# BLOOD TUMOR PREDICTION USING DATA MINING TECHNIQUES

Alaa M. El-Halees[1], Asem H. Shurrab[2]

[1]Faculty of Information Technology, Islamic University of Gaza, Gaza, Palestine
[2]M.Sc., Dept. of I.T., Faculty of Information Technology, Islamic University of Gaza, Gaza, Palestine

## *ABSTRACT*

*Healthcare systems generate a huge data collected from medical tests. Data mining is the computing process of discovering patterns in large data sets such as medical examinations. Blood diseases are not an exception; there are many test data can be collected from their patients. In this paper, we applied data mining techniques to discover the relations between blood test characteristics and blood tumor in order to predict the disease in an early stage, which can be used to enhance the curing ability. We conducted experiments in our blood test dataset using three different data mining techniques which are association rules, rule induction and deep learning. The goal of our experiments is to generate models that can distinguish patients with normal blood disease from patients who have blood tumor. We evaluated our results using different metrics applied on real data collected from Gaza European hospital in Palestine. The final results showed that association rules could give us the relationship between blood test characteristics and blood tumor. Also, it demonstrated that deep learning classifiers has the best ability to predict tumor types of blood diseases with an accuracy of 79.45%. Also, rule induction gave us an explanation of rules that describes both tumor in blood and normal hematology.*

## *KEYWORDS*

*Hematology diseases, Blood tumor, Rule induction, Association rules, deep learning.*

## 1. INTRODUCTION

Data generated from healthcare domain is vast and complex. These data contain many hidden patterns which can help to discover and predict diseases in the medical field. The prediction process of these diseases can reduce the numbers of mortalities and enhance the quality of life for the patients infected with these diseases [1]. Data mining was widely used in the healthcare domain, for example, data mining can help to detect fraud and abuse of health insurance, make customer relationship management decisions by management, identify effective treatments and best practices by physicians [2].

Hematologic diseases study the blood diseases such as leukemia, thalassemia , lymphoma …etc. The medical aspect of Hematology is concerned with the treatment of blood disorders [3]. Hematologic diseases, like any other healthcare fields, generate an enormous amount of data. Traditional statistics is not enough to analysis these data, using data mining techniques is a better alternative [4]. Many types of research were done in this field, trying to discover a new knowledge or patterns which can help the humanity to detect diseases and in best case predict it before happening, by applying different types of data mining techniques and methods.

Researchers applied data mining techniques to Hematologic diseases usually use data generated from tests like Complete Blood Count (CBC) test. CBC test measures the number of blood cells

DOI: 10.5121/hiij.2017.6202 

circulating in the bloodstream. The test is a common laboratory blood test that can be used to detect blood tumor and monitor tumor treatment [5].

The aim of our study is to use data mining techniques to classify CBC sample of a blood disease patient as normal hematology disease or blood tumor. In our study, we collected data set of CBC sampled from patients in Europe Gaza Hospital in Palestine; the data belong to the Department of Oncology. Then, we applied three data mining methods to our collected dataset, which are: association rules, rule induction and deep learning. Association rules are a method used to discovery interesting relations between variables. Association rules have been used in many applications of healthcare [6]. In our paper, we investigated which CBC test has a relation with blood tumor sample. The second method we used was rule induction. Rule induction discovers patterns hidden in data. In this paper, we used rule induction to discover patterns that associated with blood tumor and normal hematology classes. The third method we used was deep learning. Deep learning is a machine learning method that utilizes a hierarchical level of artificial neural networks to carry out the process training data. Deep learning has been used for the analysis of medical data (e.g. Ravi et al. in [7] gave a survey). We used deep learning because of its ability to detect target class more accurately than other machine learning methods especially in healthcare domain [8].

The rest of the paper is structured as follows: the second section discusses the related works, the third section addresses our material and methods, the fourth section about experiments and results, while the fifth section implies the conclusion and future works of the paper.

## 2. RELATED WORK

Because of the enormous numbers of data in medical fields, which are available today, many researchers depending on data mining techniques to get new knowledge. Some of these research done in hematology diseases, such as: Abdullah and Al-Asmari in [9] used data mining to specify the anemia type for the anemic patients through a predictive model. They used real data constructed from the Complete Blood Count (CBC) test results of the patients. These data filtered and eliminated undesirable variables, then implemented on five classification algorithms which are: Naïve Bayes, Multilayer Perception, J48 and SMO. They found that J48 decision tree and SMO performs best with 93.75% accuracy in the percentage split of 60%. Shouval et al. in [4] used data mining techniques in the field of Allogeneic Hematopoietic Stem Cell Transplantation (SCT), that predicts transplantation outcome and donor selection. They proposed to use decision trees, Artificial Neural Networks (ANNs) and Support Vector Machines (SVM). No actual experiments were done. Al-shami and Al-halees in [10] used the data mining techniques on CBC tests to detect Thalassemia disease. They conducted four type of experiments on the data with all attributes in their data set samples and then repeats the experiments after reducing some features from the dataset. They used three classifiers (Decision Tree, Naïve Bayes, and Neural Network). The accuracy results of their experiments exceeded 90%, and it showed that the critical point which can be the first indicator of the thalassemia existence is MCV $\leq$ 77.65. Also, Minnie and Srinivasan in [11] used data mining on Blood Cell Counter data to convert the raw data into transformed data that can be used for generating knowledge. They used association rules and clusters on the collected data. Saichanma et al. in [12] used data mining technique to predict abnormality in peripheral blood smear from 1,362 students by using 13 data set of hematological parameters gathered from automated blood cell counter. They found that the decision tree, which is created by the algorithm, can be used as a practical guideline for RBC morphology prediction by using four hematological parameters (MCV, MCH, Hct, and RBC). In addition, Amin and Habib in [8] compared different classification techniques using WEKA for Hematological Data. They investigated which algorithm is most suitable for user working on hematological data. Their model can predict hematological data comment and developed a mobile application that can make

diagnosis hematological data comments. The best algorithm based on the hematological data was J48 classifier with an accuracy of 97.16%. Finally, Vijayarani and Sudha in [13] developed weight based k-means algorithm for identifying leukemia, inflammatory, bacterial or viral infection, HIV infection and pernicious anemia diseases from the hemogram blood test samples data set. They found that the clustering accuracy of weight based k-means algorithm is better when compared to k means and fuzzy c means.

## 3. MATERIAL AND METHODS

### 3.1 Collected Dataset

The dataset we used in this paper was collected from Europe Gaza Hospital, Gaza Strip, Palestine. The dataset belongs to the Department of Oncology and Blood Analysis Division. The dataset contains 5350 CBS samples after cleaning with different blood diseases. We divided the dataset into two groups , group one has 1764 CBC samples of blood tumor patients we labeled them as 'Tumor' and group two which has 3586 CBS samples of patients have other blood diseases, we labeled them as 'Hematology'.

The dataset has 14 attributes represent the CBC features as in Table 1.We added one more feature which is the gender of the patient because of its importance. Name and Patient-ID dropped due to the privacy of the blood sample's owner.

Table1: Attributes of CBC sample

| No. | Symbol | Meaning |
|---|---|---|
| 1 | ***WBC*** | White Blood Cell |
| 2 | RBC | Red Blood Cell, |
| 3 | ***HGB*** | Hemoglobin |
| 4 | HCT | Hematocrit |
| 5 | MID | mid-range absolute count |
| 5 | MCV | Mean Cellular Volume |
| 6 | MCH | Mean Cellular Hemoglobin |
| 7 | MCHC | Mean Cellular Hemoglobin Concentration |
| 8 | RDW | RBC Distribution Width |
| 9 | PLT | Platelets Count |
| 10 | MPV | Platelet volume |
| 11 | GRAD | percentage of white blood cells with granules in their cytoplasm |
| 12 | LYM | Lymphocyte percent |
| 13 | Gender | Male, Female |
| 14 | Class | Tumor, Hematology |

### 3.2 Dataset Preprocessing

In the preprocessing stage, we eliminated useless attributes, refilled the missing values, removed duplicative values and removed the outlier values of the collected samples.

In addition, the collected data was imbalanced where the data have 1764 tumor patients, compared to 3586 hematology patients. To overcome this issue we used *S*ynthetic Minority Oversampling Technique method (SMOTE). For each minority data, a new synthetic data instance is generated by taking the difference between the feature vector of the example and its nearest neighbor belonging to the same class, multiplying it by a random number between 0 and 1

and then adding it to the instance [14]. After these operations, the number of records in the dataset became 7172 records where half of them tumor class and the other half hematology class.

In addition, in association rules, the data should be nominal not numerical, so we transformed the values in each attribute to three types (Low, Normal, High) based on work of The WebMD Medical Team in [15].

### 3.3 Data Mining Methods

In this paper, we used three data mining methods which are: association rules, rule induction and deep learning

Association rule mining is one of the most important and well researched techniques of data mining for descriptive task, initially used for market basket analysis. It finds all the rules existing in the transactional database that satisfy some minimum support and minimum confidence constraints. Association rules are expressed in the form of IF-THEN rules. In our experiments, we used FP-Growth method to generate frequent itemsets. Then, frequent-itemsets are converted to association rules [16]. However, the resulting rules were numerous. Therefore, rules are chosen according to the goal and taking into consideration that the selected rules are strong rules which should have a value more than certain minimum support and minimum confidence. Classification using Association rule mining is a major Predictive analysis technique that aims to discover a small set of rule in the database that forms an accurate classifier [17]. Classification Based Association used the rule of the form <features-sets> -> Class Labels. These rules ranked first by confidence and then support [18].

The second classification method we used was rule induction, which extracts a set of rules that show the relationships between the attributes of a dataset and the class label [19]. Since regularities hidden in data are expressed in terms of rules, rule induction is one of the fundamental methods of data mining. Usually, rules are expressions of the form If (attribute_1= value_ 1) and (attribute_2, value_2) …. (attribute_ n, value_n) then (class_name, class_label). In our experiments, we used covering algorithms which is a strategy for generating a rule set directly: for each class in turn find rule set that covers all instances in it (excluding instances not in the class).

The third method we used was deep learning. Deep learning is an advanced type of neural network that has a collection of algorithms used in machine learning. It uses to model high-level abstractions in data through the use of model architectures, which are composed of multiple nonlinear transformations, unlike traditional neural network which builds analysis with data in a linear way. An algorithm is considered to be deep if the input data is passed through series of nonlinearities or nonlinear transformations before it becomes output. In deep learning, the manual identification of features in data removed and, instead, it relies on whatever training process it has in order to discover the useful patterns in the input examples. This makes training easier and faster, and it can yield a better result. In deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer's output. The further the advance into the neural net, the more complex the features the nodes can recognize, since they aggregate and recombine features from the previous layer [20] .

## 4. EXPERIMENTS AND RESULTS DISCUSSION

In this section, we describe the experiments and discuss the results of applying the three classifiers on our dataset. To conduct our experiments, we used cross-validation experimental

method with n=5 where we divided our dataset to five subsets one for training and the others for testing. Then, we applied the classification step as follows:

## 4.1 Association Rules

We generated association rules from the given data set using the minimum support of 0.5 and minimum confident of 0.7. Some examples of these rules that associated with blood tumor samples are presented in Table 2. From the table, we can conclude that tumor is associated mainly by: high RDW, low HGB, low HCT, low LYM, and low HGB.

Table 2: Sample Association rules related to tumor samples.

| | |
|---|---|
| RDW = High, HGB = Low, HCT = Low | class = tumor |
| LYM = Low, HGB = Low, HCT = Low | class = tumor |
| MID = Normal, RDW = High, HCT = Low | class = tumor |
| MID = Normal, RDW = High, HGB = Low | class = tumor |
| LYM = Low, HGB = Low, HCT = Low | class = tumor |
| MID = Normal, LYM = Low, HGB = Low | class = tumor |
| LYM = Low, HGB = Low | class = tumor |
| HGB = Low, HCT = Low | class = tumor |

In addition, Table 3 gives some examples of attributes associated with normal hematology sample. From the table, we can notice that attributes related to normal hematology mainly are: normal GRAN, normal RBC, low MCV and normal MID .

Table 3: Sample Association rules related to Hematology samples.

| | |
|---|---|
| LYM = Low, GRAN = Normal, RBC = Normal, MCV = Low | class = Hematology |
| GRAN = Normal, RBC = Normal, MCV = Low | class = Hematology |
| MID = Normal, GRAN = Normal, RBC = Normal, MCV = Low | class = Hematology |
| LYM = Low, MPV = Normal, HCT = Normal, HGB = Normal | class = Hematology |

## 4.2 Rules Induction

Figure 1 gives the most important rules that resultant from apply rule induction method in our data set.

If LYM $\leq$ 1.850 and RBC $\leq$ 3.635 then tumor.
If GRAN $\leq$ 5.100 and LYM > 2.250 and MCV $\leq$ 74.700 then Hematology.
If MCV > 84.650 and RDW > 14.700 and GRAN > 5.950 then tumor.
If MCV > 78.350 and GRAN > 2.800 and LYM $\leq$ 1.850 and RDW > 14.800 then tumor.
If PLT > 374 and GRAN $\leq$ 9.900 then Hematology.
If MCV $\leq$ 86.100 and RDW $\leq$ 15.200 and RBC $\leq$ 4.415 then Hematology.
If RBC $\leq$ 4.595 and HCT > 33.450 and MCV > 90.200 then tumor.
If MCH > 25.750 and GRAN > 4.600 and HCT > 36.250 then tumor

Figure 1: Sample rules as result of using Inductive rules method

From these results, we can conclude that the most important rules that can predict tumor from CBC sample are: LYM less than 1.85 and RBC less than 3.64. Also, the rule MCV greater than 84.650 and RDW greater than 14.700 and GRAN greater than 5.950.

For normal Hematology, the most important rules are: GRAN less than 5.100 and LYM greater than 2.250 and MCV less than74.700. Also the rule: PLT greater than 374 and GRAN less than 9.900.
Table 4 gives the confusion matrix of using rule induction in blood samples. These results come with accuracy 71.66%. It has f-measure of 71.75% for tumor prediction.

Table 4: Confusion matrix of using rule induction

| | True Tumor | True Hematology |
|---|---|---|
| Predicted Tumor | 2583 | 1029 |
| Predicted Hematology | 1004 | 2558 |

### 4.3 Deep Learning

In this paper, we used the H2O Deep Learning from [21] to predict tumor from CBC samples. H2O is based on a multi-layer feed-forward artificial neural network that is trained with stochastic gradient descent using back-propagation.

To get the best results we trained the system with 20 hidden layers which have 100 neurons for each. We found that the best activation function is Rectifier Linear Unit with hidden dropout ratio of 0.5 for each hidden layer. Number of epochs used in the experiments are 10. Also, the experiment used Huber as loss function. Finally, Bernoulli distribution function was used. The rest of the parameters set as default. Using these settings accuracy of the experiment was 79.45% as see in Table 5. Also, the f-measure for tumor was 77.84%, and the f-measure for hematology was 80.83.9%.
From the experiment, we also found that the most influence attributes are: MCV, HCT, RBC and LYM.

Table 5: Confusion matrix when using deep learning

| | True Tumor | True Hematology |
|---|---|---|
| Predicted Tumor | 2590 | 477 |
| Predicted Hematology | 997 | 3110 |

## 5.0 CONCLUSION AND FUTURE WORKS

Enhancing the quality of life is the major purpose of all healthcare research. In this paper, we tried to add some knowledge to this field. We inherent our knowledge from CBC blood test characteristics. We conducted three experiments using three different type of classifiers which are: Classification based association rules, rule induction and deep learning. Then, we evaluated the results of our experiments by using accuracy and F-measure. The experiments gave different accuracy rate according to the type of blood disease and the type of the classifier. We found that the deep learning classifier has the best ability to detect tumor from blood samples disease, the problem of this technique is that it has no explanation for the results. On the other hand, rule induction has acceptable performance, but it gave us some importantly understandable rules. Also, Association rules gave us some important relations among attributes in the data sample. In our future work, we will select a big dataset to test our model on it, more classifiers also can be used.

## ACKNOWLEDGEMENTS

This research was supported by Qatar Charity under Ibhath project for research grants, which is funded by the Cooperation Council for the Arab States of the Gulf throughout Islamic Development Bank.

## REFERENCES

[1] M. Durairaj and V. Ranjani, “Data Mining Applications in Healthcare: A Study,” Int. J. Sci. Technol. Res., vol. 2, no. 10, pp. 29–35, 2013.
[2] H. C. Koh and G. Tan, “Data mining applications in healthcare,” J. Healthc. Inf. Manag., vol. 19, no. 2, p. 65, 2011.
[3] H. HemOnctoday. what is hematology. Available: http://www.healio.com/hematology-oncology/news/online/{2dd178d0-7f92-46a8-add9- 2c7d634d2cea}/what-is-hematology. 2016
[4] R. Shouval, O. Bondi, H. Mishan, a Shimoni, R. Unger, and A. Nagler, “Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in SCT.,” Bone Marrow Transplant., vol. 49, no. 3, pp. 332–7, 2014.
[5] Mayo Clinic . " Cancer blood tests: Lab tests used in cancer diagnosis" http://www.mayoclinic.org/diseases-conditions/cancer/in-depth/cancer-diagnosis/art-20046459. 2017.
[6] M. Rashid, M. Hoque, and A. Sattar, “Association Rules Mining Based Clinical Observations,” arXiv Prepr. arXiv1401.2571, 2014.
[7] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu Perez, B. Lo, and G.-Z. Yang, “Deep Learning for Health Informatics,” IEEE J. Biomed. Heal. Informatics, vol. 21, no. 1, pp. 1–1, 2016.
[8] A. Santos and D. R. Carvalho, “Deep learning for healthcare management and diagnosis,” Iberoamerican Journal of Applied Computing , Vol. 5 ,No 2pp. 15–25, 2015.
[8] M. N. Amin and A. Habib, “Comparison of Different Classification Techniques Using WEKA for Hematological Data,” Am. J. Eng. Res., no. 43, pp. 2320–847, 2015.
[9] M. Abdullah and S. Al-Asmari “Anemia types prediction based on data mining classification algorithms,” Communication, Management and Information Technology – Sampaio de Alencar (Ed.) 2017.
[10] I. H. Alshami and A. M. Alhalees, “Automated Diagnosis of Thalassemia Based on DataMining Classifiers,” Int. Conf. Informatics Appl., pp. 440–445, 2012.
[11] D. Minnie and S. Srinivasan, “Clustering the Preprocessed Automated Blood Cell Counter Data using modified K-Means Algorithms and Generation of Association Rules,” vol. 52, no. 17, pp. 38–42, 2012.
[12] S. Saichanma, S. Chulsomlee, N. Thangrua, P. Pongsuchart, and D. Sanmun, “The observation report of red blood cell morphology in Thailand teenager by using data mining technique,” Adv. Hematol., vol. 2014, pp. 4–8, 2014.
[13] S. Vijayarani and S. Sudha . "An Efficient Clustering Algorithm for Predicting Diseases from Hemogram Blood Test Samples". Indian Journal of Science and Technology, Vol 8(17), August 2015
[14] N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research 16 (2002) 321–357.
[15] The WebMD Medical Team, http://www.webmd.com/a-to-z-guides/complete-blood-count-cbc#4
[16] J. Han , J. Pei , Y. Yin, Mining frequent patterns without candidate generation, Proceedings of the 2000 ACM SIGMOD international conference on Management of data, p.1-12, May 15-18, 2000, Dallas, Texas, USA.
[17] S. Soni and O. P. Vyas, “Using Associative Classifiers for Predictive Analysis in Health Care Data Mining,” Int. J. Comput. Appl., vol. 4, no. 5, pp. 33–37, 2010.
[18] G. Chen, H. Liu, L. Yu, Q. Wei and X. Zhang. "A new approach to classification based on association rule mining". Decision Support Systems, 42(2), 674-689. 2006
[19] N. Lavrac, “Rule induction,” In Intelligent Data Analysis, M. Berthold and D. Hand pp. 1–19, 2003.
[20] Deeplearning4j Development Team. Deeplearning4j: Open-source distributed deep learning for the JVM, Apache Software Foundation License 2.0. http://deeplearning4j.org
[21] A. Candel, V. Parmar, E. LeDell, and A. Arora. "Deep Learning with H2O." http://h2o.ai/resources. Mar 2017.

**AUTHORS**

**Alaa El-Halees** is a professor in computing and Deputy Dean for the faculty of Information Technology at Islamic University of Gaza, Palestine. He holds a PhD degree in data mining from Leeds Metropolitan University, UK in 2004, Msc degree in Software Engineering from Leeds Metropolitan University, UK in 1998 and BSc in Computer Engineering from University of Arizona , USA. Alaa has more than 24 years of experience including leading a range of IT-related projects. Dr. Alaa supervises M.Sc. students in Information Technology. He also leads and teaches modules at both BSc and MSc levels in Information Technology. His research activities are in the area of data mining, in particular text mining, machine learning and e-learning, Software Engineering and computer ethics.

**Asem H. Shurrab** have a B.A. in Computer Science from Islamic University of Gaza (IUG) at 2004, He is studying MSc. in Information Technology since 2015 at IUG, he works as a teacher at public high schools.