Prediction and diagnosis of leukemia using classification algorithms

Conference Paper · October 2017

DOI: 10.1109/ICITECH.2017.8079919

Khaled A. S. Abu Daqqa
Islamic University of Gaza
1 PUBLICATION **2** CITATIONS
SEE PROFILE

Ashraf Y. A. Maghari
Islamic University of Gaza
31 PUBLICATIONS **123** CITATIONS
SEE PROFILE

Wael Al Sarraj
Islamic University of Gaza
8 PUBLICATIONS **14** CITATIONS
SEE PROFILE

2017 8th International Conference on Information Technology (ICIT)

Prediction and Diagnosis of Leukemia Using Classification Algorithms

Khaled A. S. Abu Daqqa
Faculty of Information Technology
The Islamic University of Gaza
Gaza, Palestine
k.daqqa@gmail.com

Ashraf Y. A. Maghari
Faculty of Information Technology The Islamic
University of Gaza
Gaza, Palestine
amaghari@iugaza.edu.ps

Wael F. M. . Al Sarraj
Faculty of Information Technology
The Islamic University of Gaza
Gaza, Palestine
wsarraj@iugaza.edu.ps

Abstract— **Algorithms used in data mining techniques are of great importance in the field of health care, especially in the case of getting patterns or models that are undiscovered in databases. In the area of health care, leukemia affects the blood status and can be discovered by using the Blood Cell Counter (CBC). This study aims to predict the leukemia existence by determining the relationships of blood properties and leukemia with gender, age, and health status of patients using data mining techniques. More than 4,000 patients were taken from a blood test laboratory from European Gaza Hospital at Gaza Strip. Three classification algorithms are identified for blood Cancer classification; k-nearest neighbor (k-NN), decision tree (DT) and Support Vector Machine (SVM). These three classifiers were implemented and studied thoroughly in terms of classification accuracy and F-Measure. From our experimental results, it was noticed that the decision-tree algorithm had the highest percentage of 77.30% compared with the other two techniques. In addition, the DT classifier obtains properties regarding outer attributes such as city (eastern regions) that are most vulnerable to leukemia.**

Keywords*— Leukemia, Blood Cell Count, Algorithms, k-NN, SVM, Decision Tree, Data Mining.*

I. INTRODUCTION

According to WHO (World Health Organization), 8.2 million people die from cancer related in 2012 and it is estimated that 13% of total death worldwide are caused due to cancer. 70% increase in new cases of cancer is expected over the next two decades. Over all 100 types of cancer are requiring unique treatment and diagnosis. One of these types of cancer is the blood cancer (leukemia), where it is a cancer of the tissue, which forms blood cells, white blood cells, and bone marrow. Leukemia is classified into lymphatic, and chronic lymphocytic leukemia [1]. Leukemia, anemia, infection and the other is a set of changes occur in the blood cells, where it could be detected using a complete blood count (CBC), which is the preliminary assessment of the health situation [2]. Leukemia has been classified by a large number of multiple classifier, and different algorithms, so there was widespread use of classification algorithms in leukemia. One of them is data extraction, which is considered an important for the implicit information after the data analysis process using certain algorithms, and hence getting unknown and strange pattern. This knowledge is represented using visual techniques for the user. Algorithms used in data mining techniques are of great importance in the field of health care, especially in the case of getting to know the patterns that are undiscovered previously. So the discovered patterns is expected facilitate the process of diagnosis and early detection of certain diseases and treat the injured and patients before the development disease [3, 4]. The main aim of this paper is to determine the relation of blood properties and leukemia with gender, age, and health status in patients using classification algorithms, and compare between those algorithms; to show which is the best in the detection of blood cancer. The different classification techniques and algorithms used today in many areas among others are k-nearest neighbor (k-NN), decision tree (DS) and Support Vector Machine (SVM). It's quite possible to detect so many diseases by identifying cell blood counting within the blood smear using classification techniques.

This paper is organized as follows: In Section II, we discusses background and reviews some related works in the field. Section III, presents the approach and the followed methodology to achieve the research goals. In Section IV, the results are demonstrated and discussed. In the last section, we concluded our work.

978-1-5090-6332-1/17/$31.00 ©2017 IEEE

II. BACKGROUND AND RELATED WORKS

A. BACKGROUND

In this section, we discuss the differences between the three classifiers:

First, SVM is one of the classification algorithms that applied in many different areas of science, and these areas are face detection, object recognition, text categorization, and handwriting recognition, speaker identification and health care. Moreover, different efficient implementations of SVM algorithm are facilitating and accelerate of applying these techniques in the area. [5]. The SVM takes a set of input data and doing a training set of infected instances. After training process, in case of the introduce of new data the classification of this data categorizes each one of them, according to the training data, an SVM training algorithm builds a model that assigns new instances into one the algorithm, outputs an optimal hyperplane which categorizes new instances. Therefore, the distinction between these groups is through the line between the two classes, in case there was more than one line separating them, in this case comes SVM important to find the best line between subsets and called it a hyperplane. The SVM algorithm based on finding the hyperplane that gives the largest minimum distance to the training examples. The optimal separating hyperplane maximizes the margin of the training data. Formally,

$$\min_{\beta,\beta_0} L(\beta) = \frac{1}{2}||\beta||^2 \text{ subject to } y_i(\beta^T x_i + \beta_0) \geq 1 \ \forall i, \tag{1}$$

The problem of maximizing is equivalent to the problem of minimizing a function subject to some constraints. The constraint model the requirement for the hyperplane to classify correctly all the training examples. where is known as the weight vector and as the bias and represents each of the labels of the training examples [6] .

Second, K- Nearest Neighbors (KNN): The algorithm is based on the classification of objects according to the nearest point of its known neighbor, where the set of points of their neighbors is obtained nearest neighbor through a training set of examples in the feature space (or, in the case of regression, take the average of these k label values). An object classifying any Class followed by more neighbors by nearby, and will be voted on the follows of this set or the class, where k is a positive integer [7]. In k-nn technique, the classification of a new test feature vector is specifically through the classes of its k-nn. The k-nn algorithm was executed using Euclidean distance metrics to locate the nearest neighbor [8]. The Euclidean distance metric d(x, y) between two points x and y is calculated using the Equation 2. Where N is the number of features such that $x = x_1, x_2, x_3 \ldots x_N$ and $y = y_1, y_2, y_3 \ldots y_N$. The number of neighbors (i.e., k) used to classify the new test vector varied between the range of 1 to 10, and its effects on the classification performance were determined in the form of classification accuracy with standard deviation.

$$d\,(x,y) = \sum_{i=1}^{n} \sqrt{x_i^2 + y_i^2} \tag{2}$$

Third, Decision Tree: It is a hierarchical method, which consists of rules in turn divide the independent variables into homogeneous areas. The basic idea of building DT is to get a set of rules that can be used in the prediction process through the results we get from the data and the input variables. A Decision Tree is called a regression or a classification tree if the target variables are continuous or discrete, respectively. Decision tree was used and applied in a large number of areas in the world, including health care and in the prediction and classification process [9].

A decision tree is very similar to a large extent inverted tree, because the root is at the top and the growth is down. DT representation of the data has the advantage compared with other approaches of being easy to interpret and meaningful [10].

B. RELATED WORKS

CBC has been discussed in many research papers. In the following sections, we introduce some of these researches, which are related to our research.

Some papers used data mining techniques to diagnose several types of diseases and phenomena, such as: anemia, thalassemia, diabetes, cancer, and heart diseases … etc. And many others tried to find their own formula to determine if the person is a leukemia patient or iron deficiency patient [3, 11].

Some papers [12] used classification based pattern analysis techniques for diagnosing the cancer. Several well-known classification algorithms such as SVM, DT, k-NN, and neural networks are used for diagnoses of cancer. It is established that the process of classification depends on the value of various features in the collected data. Here, the authors found that medical disease data often have some noise data as well as boundary value data. They suggested techniques to deal with such noisy data. For optimization of accuracy, they employed Ant Colony Optimization technique.

Other researches addressed the issue of applying medical data mining and using various data mining techniques for diagnosis of Acute Myeloid Leukemia cancer [13]. Various techniques employed in their work are clustering, regression, classification and a survival prediction model is constructed out of them. They discussed three important aspects; firstly, it presents significance of data mining approach in this regard, secondly it provides a comprehensive survey related to the selected task and finally it compares the correct accuracy level of various models.

Using the of gene expression data discovery of two types of leukemia [14], which could be close and are similar to a large extent; acute lymphoblastic leukemia and acute myeloid leukemia. Authors have applied neural networks algorithms to get results showing between the two types of blood cancers. It was attribute extraction of microarray genes, which has a greater change on its classification and clustering because he took as input to any network. Authors discussed the role of the feature vector in classification. Their work concludes that in order to achieve best results in learning algorithm, feature subset selection method should be applied on to the dataset.

Wide range of classification techniques and presented comparisons between each other ware discussed by David et al. [15]. In addition, they proposed a new and detailed methodology to benefit from WEKA software through a wide range of users. The main features used are: 49 data preparation and processing tools, 8 clustering algorithms, 15 attribute/subset evaluators, 76 classification/regression algorithms, 3 algorithms for finding association rules and 10 search algorithms for feature selection. Where used medical bioinformatics analyses have been used to clarify the usage of WEKA in the diagnosis of Leukemia.

However, this study is different on our case; we used the CBC data to conduct our experiments for the diagnosis of leukemia. Further, RapidMiner software is used to test the three algorithms; k-nearest neighbor (k-NN), decision tree (DT) and Support Vector Machine (SVM), to predict disease.

III. METHODOLOGY

This section describes the research methodology steps and further the various classification techniques used in this study. The research methodology is divided into five main steps as shown in Figure 1; Raw data collection, Data cleaning and processing, transformation, Data mining algorithms and finally the last step is dedicated to results visualization. There is also a final step which is called performance and knowledge that is the overall findings of the five steps. The dataset was collected from CBC tests repository of European Gaza Hospital in Gaza Strip, which contains 4000 instances including 2000 instances with leukemia disease. The dataset contains 18 attributes including the leukemia class, 12 blood properties, age, marital status, city, gender. The dataset was converted by DB tools and inputted into the popular data mining tool RapidMiner.

Poor and not arranged data in (CBC) in a particular context was forming obstacle to the researchers in the acquisition of knowledge and the discovery of undiscovered patterns from the data. Therefore, the data preprocessing operation is a very important phase to get accurate results and can improve the prediction process. The first step is selecting the most appropriate features to have a strong relationship to our mining task. Then the missing data are filled in, and detecting and removing outliers if necessary. The methodology steps are demonstrated in **Error! Reference source not found.**.

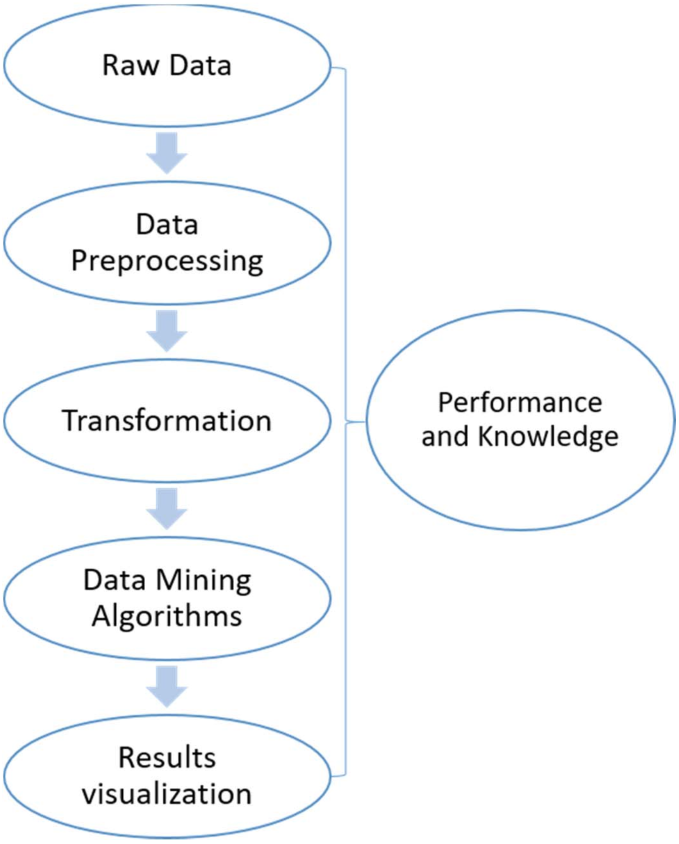

Fig. 1: The steps for methodology

Classification algorithms used in this study were applied using RapidMiner to get the precision of the algorithms. The accuracy of classification algorithms ware determined and demonstrated in TABLE **1**.

Accuracy is the proportion of the total number of predictions that were correct. It is determined by the following formula:

$$\text{Accuracy} = \frac{\text{tp+tn}}{\text{tp+tn+fp}} \tag{3}$$

Where, TP rate = positives correctly classified / total positives, FP rate = negatives incorrectly classified / total negatives.

Precision is the proportion of the predicted positive cases that were correct, as calculated using:

$$\text{Precision} = \frac{\text{tp}}{\text{tp+fp}} \tag{4}$$

Recall: Also called Sensitivity or True Positive Rate (TPR), is the proportion of positive cases that were correctly identified, as calculated using:

$$\text{Recall} = \frac{\text{tp}}{\text{tp+fn}} \tag{5}$$

As the accuracy measure is considered as useful in comparisons, it depends on the Precision and Recall, which are the best measures of evaluation [16], but there is a problem in the Recall calculating sometimes. Therefore, we decided to use percentage of the balance between Precision and Recall, which is known as the F Measure. See Equation 6:

$$\text{F Measure} = \frac{2*P*R}{P+R} \tag{6}$$

Where, P is Precision and R is Recall.

IV. RESULTS AND DISCUSSION

Three candidate classifiers are considered in this study: SVM, DT, and k-NN. SVM and Decision Tree was selected according to its accuracy and the ability to extract the classification rules which may be important in many fields and DT is particularly widely used in medical field because of its simplicity in interpreting outputs and outcomes [17]. K-NN classifier is popular in generating classification models from input records [10]. Rapid Miner 7.0.001 is used as the environment in which the three classifiers (Support Vector Machine, Decision Tree and k-Nearest Neighbor) are applied and compared.

TABLE 1: COMPARATIVE ANALYSIS OF CLASSIFICATION ALGORITHMS

Algorithms	***Accuracy in %***	***F-Measure***
Support vector machine (SVM)	76.82%	%70
K Nearest Neighbors (KNN)	72.15%	%60
Decision Tree	77.30%	%67

TABLE **1** shows clearly that the classification algorithms is not stable, and fluctuates between the datasets. From the results, it is proven that the Decision Tree algorithm has the highest accuracy in study with 77.30%. It is observed that the support vector machine (SVM) algorithm obtained 76.82% accuracy. However, for the Precision and Recall, the results have showed that SVM has the best F-Measure with 70% and F Measure for the decision tree was 67%.

There are several different Data Mining applications used in the mining work for the diagnosis of the disease, whereas Decision Tree and SVM techniques provided more accurate results than other algorithms [18].

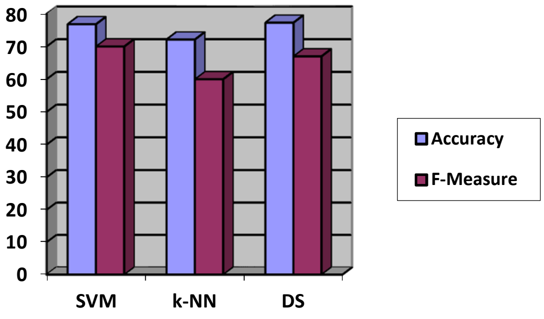

Fig. 3: Accuracy and F-measure Chart of the three classification algorithms in blood cancer

Using a decision tree, we observe that it gives some important results on blood properties that are related to leukemia, as mentioned in **Error! Reference source not found.**. We note that people who are younger than or equal to 9.5 years do not become infected with the disease. In addition, the Blood statistic (RDW) is greater than 15.55 for people over the age of 9.5 which implies that they are more susceptible to the disease. The people whose MID proportion was less than or equal to 1.150 and whose property of blood RBC is less than or equal to 2.920, are a potential of leukemia. The people whose MID proportion is greater than 1.150 and LYM blood property is greater than 1.750 are also a potential of leukemia.

While, the DT classifier obtains properties regarding outer attributes such as a city. It shows that most vulnerable pebble from eastern regions in each city as (Abasan, Khuza'a, Shijia

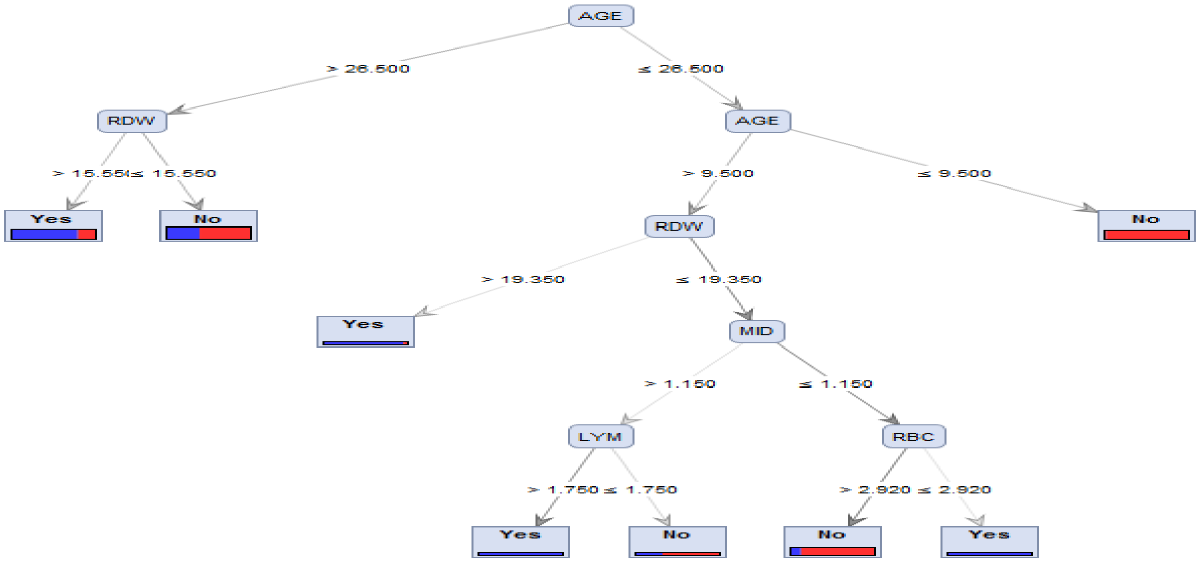

Fig. 2: Resulted decision Tree

and Beit Hanoun), are potential to leukemia.

TABLE 2: CONFUSIONMATRIX: ACCURACY: 76.82% FOR SVM

True:	**Yes**	**No**
Yes:	456	184
No:	198	810

The results of classification by precision and recall method in 1648 instances are reported as follows: By using SVM algorithm, the Performance of SVM model is equal to 76.82% as shown in TABLE **2**. Looking further in TABLE **2**, we notice that the algorithm we used, after training process, it shows that the 456 are suffering from Leukemia, while 184 (how are originally carry the disease) are not, and this is not true, here we can show the error percentage. On the other hand, for the people how are not carrying the disease, the results showed them as 810, and the algorithm showed 198 don't carry the disease, but actually they do, and this is the wrong. So, we need to calculate R and P for the algorithm, and get the results, F-Measure 70% and accuracy 76.82%.

TABLE 3: CONFUSIONMATRIX: ACCURACY: 72.15% FOR K-NN

True:	**Yes**	**No**
Yes:	366	171
No:	288	823

By using k-nn algorithm, the Performance of k-nn model is equal to 72.15% as shown in TABLE **3**. Looking further in TABLE **3**, we notice that the algorithm we used, after training process, it shows that the 366 are suffering from Leukemia, while 171 (how are originally carry the disease) are not, and this is not true, here we can show the error percentage. On the other hand, for the people how are not carrying the disease, the results showed them as 823, and the algorithm showed 288 don't carry the disease, but actually they do, and this is the wrong. So, we need to calculate R and P for the algorithm, and get the results, F-Measure 60% and accuracy 72.15%.

TABLE 4: CONFUSIONMATRIX: ACCURACY: 77.30% FOR DECISION TREE

True:	**Yes**	**No**
Yes:	408	128
No:	246	866

By using Decision Tree method, the Performance of decision tree model is equal to 77.30%. See TABLE **4**. Looking further in TABLE **4**, we notice that, after training process, it shows that the 408 are suffering from Leukemia, while 128 (who are originally carry the disease) are not, and this is not true, here we can show the error percentage. On the other hand, for the people who are not carrying the disease, the results showed them as 866, and the algorithm showed 246 don't carry the disease, but actually they do, and this is wrong. Accordingly, we calculated Recall and Precision for the algorithm, and obtained the results; F-Measure of 67% and accuracy of 77.30%.

The same experiments and comparisons were performed by [19, 20], and our results assures the results they gained in their researches, which is DS classifier get the best accuracy among others in the disease discovery process, namely SVM and k-NN.

Results and findings of the experiments showed a clear advantage of using the decision-tree algorithm which provided the highest percentage of 77.30. We may conclude that the decision-tree algorithm had the highest properties compared with the other two techniques (SVM and k-NN). In addition, DT classifier obtained properties regarding outer attributes such as city (eastern regions) that are most vulnerable to leukemia.

V. CONCLUSION

In the area of health care, leukemia affects the blood status and can be discovered by using the Blood Cell Counter (CBC). Through the results, it was noted that the correct accuracy of the classification algorithms are not stable and they differ from one to another. This research work has identified three blood Cancer Classifiers, k-nearest neighbor (k-NN), decision tree (DS), and Support Vector Machine (SVM) for this study. The three classifiers were thoroughly implemented and studied in terms of classification accuracy and F-Measure. The decision-tree algorithm had the highest percentage of 77.30% compared with the other two techniques. While, the DT classifier obtained properties regarding outer attributes such as a city which shows that most vulnerable to leukemia are eastern regions from each city. We expect that the reason is due to the presence of chemicals used by farmers. A deep empirical research work could be lunched to investigate the causes and circumstances of leukemia causes in the eastern area of Gaza strip. There is also a considerable effect of weapons used during the three last wars on Gaza 2009, 2012 and 2014. Using other classifiers such as Naive Bayes and neural networks may be considered for the future deep empirical research work on the eastern region of Gaza strip.

References

[1] WHO. (2015). Media centre. Available: http://www.who.int/mediacentre/factsheets/fs297/en/

[2] G. Cembrowski et al., "The use of serial outpatient complete blood count (CBC) results to derive biologic variation: a new tool to gauge the acceptability of hematology testing," International journal of laboratory hematology, 2015.

[3] D. Minnie and S. Srinivasan, "Clustering the Preprocessed Automated Blood Cell Counter Data using modified K-Means Algorithms and Generation of Association Rules," International Journal of Computer Applications, vol. 52, no. 17, 2012.

[4] D. Tomar and S. Agarwal, "A survey on Data Mining approaches for Healthcare," International Journal of Bio-Science and Bio-Technology, vol. 5, no. 5, pp. 241-266, 2013.

[5] J. Nayak, B. Naik, and H. Behera, "A comprehensive survey on support vector machine in data mining tasks: applications & challenges," 2015.

[6] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning, corrected ed," Berlin: Springer. Haxby, JV, Gobbini, MI, Furey, ML, Ishai, A., Schouten, JL, & Pietrini, P.(2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science, vol. 293, no. 5539, p. 24252430, 2003.

[7] R. Palaniappan, K. Sundaraj, and S. Sundaraj, "A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals," BMC bioinformatics, vol. 15, no. 1, p. 1, 2014.

[8] F. Pan, B. Wang, X. Hu, and W. Perrizo, "Comprehensive vertical sample-based KNN/LSVM classification for gene expression analysis," Journal of Biomedical Informatics, vol. 37, no. 4, pp. 240-248, 2004.

[9] J. H. Cho and P. U. Kurup, "Decision tree approach for classification and dimensionality reduction of electronic nose data," Sensors and Actuators B: Chemical, vol. 160, no. 1, pp. 542-548, 2011.

[10] GmbH. (2016). RapidMiner Documentation. Available: http://docs.rapidminer.com/

[11] E. H. Elshami and A. M. Alhalees, "Automated Diagnosis of Thalassemia Based on DataMining Classifiers," in The International Conference on Informatics and Applications (ICIA2012), 2012, pp. 440-445: The Society of Digital Information and Wireless Communication.

[12] S. S. Shrivastava, V. Choubey, and A. Sant, "Classification Based Pattern Analysis on the Medical Data in Health Care Environment," 2016.

[13] M. Durairaj and R. Deepika, "PREDICTION OF ACUTE MYELOID LEUKEMIA CANCER USING DATAMINING-A SURVEY," International Journal of Emerging Technology and Innovative Engineering, Voume1, no. 2, pp. 94-98, 2015.

[14] K. Li, M. Yang, G. Sablok, J. Fan, and F. Zhou, "Screening features to improve the class prediction of acute myeloid leukemia and myelodysplastic syndrome," Gene, vol. 512, no. 2, pp. 348-354, 2013.

[15] S. K. David, A. T. Saeb, and K. Al Rubeaan, "Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics," Computer Engineering and Intelligent Systems, vol. 4, no. 13, pp. 28-38, 2013.

[16] C. D. Manning, P. Raghavan, and H. Schütze, "Evaluation in information retrieval," in Introduction to Information Retrieval:Cambridge: Cambridge University Press, 2008, pp. 139-161.

[17] A. H. Kaji, A. M. Hanif, N. Bosson, D. Ostermayer, and J. T. Niemann, "Predictors of neurologic outcome in patients resuscitated from out-of-hospital cardiac arrest using classification and regression tree analysis," The American journal of cardiology, vol. 114, no. 7, pp. 1024-1028, 2014.

[18] H. J. Leach, D. P. O'Connor, R. J. Simpson, H. S. Rifai, S. K. Mama, and R. E. Lee, "An exploratory decision tree analysis to predict cardiovascular disease risk in African American women," Health Psychology, vol. 35, no. 4, p. 397, 2016.

[19] S. A. Sanap, M. Nagori, and V. Kshirsagar, "Classification of anemia using data mining techniques," in International Conference on Swarm, Evolutionary, and Memetic Computing, 2011, pp. 113-121: Springer.

[20] S. Kharya, "Using data mining techniques for diagnosis and prognosis of cancer disease," arXiv preprint arXiv:1205.1923, 2012.