DIAGNOSIS OF RHEUMATOID ARTHRITIS USING AN ENSEMBLE LEARNING APPROACH

Zahra Shiezadeh[1], Hedieh Sajedi[2] and Elham Aflakie[3]

[1]Department of Computer Engineering, Science and Research Branch,
Islamic Azad University, Bushehr, Iran
zshiezadeh@gmail.com
[2]Mathematics, Statistics and Computer Science School,
College of Science, University of Tehran, Tehran, Iran
hhsajedi@ut.ac.ir
[3]Rheumatology Research Center,
Shiraz University of Medical Sciences, Shiraz, Iran
aflakye@sums.ac.ir

ABSTRACT

Rheumatoid arthritis is one of the diseases that its cause is unknown yet; exploring the field of medical data mining can be helpful in early diagnosis and treatment of the disease. In this study, a predictive model is suggested that diagnoses rheumatoid arthritis. The rheumatoid arthritis dataset was collected from 2,564 patients referred to rheumatology clinic. For each patient a record consists of several clinical and demographic features is saved. After data analysis and pre-processing operations, three different methods are combined to choose proper features among all the features. Various data classification algorithms were applied on these features. Among these algorithms Adaboost had the highest precision. In this paper, we proposed a new classification algorithm entitled CS-Boost that employs Cuckoo search algorithm for optimizing the performance of Adaboost algorithm. Experimental results show that the CS-Boost algorithm enhance the accuracy of Adaboost in predicting of Rheumatoid Arthritis.

KEYWORDS

Data Mining, Adaboost, Cuckoo's Algorithm, Predictive Model, Rheumatoid Arthritis, Decision Tree.

1. INTRODUCTION

Rheumatoid arthritis (RA), is one of the arthritis that causes inflammation, pain and swelling in the joints. Usually it is chronic and can cause long term damage or deformation of the joints. One out of every hundred people is in some way affected by RA in the life [1, 2]. The cause of RA is still unknown. If RA is diagnosed in its early stages, it can be easily controlled. Usually

Jan Zizka et al. (Eds) : ICAITA, SAI, CDKP, Signal, NCO - 2015
pp. 139–148, 2015. © CS & IT-CSCP 2015 DOI : 10.5121/csit.2015.51512

RA is diagnosed when severe symptoms appear in the patient and disease need more aggressive treatment [3].

A variety of methods has been developed for the early diagnosis of RA [3, 4], such as 2010 ACR/EULAR classification criteria [5] and the van der Helm–van Mil (vHvM) score [6]. Currently, clinical experience is the basis of the diagnosis of RA using certain RA disease classification criteria. Precise and accurate assessment of RA symptoms can avoid permanent damage to the patient's joints and bones, and also have influence on patients' quality of life. Rheumatoid arthritis is an area of medicine that has been less considered from the perspective of data mining.

In recent years, data mining using electronic medical records has been very popular and is expected to improve the accuracy of diagnosis and quality treatment using data mining techniques. Developing a model that can infer predicted class is the purpose of the prediction model [7]. In this study, a predictive model for automatic detection of rheumatoid arthritis was developed using an integrated approach. The study relates to the factors that predict disease, data mining, classification techniques, and a database was created for patients in the rheumatology clinic of Shiraz university of medical sciences.

2. RELATED WORKS

In most of the data mining studies that were investigated, more attention has been paid to several medical fields, including RA [8, 9, 10], cardiovascular diseases [11, 12, 13, 14, 15, 16], cancer [17, 18, 19, 20], lung [21, 22, 23], traumatic brain injury [24, 25, 26] and diabetes [27, 28, 29].

In our study, we investigate the data mining applications in the prediction of RA diseases, and associated classification techniques. Previous researches have studied various attributes and different method to establish disease prediction models. For instance, Pinar Yildirim et al.[8] use textmining to discover similar attribute among RA patients whereas Cader et al. [30] used the 2010 and 1987 ACR/EULAR criteria for the prediction of RA patients. Huizinga et al. [31] examined nine clinical variables to construct prediction rules.

3. THE PROPOSED ALGORITHM

Adaboost algorithm was the first practical boosting algorithm and remains one of the most widely used and studied, and it is applied in numerous fields. This ensemble method is selected to apply to the database and deriving results from it [32, 33, 34, 35]. The next step is to improve the accuracy of modeling, so the Adaboost algorithm [36, 37] is combined with Cuckoo Search algorithm [38, 39] and CS-Boost algorithm is proposed. Cuckoo search (CS) is an optimization algorithm developed by Xin-she Yang and Suash Deb in 2009. In the proposed algorithm, the Decision Stump has been used as weak learners. The tree is designed as a weak learner or a base learner for bagging or boosting techniques [38], and it makes one level decision tree for classical or numerical data sets.

4. RESEARCH METHODS AND MATERIALS

Considering advantages of the CRISP-DM model for knowledge discovery, in this study the various stages of the model were implemented. The CRISP-DM model is explained as a procedure of the cross industry standard for developing data mining projects. This is one of the most widely used data mining techniques to discover knowledge, and the six phases implement the model are: The business understanding, data understanding, data preparation, and modeling, evaluation, and deployment phases. A detail of this methodology is available in [27]. In the following, we describe our implementation of each phase.

4.1. Business Understanding:

This study provides a model to predict rheumatoid arthritis among patients referred to the Rheumatology clinic. Due to the nature of rheumatic diseases, there is a wide range of overlapping symptoms in RA. A predictive model can be effective in medical knowledge in this field and promote the health of the society [41].

4.2. Data Understanding:

The cohort consisted of patients referred to the rheumatology clinic at Shiraz University of Medical Sciences during the study period. We identified 2564 patients who were admitted to the clinic with arthritis diagnoses. We constructed a new data set for the arthritis and for each patient, we have saved records consist of the demographic and clinical data. The final data set has contained more than 600 attributes such as demographic data, lab data, treatments, and physical exams, symptoms, past history and having pain, redness, and tenderness and… in the patient's joints. Finally, we categorized data values and derived new fields from existing data. These features were changed to categorical attributes for better analysis and to obtain good results. More than 72 features were selected due to preliminary feature selection and physicians' opinions. Table1 shows these features, their values and data types.

4.3. Data Preparation:

In the data preparation phase, the data were preprocessed. The preprocessing phase includes the following steps:

- *Data Cleaning*

- *Constructing New Data*: New fields are derived from existing ones. The total number of joints, MCP count pj15 to pj24, PIP count pj25 to pj40, DIP count pj41 to pj56, MTP count pj61 to pj75 and BMI that is calculated from weight and height. Figure1 shows the joint and their locations.

Table 1: Data set features and their values and data types.

Attributes	Values	Data type	Attributes	Values	Data type
Code		Integer	**Height**		Integer
Sex	{Male, Female}	Binominal	**Weight**		Integer
Age		Integer	**Disease duration**		Integer
Birth place		Nominal	**pj2[1]-pj32**	{Yes, No}	Binominal
Marital status	{Single, Married, Divorced, Widowed}	Polynomial	**Pj37-pj49**	{Yes, No}	Binominal
Education	{Diploma, High School, Associate, BS/BA, None, Intermediate, Primary School, MS/BA, Doctoral}	Polynomial	**Pj57-pj75**	{Yes, No}	Binominal
Job	{Disabled, Full Time Employed, Housewife, Part-time Employed, Retired, Student, Unemployed}	Polynomial	**ESR**		Integer

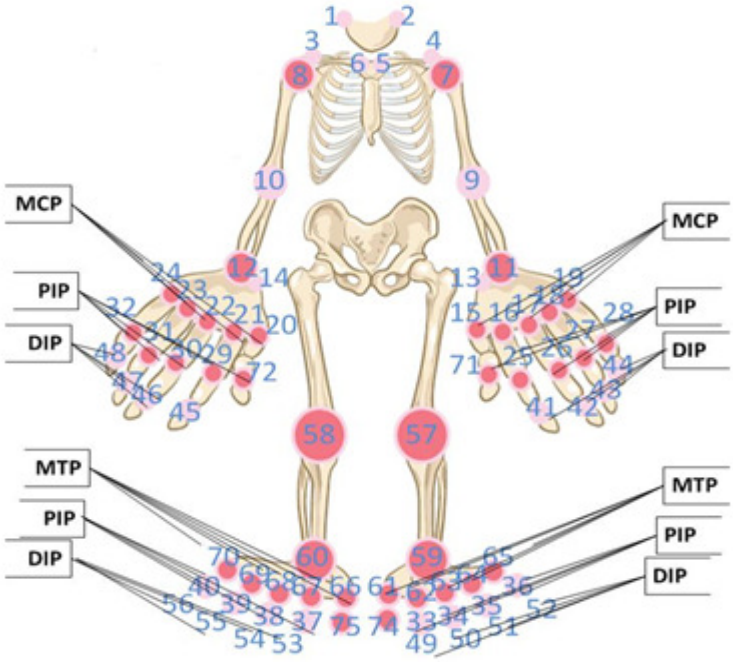

Figure 1: The joints and their location in the patient's body

[1] Patient Joint

- *Feature Selection:* At this stage, different methods for selecting effective features, in three steps have been taken. At this stage only features influencing the target field are selected as the input for modeling phase. 1) In this way, using feature selection techniques such as Chi Squared, CFS, Gain Ratio, Info Gain, OneR and Relief, the main features from each technique separately extracted [42, 43]. To increase the accuracy of assessing, 10-fold Cross Validation is used. 2) The Feature Selection nodes in SPSS Modeler software are used to select the most important features. In this way, by eliminating features that have small variances, ranked features and 11 features are selected as a subset of them. 3) The features are presented to specialists and they ranked them. The results from these three steps are integrated and finally 18 features are entered to modeling phase. Table2 shows the feature selection results.

Table 2: Features for modeling phase

Feature	Rank	Feature	Rank
Joint Count	1	MCP	10
ESR	2	pj8	11
PIP	3	pj10	12
pj9	4	pj12	13
pj11	5	Age	14
DIS	6	Duration	15
pj58	7	MTP	16
Sex	8	pj59	17
pj57	9	Marital Stat	18

4.4. Modeling:

The learning model in this study is supervised method, considering the goal field which is diagnosed by specialists as well as finding the most important factors influencing the diagnosis of rheumatoid arthritis. In fact, goal feature has two distinct values, susceptibility to rheumatoid arthritis (RA) and other rheumatic diseases (Other), so the nature of data mining tends to classification. Therefor applying classification algorithms that extract the rules and determine the relationship between individual features and goal feature is the main parts of the model. In this study, first C4.5, CHAID, ID3, W-J48 and Adaboost algorithms are implemented on the dataset. Then SVM, KNN and Adaboost algorithm with Decision Stump as a weak learner are implemented using MATLAB software on the dataset. In the implementation of these algorithms, the doctor's diagnosis was goal feature and other features that are selected in the selection phase are considered as the input features. By implementing the above steps, the ensemble algorithm, Adaboost accuracy in modeling was higher than other methods using the combination of weak classifier.

4.5. Evaluation:

The algorithms are applied to the data set using stratified 5-fold validation in order to assess the performance of classification techniques for predicting a class. Evaluation criteria in

classification problems are accuracy, sensitivity, specificity, PPV and NPV that are achieved using confusion matrix.

Table 3 : A Confusion matrix Table

Confusion Matrix	**Other**	**RA**	**Class precision**
Other	TN	FN	NPV=TN/(TN+FN)
RA	FP	TP	PPV=TP/(TP+FP)
Class recall	Specificity=TN/(TN+FP)	Sensitivity=TP/(TP+FN)	
Accuracy	(TP+TN)/(TP+TN+FP+FN)		

PPV: It denotes the percentage of RA predictions that are correct. Recall / Sensitivity: It denotes the percentage of RA labeled instances that were predicted as RA. Specificity: It denotes the percentage of Other labeled instances that were predicted as Other. Accuracy: It denotes the percentage of predictions that are correct. Table3 shows the comparison of decision tree that the Adaboost with J48 algorithm as the base learner has maximum accuracy and sensitivity. Table4 shows the comparison of classification algorithms. The Adaboost algorithm implemented with decision stamp as base learner.The proposed CS-Boost has maximum accuracy and the minimum sensitivity.

Table 3: Decision Tree Comparison

Algorithms	*Specificity (%)*	*Sensitivity (%)*	*Accuracy (%)*	*NPV (%)*	*PPV (%)*
C4.5	*60.71*	*73.61*	*70*	*47.23*	*82.81*
ID3	*64.29*	*72.22*	*70*	*47.37*	*83.87*
J48	*53.57*	*77.78*	*71*	*48.39*	*81.16*
CHAID	*35.71*	*73.61*	*63*	*34.48*	*74.65*
Adaboost	*44.83*	*88.73*	*76*	*61.90*	*79.75*

Table 4: Classification model comparison

Algorithms	*Specificity (%)*	*Sensitivity (%)*	*Accuracy (%)*	*NPV (%)*	*PPV (%)*
Decision Tree	*49*	*79*	*72*	*49*	*79*
KNN	*42*	*78*	*68*	*74*	*47*
SVM (polynomial)	*50*	*80*	*73*	*81*	*47*
Adaboost	*54*	*77*	*78*	*39*	*86*
CSBoost	*74*	*44*	*85*	*22*	*89*

Table 3 shows that the AdaBoost algorithm with J48 algorithm as the weak learner has the highest Sensitivity 88.73 percent and can diagnose RA correctly for a rheumatic patient. The ID3 algorithm on this dataset for patients with other arthritis diseases will have discretion 64.29 percent, the highest Specificity compared to other algorithms. If you have RA patients, the ID3 algorithm detects the correct model, 83.87 percent and has the highest percision. If patients have diseases other than RA, AdBoost 61.90 percent can recognize other diseases correctly that is the highest percentage among the other algorithms. Among these algorithms, AdaBoost algorithm has the highest accuracy 76 percent.

Table 4 shows that the CSBoost algorithm has the highest PPV, 89 percent. The SVM algorithm on this dataset has the highest Specificity, 80 percent compared to other algorithms. If patients have diseases other than RA, SVM 81 percent can recognize other diseases correctly that is the

highest percentage among the other algorithms. Among these algorithms, CSBoost algorithm has the highest accuracy 85 percent.

5. DISCUSSION

This study, carried out along with the 2564 patients refer to the Rheumatology clinic in Shiraz university of medical sciences, and used data mining technology to construct a rheumatoid arthritis disease predictive model. A total of 300 valid sample patients was acquired from this database, the data on the patients were collected for classification study, which included their physical exam results, symptoms, lab data results, patient history, demographic data and diagnoses. Data mining technologies adopted in this study were decision tree, c4.5, Id3, Chaid, WJ48, SVM, KNN and boosting algorithm (Adaboost). In comparison, of data mining technology, this study used sensitivity and accuracy indicators to evaluate classification efficiency of different algorithms. After comparing classification accuracy, Adaboost was the best classification algorithm in this study.

The optimum RA disease predictive model obtained in this study adopts CS-Boost as classification algorithm, 18 attributes as attribute input mode, and its classification efficiency: sensitivity indicator = 44% and accuracy indicator = 85%. 18 major influence factors were recognized for accurately predicting RA disease but education, BMI, occupation and birthplace were less important as other factors that is similar to the 2010 ACR/EULAR Classification Criteria. The research results could not be comparable with the other similar mining researches in RA such as [8, 9, 10] because they have used text mining and their search result was related to RA but were different from this study. In addition, 20 diagnosis, classification rules were extracted from this predictive model, and confirmed by three RA specialists to be conformable with the current clinical medical condition and have reference value in diagnosis and prediction of RA disease.

This study has some weak points. The research carried on patient that the treatment was started for their disease. It is suggested that in the future work the dataset can be collected from new case patients. For this research, there are 2564 records patients, but 357 of these records was suitable for this study so in the future work classification can be applied in more patients' records to get more precise and accurate results.

6. CONCLUSION

Rheumatoid arthritis, is one of the diseases that its cause is not known yet, data mining can help the medical field in order to provide early diagnosis and treatment of this disease. The aim of this study is to provide predictive models for the diagnosis of rheumatoid arthritis. The data were collected from patients referred to the rheumatology clinic of Shiraz University of Medical Sciences. Next the data is preprocessed. Decision tree algorithms for the modeling are applied such as C4.5, ID3, CHAID, J48, SVM and Adaboost. Then the Adaboost algorithm is combined with a Cuckoo Search algorithm and CSBoost algorithm is proposed. The optimum RA disease predictive model obtained in this study adopts CSBoost as classification algorithm. Comparison of the models has shown that CSBoost has the highest accuracy among them. The results indicate that elbow and knee joints, gender, number of joints and ESR test result have the most impact in the diagnosis of rheumatoid arthritis. The models can be applied in a computer software to predict rheumatoid arthritis and become a decision support for physicians.

REFERENCES

[1] Scott DL, Wolfe F, Huizinga TW,"Rheumatoid arthritis", Lancet, Volume 376 (9746),pp.1094–108, 2010, Sep 25.

[2] Majithia V, Geraci SA. "Rheumatoid arthritis: diagnosis and management". American Journal of Medicine. Volume 120- No11, pp. 936–9, 2007.

[3] Chin CY, Weng MY, Lin TC, Cheng SY, Yang YHK, Tseng VS. "Mining Disease Risk Patterns from Nationwide, Clinical Databases for the Assessment of Early Rheumatoid Arthritis Risk". PLoS ONE. Volume 10- N0-4, 2015.

[4] RACGP. "Clinical guideline for the diagnosis and management of early rheumatoid arthritis". The Royal Australian College of General Practitioners, 1 Palmerston Crescent, South Melbourne, Vic 3205 Australia, 2009

[5] Cader MZ, Filer A, Hazlehurst J, de Pablo P, Buckley CD, Raza K. "Performance of the 2010 ACR/EULAR criteria for rheumatoid arthritis: comparison with 1987 ACR criteria in a very early synovitis cohort. Annals of the Rheumatic Diseases". Volume 70- No. 6. pp. 949–55, 2011.

[6] van der Helm-vanMil AHM, le Cessie S, van Dongen H, Breedveld FC, Toes REM, Huizinga TWJ. "A prediction rule for disease outcome in patients with Recent-onset undifferentiated arthritis: How to guide individual treatment decisions". Arthritis & Rheumatism.; Volume 56- No.2, pp. 433–40. 2007 Feb.

[7] A. AZIZ, N. ISMAIL, and F. AHMAD, "MINING STUDENTS'ACADEMIC PERFORMANCE.," Journal of Theoretical & Applied Information Technology, vol. 53, no. 3, 2013.

[8] Bedran Z, Quiroz C, Rosa J, Catoggio LJ, Soriano ER. "Validation of a Prediction Rule for the Diagnosis of Rheumatoid Arthritis in Patients with Recent Onset Undifferentiated Arthritis". International Journal of Rheumatology. 548502, 2013.

[9] G. Zheng, M. Jiang, C. Lu, H. Guo, J. Zhan و A. Lu, "Exploring the biological basis of deficiency pattern in rheumatoid arthritis through text mining," BIBM Workshops;, 2011.

[10] P. Yildirim, Ç. Çeken, R. Hassanpour , M. R. Tolun, "Prediction of Similarities Among Rheumatic Diseases," J. Medical Systems, Vol.36., No 3, pp. 1485-1490, 2012

[11] Jyoti Soni, Ujma Ansari, Dipesh Sharma," Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887), Volume 17– No.8, pp. 43-48, March 2011.

[12] Chaitrali S. Danger, Sulabha S. Apte, —"Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques. International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012.

[13] Sellappan Palaniappan, Rafiah Awang," Intelligent Heart Disease Prediction System Using Data Mining Technique", 978-1-4244-1968-5/08 IEEE, 2008

[14] Jyoti Soni, Sunita Soni et al., —Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction‖; International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011

[15] Vanisree K, Jyothi Singaraju, "Decision Support System for Congenital Heart Disease Diagnosis based on Signs and Symptoms using Neural Networks", International Journal of Computer Applications (0975 – 8887) Volume 19– No.6, April 2011

[16] Shantakumar B.Patil, Y.S.Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network", European Journal of Scientific Research ISSN 1450-216X, Volume.31 No.4, pp. 642-656,2009

[17] Ng T, Chew L, Yap CW. A Clinical Decision Support Tool To Predict Survival in Cancer Patients beyond the 120 Days after Palliative Chemotherapy. Journal of Palliative Medicine; Volume 15- No 8, pp. 863–9. 2012 Aug.

[18] Shweta Kharya," USING DATA MINING TECHNIQUES FOR DIAGNOSIS AND PROGNOSIS OF CANCER DISEASE", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, 2012 April

[19] Lee SM, Kang JO, Suh YM. ,"Comparison of hospital charge prediction models of colorectal cancer patients: neural network vs. decision tree models";J Korean MED Sci., Volume 19 (5), pp. 677-81, 2004 Oct

[20] Chien CW, Lee YC, Ma T, Lee TS, Lin YC, Wang W. and Lee WJ," The application of artificial neural networks and decision tree model in predicting post-operative complication for gastric cancer patients". Hepatogastroenterology, Volume 55, pp. 1140- 1145, 2008.

[21] Wytske A. Altenburg, Mathieu H.G. de Greef, Nick H.T. Ten Hacken, Johan B. Wempe," A better response in exercise capacity after pulmonary rehabilitation in more severe COPD patients", Respiratory Medicine volume 106, pp. 694-700,2012.

[22] Behnke M, Wewel AR, Kirsten D, Jorres RA, Magnussen H.,"Exercise training raises daily activity stronger than predicted from exercise capacity in patients with COPD", Respir Med Jun;Volume 99(6):pp.711-717, 2005.

[23] Garrod R, Marshall J, Barley E, D. Jones PW.," Predictors of success and failure in pulmonary rehabilitation". EUR Respir J, Volume 27 (4), pp. 788-94,2006 April.

[24] A. Marcano-Cedeño , Paloma Chausa a, Alejandro García, César Cáceres, Josep M. Tormos, Enrique J. Gómez," Data mining applied to the cognitive rehabilitation of patients with acquired brain injury", Expert Systems with Applications, Volume 40, pp. 1054–1060,2013.

[25] Pang, B. C., Kuralmani, V., Joshi, R., Hongli, Y., Lee, K. K., Ang, B. T., et al,"Hybrid outcome prediction model for severe traumatic brain injury", Journal of Neurotrauma,Volume 24(1),pp. 136–146,2007.

[26] Rughani, A. I., Dumont, T. M., Lu, Z., Josh Bongar, M. S., Horgan, M. A., Penar, P. L.,et al," Use of an artificial neural network to predict head injury outcome." ,Journal of Neurosurgery, Volume 113 (3), pp. 585–590,2011

[27] S.Priya, R.R.Rajalaxmi," An Improved Data Mining Model to Predict the Occurrence of Type-2 Diabetes using Neural Network", ICON3C,pp.26-29,2012

[28] B.M. Patil, R.C. Joshi, Durga Toshniwal, "Hybrid Prediction Model for Type-2 Diabetic patients", Expert Systems with Applications, Science direct, pp. 8102-8108,2010.

[29] Casanova R, Saldana S, Chew EY, Danis RP, Greven CM, Ambrosius WT."Application of Random Forests Methods to Diabetic Retinopathy Classification Analyses". PLoS ONE 9(6): e98587. doi:10.1371/journal.pone.0098587,2014.

[30] Cader MZ, Filer A, Hazlehurst J, de Pablo P, Buckley CD, Raza K. Performance of the 2010 ACR/EULAR criteria for rheumatoid arthritis: comparison with 1987 ACR criteria in a very early synovitis cohort. Annals of the Rheumatic Diseases. 2011; 70(6):949–55.

[31] Huizinga TWJ, van der Helmvan Mil AHM. Prediction and prevention of rheumatoid arthritis. Revista Colombiana de Reumatología. 2007; 14:106–14

[32] Freund, Y., "An Adaptive Version of the Boost by Majority Algorithm," Machine Learning, 43, 293-318, 2001.

[33] Freund, Y., and Schapire, R., "A Short Introduction to Boosting," Journal of the Japanese Society for Artificial Intelligence, 14, 771-780, 1999.

[34] Freund, Y., and Schapire, R., "A Decision-Theoretic Generalization of On-LineLearning and an Application toBoosting," Journal of Computer and System Sciences, 55(1), 119–139, 1997.

[35] Friedman, J. H., "Recent Advances in Predictive (Machine) Learning," Journal of Classification, 23, 175-197, 2006.

[36] E. AhmedSharaf , M. A. Moustafa, M. Harb و A. Emara, "ADABOOST ENSEMBLE WITH SIMPLE GENETIC ALGORITHM FOR STUDENT PREDICTION MODEL," International Journal of Computer Science & Information Technology (IJCSIT), Volume 5(2), pp. 73-85, 2013.

[37] M. Billah, Symptom Analysis of Parkinson Disease using SVM-SMO and Ada-Boost Classifiers, BRAC University, Dhaka, Bangladesh, 2014.

[38] X.-S. Yang, S. Deb, "Cuckoo Search via L´evy Flights," IEEE World Congress on Nature & Biologically Inspired Computing;, pp. 210-215, 2009.

[39] R. Rajabioun, "Cuckoo Optimization Algorithm," Applied Soft Computing Journal;, Vol.11, pp. 5508, 2011.

[40] L. Reyzin, R. E. and Schapire, "How Boosting the Margin Can Also Boost Classifier Complexity;," Proceedings of the 23rd international conference on Machine Learning;, pp. 753-760;, 2006.

[41] Linda Miner et al., "Practical Predictive Analytics and Decisioning Systems for Medicine", ISBN: 978-0-12-411643-6, 2015.

[42] P. Saengsiri,S.N.Wichian,P.Meesad,andU.Herwig,"Comparison of hybrid feature selection models on gene expression data," in 8th International Conference on ICT and Knowledge Engineering, pp.13-18, 2010.

[43] Shoushan Li et al , "A frame work of feature Selection Methods for Text categorization" , Proceedings of 47th Annual meeting of ACL & 4th ICCNLP of AFNLP , pp 692-700, 2009.