IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 8, August 2017

Prediction Thalassemia Based on Artificial Intelligence Techniques: A Survey

Fatemeh Yousefian[1], Touraj Banirostam [2], Azita Azarkeivan[3]

Dept of Computer Engineering, Islamic Azad University, Central Tehran Branch, Tehran, Iran[1]

Dept of Computer Engineering, Islamic Azad University, Central Tehran Branch, Tehran, Iran[2]

High Institute for Research and Education in Iranian Blood Transfusion Organization Transfusion Medicine, Thalassemia Clinic, Tehran, Iran[3]

Abstract: Thalassemia is a type of genetic disease that can be observed in many areas of the world. The first step is a CBC test to diagnose a person with thalassemia. In this paper, the used data mining methods to diagnose thalassemia are studied and evaluated. The effective parameters in thalassemia diagnosis are the available variables in the CBC test in people that among these parameters, RBC, HGB, MCV and HTC have a significant effect on the disease diagnosis. Based on the available values in the CBC test and using artificial intelligence algorithms, the patient with thalassemia is diagnosed. Artificial intelligence algorithms are used to analysis laboratory data properly, which leads to increase accuracy in the diseases diagnosis, which has a significant impact on the treatment process and improvement of patient health.

Keywords: Thalassemia, Data mining, Classification, Artificial Intelligence Techniques.

I. INTRODUCTION

According to conducted research, applied techniques for artificial intelligence in the diagnosis and prediction of diseases such as diabetes, using the UTA feature selection method and multi-layered perceptron network, fuzzy classification are based on the ant colony algorithm, an intelligent system for diagnosis of diabetes based on Linear Discriminant Analysis and adaptive networks based on the fuzzy inference system, Decision tree, support vector machine and Bayesian network [4-10], Prostate cancer prediction using neural network, Logistic and regression [11-13], Classification of heart disease like Coronary Artery Disease [14], Early diagnosis of Alzheimer using Artificial Neural Network [15], Diagnosis of brain disease based on the fuzzy discrete-secret Markov model [16], classification of cervical cancer using an artificial neural network [17], using a discrete method for diagnosis of optic nerve disease[18], diagnosis of ovarian cancer based on fuzzy neural network [19]. The outline of this article is as follows: In part II, the effective parameters for the diagnosis of thalassemia are introduced. In the following, the used methods to identify thalassemia will be described and the results will be presented. Section VI compares and evaluates the results for each method. In last, the obtained result from this study will be presented.

II. EFFECTIVE PARAMETERS

CBC test is the first and easiest way to detect thalassemia. Doctors recognize the type of Anemia based on the variables in the test. Therefore, in all researches in the field of diagnosis of thalassemia using artificial intelligence techniques, the used parameters in the CBC test are used as a feature. Using the feature selection methods, the most effective parameter is identified among all existing variables that would make it easy to make decisions. Available parameters in CBC test include : Red Blood Cell (RBC), Hemoglobin (Hb), Hematocrit (Ht) , Mean Corpuscular Volume (MCV), White Blood Cell (WBC); Hematocrit is a percentage of the total blood volume that made up of red blood cells (HCT), Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Red Cell Distribution Width (RDW), Platelet (PLT). In all of these studies, these parameters have been used to diagnose thalassemia types.

III. COMPARISON K-NEAREST NEIGHBOR AND MULTILAYER PERCEPTRON WITH SUPPORT VECTOR MACHINE FOR SCREENING THALASSEMIA

According to an article presented by Amendolia et.al. in 2003, they studied screening for thalassemia using two the nearest neighbor and support vector machine methods. The used dataset in this article was collected from 304 students aged 14 to 15 years old in North Sardinia. Students with Anemia caused by iron deficiency were excluded from the dataset. Among 304 samples, 27 samples were with beta thalassemia and 277 healthy samples and alpha-thalassemia.

DOI10.17148/IJARCCE.2017.6847

The used features include: RBC, Hemoglobin (Hb), Hematocrit (Ht) and Mean Corpuscular Volume (MCV). 108 records belong to test data that contains 55 healthy samples, 44 alpha thalassemia sample and 9 beta thalassemia samples. Training set records include 141 healthy samples, 37 alpha thalassemia sample and 18 beta-thalassemia samples in total that in general, there are 196 records in the training set [20]. In this paper, SVM-Light software, version 4 was used to classify support vector machine. In this method, the LOO (leave-one-out) algorithm is used to maximize the information from data. In this algorithm, at each repetition, one of the data is deleted, and then the remaining data put in the training process and finally the class predicts the deleted pattern. This action is repeated until it reaches the minimum value of the mean squared error. The parameter C represents the number of deleted patterns in each repetition which determines the best value of C based on experience and repetition. Based on different values of C, the best value of C is equal to10 that the results are shown in Table 1 [20].

TABLE 1 RESULTS FROM THE SVM METHOD FOR THE DIAGNOSIS OF PATIENT SAMPLE AND HEALTHY SAMPLES [20].

Class	Healthy	Patient	The number of healthy samples was diagnosed	The number of patient samples was diagnosed	total	accuracy
Healthy	94.55%	5.45%	51	4	55	94.55%
Patient	16.98%	83.2%	45	8	53	83.2%
			96	12	108	88.89%

In the second method, the nearest neighbour method was used. In this method, according to the 23 nearest neighbours based on polling is conducted that the accuracy is 85%. The more detailed information is shown in Table 2.

TABLE 2 THE RESULTS OF THE KNN METHOD IN DETERMINING THE PATIENT AND HEALTHY SAMPLE [20].

Class	Healthy	Patient	The number of healthy samples was diagnosed	The number of patient samples was diagnosed	total	accuracy
Healthy	92.73%	7.27%	51	4	55	92.73%
Patient	22.64%	77.26%	45	8	53	77.36%
			96	12	108	85.19%

given the results obtained in the above table, it can be said that the SVM method is more accurate than the KNN method, but in the KNN method, difference between the two classes is more efficient.

The proposed method by Amendolia et.al. has used SVM to classify the first layer, i.e. diagnosis of a healthy patient and unhealthy patient sample and in the second layer, each KNN method has been used to diagnose the type of healthy sample i.e. alpha thalassemia.

TABLE 3 RESULTS FROM THE KNN FOR THE DIAGNOSIS OF ALPHA-THALASSEMIA AND BETA-THALASSEMIA [20]

	alpha-thalassemia	beta-thalassemia	The number of healthy samples was diagnosed	The number of patient samples was diagnosed	total	accuracy
alpha-thalassemia	92.11%	7.89%	35	3	38	92.11%
beta-thalassemia	11.11%	88.89%	8	1	9	88.89%
			43	4	47	91.49%

In other words, in the second layer, the type of healthy sample is diagnosed based on normal and alpha thalassemia. Table 10 shows the results of the SVM and KNN method for the diagnosis of alpha thalassemia and beta-thalassemia. Amendolia et.al. also use the MLP method to classify. The results of a healthy and unhealthy classification using the three KNN, MLP, and SVM methods are shown in Figure 1.

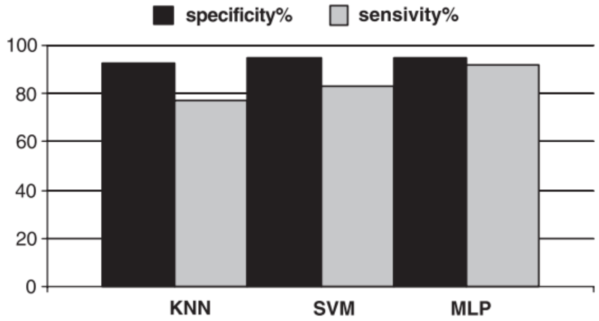

Fig. 1 Comparison of three methods, MLP, KNN, and SVM in the diagnosis of patient and healthy patterns [20]

Based on the above diagram, it is concluded that the MLP method has a better performance in diagnosing patient and healthy patterns. In the next step, the three above-mentioned methods are compared to diagnose alpha thalassemia and beta-thalassemia, as shown in Figure 2.

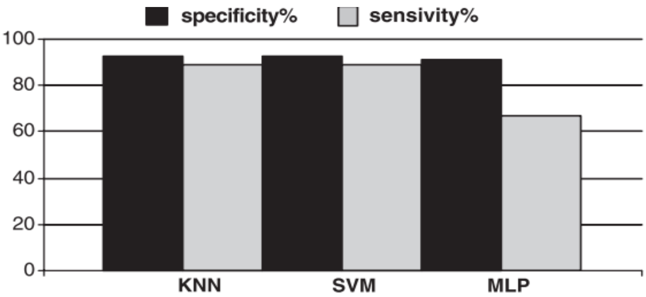

Fig. 2 Comparison of three methods , KNN, MLP and SVM for the detection of alpha thalassemia and beta-thalassemia [20]

As the above chart, the KNN and SVM methods have better performance than MLP for diagnosing alpha thalassemia and beta-thalassemia. Hence, the hybrid method is used for categorization. Hence, the hybrid method is used for categorization.

IV. DIAGNOSIS OF BLOOD DISORDERS AND CANCER USING ARTIFICIAL NEURAL NETWORKS

Dr. Payandeh et.al. diagnose 5 types of blood disorders including Anemia megaloblastic, thalassemia, Idiopathic Thrombocytopenic Pupura (ITP), AnemiaChronic myelogenous leukemia and lymph proliferative using a multi-layer perceptron network, together with a back-propagation error algorithm. The used dataset is collected here from 450 patients from Kermanshah Hospital. The collected parameters from each patient are: White Blood Cell (WBC), RBC, HGB, MCV, Mean Corpuscular Hemoglobin (MCH), Mean Corpuscular Hemoglobin Concentration (MHC), Red Cell Distribution Width (RDW), Platelet (PLT)), Neutrophil (NEUT), Lymphocytes and Leukocytes (LUC).
The values of these parameters are the input of the neural network and based on this, it categorizes each pattern into the expressed five classes. Each class represents one of the five desired diseases [2].

TABLE 4 SELECTIVE MLP SPECIFICATIONS FOR THALASSEMIA CLASSIFICATION [2].

Hidden layers number	1
Number of hidden layer neurons	15
Layer transfer function	Tan-sigmoid
Test error	0/0044
Repetition number	67

The obtained results for each class are shown in Table 5.

TABLE 5 RESULT MLP [2].

Class	Actual Class Prediction MLP	**Yes**	**No**	**Accuracy**
Anemia	Yes	2	0	100%
	No	0	316	
Thalassemia	Yes	0	1	99.7%
	No	0	317	
ITP	Yes	2	5	98.4
	No	0	313	
Leukemia	Yes	0	6	98.1%
	No	0	312	
Lymphoproliferative	Yes	2	2	99.4%
	No	0	314	

Based on the results of the data test, the network has a good performance in diagnosing 5 types of blood disorders.

DOI10.17148/IJARCCE.2017.6847

In the layers, the used function is the tan-sigmoid function. According to the results of changing the number of neurons in each layer and the number of layers, the best answer is from a network with one hidden layer and 15 neurons per layer. The exact MLP specification is shown in Table 4.

V. AUTOMATIC DIAGNOSIS OF THALASSEMIA BASED ON CLASSIFICATION METHODS

Al-Shami and Alhallis used three classification methods to diagnose thalassemia. The used methods are Decision tree, neural network and Bayesian. The used dataset in this study includes 49620 samples. The characteristics of the dataset include the effective parameters in the CBC test and gender, as shown in Table 6.

TABLE 6 PARAMETERS EFFECTIVE IN CBC TEST [21]

Feature name	Feature value for Male	Feature value for Female	Data type
WBC	4300- 10800	4300-10800	Real
RBC	4/7 -6/1	2/4- 5/4	Real
HT	12-18	12-16	Real
HCT	37-54	33-57	Real
MCV	80-98	80-98	Real
MCH	24-30	24-30	Real
MCHC	24-30	24-30	Real
RDW	11/5 -14/5	11/5 -14/5	Real
PLT	150000-450000	150000-450000	Real
Age	-	-	Integer
Gender	-	-	Binomial

Each pattern belongs to one of 7 thalassemia classes, thalassemia major, thalassemia minor, thalassemia intermedia, thalassemia minor - iron deficiency, iron deficiency, and etc .

The obtained results of the classification in terms of all features are expressed as. The Bayesian method with an average accuracy of 93.7%, decision tree with 93.64% and neural network with 95.71%.

Based on research by Seera et.al., MCV <80 and MCH <26 parameters show minor thalassemia or iron deficiency, so one of these parameters can be eliminated. Here the MCH parameter is omitted.

TABLE 7 COMPARISON OF CLASSIFICATION RESULTS WITH INITIAL DATA WITH FEATURE SELECTION [21]

Method name			Neural Networks	Bayesian	decision tree
Obtained result from first stage			95.71%	93.7%	93.64%
Neural Networks	**Results from Feature Selection**	95.48%	0		
Bayesian		94.32%		1%	
Decision tree		93.65%			0

According to the observations and experimental results obtained from Amendolia et.al. the effective parameters for the diagnosis of thalassemia include RBC, HB, HCT and MCV. As a result, the four parameters is used for categorization here, and the results indicate reduction of dimensions with the results in the previous step based on Table 7.

Given the accuracy of each method, it is concluded that the neural network has better results than the other two methods. Bayesian method with lower attribute values will have better results and the decision tree is understandable and interpretable due to the visual nature of the tree.

On the other hand, because making decision is conducted on the basis of a feature at any moment, the effect of each feature and its value can be found in the classification. Figure 3 is the created decision tree for the classification of thalassemia.

Based on created decision tree in Fig. 3, it can be said that MCV is an important and crucial parameter in thalassemia classification and Also, if MCV> 77.65 and age is older than 12.5, there is no thalassemia in person.

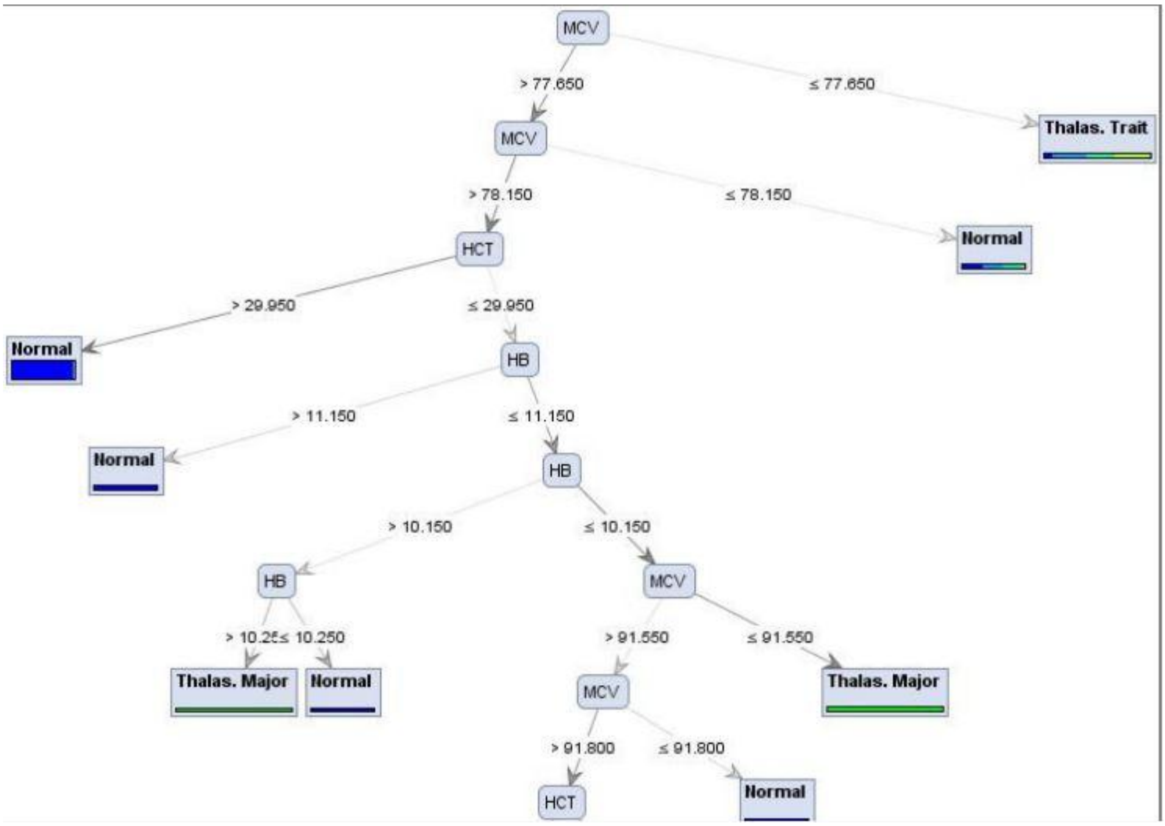

Fig. 3 Decision tree of Thalassemia diagnosis [21]

VI. EVALUATION

Here, the described methods in the previous section or using effective parameters in each method and the accuracy degree of the results are based on relation 1.

Accuracy=1- 〖Error〗 _rate (1)

In [20], the parameters of RBC, HGB, MCV and HTC were used to diagnose alpha thalassemia and beta-thalassemia. Amendolia et al. separated the patient samples from healthy samples using the three methods of MLP, KNN, and SVM and then categorized the available samples in the patient's class, alpha thalassemia and beta-thalassemia.

Based on the results of these three methods, MLP has the highest level of accuracy in the diagnosis of healthy and diseased samples and alpha thalassemia and thalassemia. In [2], in addition to thalassemia, they classified other disease such as Anemia, ITP, leukemia and lymphoproliferative tuberculosis. This requires the use of other parameters such as MCH, MCHC, RDW, PLT NEUT, LIC, RBC, HGB, MCV and WBC.

Dr.Payandeh et.al, identify the accuracy degree of Anemia class with 100% using MLP. Overall, their proposed MLP has had a good performance in the classification. In [2], in addition to laboratory parameters (RBC, HGB, MCV, MCH, MCHC, RDW, PLT, WBC and HTC), age and gender characteristics were used to diagnose the types of thalassemia and Anemia.

Al-Shami and Alhallis identify five major disease, thalassemia-major, thalassemia-intermedia, thalassemia-minor, thalassemia minor caused by iron deficiency, iron deficiency using Decision tree, Neural Network and Bayesian. Based on the results of these three methods, the neural network with 93% accuracy has the best performance in classification. Based on the results of these three methods, the neural network with 93% accuracy has the best performance in classification. According to the decision tree derived from these parameters, Al-Shami and Alhallis found that the best and most effective attribute are the four factors, RBC, HGB, MCV and HTC in thalassemia classification. Table 15 shows the used parameters in each method and its accuracy.

DOI10.17148/IJARCCE.2017.6847

TABLE 8 EVALUATION OF THE COMPARISON OF THE USED METHODS FOR DIAGNOSIS OF THALASSEMIA

Author name	Feature													Class	Method	Accuracy
	RBC	HGB	MCV	MCH	MCHC	RDW	PLT	NEUT	LUC	WBC	HTC HT	Age	Gender			
Amendolia et.al.	■	■	■								■			Alpha-thalassemia Beta thalassemia healthy	KNN MLP SVM	93.2%
Dr. Payandeh et.al.	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨				Anemia Thalassemia (ITP) Cancer Blood lymphoproliferative	MLP	99.12%
Elshami and Alhalees	■	■	■	■	■	■	■			■	■	■	■	Major Minor Intermedia Minor - Iron Deficiency iron-Deficiency Healthy and Etc.	NB	94.35%
															DT	93.64%
															NN	95.71%
	▨	▨	▨								▨	▨	▨		NB	94.32%
															DT	93.65%
															NN	95.48%

VII. CONCLUSION

Beta thalassemia is the most common monogenic disease in the Mediterranean countries. Diagnosing this disease is one of the major issues in the Hematology. This study describes the effective parameters for diagnosis of thalassemia based on the CBC test. The classification performance such as KNN, MLP, NN DT and SVM were also evaluated. Based on the results of the above methods, neural network is the best method for diagnosing thalassemia and other blood diseases using parameters such as, MCH, MCHC, RDW, PLT NEUT, LIC RBC, HGB, MCV and WBC. the most effective factors to diagnose thalassemia are RBC, HGB, MCV, and HTC, which is diagnosed the person with thalassemia by examining these parameters. Early diagnosis of thalassemia helps to decide on its treatment and will prevent the severity of the disease and its consequences.

REFERENCES

[1] Z. Rahimi, G. H. Bahrami, H. Naamaee, and M. Rezaee, "Xmnl polymorphism in 5 genes area and its relation with Hbf degree in patients with beta-thalassemia major and intermedia in Kermanshah," Kermanshah Medical Journal, 2007.

[2] M. Payandeh, M. Aeinfar, V. Aeinfar and M. Hayati, “A New Method for Diagnosisand Predicting Blood Disorder and Cancer Using Artificial Intelligence“, IJHOSCR, Vol. 3, 2009.

DOI10.17148/IJARCCE.2017.6847

[3] P. A. Maiellaro, R. Cozzolongo, P. Marino, "Artificial Neural Networks for the Prediction of Response to Interferon Plus Ribavirin Treatment in Patients with Chronic Hepatitis C," Current Pharmaceutical Design, vol. 10, pp. 2101-2109, 2004.

[4] M. F. Ganji, and M. S. Abadeh,“A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis,“ Expert Systems with Applications, vol. 38, pp. 14650–14659, 2011.

[5] M. Seera, and CH. P. Lim, “A hybrid intelligent system for medical data classification,” Expert Systems with Applications, vol.41, pp. 2239–2249, 2014.

[6] N. Dogantekin, A. Dogantekin, D. Avci, and L. Avci, ”An intelligent diagnosis system for diabetes on Linear Discriminant Analysis and Adaptive Network Based Fuzzy Inference System: LDA-ANFIS,” Digital Signal Processing, vol.20, pp. 1248–1255, 2010.

[7] A. AlJarullah, “Decision Tree Discovery for the Diagnosis of Type II Diabetes,” International Conference on Innovations in Information Technology, 2011.

[8] M. Pradhan, and R. K. Sahu, “Predict the onset of diabetes disease using Artificial Neural Network (ANN),” International Journal of Computer Science & Emerging Technologies, vol. 2, pp 2044-6004, 2011.

[9] V. A. Kumari, and R. Chitra, “Classification Of Diabetes Disease Using Support Vector Machine,” International Journal of Engineering Research and Applications, vol. 3, pp. 2248-9622, 2013.

[10] Y. Guo, G. Bai, and Y. Hu, ”Using Bayes Network for Prediction of Type-2 Diabetes,” The seventh International Conference for Internet Technology and Secured Transactions, 2012.

[11] W. M. Kattan, Editorial Comment on: Development, Validation, and Head-to-Head Comparison of Logistic Regression-Based Nomograms and Artificial Neural Network Models Predicting Prostate Cancer on Initial Extended Biopsy. Eur Urol, 2008; 54(3): 611. Epub 2008 Jan 15.

[12] Kawakami S, Numao N, Okubo Y, et al. Development, Validation, and Head-to-Head Comparison of Logistic Regression-Based Nomograms and Artificial Neural Network Models Predicting Prostate Cancer on Initial Extended Biopsy. Eur Urol, 2008; 54(3): 601-11. Epub 2008 Jan 15.

[13] F. K. H. Chun, M. Graefen, A. Briganti, et al. "Initial Biopsy Outcome Prediction Head-to-Head Comparison of a Logistic Regression-Based Nomogram versus Artificial Neural Network," Eur Urol, 2007; 51: 1236- 43.

[14] I. Kurt, M. Ture, A. T. Kurum, "Comparing Performances of Logistic Regression, Classification and Regression Tree, and Neural Networks for Predicting Coronary Artery Disease," Expert Systems with Applications, vol. 34, pp. 366- 374, 2008.

[15] M. Di Luca1, E. Grossi, B. Borroni, et al. "Artificial Neural Networks Allow the Use of Simultaneous Measurements of Alzheimer Disease Markers for Early Detection of the Disease," J Transl Med, vol. 3, pp. 1: 30. 2005.

[16] H. Uğuz, A. Öztürk, R. Saraçoğlu, et al. "A Biomedical System Based on Fuzzy Discrete Hidden Markov model for the diagnosis of the brain diseases," Expert Systems with Applications: An International Journal, vol. 35, pp. 1104-1114, 2008.

[17] X. Qiua, N. Taob, Y. Tana, and et al, "Constructing of the Risk Classification Model of Cervical Cancer by Artificial Neural Network," Expert Systems with Applications: An International Journal archive, vol. 32, pp. 1094-1099, 2007

[18] C. L. Chang, C. H. Chena, "Applying Decision Tree and Neural Network to Increase Quality of Dermatologic Diagnosis," Expert Systems with Applications, vol. 36, pp. 4035-4041, 2009.

[19] M. Mataria, G. M. Janech, J. Almeida, et al, "Prediction of Progression of Diabetic Nephropathy in a Small Set of Patients by Artificial Neural Networks and Proteomic Analysis," American Journal of Kidney Diseases, vol. 51, B67-B67, 2008.

[20] S. R. Amondelia, G. Cossu, M. L. Ganadu, B. Golosio, G. L. Masala and 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, Vol. 69, 2003.

[21] E. H. Elshami and A. M. Alhalees, “Automated Diagnosis of Thalassemia Based on DataMining Classifiers”, The International Conference on Informatics and Applications, The Society of Digital Information and Wireless Communication, 2012.

DOI10.17148/IJARCCE.2017.6847