Automated Diagnosis of Thalassemia Based on DataMining Classifiers

[1] Eyad H. Elshami, [2] Alaa M. Alhalees
Faculty of Information Technology, Islamic University-Gaza, Palestine
P.O.Box 108, Rimal, Gaza, Palestine
[1] eshami@iugaza.edu.ps, [2] alhalees@iugaza.edu.ps

ABSTRACT

Thalassemia is a genetic disease that is commonly found in many parts of the world. It leads to death in most of its major cases so we must control it by determining the persons who trait the thalassemia genes. Complete Blood Control (CBC) is the first and the simplest test which can narrow to the existence of thalassemia. This paper presents an investigation for thalassemia existence by using data mining classifiers depending on CBC. Three data mining classifiers were used in this investigation. Each of the classifiers used to differentiate between thalassemia traits patients- with its different levels-: iron deficiency patients, normal persons, and the patient who suffer from other blood diseases. The experimental results of this investigation were bright with accuracy exceeding 90% and it showed that the critical point which can be as first indicator for the thalassemia existence is MCV ≤ 77.65.

KEYWORDS

Thalassemia Diagnosis; Classification; Decision Tree; Naïve Bayes; Neural Network.

1. INTRODUCTION

Thalassemia (also spelled thalassaemia) is called "anemia Mediterranean" because it is more popular in the Mediterranean basin. Thalassemia is an inherited blood disease, it influences the blood manufacture, and so the hemoglobin in red blood cells unable to do its job [6].

Normal hemoglobin is composed of two chains each of α and β globin. Thalassemia patients produce a deficiency of either α or β globin. The thalassemia is classified according to which chain of the hemoglobin molecule is affected, α-Thalassemia or β-Thalassemia [1], [4]. As a result, person with thalassemia usually has a reduced number of RBCs (Red Blood Cells) in the bloodstream (anemia), which affects the oxygen transportation to the body tissues. In addition, thalassemia can cause RBCs to be smaller than the normal or decrease hemoglobin in the RBCs to below-normal levels [6].

However, thalassemia always inherited to the children through the parents' genes, it also can appear according to genes mutation. A child normally does not develop symptoms unless both of the parents carry the thalassemia gene and the child is said to have thalassemia trait, if only one parent passed the thalassemia gene for the child.

The thalassemia disease classified into three categories: Thalassemia Major (*Thal-M*), Thalassemia Intermediate (*Thal-I*), and Thalassemia Trait or Minor (*Thal-T*). Patients with *Thal-M* and *Thal-I* require repeated blood transfusions during whole their lives and other treatments, while *Thal-T* usually do not require any specific therapy [11].

Since thalassemia is a genetic disease so we can control it by eliminate the marriage between the both people who have carried the genes. Complete Blood Count (CBC) test is the first and the simplest test can perform on the way to indicate the thalassemia's genes existence [7].
In this paper we tried to use the data mining techniques to predict if the person carries the thalassemia genes without going toward the costly test to determine if the person suffers from iron deficiency or thalassemia.

"Data mining is the search for new, valuable, and nontrivial information in large volumes of

data" [15]. It is considered as a cooperative effort of humans and computers. The best results can be achieved by balancing the knowledge of human experts in defining the problems and goals with the search capabilities of computers [15].

The remaining parts of the essay are section II Literature work, section III describes the proposed data mining classification techniques and the dataset, section IV experiments and discussion, and finally we will draw the conclusion and the future work.

2. RELATED WORKS

There are related works using data mining techniques to diagnose several types of diseases and phenomena, such as: thalassemia, diabetes, cancer, heart diseases … etc. And many other tried to find their own formula to determine if the person is a thalassemia patient or iron deficiency patient [3],[9].

Wongseree et.al [10] investigated thalassemia classification by using a neural network and a decision tree, which is evolved by genetic programming, in thalassemia classification. The aim is to differentiate between thalassemic patients, persons with thalassemia trait and normal subjects by inspecting characteristics of red blood cells, reticulocytes and platelets. But they need in the proposed model more blood testing like Platelet and Reticulocyte.

El-Sebakhy and Elshafei in [5] proposed thalassemia screening using unconstrained functional networks classifier and compare the performance of the proposed model with both multilayer perceptron (MLP) and support vector machine (SVM), and the results showed that using unconstrained functional networks classifier takes much less computations. They tried to assign patients to either "normal" group (have no thalassemias) or "a" or "h" groups (having thalassemias).

Amendolia et.al. in [2] investigated the use of artificial neural networks (ANNs) for the classification of thalassemic pathologies using the hematologic parameters resulting from hemochromo cytometric analysis only. Different combinations of ANNs are reported, which allow thalassemia carriers to be discriminated from normals with 94% classification accuracy, 92% sensitivity, and 95% specificity. On the basis of these results, an automated system that allows real-time support for diagnoses is proposed. The automated system interfaces a hemochromo analyzer to a simple PC.

All the previous works tried to diagnosis thalassemia focus on how to differentiate between the persons with thalassemia trait and normal persons, and most of them just try to use one data mining technique they consider it the best one without any comparison with the other techniques in the domain. In this study, we will used more than one classifier to get most significance one, and try to differentiate between the normal persons, thalassemia patients with its different types (Major, Intermediate, Trait), Iron Deficiency patients and the other patients who suffer from other blood diseases.

3. MATERIAL AND METHODS

The used dataset in this study represents the blood samples of the subjects of the essential premarital tests screening program of β-thalassamia in Gaza strip, which was implemented and started since the 9th of September 2000 under full control and supervision of the Thalassemia Center, Palestine Avenir foundation, and the premarital tests are being performed according to defined and approved working protocol.

3.1 Dataset And Preprocessing:

The dataset consists of 46920 samples. Its attributes represents the CBC features as in TABLE 1, some features; such as the sex, age, and some others features which are dropped due the privacy of the blood sample's owner, and finally it contain diagnoses attribute which represent the target label of the sample, it has seven different labels: *Normal*,

Thalassemia Major (Thal-M), *Thalassemia Intermediate (Thal-I)*, *Thalassemia Trait (Thal-T)*, *Iron Deficiency (Iron Def)*, *Thalassemia Trait/Iron Deficiency (Thal-T/Iron Def.)*, and *Other* which represented any other blood diseases.

In the preprocessing of the dataset we eliminate useless attributes, refill the missing values and remove/refill the outlier values on the outlier samples. TABLE 2 represent the dataset attributes which we used in our investigation.

TABLE 1: CBC test features [11]

Shortcut	Term	Male N. Value	Female N. Value
WBC	White Blood Cell	4300 – 10800	
RBC	Red Blood Cell, 10^6 cells/mcL	4.7 – 6.1	4.2 – 5.4
HB	Hemoglobin	12 – 18 mg/d	12 – 16 mg/d
HCT	Hematocrit	37 – 54%	33 – 57%
MCV	Mean Cellular Volume	80 – 98%	
MCH	Mean Cellular Hemoglobin	24 – 30%	
MCHC	Mean Cellular Hemoglobin Concentration	24– 30%	
RDW	RBC Distribution Width	11.5 – 14.5 %	
PLT	Platelets	150000 – 450000	

3.2 Classification Methods:

Three candidate classifiers are considered in this study: Decision Tree, Naïve Bayes, and Neural Network. *Decision Tree* is a tree-structured plan of a set of attributes to test in order to predict the output. It is a type of tree-diagram used in determining the optimum course of action, in situations having several possible alternatives with uncertain outcomes, and it was selected according to its accuracy and the ability to extract the classification rules which may be important in a lot of cases [12]; *Naïve Bayes* is a technique for estimating probabilities of individual variable values, given a class, from training data and to then allow the use of these probabilities to classify new entities. Bayes Classification is a term in Bayesian statistics dealing with a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions, and it was selected according to its high accurate results in a lot of domains [13]; finally *Neural Network* is an interconnected group of artificial neurons that uses a mathematical or computational model for information processing based on a connectionist approach. Multilayer perceptron neural network with Backpropagation learning algorithm will be used in this study [14]. *Rapid Miner 5.0.010* will be used as the environment in which the three classifiers (Decision Tree, Naïve Bayes, and Neural Network) will be applied and compared.

TABLE 2: Dataset attributes

Attribute	Data type	Attribute role
SEX	Binominal	Regular
AGE	Integer	Regular
WBC	Real	Regular
RBC	Real	Regular
HB	Real	Regular
HCT	Real	Regular
MCV	Real	Regular
MCH	Real	Regular
MCHC	Real	Regular
RDW	Real	Regular
PLT	Real	Regular
Diagnoses	Nominal	Label

4 EXPERIMENTS AND RESULTS DISCUSSION

In this investigation, the experiment using the data mining classifiers will be divided into two parts: the experiment with full and reduced features. The results from these two parts and a detailed classification accuracy analysis emphasizing on the classification errors will be presented in following Sections. Four experiments were conducted in each type: the first one is to measure the performance of the decision tree classifier; the second one is to measure the performance of the naïve bayes classifier; the third one to measure the performance of the neural network; and finally the fourth one to get the most significance classifier between the three classifiers by using the T-test with $\alpha = 0.05$. We used cross-validation to measure the classifiers performance. The feed-forward back-propagation neural network classifier was adjusted with 500

training cycles, learning rate 0.3, and momentum 0.2.

4.1 Experiments With Full Features:

In these experiments we used the whole record's attributes of each sample as in Table 3. The decision tree classifier gives a result with general accuracy of 93.64%±0.11%, the naïve bayes classifier gives a result with general accuracy of 93.7% ±0.31%, and finally the neural network classifier gives a result with general accuracy of 95.71% ±0.71% as shown in Figure 1 and Table 3.

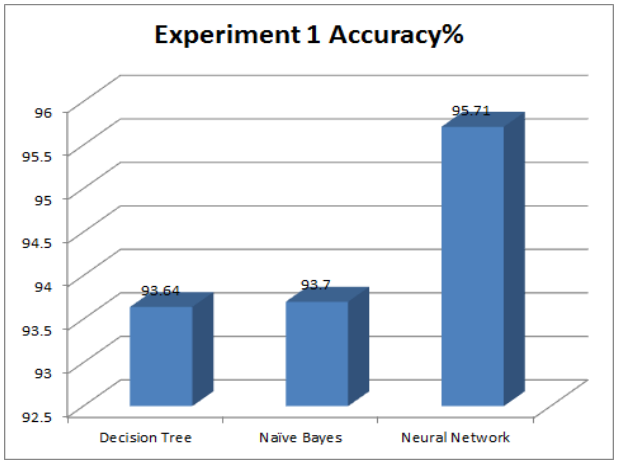

Figure 1: Experiment1 classifiers accuracy values

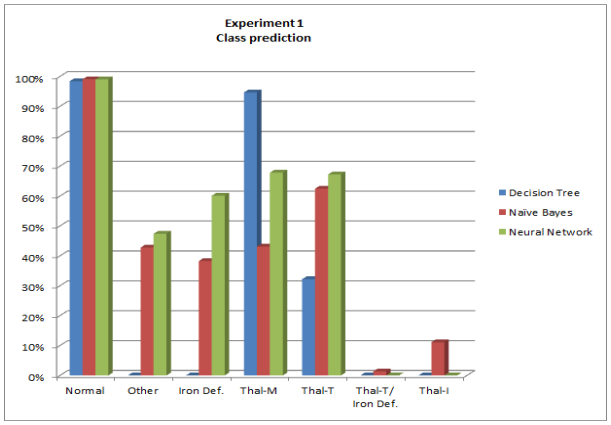

Figure 2: Experiment1 classes prediction accuracy

4.2 Experiments With Reduced Features

According to. Sirdah et al. in [9], MCV<80 fl and/or MCH<26pg are usually the common features of blood samples from thalassemia trait and iron deficient subjects, Amendolia et al. in [2] have empirically chosen the RBC, HB, HCT and MCV as the input for their classifiers. In this type of experiment we combined all of these features in this experiments type.

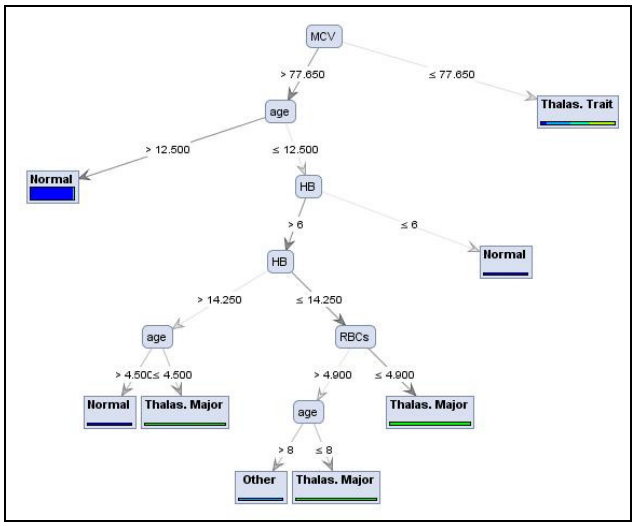

Figure 3: Decision tree form expriment1

In our experiments we used the whole record's attributes of each sample as in Table 2. The Decision Tree classifier gives a result with general accuracy: 93.65%±0.18%, while the Naïve Bayes classifier gives a result with general accuracy: 94.32%±0.32 and the Neural Network classifier give a result with general accuracy: 95.48%±0.21% as shown in Figure 4 and Table 4

TABLE 3: T-test for the three classifiers in experiment1

		NN	NB	DT
		0.956 ± 0.003	0.937 ±0.005	0.936 ± 0.001
NN	0.956 ± 0.003		**0.000**	**0.000**
NB	0.937 ± 0.005			0.654
DT	0.936 ± 0.001			

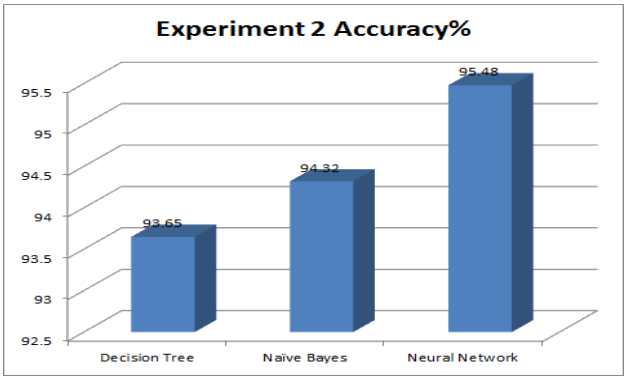

Figure 4: Experiment2 classifiers accuracy values

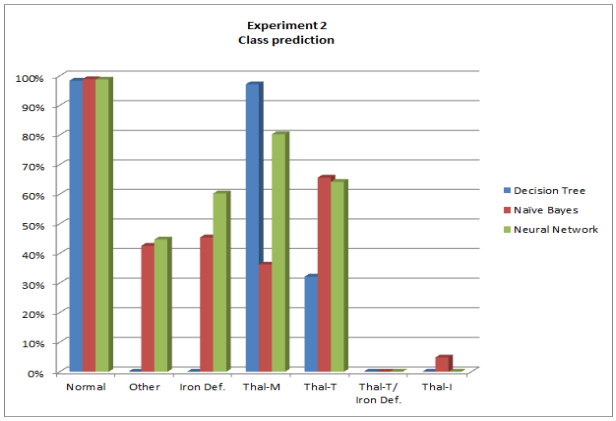

Figure 5: Experiment2 classes prediction accuracy

4.3 Result And Discussion:

From the previous charts and the t-test tables we can get the following results: the Neural Network classifier has more significance result in the three experiments for the both types of experiments, the features in the second type of experiments is the most effective features in the dataset since the three classifiers give more accurate prediction in the second experiments, the Naïve Bayes classifier has more accurate when small number of attributes belong to specific class label, from the Decision Tree classifier see that the single-point models of whole blood sample group for MCV indicator and its normal range value started at 77.65 not 80; due to the lack of Thal-M and Thal-T/Iron Def. all the classifiers almost failed in that classification because it seems as an outlier.

TABLE 4: T-test for the three classifiers in experiment2

		NN	NB	DT
		0.955 ±0.003	0.943 ± 0.004	0.936 ± 0.001
NN	0.955 ± 0.003		**0.000**	**0.000**
NB	0.943 ± 0.004			**0.000**
DT	0.936 ± 0.001			

5 CONCLUSION AND FUTURE WORK

Thalassemia is a genetic disease that causes a reduction in the life span of a red blood cell. The disease is a result of an abnormality in the genes that regulate the formation of hemoglobin (HB) of the red blood cell. In order to make the diagnosis, the blood characteristic must be analyzed.

A complete blood count (CBC) is the primary screening test for a laboratory diagnosis of thalassemia. Data mining is important automated systems which can help get automated classifier for the new incoming CBC test samples.

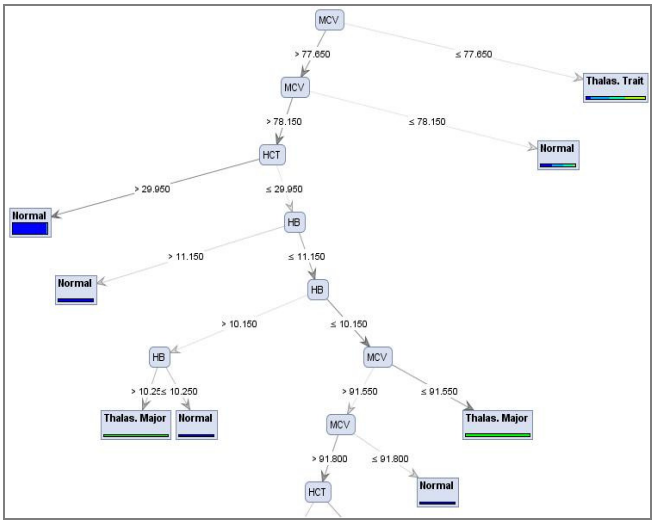

Figure 6: Decision tree form expriment2

Different experiments were done to get more significance classifier; these experiments show that the neural network classifier is the more significance classifier to differentiate between the our study classification classes (Normal, Tha-M, Thal-I, Thal-I, Iron Def, Thal-T/Iron Def, Other), also the experiments we can support the knowledge which says that MCV is the main feature to indicate the thalassemia existence but indicator value less than 77.65. Finally, if the MCV is greater than 77.65 and age greater than 12.5 then there is no existence of thalassemia.

The results suggest more investigation is required to determine that if the classification classes were reduced, into Thalassemia trait, iron deficiency, and normal, then examine if the classifiers can provide us with more accurate results.

6 ACKNOWLEDGMENT

The author is extremely grateful to the Thalassemia Center, Palestine Avenir foundation, for their collaborative and providing us with the dataset.

7 REFERENCES

1. Akay, A., Dragomir, A., Yardimci, A., Canatan, D., Yesilipek, A., Pogue, B.: A Data-Mining Approach for Investigating Social and Economic Geographical Dynamics of β-Thalassemia's Spread. IEEE Transaction Information Technology In Biomedicine, vol. 13, pp.774–783, (2009).
2. Amendolia, S., Brunetti, A., Carta, P., Cossu, G., Ganadu, M., Golosio, B., Mura, G., Pirastru, M.:A Real-Time Classification System of Thalassemic Pathologies Based on Artificial Neural Networks. Med Decis Making, vol. 2, pp. 18-26, (2002).
3. Chen, D., Li, P., Rong, K. B.: The application research of operative technology in detecting thalassemia. Chinese Journal of Birth Health & Heredity, vol. 16, pp. 39–41, (2008).
4. Dogan, S., Turkoglu, I.: Iron-Deficiency Anemia Detection from Hematology Parameters by Using Decision Trees. International Journal of Science & Technology, vol. 3, pp.85–92, (2008).
5. El-Sebakhy, E., Elshafei, M.: Thalassemia Screening Using Unconstrained Functional Networks Classifier. In Proc. IEEE International Conference on Signal Processing and Communications (ICSPC 2007), vol. 1, pp. 1027 – 1030, (2007).
6. Galanello, R., Eleftheriou, A., Traeger-Synodinos, J.: Prevention of Thalassaemias and other Haemoglobin Disorders. Thalassaemia International Federation (TIF), Cyprus, vol. 1, pp., (2003).
7. Sun, Y. J., Li, H. J., Wang X. Y.: The diagnostic value about the MCV and red blood cell fragility test in thalassemia screening. Chinese Journal of Birth Health & Heredity, vol. 15, no. 8, pp. 115–116, (2007).
8. Nadarjan, V., Sthaneshwar, P., Jayaranee, S.: RBC-Y/MCV as adiscriminant function for differentiating carriers of thalassaemia and HbE from iron deficiency. International Journal of Laboratory Hematology, vol. 32, pp. 215–221, (2009).
9. Sirdah, M., Tarazi, I., Alanajjar, E., Alahaddad, R.: Evaluation of the diagnostic reliability of different RBC indices and formulas in the differentiation of the b-thalassaemiamin or from iron deficiency in Palestinian population. International Journal of Laboratory Hematology, vol. 30, pp. 324–330, (2008).
10. Wongseree, W., Chaiyaratana, N., Vichittumaros, K., Winichagoon, P., Fucharoen, S.: Thalassaemia classification by neural networks and genetic programming. Information Sciences, vol. 177, pp.771–786, (2007).
11. National Heart Lung Blood Institute site, http://www.nhlbi.nih.gov/health/dci/Diseases/Thalassemia/Thalassemia_WhatIs.html
12. Safavian, S.R., Landgrebe, D.: A survey of decision tree classifier methodology. Systems, Man and Cybernetics, IEEE Transactions on, vol. 21/3, pp. 660 - 674, (1991).
13. Irina, R.: An empirical study of the naive bayes classifier. In Proc. of IJCAI-01 workshop on Empirical Methods in AI, International Joint Conference on Artificial Intelligence, vol., pp. 41–46., (2001).
14. Hecht-Nielse R.: Theory of the backpropagation neural network. In Proc. of 1989 IJCNN., International Joint Conference on , vol.1, pp.18-22, (1989).
15. Kantardzic, M.: Data Mining: Concepts, Models, Methods, and Algorithms. ebook ed., John Wiley & Sons ©, (2003).