].

2.1.4. PCR

Polymerase chain reaction (PCR) is a molecular test to diagnose genetic disorders. Types of PCR tests are used to diagnose α-thalassemia disease [5]

2.2. The Proposed Algorithms

This study suggested the following algorithms:

2.2.1. C4.5 Algorithm

C4.5 decision tree is known as J48 in Weka software. C4.5 builds decision trees based on the training data set. This algorithm makes a decision tree for data sets using recursive classification. The decision grows up using depth-first method. C4.5 algorithm consisted of probable tests that can divide data sets and select the informative

tests. The test is considered equals to the number of defined amounts of features in every discretized feature [18].

2.2.2. KNN Algorithm

This algorithm requires approximately large training set with certain distance. KNN can be useful to solve the problem of classes simultaneously as it performs well and provides the best choice for K value that has the best efficiency for its classifier, too [13]. In order to choose the most appropriate K, according to trial and error method, different numbers were evaluated (here, up to k = 10 has been tested) that in the first group (with all features) k = 5 and in the second group (with reduced features) k = 4 were selected as they had the highest accuracy in this paper.

2.2.3. Random Forest Algorithm

Random Forest consists of a series of classifiers with a tree structure which contains a large number of decision trees, prediction is done alone by any of these trees. Random trees randomly select the samples and classify them with each tree in the forest and output is selected by the majority voting. Two advantages of this algorithm are high accuracy and its efficiency on large data sets [12].

2.2.4. Multi-Layer Perceptron Algorithm

Back-propagation method uses for weight training of a multi-layered network. The output value for each unit is calculated for each sample to reach the output layer. This algorithm uses the sample to sample rule for adjusting the connection weight. In this method, by using gradient descent method, it is tried to minimize the squared error between the network outputs and the function. Gradient descent method tries to reach a satisfactory hypothesis by reducing the error [10]. As the number of hidden layers increases, construction time and complexity of the model increases, so based on trial and error method it has been tried to choose the minimum layer with the highest accuracy (here, up to 5 hidden layers has been tested). In this study for the first group with all the features number of hidden layers was 3 and for the second group with reduced features was 4 which had the best accuracy and the least error.

2.2.5. Naive Bayes Algorithm

Naive Bayes algorithm predicts the class based on the Bayesian theory. Naive Bayes classifier can predict and calculate the most probable output of unclear set by using available test sample sets [9].

2.2.6. Vote Algorithm

In recent years, combining multi-classifier algorithms instead of using one algorithm has been taken into consideration. The basis of these algorithms is on combining multiple classifiers which aims to increase the accuracy and reduce the algorithm error in comparison with using only one algorithm. This aim can be considered as a solution for detecting a simple pattern. Another application of this type of algorithm is constructing a hybrid model that is derived from the combination of votes obtained from different classifiers which its result is better than one algorithm alone. The basic assumption is that increasing the efficiency of the hybrid model is easier than increasing the efficiency of an algorithm alone. The main reason for combining several algorithms is to increase its accuracy and reduce the error. Using several different classifiers can increase the efficiency of the hybrid algorithm. The combination of algorithms can be done using the training set, collection of different characteristics or random selection [14].
The easiest way to combine the classifiers is majority voting, in which outputs of several classifiers are combined together then the output with the most votes will be selected as the final decision [14]. In the present study, the outputs of three classifiers (KNN, Naive Bayes and J48) were combined to produce the final output.

2.3. Methodology

The total data set has been selected based on CBC, hemoglobin electrophoresis, ferritin, PCR tests, age and gender. As the normal range of these tests is different in children under 12 years, the tests of the subjects over 12 years were used in this study. the samples were α-thalassemia trait, β-thalassemia trait, IDA and the normal subjects. Subjects with iron deficiency: their ferritin's levels was less than 12 ng / mL, normal hba2, normal or reduced RBC, and reduced MCV (MCV < 80 fl) and reduced MCH (MCH < 27 pg). They took iron supplements for a period of time and in the next test they blood characteristics were normal that results of their diseases period were used in this study. Samples of beta thalassemia trait: people who had Hba2 ≥ 3.5%, reduced MCV and MCH were selected and those who had iron deficiency at the same time were eliminated. α-thalassemia trait samples (cis and trans): subjects with reduced MCV and MCH, having Hba2 less than 3.5%, their ferritin test results were normal, or they were cured after taking iron supplements and also their PCR test showed that they have α-thalassemia trait (cis and trans). Normal subjects: selection of normal people was according to normal ranges of CBC test characteristics.

These subjects have been selected to detect IDA, β-thalassemia trait, α-thalassemia trait (cis and trans) and normal subjects. Here, α-thalassemia trait is selected as patients with hemoglobin Bart's hydrops fetalis die before birth or shortly after birth [5] and those who suffer from hemoglobin H are detectable by Microcytic, Hemolytic Anemia, Hypochromic and often Hepatosplenomegaly and mild jaundice [19]. In silent carriers, the range of indices of red blood cells is often similar to normal subjects [5] that are not in the scope of this study but all the blood characteristics of α-thalassemia trait (cis and trans) are similar to β-thalassemia trait and IDA (all these cases have mild anemia) [5].

As the blood characteristics of these three kinds of anemia are similar to each other, and specialists are able to detect them by checking some tests. in this study classification was according to only CBC test data, age and gender then data were discretized in this way: in pre-process tab, at the first supervised option and then discretize option was selected. Classification in two groups was conducted using 10 fold cross validation. the study consisted of two groups: The first group with all the blood features, age and gender, and the second group with reduced blood features. To analyze the data J48 (C4.5), Naive Bayes, Random Forest, KNN (IBK), MLP algorithms and vote algorithm were used.

3. Results and Discussion

The results of algorithms analysis is shown based on the accuracy of prediction and mean absolute error for both groups with all the features and reduced features in this section. To assess the accuracy of the predictions, 10 Fold Cross Validation is used. Then, the results of these two groups is presented in Sections 4.1 and 4.2.

3.1. Analysis with all the Features

The accuracy and error of each algorithm can be determined by Weka software. Here the results of diagnosis of each classifier are evaluated and the best classifier to diagnose and predict IDA, β-thalassemia trait and α-thalassemia trait (cis and trans) is selected. In this study, six algorithms were suggested including: J48 (C4.5), Naive Bayes, Random Forest, KNN, MLP and Vote algorithm with combining KNN, Naive Bayes and J48. All the features including, CBC test, age and gender were analyzed with these algorithms. In Table 1,the results of comparison of proposed algorithms with all the features is shown.

Table 1. Comparison of proposed algorithms with all the features

Algorithms	Results with all the Features	
	Accuracy	**Mean Absolute Error**
J48	95.3342	**0.0322**
Naive Bayes	95.082	**0.0255**
Random forest	95.9647	**0.0391**
IBK	95.9647	**0.0377**
MLP	95.9647	**0.0298**
Vote	96.343	**0.0183**

According to Table 1 and based on the available data, the accuracy can be regarded as the basis to compare the algorithms and to the diagnose IDA, β-thalassemia trait and α-thalassemia trait (cis and trans) and choose the most efficient model. To analyze all the features with proposed algorithms, the vote algorithm had the best accuracy of 96.343 and also the lowest mean absolute error rate of 0.0183. In Fig. 1 and 2 the comparison of the accuracy and mean absolute error of algorithms using all features are shown separately.

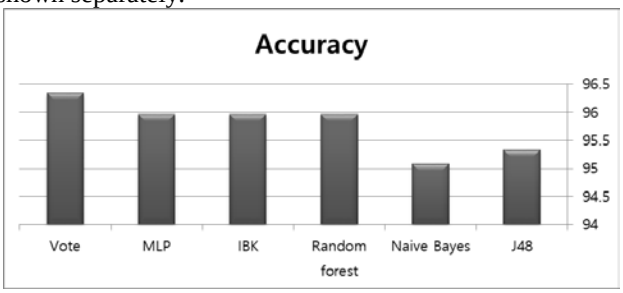

Fig. 1. The Comparison of the Accuracy of the Algorithms with all the Features

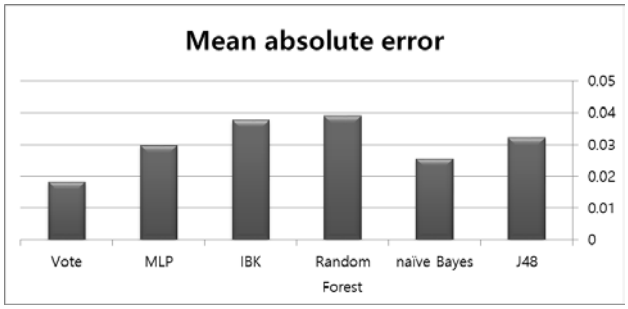

Fig. 2. Comparing the Mean Absolute Error of Algorithms with all the Features

3.2. Analyzing with the Reduced Features

In previous studies, four blood features were considered as the most important features to diagnose anemia including RBC, Hb, HCT and MCV [20]. In the present study, the reduced features by the same algorithms were evaluated. Then, the results of each method have been assessed and the best algorithm to diagnose and predict IDA, β-thalassemia trait and α-thalassemia trait (cis and trans) is selected. The accuracy and the mean absolute error in checking with reduced features are shown in Table 2.

Table 2. Comparison of proposed algorithms with reduced features

Algorithms	Results with Reduced the Features	
	Accuracy	Mean Absolute Error
J48	95.4603	0.0346
Naive Bayes	96.0908	0.0335
Random forest	95.3342	0.0331
IBK	95.7125	0.0336
MLP	95.4603	0.0363
Vote	96.2169	0.0189

As Table 2 shows, the algorithm which has the highest accuracy, is selected as the most efficient model. In the suggested models, in checking reduced features with five separate algorithms, naive Bayes algorithms with the accuracy of 96.9466 had higher accuracy, but the vote algorithm had the highest accuracy of 96.2169. Also, this algorithm had the lowest Mean absolute error of 0.0189. In Fig. 3 and 4 comparison of the accuracy and mean absolute error of the reduced features are shown separately.

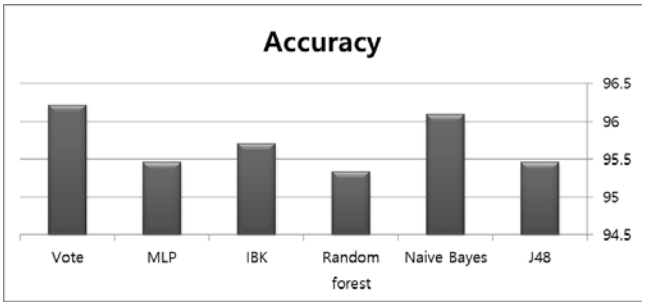

Fig. 3. Comparison of the accuracy of the algorithms with reduced features

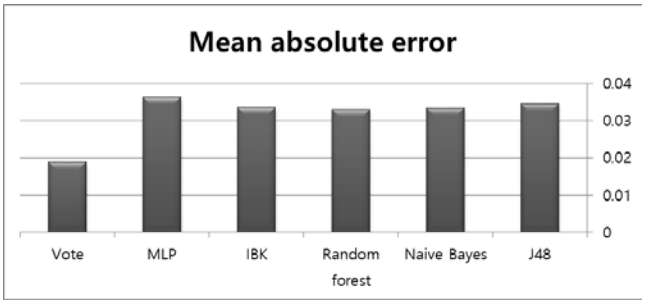

Fig. 4. Comparison of the mean absolute error of algorithms with reduced features

The objective of this study was the diagnosis of IDA, β-thalassemia trait, α-thalassemia trait (cis and trans) and normal subjects using only the CBC test and Weka software. To provide an automatic model to identify these kinds of anemia, subjects who suffers from anemia were studied and their data were analyzed. The samples were studied in two groups, the first group with all features and the second group with reduced features. Some of the best data mining algorithms were used on the data by Weka software. To analyze data sets five algorithms and a vote algorithm were used. To analyze all the features, vote algorithm was the best with the accuracy of 96.343 in detection, and also to analyze reduced features, vote algorithm had the accuracy of 96.2169, which analyzing the algorithms was by 10 fold cross validation method. The results showed that using just a CBC test will help specialists to diagnose these types of anemia and the diagnosis is more accurate using vote algorithm than using just one single algorithm.

4. Conclusion

The basis of this study was the automatic detection of three types of anemia. Samples were classified into four groups including IDA, β-thalassemia trait, α-thalassemia trait (cis and trans) and healthy subjects. After examining the blood characteristics of the CBC test of the samples by the proposed algorithms, it is concluded that if these three types of anemia are pure, using data mining techniques and only CBC test can detect them with a high accuracy and also reducing the features had no noticeable difference in the diagnosis process. The purpose of using the vote algorithm was to increase the accuracy and reduce the error that Among the proposed algorithms, vote algorithm showed better results than the other algorithms.

Acknowledgements

The authors wish to thank to staff of the Hafezie, Farhangian and Reference laboratories, particularly Masoud Goodarzi, the head of reference laboratory. Special thanks to Dr. Reza Alibakhshi and Dr. Reza Akramipour for their support to improve this paper.
Conflict of interest statement
None Declared.

References

[1] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J. F. Chang, L. Hua, Data mining in healthcare and biomedicine: a survey of the literature, Journal of Medical Systems 36 (2012) 2431-2448.
[2] M.B. Zimmermann, R.F. Hurrell, Nutritional iron deficiency, The Lancet 370 (2007) 511-520.
[3] F. Kutlar, Diagnostic approach to hemoglobinopathies, Hemoglobin 31 (2007) 243-50.
[4] H. Najmabadi, R. Karimi-Nejad, S. Sahebjam, F. Pourfarzad, S. Teimourian, F. Sahebjam, et al, The beta-thalassemia mutation spectrum in the Iranian population, Hemoglobin 25 (2001) 285-29.
[5] J.S. Waye, B. Eng, Diagnostic testing for α-globin gene disorders in a heterogeneous North American population, International Journal of Laboratory Hematology 35 (2013) 306-313.

[6] R. Galanello, R. Origa, Beta-thalassemia, Orphanet Journal of Rare Diseases 5 (2010) 1-15.
[7] M.N. Amin, M.A. Habib, Comparison of different classification techniques using WEKA for hematological data, American Journal of Engineering Research 4 (2015) 55-61.
[8] S. Saichanma, S. Chulsomlee, N. Thangrua, P. Pongsuchart, D. Sanmun, The observation report of red blood cell morphology in Thailand teenager by using data mining technique, Advances in Hematology 2014 (2014) 1-5
[9] D. Setsirichok, T. Piroonratana, W. Wongseree, T. Usavanarong, N. Paulkhaolarn, C. Kanjanakorn, M. Sirikong, C. Limwongse, N. Chaiyaratana, Classification of complete blood count and haemoglobin typing data by a C4. 5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening, Biomedical Signal Processing and Control 7 (2012) 202-212.
[10] W. Wongseree, N. Chaiyaratana, K. Vichittumaros, P. Winichagoon, S. Fucharoen, Thalassaemia classification by neural networks and genetic programming, Information Sciences 177 (2007) 771– 786.
[11] S.R. Amendolia, G. Cossu, M.L. Ganadu, B. Golosio, G.L. Masala, G.M. Mura, A comparative study of k-nearest neighbour, support vector machine and multi-layer perceptron for thalassemia screening, Chemometrics and Intelligent Laboratory Systems 69 (2003) 13-20.
[12] O. Okun, H. Priisalu, Random forest for gene expression based cancer classification: overlooked issues, In Iberian Conference on Pattern Recognition and Image Analysis, Springer Berlin Heidelberg (2007) 483-490.
[13] P. Paokanta, M. Ceccarelli, S. Srichairatanakool, The efficiency of data types for classification performance of machine learning techniques for screening β-Thalassemia, In 2010 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL), IEEE (2010) 1-4.
[14] C.F. Tsai, J.W. Wu, Using neural network ensembles for bankruptcy prediction and credit scoring, Expert Systems With Applications 34 (2008) 2639-2649.
[15] A.D. Flouris, K.P. Poulianiti, M.S. Chorti, A.Z. Jamurtas, D. Kouretas, E.O. Owolabi, M.N. Tzatzarakis, A.M. Tsatsakis, Y. Koutedakis, Acute effects of electronic and tobacco cigarette smoking on complete blood count, Food and Chemical Toxicology 50 (2012) 3600-3603.
[16] G. Hussein, M. Fawzy, T. El Serafi, E.F. Ismail, D. El Metwally, M.A. Saber, M. Giansily, J.F. Schved, S. Pissard, P. Aguilar Martinez, Rapid detection of b-thalassemia alleles in Egypt using naturally or amplified created restriction sites and direct sequencing: a step in disease control, Hemoglobin 31 (2007) 49-62.
[17] S. Dogan, I. Turkoglu, Iron-deficiency anemia detection from hematology parameters by Using decision trees, International Journal of Science & Technology 3 (2008) 85-92.
[18] Y. Zhao, Y. Zhang, Comparison of decision tree methods for finding active objects, Advances in Space Research 41 (2008) 1955-1959.
[19] R. Origa, M.C. Sollaino, N. Giagu, S. Barella, S. Campus, C. Mandas, P. Bina, L. Perseu, R. Galanello, Clinical and molecular analysis of haemoglobin H disease in Sardinia: haematological, obstetric and cardiac aspects in patients with different genotypes, British Journal of Haematology 136 (2007) 326-332.
[20] S.R. Amendolia, A. Brunetti, P. Carta, G. Cossu, M.L. Ganadu, B. Golosio, G.M. Mura, M.G. Pirastru, A real-time classification system of thalassemic pathologies based on artificial neural networks, Medical Decision Making 22 (2002) 18-26.