International Journal of Innovative Computing, Information and Control
Volume **11**, Number **3**, June **2015**
ICIC International ©2015 ISSN 1349-4198
pp. **965–971**

A NEW ALGORITHM OF ENSEMBLE LEARNING FOR MEDICAL KNOWLEDGE-BASED SYSTEMS AND KNOWLEDGE-BASED SYSTEMS: HYBRID BAYESIAN COMPUTING (MULTINOMIAL LOGISTIC REGRESSION CASE-BASED C5.0-MIXED CLASSIFICATION AND REGRESSION TREE)

Patcharaporn Paokanta[1] and Somdet Srichairatanakool[2]

[1]College of Arts, Media and Technology
[2]Department of Biochemistry
Faculty of Medicine
Chiang Mai University
No. 239, Huay Kaew Road, Muang District, Chiang Mai 50200, Thailand
Patcha535@gmail.com; Patcharaporn.p@cmu.ac.th; Ssrichai@med.cmu.ac.th

Received April 2014; revised August 2014

Abstract. *This paper attempts to answer the question "How to construct and apply the novel algorithm based on Ensemble Learning approach called Bayesian Mixed Probability Distributions-CBR-C5.0-CART for Medical Knowledge-Based Systems and Knowledge-Based Systems (KBSs)?" The finding of this study is the new algorithm of Bayesian-Mixed Probability Distributions-C5.0-CART which is developed for the inference engines of KBSs. The proposed algorithm is applied to Thalassemia data set including F-cell, HbA_2, and Inclusion Body of Thalassemia patients. These are collected from medical practitioner and scientist who are the experts in Thalassemia diagnosis. In the future, this algorithm and a new collected data set will be combined with graph theory to generate the new theory called Ramsey Graph Bayesian-Mixed Probability Distributions for Digital Images Processing and Images Processing.*

Keywords: Ensemble learning, Bayesian artificial intelligence, Bayesian-mixed probability distributions, Decision trees, Multinomial logistic regression, Markov Chain Monte Carlo (MCMC), C5.0, Classification and Regression Tree (CART), Case-Based Reasoning (CBR), Medical Knowledge-Based Systems, Knowledge-Based Systems (KBSs)

1. **Introduction.**

Ensemble Learning approach is applied dramatically to Knowledge-Based Management and Knowledge-Based Systems in particular Medical KBSs in which the specific knowledge and corrected diagnostic results play an important role. The inference engine of a Medical KBS and KBS is one of the three main components of KBS which includes facts, rules and inference engine. In the previous studies of authors', several algorithms for Medical Knowledge Discovery are developed including the following.

The study of DBNs-BLR (MCMC)-GAs-KNN: a novel framework of hybrid system for Thalassemia Expert System presents the accuracy percentages of Beta-Thalassemia data set including F-Cell, HbA_2 and Sub-types of Beta-Thalassemia [1]. These indicators in three main types including Beta-Thal (Major), HbH and HbE are used for C5.0 and CART to induct the rules for the implementation of Thalassemia Expert System. The results of using both machine learning techniques show the different rules [2] and these rules of Beta-Thalassemia are combined with Case-Based Reasoning (CBR) which is one of Data and Knowledge Engineering (DKE) and this implemented technology is called Data and Knowledge Engineering Technology (DKET). Moreover, in this study, Knowledge Management theories are used to describe the specific knowledge and knowledge workers as the conceptual frameworks also [3]. On this data set, Fuzzy C-Means, K-Means clustering and Fuzzy C-Means-GAs are used to identify the groups of Beta-Thalassemia and some subtypes of Beta Thalassemia cannot be detected by Fuzzy C-Mean-GAs [4]. From the result of this study, the Beta-Thalassemia data set is improved by cutting the records that Fuzzy C-Means-GAs cannot be detected from 127 records to 60 records and some indicators such as symptoms are collected with 22 total laboratory and symptom indicators. The result of this study reveals the satisfying accuracy percentage with 94.90 by using Fuzzy C-Means-GAs and Case-Based Reasoning (CBR) [5]. Finally, there are not only CBR and Fuzzy-Based Reasoning (FBR) that are selected for inference engines implementations but Bayesian-Based Reasoning (BBR) also. In the study of Hybrid Bayesian-Based Reasoning: Multinomial Logistic Regression Classification and Regression Tree for Medical Knowledge-Based Systems and Knowledge-Based Systems, the Bayesian-Mixed Probability Distributions are generated and improved the estimated parameters by using Markov Chain Monte Carlo with Methopolis Hasting and Gibbs algorithms. The obtained satisfying model is the model with Markov Chain error 0.0112-0.2473 for 500,000 iterations [6].

According to the literature reviews of the related studies, especially in the last paper, although the result and related theories of using Hybrid Bayesian-Based Reasoning: Multinomial Logistic Regression Classification and Regression Tree are described, these algorithms including the rules of C5.0 and CART, and MLR-Based Bayesian-Mixed Probability Distributions do not be proposed with the hybridization of CBR as well as their results. As this reason, this new algorithm is illustrated in this paper as the form of suedocode and the obtained results. After this section, the organization of this paper includes the review of Ensemble Learning, the Bayesian-Based Reasoning: Multinomial Logistic Regression Case-Based C5.0-Mixed Classification and Regression Tree in the second and third section; the proposed algorithms and the experimental results are revealed in the fourth section. Finally, conclusions are shown in the fifth section.

2. **Ensemble Learning.** KM for organizations (H. Gregory, 2010) [7] is the activity focused on the collecting of experiences related to strategies of organizations. Each activity is completed by the integration of technology and organizational structure, the planning strategies with intelligence based on using knowledge to create the new knowledge. These use Artificial Intelligence which consists of organization, human and computer for discovering, storing and applying knowledge for learning to solve problems and making the decisions. Knowledge Based Management (KBM) is the activity for collecting knowledge in the forms of individual and other experiences and problem bases from experiences of working processes also.

DKET is the technology for discovering problems, solutions and solving these problems by using the discovered solutions (P. Paokanta et al., 2014) [3], as same as Intelligent System, Decision Support System, Expert System, Hybrid System, Information System, etc. KBS is one popular KBM which is implemented for medical systems. A main part of KBS components is the inference engine which is usually implemented by DKET. In the authors' perspective, DKET can be separated as two types including, qualitative and quantitative DKET techniques. The qualitative DKET techniques include the interviewing techniques, questionnaire, flow chart, E-R diagram, EER diagram, Fish-bone diagram, Analytic network, etc. On the other hand, the quantitative DKET techniques are well known as Knowledge Discovery (KD) techniques such as Machine learning, Statistics, and Soft Computing in which these KD algorithms are known as Reasoning methods.

Among the reasoning methods for implementing inference engine, DKET approach such as Artificial Neural Network-Based Reasoning (ANNBR), Case-Based Reasoning (CBR), Fuzzy-Based Reasoning (FBR) [5,8], Ensemble Learning is the combination of methods and/or training data sets and/or algorithms and/or methodologies for improving their results. Due to the previous studies, multiple methods obtain more accuracy than a single method [9]. Ensemble Learning can be separated as two main types including parallel and hierarchy architecture [10].

Parallel architecture: Ensemble Learning in parallel architecture is the multiple methods and/or training data sets and/or algorithms and/or methodologies processed together. It means that one process can work together with another process at the same time. The Ensemble Learning schemes in this type are voting, sum, mean, median, fuzzy integrals, etc.

3. **Bayesian-Based Reasoning: Multinomial Logistic Regression Case-Based C5.0-Mixed Classification and Regression Tree.** In this paper, Bayesian-Mixed Probability Distributions are constructed to estimate the parameters for inference engines of KBSs by selecting the facts from the results of C5.0 and CART which are the Decision Trees algorithms.

Because of the publications of authors for Bayesian-Based Reasoning: Multinomial Logistic Regression Classification and Regression Tree in [2,6], in this paper, authors present only the theoretical review of CBR. For more detail of the remaining algorithms can search from the proposed references.

In the case that the same problems such as medical diagnostic problems and Judgment problems occur usually, as well as solutions are found, these are stored in the Case-Base of the inference engines in the KBSs. Case-Base is a branch of Artificial Intelligence approach. Case has two components including problem and solution. The concept of CBR is the process to discover the solution by using the previous cases to train in the problem solving process. This method is the iterated learning. It is the transferring process of Tacit Knowledge from expert to Explicit Knowledge, Tacit Knowledge to Embedded Knowledge and Explicit Knowledge to Embedded Knowledge. When the experts resigned from organization, their knowledge is still in organizations as the Knowledge-Based Systems. Case-Base can be designed as two types including

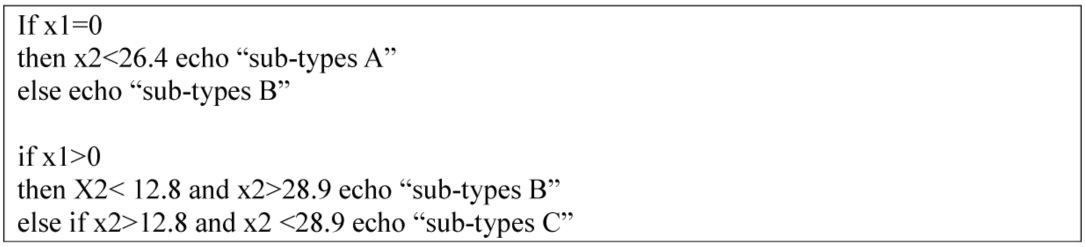
1. Case-Based Interpretation (To classify groups of cases)
2. Case-Based Problem Solving (To store problems and their solutions)

In this study, CBR is developed for the inference engine of KBS based on classical approach in which the indentified corrected types of Beta-Thalassemia are selected by mapping the results of C5.0, CART and expert opinions to discover the new Thalassemia knowledge before using the Hybrid Bayesian computing (Multinomial Logistic Regression (MCMC)-Mixed Probability Distributions).

4. **Algorithms and Experimental Results.** In this section, the algorithms and experimental results of the Bayesian-Based Reasoning: Multinomial Logistic Regression Case-Based C5.0-Mixed Classification and Regression Trees are demonstrated as below.

4.1. **Induction rules (C5.0) algorithm.** The algorithm of induction rules C5.0 algorithm shows in Figure 1.

Figure 1 presents that the induction rules (C5.0) algorithm can classify sub-types of Thalassemia as three sub-types, A, B and C.

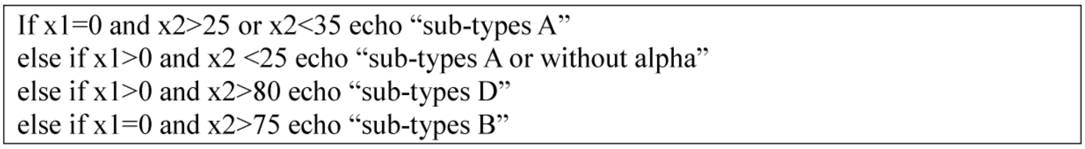


```
If x1=0
then x2<26.4 echo “sub-types A”
else echo “sub-types B”

if x1>0
then X2< 12.8 and x2>28.9 echo “sub-types B”
else if x2>12.8 and x2 <28.9 echo “sub-types C”
```

Figure 1. Induction rules (C5.0) algorithm

4.2. **Induction rules (CART) algorithm.** The algorithm of induction rules CART algorithm shows in Figure 2.



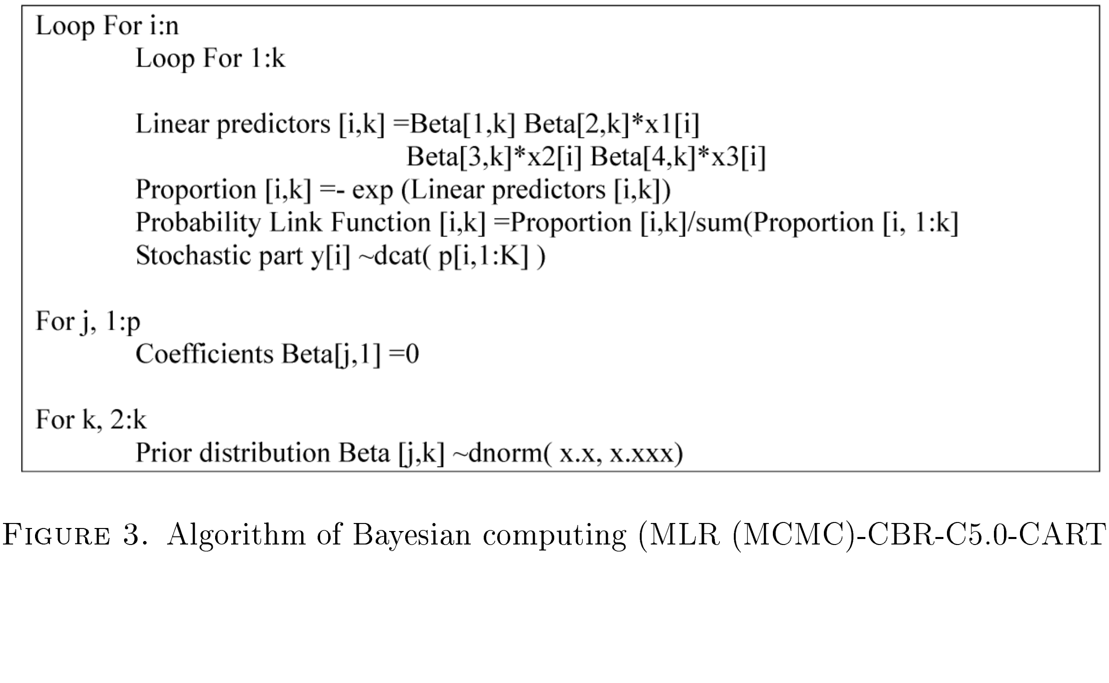
```
If x1=0 and x2>25 or x2<35 echo “sub-types A”
else if x1>0 and x2 <25 echo “sub-types A or without alpha”
else if x1>0 and x2>80 echo “sub-types D”
else if x1=0 and x2>75 echo “sub-types B”
```

Figure 2. Induction rules (CART) algorithm

Figure 2 presents the induction rules (CART) algorithm which obtains four sub-types of Thalassemia, A, B, D and A or without Alpha. The obtained results of using C5.0 and CART obtain the different sub-types. CART can detect sub-types D and on the other hand C5.0 cannot detect it. These algorithms are implemented from the results of using C5.0 and CART [2].

4.3. **Algorithm of Bayesian computing (MLR (MCMC)-CBR-C5.0-CART).** After the obtained algorithms of C5.0 and CART are implemented in the form of Web-KBS, the results of these algorithms are selected by the inference engine developed as CBR. Finally, the results of using C5.0, CART and CBR are used to be the input training by the algorithm of Bayesian Computing (MLR (MCMC)-CBR-C5.0-CART) shown in Figure 3.

Figure 3 presents that algorithm of Bayesian Computing (MLR (MCMC)-CBR-C5.0-CART). This algorithm generates the coefficients for Multinomial Logistic Regression (MLR) based on MCMC method through using WinBugs software. The mixed probability



```
Loop For i:n
        Loop For 1:k

        Linear predictors [i,k] =Beta[1,k] Beta[2,k]*x1[i]
                                   Beta[3,k]*x2[i] Beta[4,k]*x3[i]
        Proportion [i,k] =- exp (Linear predictors [i,k])
        Probability Link Function [i,k] =Proportion [i,k]/sum(Proportion [i, 1:k]
        Stochastic part y[i] ~dcat( p[i,1:K] )

For j, 1:p
        Coefficients Beta[j,1] =0

For k, 2:k
        Prior distribution Beta [j,k] ~dnorm( x.x, x.xxx)
```

Figure 3. Algorithm of Bayesian computing (MLR (MCMC)-CBR-C5.0-CART)

distributions between Normal distribution and Category distribution are generated for linear logistic model.

4.4. **Classification by hybrid Bayesian computing (MLR (MCMC)-CBR-C5.0-CART).** Due to the results of using Ensemble Learning called Hybrid Bayesian Computing (MLR (MCMC)-CBR-C5.0-CART), the estimated parameters are applied to classify sub-types of Thalassemia through the inference engine of Web-KBS. The obtained results of each algorithm show in Table 1.

TABLE 1. Classification performance of hybrid Bayesian computing (MLR (MCMC)-CBR-C5.0-CART) for Thalassemia Web-KBSs and KBSs

Measurements	*C5.0 -Expert*	*CART -Expert*	*CBR -C5.0-CART*	*CBR-C5.0 -CART-Expert*	*MLR (MCMC)-CBR -C5.0-CART-Expert*
Corrected Items	*127*	*98*	*114*	*97*	*97*
Uncorrected Items	*20*	*29*	*13*	*30*	*30*
Total	*127*	*127*	*127*	*127*	*127*
Accuracy Percentages	*84.25*	*77.17*	*89.76*	*76.38*	*76.38*

Table 1 reveals the classification performance of C5.0, CART, CBR-C5.0-CART and Hybrid Bayesian Computing (MLR (MCMC)-CBR-C5.0-CART) based on with or/and without the expert opinions. The result of Hybrid Bayesian Computing (MLR (MCMC)-CBR-C5.0-CART) for Thalassemia Web-KBS is as same as CBR-C5.0-CART-Expert with 76.38 accuracy percentages. On the other hand, C5.0 and CART-Expert obtain 84.25 and 77.17 accuracy percentages, respectively. Besides, the results of the combination of C5.0 and CART using CBR for the inference engine obtain 89.76 accuracy percentages in the case without expert opinions, but the case of expert opinions, this algorithm reaches 76.38 accuracy percentages.

5. **Conclusions.** The Ensemble Learning algorithm, the Bayesian-Based Reasoning: Multinomial Logistic Regression Case-Based C5.0-Mixed Classification and Regression Tree is proposed to classify the sub-types of Thalassemia. This is the new methodology and algorithm for managing the specific knowledge such as medical knowledge. Although the applications of MLR (MCMC), C5.0, CBR and CART are widely implemented, they are developed in the single algorithm. For example, the study of S. R. Amendolia et al. show Thalassemia screening indicators using Principle Component Analysis (PCA) in which the selected features are Red Blood Cell, Hb, Ht and MCV. They compare the study of K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) for Thalassemia screening and the specificity of MLP is better than SVM with 95 percentages, even though the sensitivity of MLP and SVM are 92 and 83 percentages, respectively [11].

In addition, the study of W. Wongseree et al. present Thalassemia classification by neural networks and Genetic Programming (GP) in which the data sets are a mature Red Blood Cell, a reticulocyte and a platelet. 20 indicators are collected to classify by using MLP and GP with the 90 percentages classification accuracy for 13 input features and 82 percentages for 15 input features [12].

The other related study, T. Piroonratana et al. propose the classification of hemoglobin typing chromatograms by neural networks and decision trees for thalassemia screening and the best result is 97.24 percentages by using C4.5 for classifying 13 classes based on discrete attributes of HPLC results. On the other hand, the results of Random forests and MLP obtained 96.00 and 93.44 percentages, respectively [13].

Moreover, the study of D. Setsirichok et al. present classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naive Bayes classifier and a multilayer perceptron for thalassaemia screening. The obtained results of using naive Bayes classifier and a multilayer perceptron to classify 18 classes with 8 input features are 93.23, 92.60 percentages, respectively [14].

According to the previous study, the different summarization between them and this study including the following.

1. The data sets: This study uses 24 input features including 12 variables collected from laboratory and 10 symptom variables of – Thalassemia patients which differ from the data sets of the previous studies. Moreover, the results of C5.0 and CART are used to input for improving the performance of methodology also.

2. The purpose of the study: The attempt of this study is to discover the suitable methodology or the new algorithms for Thalassemia WEB-KBSs and KBSs based on the incomplete information (Classify based on different data sets).

3. Algorithms: In this study, several Aritificial Intelligence and DKE and DKET such as Statistical-Based Reasoning (SBR), ANNBR, FBR, EBR are used to discover the appropriate results and algorithms. Moreover, the Ensemble Learning is applied also.

In conclusion, the objective of this study is to discover the Novel suitable Knowledge Discovery methodologies and algorithms for Web-Knowledge-Based Systems (Web-KBSs) and KBSs implementation by using Data and Knowledge Engineering (DKE) which is the technique to discover problems, solutions and solving these problems by using the discovered solutions (P. Paokanta et al., 2014) [3] in the form of using DKET, Knowledge Management theories (KM Processes, Mental model and Systems Thinking).

The results of using these to construct the novel systematic framework based on Systems Thinking and Knowledge processes for discovering the appropriate methodology for developing the inference engines of Web-Knowledge-Based Systems (Web-KBSs) and KBSs present that the using Machine Learning methods and Ensemble Learning such as Hybrid Statistical-Based Reasoning (SBR) and Hybrid Artificial Neural Network (ANNBR), Binomial Logistic Regression (Maximum Likelihood)-Fuzzy C-Means GAs-CBR-C5.0-CART, CBR-C5.0-CART, CBR-C5.0-CART-Expert, Bayesian-Based Reasoning: Multinomial Logistic Regression Classification and Regression Tree, and Bayesian-Based Reasoning: Multinomial Logistic Case-Based C5.0-Mixed Regression Classification and Regression Tree obtain the satisfied performance with 100.00 accuracy percentage.

This proposed methodology generates the solutions through using If-Then rules in the SBR-inference engine mechanism which evaluates the suitable decision making results and improves the classification performance of Medical-KBSs and KBSs. Moreover, these systems can update and collect the appropriate data sets and the best results also. In the future, the Ramsey-Graph Bayesian-Mixed Probability Distributions are implemented for discovering the new algorithms or improving the performance of Thalassemia Web-KBSs and KBSs. Moreover, Thalassemia digital images are collected for this novel algorithm also.

Acknowledgment. This work is partially supported by Thalassemia Clinic and Faculty of Medicine, Chiang Mai University, Thailand. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the quality of this paper.

REFERENCES

[1] P. Paokanta, DBNs-BLR (MCMC)-GAs-KNN: A novel framework of hybrid system for Thalassemia expert system, *Lecture Notes in Computer Science*, vol.7666, no.4, pp.264-271, 2012.

[2] P. Paokanta, M. Ceccarelli, N. Harnpornchai, N. Chakpitak and S. Srichairatanakool, Rule induction for screening Thalassemia using machine learning techniques: C5.0 and CART, *ICIC Express Letters*, vol.6, no.2, pp.301-306, 2012.

[3] P. Paokanta, N. Harnpornchai and M. Ceccarelli, A knowledge creation innovation for web-knowledge-based system using knowledge management, and data and knowledge engineering technology, *Proc. of the 15th European Conference on Knowledge Management*, Santarem, Potugal, vol.3, pp.1255-1264, 2014.

[4] P. Paokanta, N. Harnpornchai, N. Chakpitak, S. Srichairatanakool and M. Ceccarelli, Knowledge and data engineering: Fuzzy approach and genetic algorithms for clustering – Thalassemia of knowledge based diagnosis decision support system, *ICIC Express Letters*, vol.7, no.2, pp.479-484, 2013.

[5] P. Paokanta, A new methodology for web-knowledge-based system using systematic thinking, KM process and data & knowledge engineering technology: FBR-GAs-CBR-C5.0-CART, *International Journal of Engineering and Technology*, vol.5, no.5, pp.4320-4325, 2013.

[6] P. Paokanta, Chapter 17: A novel hybrid Bayesian-based reasoning: Multinomial logistic regression classification and regression tree for medical knowledge-based systems and knowledge-based systems, *Case Studies in Intelligent Computing – Achievements and Trends*, pp.363-378, 2014.

[7] H. Gregory, *An Introduction to Knowledge Management*, The Wasson Center, www.slideshare/hgregory.com, 2010.

[8] P. Paokanta, M. Ceccarelli and S. Srichairatanakool, The efficiency of data types for classification performance of machine learning techniques for screening – Thalassemia, *Proc. of the 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies*, Rome, Italy, pp.1-4, 2010.

[9] D. Opitz and R. Maclin, Popular ensemble methods: An empirical study, *Journal of Artificial Intelligence Research*, vol.11, pp.169-198, 1999.

[10] M. Sewell, *Ensemble Learning, Research Note*, Department of Computer Science, University College London, London, U.K., 2011.

[11] S. R. Amendolia, 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, no.1, pp.13-20, 2003.

[12] W. Wongseree, N. Chaiyaratana, K. Vichittumaros, P. Winichagoon and S. Fucharoen, Thalassaemia classification by neural networks and genetic programming, *Information Sciences*, vol.177, pp.771-786, 2007.

[13] T. Piroonratana, W. Wongseree, A. Assawamakin, N. Paulkhaolarn, C. Kanjanakorn, M. Sirikong, W. Thongnoppakhun, C. Limwongse and N. Chaiyaratana, Classification of haemoglobin typing chromatograms by neural networks and decision trees for thalassaemia screening, *Chemometrics and Intelligent Laboratory Systems*, vol.99, pp.101-110, 2009.

[14] D. Setsirichok, T. Piroonratana, W. Wongsereea, T. Usavanarong, N. Paulkhaolarn, C. Kanjanakorn, M. Sirikong, C. Limwongse and N. Chaiyaratana, Classification of complete blood count and haemoglobin typing data by a C4.5 decision tree, a naive Bayes classifier and a multilayer perceptron for thalassaemia screening, *Biomedical Signal Processing and Control*, vol.7, pp.202-212, 2012.