A STUDY ON ASSOCIATION RULE MINING FOR VARIOUS HEART DISEASES MEDICAL DATA

Prof. Anish Kumar Choudhary,

Chameli Devi Institute of Professional Studies, Indore

ABSTRACT

This paper describes our experience on discovering association rules in medical data to predict heart disease. Heart disease is the leading causes of mortality accounting for 32% of all death, a rate is high as in Canada (35%) and USA. Association rule mining a computational intelligence approach is used to identify the factors that contribute to heart disease and Uci Cleveland data set, a biological data base is considered along with the rule generation algorithm – Apriori. Analyzing the information available on sick and healthy individuals and taking confidence as indicator. Females are seen to have more chance of being free from coronary heart disease than males. It is also seen that factors such as chest pain being asymptomatic and the presence of exercise- induced angina indicate the likely of existence of heart disease for both men and women. On the other hand, the result showed that when exercise induced angina (chest pain) was false, it was a good indicator of a person being healthy irrespective of gender. This research has demonstrated the use of rule mining to determine interesting knowledge.

Keywords: *Heart Disease, Apriori, Association Rule Mining, Computational Intelligence, Uci Cleveland*

I. INTRODUCTION

Heart is a muscular organ situated near the middle of chest, it is responsible for pumping blood to the other part of the body and together with network of blood vessels and blood from the human body's cardiovascular system, disruptions to this circulation of blood can result in serious health problem including death. Throughout history, humans have been affected by life-threatening diseases. Of the various life-threatening diseases, heart disease has received a great deal of attention from medical researchers.

Heart disease is the major cause of deaths. The World Health Organization (WHO) has estimated that 12 million deaths occur worldwide, every year due to the Heart diseases. In 2008, 17.3 million people died due to Heart Disease. Over 80% of deaths in world are because of Heart disease. WHO estimated by 2030, almost 23.6 million people will die due to Heart disease as written in [10]. Prediction by using data mining techniques gives us accurate result of disease.

Computational intelligence concepts have recently been used in discovering the relationships between different diseases and patient attributes (Huang, Li, Su, Watts, & Chen, 2007; Ishibuchi, Kuwajima, Nojima, 2007; Karabatak & Ince, 2009; Shin et al., 2010; Wang & Hoy, 2005). So, this research also uses the computational intelligence approach. Particularly, this research presents rule extraction experiments on heart disease data using rule mining algorithms – Apriori. It also highlights the efficiency of these algorithms for this diagnostic task. A

considerable issue in a research on heart disease diagnosis is the privacy issue related to medical data. So, Cleveland dataset (UCI, 2009), a publicly available dataset and widely popular with data mining researchers, has been used. For heart disease, diagnostic systems are time consuming, costly and prone to errors. Patients suffering from heart disease need to be under constant observation as improper treatment can be fatal. Proper identification of the disease and early treatment are essential. The World Health Organization (WHO) identified the potential of data mining for improving the problems in this medical domain as early as 1997 (Gulbinat, 1997). In the WHO research, emphasis was placed on the usefulness of knowledge detection from medical data repositories that could benefit medical diagnosis and prediction, patient health planning and progress, healthcare system monitoring and assessment, hospital and health services management, and disease prevention. This paper is motivated by these views and the aforementioned issues, and proposes a set of computational intelligence based approaches for diagnosing heart disease.

II. PROBLEM STATEMENT

Many hospital information systems are designed to support patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but they are largely limited. They can answer simple queries like "What is the average age of patients who have heart disease?", "How many surgeries had resulted in hospital stays longer than 10 days?", "Identify the female patients who are single, above 30 years old, and who have been treated for cancer."However, they cannot answer complex queries like "Identify the important preoperative predictors that increase the length of hospital stay", "Given patient records on cancer, should treatment include chemotherapy alone, radiation alone, or both chemotherapy and radiation?", and "Given patient records, predict the probability of patients getting a heart disease." Clinical decisions are often made based on doctors‟ intuition and experience rather than on the knowledge-rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. Wu, et al proposed that integration of clinical decision support with computer- based patient records could reduce medical errors, enhance patient safety, decrease unwanted practice variation, and improve patient outcome [7]. This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

III. RESEARCH METHODOLOGY

This research paper exhibits a critical analysis of well-known data mining algorithm that could prove to be beneficial for the medical practitioners and analysts for accurately predicting the heart disease diagnosis [6]. The methodology used for this research study includes the Association Rules available in *STATISTICA* Data Miner (SDM) are used in this project. It uses the so called Apriori algorithm. It needs predefined "threshold" values for association detection.
Thresholds are :

- Minimum Support
- Minimum Confidence
- Minimum Correlation

3.1 Applying Apriori Algorithm Over Medical Data

This data mining algorithm could be used for finding the frequent item sets from a transactional dataset, and then generate association rules. However, under several circumstances finding item sets is not trivial due to the combinational explosion. One the frequent item sets are obtained, they automatically generate an association rule that is either equal or greater than the minimum number of users confidence. Apriori is a seminal algorithm for finding frequent item sets using candidate generation [3]. It is characterized as a level-wise complete search algorithm using anti-monotonicity of item sets, "if an item set is not frequent, any of its superset is never frequent". In this algorithm the system assumes that the items existing within a transaction are stored in lexicographic order. The algorithm then lets the set of frequent item set to be of size k be Fk and their candidates be of size Ck. Then in the next step the algorithm searches for a frequent number of the item sets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement. It then iterates on the following three steps and extracts all the frequent item sets.

1. Generate Ck+1, candidates of frequent item sets of size k +1, from the frequent item sets of size k.
2. Scan the database and calculate the support of each candidate of frequent item sets.
3. Add those item sets that satisfies the minimum support requirement to Fk+1.

1. Function Apriori generates Ck+1 from Fk in the following two step process:

1. Join step: Generate RK+1, the initial candidates of frequent item sets of size k + 1 by
2. taking the union of the two frequent item sets of size k, Pk and Qk that have the first k−1
3. elements in common.

4. RK+1 = Pk ∪ Qk ={item1 , item2, . . . , itemk−1, itemk , itemk" }

5. Pk = {item1 , item2, . . . , itemk−1, itemk }

6. Qk = {item1 , item2, . . . , itemk−1, itemk" }

7. where, item1 < item2 < · · · < itemk < itemk"_ .

2. Prune step: Check if all the item sets of size k in Rk+1 are frequent and generate Ck+1 by removing those that do not pass this
8. requirement from Rk+1. This is because any subset of size k of Ck+1 that is not frequent cannot be a subset of a frequent item set of size
9. k + 1. Function subset finds all the candidates of the frequent item sets included in transaction t. Apriori, then, calculates frequency only for those candidates generated this way by scanning the database. It is evident that Apriori scans the database at most kmax+1 times when the maximum size of frequent item sets is set at kmax.

IV. IMPLEMENTATION AND ANALYSI

1 DATA SET: As mentioned earlier, we use the publicly available UCI heart disease dataset in our research. The heart disease dataset consists of a total of 76 attributes, however majority of the studies use a maximum of 14 attributes (, 2010; UCI, 2009) as these are considerably linked to the heart disease. These 14 attributes are as follows: (, 2010; UCI, 2009).
1. Age: numeric;

2. Sex: nominal – 2 values: male, female;
3. Chest pain type: nominal – 4 values: typical angina (angina), atypical angina (abnang), non anginal pain (notang), asymptomatic (asympt).
4. Trestbps: numeric, indicates resting blood pressure on admission;

5. Chol:: numeric, indicates Serum cholesterol in mg/dl;

6. Fbs: nominal – 2 values: True, False, indicates whether fasting blood sugar is greater than 120 mg/dl;

7. Restecg: nominal – 4 values: normal (norm), abnormal (abn): ST–T wave abnormality, ventricular hypertrophy (hyp) – indicates resting electrocardiographic outcomes;
8. Thalach: numeric, indicates maximum heart rate achieved;

9. Exang: nominal – 2 values: yes, no – highlights existence of exercise induced angina;

10. Oldpeak: numeric: ST depression induced by exercise relative to rest;

11. Slope: nominal – 3 values: upsloping, flat, downsloping – the slope characteristics of the peak exercise ST segment;
12. Ca: numeric – number of fluoroscopy colored major vessels (0–3);

13. Thal: nominal – 3 values: normal, fixed defect, reversible defect- the heart status;

14. The class attribute: value is either healthy or existence of heart disease (sick type: 1, 2, 3, and

4.2 Association Rule Mining on Heart Disease Data

While most existing works have considered the Cleveland database as a classification problem, we view, in this research, the dataset as a knowledge extraction problem and explore the use of association rule mining. Two experiments have been performed.
The experiments set out extracting rules to indicate healthy and sick conditions. In the medical domain, the gender of a person has been found to be an important factor influencing heart disease (Andersen & Haraldsdottir, 2009; Barrett-Connor, Cohn, Wingard, & Edelstein, 1991; Dalaker, Smith, Arnesen, & Prydz, 2009; Ferrara et al., 2008; Flint et al., 2010; Haley, Roth, Howard, & Safford, 2010; Jeppesen, Hein, Suadicani, & Gyntelberg, 1998; Pencina, D"Agostino, Larson, Massaro, & Vasan, 2009; Schenck-Gustafsson, 2009; Tucker et al., 2009). Details of these two experiments are provided in the following sub-sections

Table 4.1 Rule Extraction for Healthy Class through the Apriori Algorithm

Algorithms	Rules
Apriori	Healthy rules:
	If{Sex=femaleandexercise_induced_angina=Noandthal=normal}=> class **healthy** (conf., 0.8985).
	If{Sex=femaleandnumber_of_vessels_colored=0andthal=normal}=> class **healthy** (conf., 0.8611).
	If {exercise_induced_angina = No and thal = normal and number_of_vessels_colored=0 and slope=upsoping} => class **healthy** (conf., 0.8571).

Table 4.2 Rule Extraction for Sick Class Through the Apriori Algorithm

Algorithms	Rules
Apriori	Sick rules:
	If{chestpaintype=asymptomicandthal=reversibledefect}=>class**sick**(conf., 0.91).
	If{chestpaintype=asymptomicandexercise_induced_angina=Yes}=>class**sick** (conf., 0.875).
	If { chest pain type=asymptomic and slope=flat} => class **sick** (conf., 0.8095).

Frequent i tem sets com puted (cl evel ad_heart)
Min. support = 20.0%, Min. confidence = 20.0%, Min. correl ation = 20.0
Max. size of body = 10, Max. size of head = 10

	Frequent i tem sets	Frequency	Support(%)
11	fasti ng blood sugar <120 m g	258.000	85.1485
9	norma	232.000	76.5676
1	M ale	206.000	67.9868
4	0	206.000	67.9868
3	No	204.000	67.3267
76	norm al, fasti ng blood sugar <120 m	200.000	66.0066
52	0, fasti ng blood sugar <120 m g	180.000	59.4059
43	No, fasti ng blood sugar <120 m g	175.000	57.7557
26	M ale, fasti ng blood sugar <120 m	173.000	57.0957
50	0, norm a	172.000	56.7656
41	No, norm a	168.000	55.4455
5	health	164.000	54.1254
37	No, 0	153.000	50.4950
150	0, norm al, fasti ng blood sugar <120 m	152.000	50.1650
56	healthy, norm	151.000	49.8349
2	showi ng probab	148.000	48.8448
24	M ale, norm	147.000	48.5148
6	asym ptom ati	144.000	47.5247
47	0, health	144.000	47.5247
15	upsl opin	142.000	46.8646
137	No, norm al, fasti ng blood sugar <120 m	142.000	46.8646
38	No, heal th	141.000	46.5346
57	healthy, fasti ng blood sugar <120 m	141.000	46.5346

8	fl a	140.000	46.2046
10	si ck	139.000	45.8745
128	No, 0, fasti ng blood sugar <120 m g	135.000	44.5544
127	No, 0, norm a	134.000	44.2244
144	0, heal thy, norm	131.000	43.2343
19	M ale,	130.000	42.9042
131	No, heal thy, norm	130.000	42.9042
18	M ale, N	129.000	42.5742
155	healthy, norm al, fasti ng blood sugar <120 m	129.000	42.5742
65	asym ptomatic, fasti ng blood sugar <120 m	126.000	41.5841
79	normal, upsl opin	125.000	41.2541
145	0, healthy, fasti ng blood sugar <120 m	125.000	41.2541
111	M ale, normal, fasti ng blood sugar <120 m	124.000	40.9240
126	No, 0, heal th	124.000	40.9240
36	owing probable, fasti ng blood sugar <120 m	122.000	40.2640
85	fasti ng blood sugar <120 m g/dl, upsl op	122.000	40.2640
73	fl at, fasti ng blood sugar <120 m g	121.000	39.9339
132	No, healthy, fasti ng blood sugar <120 m	120.000	39.6039
12	reversabl e defe	117.000	38.6138
55	0, upsl opin	117.000	38.6138
80	si ck, fasti ng blood sugar <120 m g/	117.000	38.6138
199	No, 0, normal, fasti ng blood sugar <120 m	117.000	38.6138
46	No, upslopin	116.000	38.2838
25	M ale, si ck	114.000	37.6237
196	No, 0, heal thy, norm	113.000	37.2937
208	healthy, norm al, fasti ng blood sugar <120 m	113.000	37.2937

Figure 4.1: Frequent item Sets Computed

In the experiments, all healthy individuals were regarded to be in one class and sick individuals to be in another class. Popular association rule mining algorithm, Apriori was used for the experiments. Results of the experiment are shown in figure 4.1 – 4.5. Rules with confidence levels above 80%, with accuracy levels above 99% and confirmation levels above 79% were selected. As there can be many such rules, only the rules containing the „sick" or „healthy" class in the right-hand side (RHS) were considered. If no such rules were available, rules containing the „sick" or „healthy" class in the left-hand side (LHS) were reported.

Summary of association rules (clevelad_heart)
Min. support = 20.0%, Min. confidence = 20.0%, Min. correlation = 20.0% Max. size of body = 10,
Max. size of head = 10

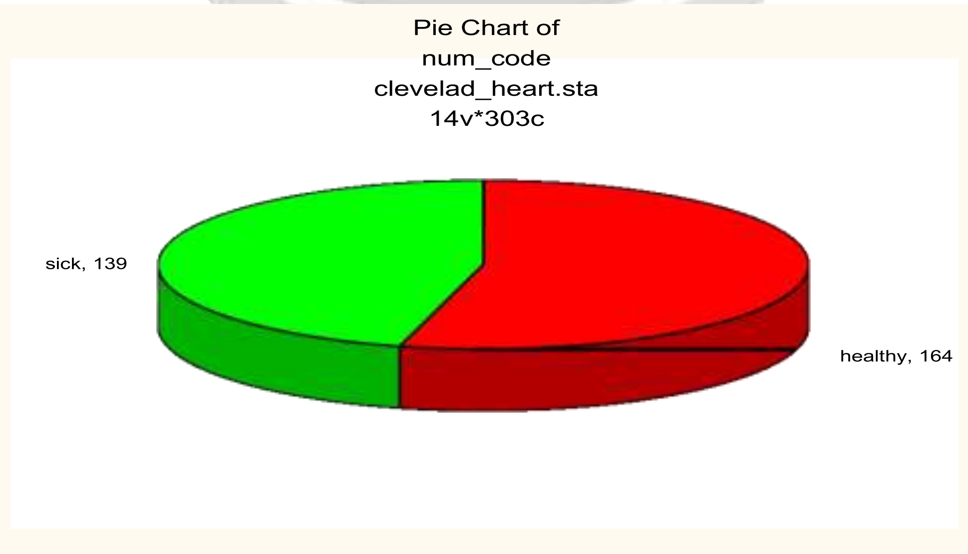
	Body	==>	Head	Support (%)	Conf idence(%)	Correlat ion(
1329	No, normal, Femal	==>	healt hy	20. 46205	89. 85507	58. 283
1383	0, normal, Femal	==>	healt hy	20. 46205	86. 11111	57. 056
1460	No, 0, normal, ups lopin	==>	healt hy	25. 74257	85. 71429	63. 848
850	No, Female	==>	healt hy	21. 12211	85. 33333	57. 706
1122	normal, non-anginal pair	==>	healt hy	20. 79208	85. 13514	57. 187
1496	o, 0, normal, fasting blood sugar <120 mg/dl, upslopi	==>	healt hy	21. 78218	84. 61538	58. 354
1269	No, 0, norma	==>	healt hy	37. 29373	84. 32836	76. 226
949	0, F emale	==>	healt hy	21. 12211	84. 21053	57. 325
1433	normal, f ast ing blood sugar < 120 mg/ dl, Fem	==>	healt hy	21. 12211	84. 21053	57. 325
1333	No, normal, upsloping	==>	healt hy	29. 37294	83. 96226	67. 501
1387	0, normal, upsloping	==>	healt hy	29. 04290	83. 80952	67. 060
1457	No, 0, normal, f ast ing blood sugar < 120 mg	==>	healt hy	32. 01320	82. 90598	70. 025
1287	No, 0, upsloping	==>	healt hy	27. 06271	82. 82828	64. 353
1488	0, normal, fast ing blood sugar < 120 mg/dl, upslopi	==>	healt hy	24. 75248	82. 41758	61. 392
1125	normal, F emal	==>	healt hy	23. 10231	82. 35294	59. 287
1476	No, normal, fasting blood sugar <120 mg/dl, upslopi	==>	healt hy	24. 42244	82. 22222	60. 909
1463	No, 0, f ast ing blood sugar < 120 mg/ dl, upslopi	==>	healt hy	23. 10231	81. 39535	58. 942
846	No, non-anginal pain	==>	healt hy	20. 13201	81. 33333	55. 001

751	No, 0	==>	healt hy	40. 92409	81. 04575	78. 280
1135	normal, upsloping	==>	healt hy	33. 33333	80. 80000	70. 541
857	No, upsloping	==>	healt hy	30. 69307	80. 17241	67. 426
956	0, ups loping	==>	healt hy	30. 69307	79. 48718	67. 137
1437	normal, f ast ing blood sugar < 120 mg/ dl, upslopi	==>	healt hy	28. 05281	79. 43925	64. 166
1278	No, 0, f ast ing blood sugar < 120 mg/	==>	healt hy	35. 31353	79. 25926	71. 910
537	non-anginal pain	==>	healt hy	22. 44224	79. 06977	57. 258
1396	0, f ast ing blood sugar < 120 mg/dl, upslopin	==>	healt hy	26. 40264	78. 43137	61. 854
1342	No, fasting blood s ugar < 120 mg/ dl, upslopir	==>	healt hy	25. 74257	78. 00000	60. 907
1173	f asting blood sugar < 120 mg/ dl, Fema	==>	healt hy	21. 78218	77. 64706	55. 900
1321	No, normal, f asting blood sugar < 120 mg/	==>	healt hy	36. 30363	77. 46479	72. 081
797	No, norma	==>	healt hy	42. 90429	77. 38095	78. 318
900	0, norma	==>	healt hy	43. 23432	76. 16279	77. 998
567	upsloping	==>	healt hy	34. 98350	74. 64789	69. 460
1375	0, normal, fast ing blood sugar < 120 mg/	==>	healt hy	37. 29373	74. 34211	71. 570
546	Female	==>	healt hy	23. 76238	74. 22680	57. 085
1181	f asting blood sugar < 120 mg/ dl, upslopin	==>	healt hy	29. 70297	73. 77049	63. 626
1197	Male, No, C	==>	healt hy	22. 11221	72. 82609	54. 545
151	0	==>	healt hy	47. 52475	69. 90291	78. 344
925	0, f ast ing blood sugar < 120 mg/	==>	healt hy	41. 25413	69. 44444	72. 753
86	No	==>	healt hy	46. 53465	69. 11765	77. 087
1222	Male, 0, norma	==>	healt hy	22. 77228	69. 00000	53. 879
1206	Male, No, norma	==>	healt hy	22. 44224	68. 68687	53. 366
823	No, fasting blood s ugar < 120 mg/	==>	healt hy	39. 60396	68. 57143	70. 833
720	Male, upsloping	==>	healt hy	20. 46205	65. 26316	49. 671
332	norma	==>	healt hy	49. 83498	65. 08621	77. 412
732	showing probable, 0	==>	healt hy	20. 13201	64. 89362	49. 129
1096	normal, f ast ing blood sugar < 120 mg/	==>	healt hy	42. 57426	64. 50000	71. 228

Figure 4.2: Association Rules to Indicate Healthy Condition

The rules for the „healthy" class were attributed to the female gender indicating that, based on this particular dataset, females have more chance of being free from coronary heart disease. Also if the results showed that when exercise induced angina (chest pain) was false, it was a good indicator of a person being healthy, irrespective of gender (exercise induced angina = false has appeared in the LHS of all the high confidence rules). The number of coloured vessels being zero and then (heart status) being normal were also shown to be good indicators of health. Rules mined for the „sick" class, on the other hand, showed that chest pain type being asymptomatic and than being reversed were probable indicators of a person being sick (both the high confidence rules have these two factors in LHS).

Figure 4.3: A Pie Chart to Indicate Healthy and Sick Proportion



V. CONCLUSION

This research has presented a rule extraction experiment on heart disease data using rule mining algorithms (Apriori). Further rule-mining-based analysis was undertaken by categorizing data based on gender and significant risk factors for heart disease were found for both men and women. Interestingly, it is found from the set of healthy rules, being „female‟ is one of the factors for a healthy heart condition. In other words, the results indicated females to have more chance of being free from coronary heart disease than males. This is supported by existing medical research as well. Research, for example, has identified that before the start of menopause, women have lower rates of coronary heart disease compared to their male counterparts of the same age (Castelli, 2007).

Overall, this research has demonstrated the use of rule mining to determine interesting knowledge. In medical literature, doctors are in discrepancies about the factors highlighted. This research has focused on the application of computational intelligence, in particular, association rule mining-based classifiers, to identify the key factors behind the disease, as well as considered gender diversity. The proposed work can be further enhanced and expanded for the automation of Heart disease prediction. In the future studies that researcher can use real data from Health care organizations and agencies and they use the available techniques for achieving optimum accuracy.

REFERENCES

[1] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. In *ACM SIGMOD Conference*, pages 207–216, 1993.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *VLDB Conference*, 1994.

[3] Roberto Bayardo and Rakesh Agrawal. Mining the most interesting rules. In *ACM KDD Conference*, 1999. UCI. 2009. Heart disease dataset <http://archive.ics.uci.edu/ml/machine-learningdatabases/heart- disease/cleve.mod> Accessed 5.03.09.

[4] UCI. 2010. Cleveland Heart disease data details. <http://archive.ics.uci.edu/ml/ machine-learning- databases/heart-disease/heart-disease.names> Accessed 8. 02.10. Vijaya, K., Khanna Nehemiah, H., Kannan, A., & Bhuvaneswari, N. (2010). Fuzzy neuro genetic approach for predicting the risk of cardiovascular diseases.

[5] International Journal of Data Mining, Modelling and Management, 2, 388–402. Wang, Z., & Hoy, W. (2005). Is the Framingham coronary heart disease absolute risk function applicable to Aboriginal people. Medical Journal of Australia, 182, 66–69.

[6] Kaur, H., Wasan, S. K.: "Empirical Study on Applications of Data Mining Techniques in Healthcare", Journal of Computer Science 2(2), 2006: pp. 194-200.

[7] Larose, Daniel T. *Discovering knowledge in data: an introduction to data mining*. Wiley. com, 2005.

[8] Obenshain, M.K: "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, 25(8), 2004: pp. 690–695.

[9] Kemal Polat and Salih Gunes," A new feature selection method on classification of medical datasets: Kernel F-score feature selection", Journal of Expert Systems with Applications, Vol. 36, PP.10367–10373, 2009.

[10] N.A. Setiawan, *et al*," A Comparative Study of Imputation Methods to Predict Missing Attribute Values in Coronary Heart Disease Data Set**"**, Journal in Department of Electrical and Electronic Engineering,Vol.21,
PP. 266–269, 2008

[11] Pasi Luukka and Jouni Lampinen," A Classification Method Based on Principal Component Analysis and Differential Evolution Algorithm Applied for Prediction Diagnosis from Clinical EMR Heart Data Sets" , Journal of Computer Intelligence in Optimization Adaption, Learning and Optimization, Volume 7, pp 263-283, 2010

[12] Resul Das and Ibrahim Turkoglu, *et al*, "Effective diagnosis of heart disease through neural networks ensembles", Journal of expert system with applications, Vol.36, PP. 7675–7680, 2009