Anemia types prediction based on data mining classification algorithms

Article · November 2016

CITATIONS
33

READS
9,457

2 authors:

Manal Abdullah
King Abdulaziz University
84 PUBLICATIONS **685** CITATIONS
SEE PROFILE

Salma Al-Asmari
King Khalid University
3 PUBLICATIONS **45** CITATIONS
SEE PROFILE

Communication, Management and Information Technology – Sampaio de Alencar (Ed.)
© 2017 Taylor & Francis Group, London, ISBN 978-1-138-02972-9

Anemia types prediction based on data mining classification algorithms

Manal Abdullah
Department of Computer Science, Faculty of Computing and Information Technology, King Abdul-Aziz University, Jeddah, Saudi Arabia

Salma Al-Asmari
Department of Computer Science, Faculty of Computing, King Khalid University, Abha, Saudi Arabia

ABSTRACT: Medical Data Mining domain concerned with prediction knowledge as a method to extract desired outcomes from data for specific purposes. Anemia is one of the most common hematological diseases and in this study concentrate on the most five common types of anemia. This paper specifies the anemia type for the anemic patients through a predictive model conducting some data mining classification algorithms. The real data of dataset constructed from the Complete Blood Count (CBC) test results of the patients. These data filtered and eliminated undesirable variables, then implemented on some classification algorithms such as Naïve Bayes, Multilayer Perception, J48 and SMO using WEKA data-mining tool. Several experiments has proven that J48 decision tree algorithm gives the best potential classification of anemia types. WEKA experimenter proves J48 decision tree algorithm has the best performance with accuracy, precision, recall, True Positive rate, False Positive rate and F-measure.

Keywords: Anemia, Medical Data mining, classification algorithms, naïve Bayes, J48 decision tree, Support vector machine, SMO

1 INTRODUCTION

Data mining concept is sorting the data to identify patterns and find relationships between these data. It is techniques are appropriate for simple or structured datasets such as relational databases, transactional databases. Different approaches of data mining proposed to improve the challenges of storing and processing all types of data (Kaur et al., 2015 & Kishore et al., 2015).

Data mining has three basic mechanisms Clustering (Classification), Decision Rules and Analysis. Classification analyzes a set of data and produces a set of decision rules, which used to classify the data sets. In the artificial intelligence, machine learning or database systems data mining process is starting by extract the information from dataset then convert it to meaning full structure. This means that it determines patterns in datasets and embracing methods. There are many classes in data mining where the most common one is classification, which is used to predict set of relationship between data. In healthcare, it is significant to invest the development in computer technology to enhance processing the medical data such as data mining classification algorithms and tools. This paper will utilize the WEKA tool for data mining (Shouval et al., 2014). As data mining tool, WEKA name is derived from Waikato Environment for Knowledge Analysis. It is an open source data-mining tool that provides an efficient framework for implementing several classification algorithms. This tool provides processing the datasets and filtering out and remove irrelevant (not useful) data and the dataset can be incision into test and training sets. It supports perform classification algorithms then transforming all the dataset into appropriate pattern as a machine learning form. WEKA also can upload different file formats such as ARFF, CVS, C4.5 and different databases Garner (1995).

There are growing researches interest in using data mining in the medical domain. Developing in this new approach, called medical data mining, concerned with developing systems that determine and predict knowledge from data generating from medical environments. The data mining in the medical domain specifically the hospital database, including the data, which is huge in amounts, complex in contents, with heterogeneous types, hierarchical and varying in quality. Among last years, the information on laboratories keeps on enhancing and developing. The specific patterns of information can predicated through using data mining methodologies to enhance conducting researches and evaluation of reports. The data mining classification depends on similarities existing in the data. The classification algorithms used to prove

the results is acceptable to the doctors or the end user. Medical data mining uses many algorithms such as Decision Trees, Neural Networks, Naïve Bayes and others.

This paper identifies set of attributes associated with the patient CBC test result that give the anemia type, and improve the quality of prediction by identifying the anemic patients, so that can help doctors immediately improving their performance. This paper investigates the accuracy of some classification algorithms in predicting some anemia types. It is also utilizing WEKA tool for conducting classification, decision rules and analyzing the results. The evaluation of data using classification algorithms takes a set of classified data as training set and use it for training the algorithms. Then classifies the test data based on the decision rules extracted from the training set for predicting anemia diseases. The use of WEKA Experimenter conducted to specify which classification algorithm gives best performance in terms of accuracy, precision, recall, True Positive rate, False Positive rate and F-measure. The main objectives of this work are: using predictive attributes for producing data and performing data mining algorithms to get the best prediction of the anemia types using the patient Complete Blood Count (CBC) data results.

2 RELATED WORKS

There are many works that used different data mining algorithms to classify several types of diseases, such as anemia disease for specific types based on Data Mining algorithms Elshami & Alhalees (2012). In addition, many other researchers tried to find their own method. A person with anemia probably unaware of the problem because symptoms may not appear. Millions of people may have anemia and their health exposed risk. Therefore the disease is significant, several studies carried out in this domain mentioned in the literature (Yilmaz et al., 2013). (Sanap et al., 2011) developed a system using the classification technique: C4.5 decision tree algorithm and SMO support vector machine WEKA. They implemented a number of experiments using these algorithms. The anemia classification using decision tree that given clear results depend on CBC reports. (Amin et al., 2015) have compared between naïve Bayes, J48 classifier and neural network classification algorithms using WEKA and working on hematological data to specify what the best and appropriate algorithm. The proposed model can predict hematological data and the results showed that the best algorithm is J48 classifier with high accuracy and naïve Bayes is the lowest average in average errors. The study of (Sanap et al., 2011) and (Amin et al., 2015) proved that the C4.5 algorithm (as J48 in WEKA) results gives high accuracy more than other classifiers. Dogan & Turkoglu (2008) based on the biochemistry blood parameters they designed a system to help physicians in the diagnosis of Anemia. The system designed using the decision tree algorithm. The system used the characteristics of the hematology and classify the results into positive or negative Anemia. The results of this system accorded with physicians' decision. Siadaty & Knaus (2006) selected decision trees as a common and simple classifier and also has low computational complexity. The problem was the needed time to build a decision tree for large dataset is come to be intractable. They solved the problem by developing a parallel model of ID3 algorithm. It is a thread-level parallelism decision tree and do the computations independently. The experiment done on anemic patient's data set. (Kishore et al., 2015) presented set of the basic classification algorithms, which groupof essential types of classification methods such as decision trees, Bayesian networks, k-nearest neighbor and support vector machine classifier. The study shows a comprehensive review of diverse classification algorithms in data mining. This research presents an investigation for five types of anemia disease by using naïve Bayes, Multilayer perception, J48 decision tree and support vector machine data mining algorithms depending on CBC data. The best one of classification algorithms depends on specifically in the problem domain Kesavaraj & Sukumaran (2013).

3 ANEMIA CLASSIFICATION

3.1 *What is anemia?*

It is a medical condition indicates to the reduction of hemoglobin or red cell concentration in the human blood. A Complete Blood Cell (CBC) count test conducted for patients in laboratory. The anemia disease types identified using this information: age, gender, hemoglobin, Hematocrit and other attribute values when it is lower a normal range Green (2012). Anemia types classification according to CBC test values illustrated in Fig. 1 (Sanap et al., 2011).

3.2 *The anemia classification*

Anemia disease categorized into different types based on the CBC test values. In this model Anemia types nomenclature illustrated (see Table 1) and classified according to MCV (Mean corpuscular volume) value into the three essential kinds of microcytic (MCV<80) ft, normocytic

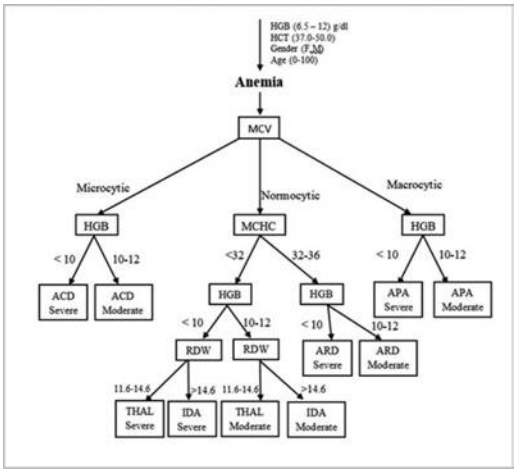

Figure 1. Anemia types classification.

Table 1. ANEMIA types nomenclature.

ACD	Anemia of chronic disease
IDA	Iron deficiency anemia
ARD	Anemia of renal disease
THAL	Thalassemia
APA	Aplastic Anemia

(MCV = 80–100) ft, and macrocytic (MCV>100) ft anemia, and classified using MCHC (Mean corpuscular hemoglobin concentration) into normochromic (MCHC = 32–36) g/dl and hypochromic (MCHC<32) g/dl anemia. RDW (Red Cell Distribution Width) used to measure the anemia and it is high if (RDW>14.6) and normal if (RDW = 11.6–14.6) Green, (2012) and (Sanap et al., 2011).

4 THE PROPOSEDMETHOD

4.1 *Experimental setup*

In the context of classification the anemia types. A number of attributes are considered to predict the type of anemia for the anemic patient. These influencing attributes are categorized as an input. The data is taken from Complete Blood Count (CBC) test results, which are conducted by collecting blood samples from 41 anemic patients (41 instances) and constructing ANEMIA dataset. The dataset consists of 7 attributes and defined in Table 2 along with their values.

Then data is transformed into a standard file format. CSV, which is supported by the WEKA tool to construct ANEMIA dataset, filtered and eliminating out irrelevant data using specific techniques. The CBC data contain 34 irrelevant attributes that are removed. The relevant attributes are shown in Table 2. The attributes are verities between nominal and numeric values and each has its own determined category.

Table 2. Attributes of ANEMIA dataset.

Attribute	Attribute value	Attribute category
Age	0–12	Child
	>12	Adult
Gender	Female	F
	Male	M
MCV	<80	Microcytic
	80–100	Normocytic
	>100	Macrocytic
HCT	<37	Low
	37.0–50.0	Normal
HGB	<10	Severe
	10–12	Moderate
MCHC	<32	hypochromic
	32–36	normochromic
RDW	>14.6	High
	11.6–14.6	Normal

The classification algorithms performed for predicting and classifying five most common Anemia types based on rules that shown in Table 3. The analysis of identifying anemia types are conducted using the WEKA tool. (Siadaty et al., 2006, Sanap et al., 2011 and Shashidhara, 2012):

The implementation of the proposed method starts by collecting CBC results and build our own dataset. Then data are preprocessed to extract and filter the attributes of importance. Data are converted to CSV format to be able using by WEKA classifier software. CSV file format is selected to allows data to be saved in a table structured (spreadsheet) format.After the classification and generated results, evaluated using the WEKA experimenter and the Knowledge Flow Model.

4.2 *The proposed algorithms for classification*

In this method, various data mining algorithms are used for predicting the anemia type for patients. During this study, classification algorithms used for prediction and the dataset are tested then analyzed with four candidate algorithms which are: Naïve Bayes, neural network (multilayer perception), Decision Tree (J48) and Support Vector Machine (SMO). The Naive Bayes algorithm implements the principle of conditional probabilities that computes a probability by calculating the rate of values and combinations of values in the specific data. This algorithm determines the probability of an event happen given the probability of another event that has already happened. Naïve Bayes algorithm use

Table 3. Anemia classification rules.

The rule	Decision*
IF (MCV = microcytic AND HGB = 10–12) then	ACD, moderate
Else if (MCV = microcytic AND HGB = <10) then	ACD, severe
Else if (MCV = normocytic AND MCHC <32 AND RDW = 11.6–14.6 AND HGB = 10–12) then	THAL, moderate
Else if (MCV = normocytic AND MCHC <32 AND RDW = 11.6–14.6 AND HGB = <10) then	THAL, severe
Else if (MCV = normocytic AND MCHC <32 AND RDW = 11.6–14.6 AND HGB = 10–12) then	IDA, moderate
Else if (MCV = normocytic AND MCHC <32 AND RDW = 11.6–14.6 AND HGB = <10) then	IDA, severe
Else if (MCV = normocytic AND MCHC = 32–36 AND HGB = 10–12) then	ARD, moderate
Else if (MCV = normocytic AND MCHC = 32–36 AND HGB = <10) then	ARD, severe
Else if (MCV = macrocytic AND HGB = 10–12) then	APA, moderate
Else if (MCV = macrocytic AND HGB = <10)	APA, severe

*The decision includes (Anemia type and severity grade).

kernel density estimators that improve implementation if the normal assumption clearly correct; it can also deal with numeric attributes using supervised discretization Vijayarani & Muthulakshmi (2013). The second algorithm is a neural network in WEKA named (multilayer perception). It is a feed forward neural network multilayer model that can map set of the input data (each one is a neuron) into a set of suitable outputs. The input node is an element with a nonlinear activation function. The multilayer perception consists of multiple one or more of hidden layers of nodes called (hidden neurons) in a directed chart, with each layer completely connected to the next layer (Prakash et al., 2015). The J48 decision tree algorithm is used also for automatic processing and canchoose related aspects from training data. It can cut the meaningless approaches into effective process, especially when dealing with continuous attributes. It split the values based on the thresholding to specify what is upper than, less than or equal to the threshold value. J48 algorithm contains the capability of dealing with training data with missing values of some attributes (Ahmad et al., 2011). Support Vector Machines named (SMO) in WEKA used as a supervised learning method which analyzing data and recognizing patterns. It is not probable classifier, which process set of input data and

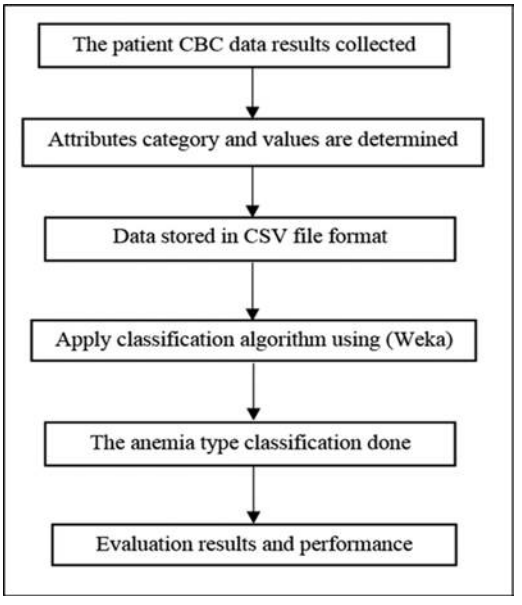

Figure 2. Flowchart of proposed method.

classify it into what is the two probable classes and give the output. The SVM algorithm has the same functional form of neural networks and radial basis functions Kesavaraj & Sukumaran (2013). It is generally used to a two class classification problem, its detect the plane and gives the greatest separation between the two classes. The SVM algorithm discovers the optimal plane with a maximum distance to the nearby point of the two classes. A set of instances that are closest to the optimal plane, explains the support vector and specify the margins of each class (Shouval et al., 2014). See the description of the proposed methodology illustrated as flowchart shown in Fig. 2.

5 RESULTS AND DISCUSSION

Evaluation of data is done by using 41 instances in the dataset using Naïve Bayes, neural network in WEKA (multilayer perception), J48 decision tree algorithms, and support vector machine in WEKA (SMO) with the test option: several percentages splits (20%, 40%, 60%) of the dataset see Table 4.

The results in Table 4 of evaluation an ANEMIA dataset using WEKA through different experiments 20%, 40%, 60% percentage split data. The table include the result through accuracy (correctly classified instances), mean absolute error, weighted average ROC and F-measure. Fig. 3 show the SMO algorithm results using 60% training set data.

Table 4. Simulation result of algorithms using 20%, 40%, 60% training set data.

Algorithm	Training Set	Accuracy*%	Mean absolute error%	Weighted av. ROC	F-Measure
Naïve Bayes	20%	30.303	0.458	0.507	0.257
	40%	60	0.3372	0.708	0.587
	60%	68.75	0.2645	0.825	0.68
Multilayer Perception	20%	39.3939	0.3744	0.775	0.383
	40%	72	0.2198	0.852	0.716
	60%	87.5	0.1372	0.921	0.859
J48 Decision tree	20%	27.2727	0.3207	0.855	0.218
	40%	88	0.1689	0.868	0.878
	60%	93.75	0.1743	0.97	0.935
SMO	20%	39.3939	0.4108	0.677	0.396
	40%	84	0.2578	0.902	0.83
	60%	93.75	0.2361	0.96	0.912

*Correctly Classified Instances.

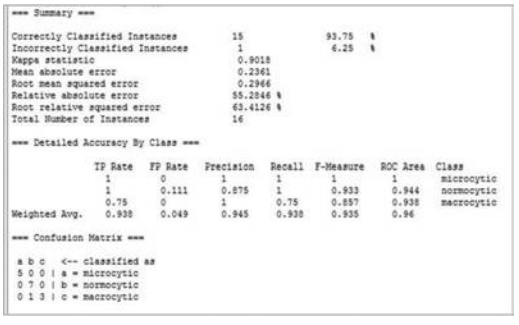

Figure 3. Support vector machine (SMO) algorithm output using 60% training set data.

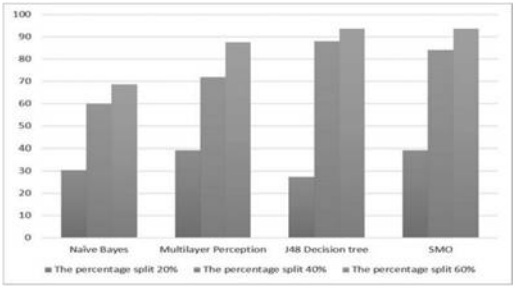

Figure 4. Comparing algorithms accuracy using the percentage split 60%.

The test that using percentage split is conducted by deciding a specific percent of data for training and the rest of data for testing. In this experiment the percentage split are chosen as 20%, 40% and 60%, where the partitions is conducted randomly. The percentage split 20%: the data will split into 20% will used as training set data and the rest 80% will used as testing set data. The same process done with other percentages 40% and 60%.

The accuracy (Correctly Classified Instances) rate of the results using different splitting percentages increased in naïve Bayes, J48, multilayer perception and SMO. The accuracy increasing with the training set average respectively. All statistic results provide an important comparison of the accuracy between all algorithms done and finally it have been investigated that J48 decision tree and SMO algorithms implement best results with accuracy 93.75% when using the percentage split 60%. The accuracy measure of all the algorithms using 60% training set are illustrated in Fig. 4.

The results shown in the Table 5 are the performance of naïve Bayes, neural network (multilayer perception), J48 decision tree and SMO using WEKA experimenter. The data mining measures in the table illustrates more useful and precise evaluation of algorithm's performance, especially when dealing with datasets: recall (sensitivity), precision, F-measures, true positive rate and false positive rate, which computed as follows:

Recall (sensitivity) = True Positive rate/(True Positive rate + False Negative rate).

Precision = True Positive rate/(True Positive rate + False Positive rate).

F-measure = (2 *recall *precision)/(recall + precision).

The True Positive rate is the number of positive instances classified correctly, The False Negative rate is the number of positive instances (records) classified negatively; False Positive rate is the number of negative instances classified positively (Huang et al., 2012).

In the context of using WEKA experimenter a snapshot of using F-measure illustrated in Fig. 5, using the precision in Fig. 6 and using the TP rate

Table 5. Comparison of classification algorithms.

Algorithm	TP Rate	FP Rate	Precision	F-Measure	Recall
Naïve Bayes	0.92	0.10	0.93	0.91	0.92
Multilayer Perception	0.92	0.10	0.95	0.91	0.92
J48 Decision tree	0.93	0.05	0.97	0.93	0.93
SMO	0.90	0.40	0.85	0.84	0.90

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  F_measure
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       29/12/15 01:44 μ

Dataset                   (1) bayes.N | (2) func (3) func (4) tree
-----------------------------------------------------------------
ANEMIA                    (10)   0.91 |   0.91     0.84     0.93
-----------------------------------------------------------------
                                (v/ /*) |  (0/1/0)  (0/1/0)  (0/1/0)

Key:
(1) bayes.NaiveBayes
(2) functions.MultilayerPerceptron
(3) functions.SMO
(4) trees.J48
```

Figure 5. Comparing algorithms with use the WEKA experimenter using F-measure.

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  IR_precision
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       29/12/15 01:58 μ

Dataset                   (1) bayes.N | (2) func (3) func (4) tree
-----------------------------------------------------------------
ANEMIA                    (10)   0.93 |   0.95     0.85     0.97
-----------------------------------------------------------------
                                (v/ /*) |  (0/1/0)  (0/1/0)  (0/1/0)

Key:
(1) bayes.NaiveBayes
(2) functions.MultilayerPerceptron
(3) functions.SMO
(4) trees.J48
```

Figure 6. Comparing algorithms with use the WEKA experimenter using precision.

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  True_positive_rate
Datasets:   1
Resultsets: 4
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       30/12/15 01:23 μ

Dataset                   (1) bayes.N | (2) func (3) tree (4) func
-----------------------------------------------------------------
ANEMIA                    (10)   0.92 |   0.92     0.93     0.90
-----------------------------------------------------------------
                                (v/ /*) |  (0/1/0)  (0/1/0)  (0/1/0)

Key:
(1) bayes.NaiveBayes
(2) functions.MultilayerPerceptron
(3) trees.J48
(4) functions.SMO
```

Figure 7. Comparing algorithms with use the WEKA experimenter using true positive rate.

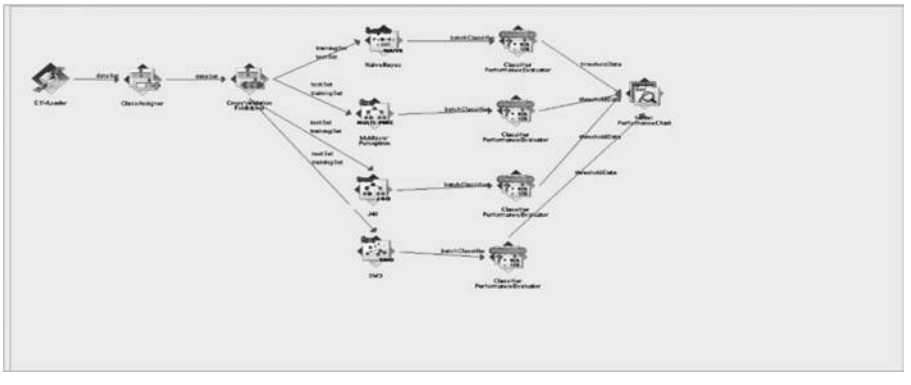

Figure 8. Knowledge flow model using WEKA.

in Fig. 7. In these experiments, it has shown that J48 decision tree performs best among four algorithms with F-Measure 93%, Sensitivity is 93%, true positive rate is 93%, Precisions 97% and it is the lowest in the false positive rate 0.05.

The comparative performance based on the accuracy among four algorithms also conducted by using knowledge flow model shown in Fig. 8, which shows the membership tree structure using 10 folds validation test.

The performance chart of knowledge flow modelconducted for the experiment algorithms Naive Bayes, Multilayer Perceptron, J48 and SMO. It is another important performance measures in WEKA.The performance represented by the Region of meeting Curve (ROC) for each

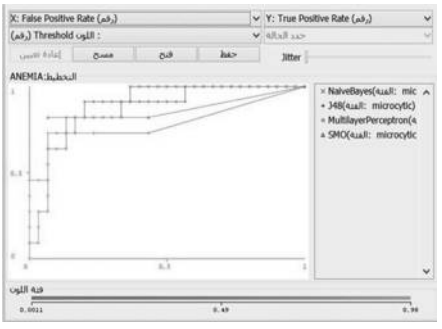

Figure 9. Performance chart of (ROC) curve.

algorithm based on 10 folds validation test. From the Fig. 9, it is clearly shown that J48 decision tree has the highest weighted average ROC0.97.

6 CONCLUSION AND FUTURE WORK

This paper used many classification algorithms to get the best prediction of Anemia types based on a dataset of 41 patients. The proposed model is designed depending on five most common anemia types then classifying and analyzing the anemia type for anemic patients' dataset.

The dataset constructed from results of complete blood count test CBC. The experiment conducted by using four data mining classification algorithms where J48 decision tree and SMO performs best with 93.75% accuracy in the percentage split 60%.

When comparing the selected algorithms through utilizing of WEKA experimenter is proved that the J48 decision tree algorithm gives the best performance with F-Measure, Sensitivity, The true positive rate, Precisions and the lowest value in the false positive rate. Therefore, J48 proved to be potentially the most effective and efficient classification algorithm. In the same context, based on anemia model the performance chart by Region under meeting Curve (ROC) shown that the highest weight for J48 decision tree.

In future, use more of the data mining algorithms to classify all types of anemia diseases on different datasets to find the accuracy and predictions of preferred results.

REFERENCES

Ahmad, A., Mustapha, A., Zahadi, E. D., Masah, N., & Yahaya, N. Y. (2011). Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus Digital Information Processing and Communications (pp. 537–545): Springer.

Amin, M. N., & Habib, M. A. Comparison of Different Classification Techniques Using WEKA for Hematological Data.

Dogan, S., & Turkoglu, I. (2008). Iron-deficiency anemia detection from hematology parameters by using decision trees. International Journal of Science & Technology, 3(1), 85–92.

Elshami, E. H., & Alhalees, A. M. (2012). Automated Diagnosis of Thalassemia Based on Data Mining Classifiers. Paper presented at the The International Conference on Informatics and Applications (ICIA2012).

Garner, S. R. (1995). Weka: The waikato environment for knowledge analysis. Paper presented at the Proceedings of the New Zealand computer science research students conference.

Green, R. (2012). Anemias beyond B12 and iron deficiency: the buzz about other B's, elementary, and nonelementary problems. ASH Education Program Book, 2012(1), 492–498.

Huang, F., Wang, S., & Chan, C.-C. (2012). Predicting disease by using data mining based on healthcare information system. Paper presented at the Granular Computing (GrC), 2012 IEEE International Conference on.

Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. Procedia Computer Science, 57, 500–508.

Kesavaraj, G., & Sukumaran, S. (2013). A study on classification techniques in data mining. Paper presented at the Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on.

Kishore, C. R., Rao, K. P., & Murthy, G. Performance Evaluation of Entorpy and Gini using Threaded and Non Threaded ID3 on Anaemia Dataset. Life, 6(10), 10–12.

Prakash, V. A., Ashoka, D., & Aradya, V. M. (2015). Application of Data Mining Techniques for Defect Detection and Classification. Paper presented at the Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014.

Sanap, S. A., Nagori, M., & Kshirsagar, V. (2011). Classification of anemia using data mining techniques Swarm, Evolutionary, and Memetic Computing (pp. 113–121): Springer.

Shashidhara, M. Classification of Women Health Disease (Fibroid) Using Decision Tree algorithm.

Shouval, R., Bondi, O., Mishan, H., Shimoni, A., Unger, R., & Nagler, A. (2014). Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in SCT. Bone marrow transplantation, 49(3), 332–337.

Siadaty, M. S., & Knaus, W. A. (2006). Locating previously unknown patterns in data-mining results: a dual data-and knowledge-mining method. BMC Medical Informatics and Decision Making, 6(1), 13.

Vijayarani, S., & Muthulakshmi, M. (2013). Comparative Analysis of Bayes and Lazy Classification Algorithms. International Journal of Advanced Research in Computer and Communication Engineering, 2(8), 3118–3124.

Yilmaz, A., Dagli, M., & Allahverdi, N. (2013). A fuzzy expert system design for iron deficiency anemia. Paper presented at the Application of Information and Communication Technologies (AICT), 2013 7th International Conference on.