A Study of Data Mining Techniques Accuracy for Healthcare

Hilal Almarabeh
King Saud Bin Abdulaziz University for Health Sciences, College of Sciences and Health Profession
Riyadh, Kingdom of Saudi Arabia

Ehab F. Amer
King Saud Bin Abdulaziz University for Health Sciences, College of Sciences and Health Profession
Riyadh, Kingdom of Saudi Arabia

ABSTRACT
Data mining is the analysis of large datasets to discover patterns and use those patterns to predict the likelihood of the future events. Data mining is becoming a very important field in healthcare sectors and it holds great potential for the healthcare industry. This paper presents an overview of current research being carried out using data mining techniques in different medical areas such as heart disease, diabetes, breast and lung cancer and skin disease by using different data mining techniques to find the best method of prediction and accuracy.

General Terms
Data Mining

Keywords
Data mining, Healthcare, Disease diagnosis, Data mining techniques, Accuracy.

1. INTRODUCTION
Nowadays, data mining is playing a vital role in healthcare and one of the most motivating areas of research with the objective of finding meaningful information from huge datasets [1]. This stored information is much useful for decision making process in healthcare and being to be potential with the help of knowledge discovery in database (KDD) and understand the cause of disease and providing better and cost effective treatment to patients. Data mining techniques used in healthcare play a significant role for detecting unknown information and diagnosis of the diseases. Different data mining techniques such as classification, clustering and association are used in healthcare organizations to exceed the efficiency for making decision concerning in patient heath. Various studies highlighted that data mining techniques help the data holder to analyze and discover unsuspected relationship among their data which in turn helpful for making decision [2]. Data miming also used for both analysis and prediction of various disease [3, 4]. Some researches work proposed an enhancement in available data methodology in order to improve the result [5-7] and some studies develop new methodology [8, 9] and proposed framework in order to improve the healthcare system [10, 11]. There are many of research studies available concerning in data mining in healthcare with their benefits and drawbacks.

2. DATA MINING IN HEALTHCARE
Nowadays, healthcare organizations generates a huge amount of data about patients, diseases, electronic medical record, medical devices, hospital management, prescriptions, and others, and the large data needs to be processed for knowledge extraction that enables support for cost-savings and decision making [12]. A great number of diseases are strongly associated with a symptom which makes it sophisticated for the physicians to predict the precise diseases on one go. Data mining is a best tool in predicting the disease which is almost valuable. Although the predicting is not considerably accurate, it assigns a valuable results to the physicians about the disease, so, data mining is an embracing to physicians to visualize the diseases in advance stages to produce the best treatment.

3. PERFORMANCE OF DATA MINING ALGORITHMS FOR DISEASE DIAGNOSE
The main idea of data mining techniques for disease diagnosis is to get the best performance such as efficiency and accuracy for prediction. This section illustrates the accuracy of data mining algorithms for different diseases such as heart disease, breast cancer, lung cancer, diabetes and skin disease.

3.1 Heart Disease
Heart is the most important organ in our body, it is responsible to pumps blood to the whole body. All heart disease concern to category of cardiovascular disease such as Coronary heart disease, Angina pectoris, Congestive heart failure, Cardiomyopathy, Congenital heart disease, Arrhythmias and Myocarditis [13]. There are number of factors which increase the chance of heart disease such as smoking, obesity, inactive physical exercises, high blood pressure, hypertension, and other. Many studies have been done on prediction heart disease by applying different data mining techniques to predict the accuracy of heart disease. Table.1. shows the effectiveness of data mining techniques used for heart disease form different research studies and Figure.1 represents the accuracy result analysis of different data mining techniques.

Table 1. Accuracy of Classifiers for Heart Disease

Ref#	Year	Techniques	Accuracy
[14]	2007	Neural Network	91%
[15]	2010	DT and GA Feature Reduction	99.2%
[16]	2012	G4.5 Classifier	74.20%
[14]	2007	SVM	92.1%
[9]	2012	K-NN	61.39%
[17]	2010	Multilayer NN	89.7%
[16]	2012	Naive Bayes	96.5%
[18]	2009	GSVM	95%
[19]	2010	Bayesian NN	78.43%
[20]	2015	Fuzzy Logic and DT	69.51%

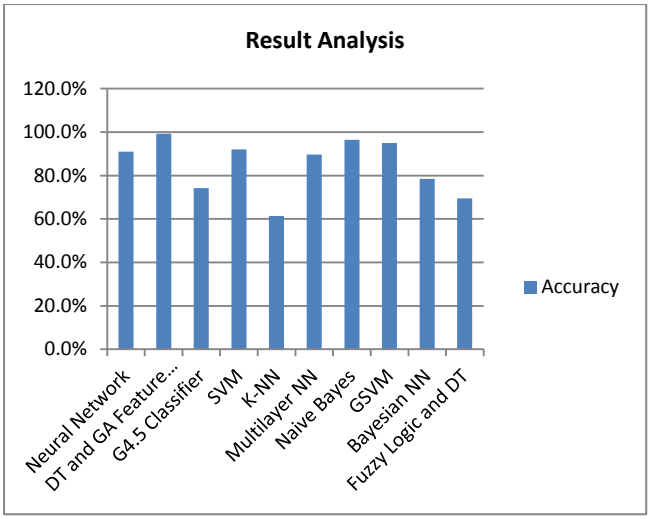

Figure.1. Accuracy Level of Different Data Mining Techniques for Heart Disease

It is observed form Fig.1. that the high accuracy belongs to Decision Tree (DT) and Genetic Algorithm (GA) Feature Reduction (99.2%) which has the best accuracy among other data mining techniques for heart disease prediction. Although K-NN algorithm is simple to use but it has the lowest accuracy (61.39%) and this means it does not give a good result for prediction.

3.2 Cancer Disease

Cancer is a deadly disease which characterized by the uncontrolled growth of abnormal cells in the body. Detection cancer in early stages is difficult, but early detection of cancer is curable. Men are more prone to lung, prostate, stomach and liver cancer, while women are more prone to breast, colorectal, lung, cervix uteri, and stomach cancer [21]. There are too many factors that cause a cancer such as the immune system, tobacco, alcohol and other. Breast cancer is being extremely injurious to all women, may be a reason for lost of breast or their life. Lung cancer is one of the leading cause of cancer deaths in both women and men. Manifestation of Lung cancer in the body of the patient reveals through early symptoms in most of the cases [22]. Many studies have been done on prediction breast and lung cancer diseases by applying different data mining techniques. Table.2. and Table.3 represents the accuracy result analysis of different classifiers used to predict breast and lung cancer consecutively from different research studies. Figure.2. and Figure.3 represents the accuracy result analysis of data mining techniques for cancer and lung disease

Table 2. Accuracy of Classifiers for Breast Cancer

Ref#	Year	Techniques	Accuracy
[23]	2014	MLP BPN	95.71%
[24]	2012	J48	95.1359%
[25]	2015	Neural Network	98.09%
[26]	2016	C4.5	95.13%
[26]	2016	SVM	97.13%
[26]	2016	K-NN	95.27%
[27]	2016	LMT	96.18%
[27]	2016	Bayes Net	97.24%
[28]	2013	Naive Bayes	96.79%
[29]	2015	Decision Tree	96.50%

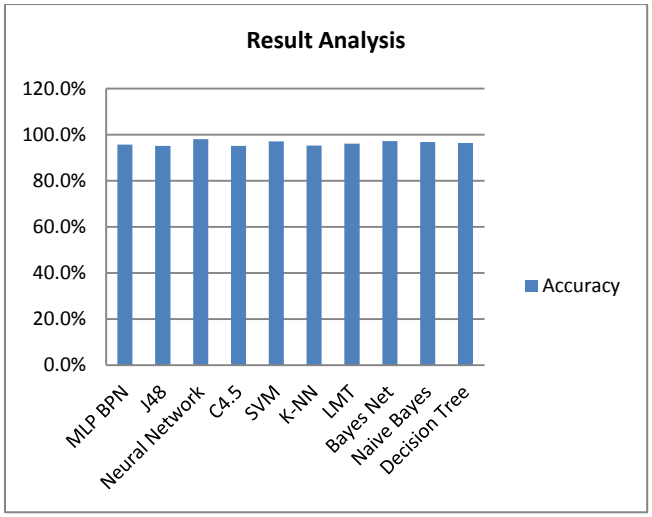

Figure 2: Accuracy Level of Different Data Mining Techniques for Breast Cancer Diseases Diagnosis

It is observed formTable.2 and Fig.2. that the high accuracy belongs to Neural Network (98.09%) which has the best accuracy among other data mining techniques for breast cancer disease prediction. Bayes Net and SVM follow that which has (97.24%) and (97.13) consecutively, and C4.5 (95.13) and J48 (95.1359) are same result and have the lowest accuracy result.

Table 3. Accuracy of Classifiers for Lung Cancer

Ref#	Year	Techniques	Accuracy
[30]	2013	Naive Bayes	89.03%
[31]	2014	Decision Table	76.2%
[31]	2014	J84	77.5%
[32]	2015	ANT Colony	78%
[33]	2014	ANN	83.5%
[34]	2016	SVM	95.12%
[35]	2016	EKNN	97%
[36]	2010	Bayes Network	77%
[37]	2011	J48	91.4%
[38]	2013	Neural Network	93.3%

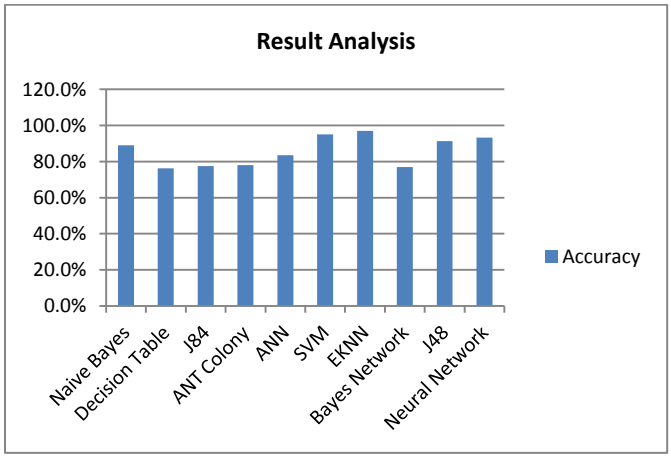

Figure 3: Accuracy Level of Different Data Mining Techniques for Lung Cancer Diseases Diagnosis

It is observed form Figure.3. that the high accuracy belongs to Enhanced K-NN algorithm(79%) and SVM(95.12%) which have the best accuracy among other data mining techniques for Lung cancer prediction. Although Bayes Network and J48

have a good accuracy to predict breast cancer but they have the lowest accuracy prediction for lung cancer (77%) and (77.5), respectively.

3.3 Diabetes Mellitus Diagnosis

Diabetes is often called a modern-society disease because widespread lack of regular exercise and rising obesity rates are some of the main contributing factors for it. Diabetes is a very serious disease that, if not treated properly and on time, can lead to very serious complications, including death. This makes diabetes one of the main priorities in medical science research, which in turn generates huge amounts of data [39]. Many studies have been done on prediction diabetes diseases by applying different data mining techniques. Table.4. represents the accuracy of different classifiers used to predict diabetes disease from different research studies. Figure.4 represents the accuracy result analysis of different data mining techniques for diabetes disease.

Table 4. Accuracy of Classifiers for Diabetes Disease

Ref#	Year	Techniques	Accuracy
[40]	2013	PCa with NN	71%
[41]	2014	FCM - Weka Tool	94.3 %
[42]	2014	BLR- Tanagra Tool	75%
[43]	2015	Naive Bayes-Weka Tool	76.95%
[44]	2015	ANN	89%
[45]	2016	J4.8 Weka Tool	99.87
[46]	2014	GA with Fuzzy Logic	80.5%
[47]	2014	Bayesian Network	90.4%
[48]	2014	C4.5 Weka Tool	86%
[49]	2016	Machine Learning Ensemble LDA,KNN and RPCT	94.27%

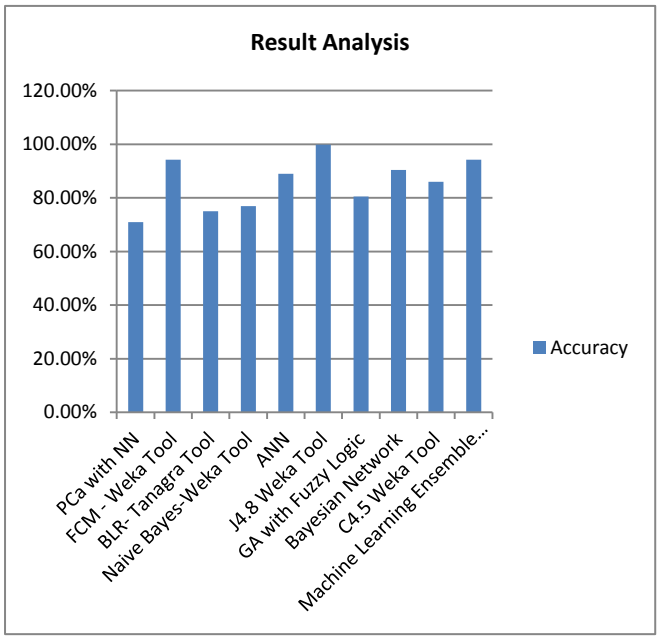

Figure 4: Accuracy Level of Different Data Mining Techniques for Diabetes Diagnosis

It is observed form Fig.4. that the high accuracy belongs to J4.8 (99.87%) by using weka tool, followed by machine FCM (94.3%) and learning ensemble (94.27%) which have the best accuracy among other data mining techniques for diabetes prediction. BLR algorithm by using Tangar tool has the lowest accuracy (75%) followed by Naive Bayes by using Weka tool (76.95%).

3.4 Skin Disease Diagnosis

Skin diseases or dermatological are becoming more and more in these days and many of these disease are dangerous if not treated at an early stages. Skin provides a protection against fungal infection, bacteria, allergy, viruses and controls temperature of body. Many of the skin disease such as acne, alopecia, ringworm, eczema affect the body look [50]. Many studies have been done on prediction skin diseases by applying different data mining techniques. Table.5. represents the accuracy of different classifiers used to predict diabetes disease from different research studies. Figure.5 represents the accuracy result analysis of different data mining techniques for skin disease.

Table 5. Accuracy of Classifiers for Skin Disease

Ref#	Year	Techniques	Accuracy
[51]	2015	Adaboost	65%
[51]	2015	BayesNet	80%
[51]	2015	J48	90%
[51]	2015	MLP	95%
[51]	2015	Naive Bayes	85%
[52]	2016	ANN	97.17%
[52]	2016	SVM	94.04%
[53]	2009	DT	92.62%
[54]	2012	Weighted K-NN	95.2381%
[55]	2009	DT and NN	92.62%

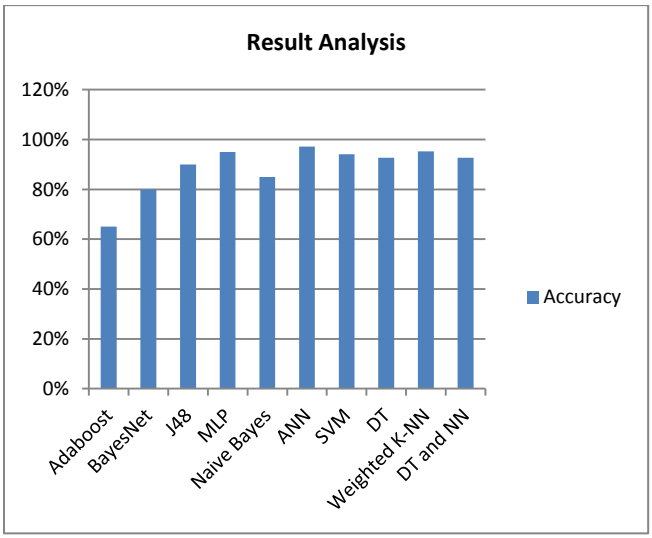

Fig 5: Accuracy Level of Different Data Mining Techniques for Skin Diagnosis

It's observed from Figure.5. , the ANN technique has the best accuracy (97.17%) among the other algorithms, and Weighted K-NN, MLP and SVM are slightly less than ANN. The lowest accuracy among all techniques belongs to Adaboost with (65%). From this result the ANN has the best accuracy for skin disease prediction.

4. CONCLUSION

Data mining has a significant importance in healthcare organizations. The obtained knowledge with the use of data mining techniques can be used to make successful and effective decisions that will improve and progress of healthcare organizations. This paper illustrates different data mining techniques for healthcare prediction. Techniques have used in heart, breast and lung cancer, diabetes and skin disease. Comparisons are made based on the accuracy among these techniques. The analysis results show that there is no single classifier which produce best result to all diseases. For heart disease, DT and GA has the best accuracy (99.2%), in breast cancer, NN (98.09%), lung cancer has (97%) when applied EKNN, in diabetes J4.8 has (99.87%) accuracy by using weka tools, and in skin disease the best accuracy is ANN (97.17%). Therefore, availability and quality are the most important factors in data mining. Healthcare organizations need to provide a strong data quality before doing any research. Providing a quality and enough data for research, data mining will achieve better discovery of knowledge in hidden in the medical data.

5. REFERENCES

[1] Thenmozhi, K. and P. Deepika, *Heart disease prediction using classification with different decision tree techniques.* Int. J. Eng. Res. Gen. Sci, 2014. **2**(6).

[2] Han, J., J. Pei, and M. Kamber, *Data mining: concepts and techniques*. 2011: Elsevier.

[3] Gupta, S., D. Kumar, and A. Sharma, *Data mining classification techniques applied for breast cancer diagnosis and prognosis.* Indian Journal of Computer Science and Engineering (IJCSE), 2011. **2**(2): p. 188-195.

[4] Kumari, M. and S. Godara, *Comparative study of data mining classification methods in cardiovascular disease prediction 1*. 2011.

[5] Ha, S.H. and S.H. Joo, *A hybrid data mining method for the medical classification of chest pain.* International Journal of Computer and Information Engineering, 2010. **4**(1): p. 33-38.

[6] Kavitha, K., K. Ramakrishnan, and M.K. Singh, *Modeling and design of evolutionary neural network for heart disease detection.* International Journal of Computer Science Issues, 2010. **7**(5): p. 272-283.

[7] Parvathi, R. and S. Palaniammalì, *An improved medical diagnosing technique using spatial association rules.* European Journal of Scientific Research ISSN, 2011: p. 49-59.

[8] Chao, S. and F. Wong. *An incremental decision tree learning methodology regarding attributes in medical data mining*. in *Machine Learning and Cybernetics, 2009 International Conference on*. 2009. IEEE.

[9] Habrard, A., M. Bernard, and F. Jacquenet. *Multi-relational Data Mining in medical databases*. in *Conference on Artificial Intelligence in Medicine in Europe*. 2003. Springer.

[10] Duan, L., W.N. Street, and E. Xu, *Healthcare information systems: data mining methods in the creation of a clinical recommender system.* Enterprise Information Systems, 2011. **5**(2): p. 169-181.

[11] Kumar, D.S., G. Sathyadevi, and S. Sivanesh, *Decision support system for medical diagnosis using data mining.* International Journal of Computer Science Issues, 2011. **8**(3): p. 147-153.

[12] Werts, N. and M. Adya, *Data Mining in Healthcare: Issues and a Research Agenda*. AMCIS 2000 Proceedings, 2000: p. 98.

[13] Purusothaman,G. and P. Krishnakumari, *A survey of data mining techniques on risk prediction: Heart disease*. Indian Journal of Science and Technology, 2015. **8**(12).

[14] Y. W. Xing, J. Wang, Z. H. Zhao, and Y. H. Gao., "Combination data mining methods with new medical data to predicting outcome of coronary heart disease," presented at the International Conference on Convergence Information Technology, 2007.

[15] M. Anbarsi, E. Anupriya, and N. Iyengar, "Enhanced prediction of heart disease with feature subset selection using genetic algorithm," International Journal of Engineering Science and Technology, vol. 2, no.10, pp. 5370-5376, 2010.

[16] R. Alizadehsani, J. Habibi, B. Bahadorian, H. Mashayekhi, A. Ghandeharioun, and R. Boghrati, et al., "Diagnosis of coronary arteries stenosis using data mining," J Med Signals Sens, vol. 2, pp. 153-9, Jul 2012.

[17] K. Srinivas, G. R. Rao, and A. Govardhan, "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," presented at the 5th International Conference on Computer Science and Education, 2010.

[18] Avci, "A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier", Expert Systems with Applications, Elsevier, vol. 36, (2009), pp. 10618-10626.

[19] Y. Kangwanariyakul, C. Nantasenamat, T. Tantimongcolwat and T. Naenna, "Data Mining of Magneto cardiograms For Prediction of Ischemic Heart Disease", EXCLI Journal, (2010).

[20] Jaekwon Kim, MS, Jongsik Lee, PhD, and Youngho Lee, PhD, , Data-Mining-Based Coronary Heart Disease Risk Prediction Model Using Fuzzy Logic and Decision Tree, Health inform Res, Jul(2015), Vol.21, no.3.

[21] Satyam Shukla, Dharmendra Lal Gupta, Bakshi Rohit Prasad, Comparative Study of Recent Trends on Cancer Disease Prediction Usind Data Mining Techniques, Internationl Journal of Database Theory and Application, Vol.9, no.6, (2016), pp.107-118.

[22] V.Krishnaiah , Dr.G.Narsimha, Dr.N.Subhash Chandra, Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques, International Journal of Computer Science and Information Technologies, Vol. 4 , no.1. , (2013), pp. 39 - 45.

[23] Soumadip Ghosh, Sujoy Mondal, Bhaskar Ghosh, A comparative study of breast cancer detection based on SVM and MLP BPN classifier, First International Conference on Automation, Control, Energy and Systems (ACES), IEEE, (2014).

[24] Gouda I. Salama , M.B.Abdelhalim , and Magdy Abd-elghany Zeid, Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers, International Journal of Computer and Information Technology Vol.1, no.1, September (2012), pp. 2277 – 0764.

[25] Zahra Nematzadeh, Roliana Ibrahim, Ali Selamat, Comparative studies on breast cancer classifications with k-fold cross validations using machine learning techniques, Control Conference (ASCC), 2015 10th Asian, IEEE, June (2015).

[26] Hiba Asria ,Hajar Mousannif, Hassan Al Moatassime, Thomas Noel, Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis, The 6th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS 2016), pp. 1064 – 1069.

[27] Mohammed Abdullah Hassan Al-Hagery, Classifiers' Accuracy Based on Breast Cancer Medical Data And Data Mining Techniques, International Journal of Advanced Biotechnology and Research (IJBR), Vol.7, Ino.2, (2016), pp760-772.

[28] G. Ravi Kumar, Dr. G. A. Ramachandra, K.Nagamani, An Efficient Prediction of Breast Cancer Data using Data Mining Techniques, International Journal of Innovations in Engineering and Technology (IJIET), Vol. 2 , no. 4, August (2013).

[29] Subrata Kumar Mandal, Performance Analysis Of Data Mining Algorithms For Breast Cancer Cell Detection Using Naïve Bayes, Logistic Regression and Decision Tree, International Journal Of Engineering And Computer Science , Vol.6, no.2 , Feb. (2017), pp. 20388-20391.

[30] A.Priyanga, S.Prakasam, Ph.D, Effectiveness of Data Mining - based Cancer Prediction System (DMBCPS), International Journal of Computer Applications , Vol. 83 – No 10, December (2013), pp. 0975 – 8887.

[31] Thangaraju , Barkavi , Karthikeyan, Mining Lung Cancer Data for Smokers and Non-Smokers by Using Data Mining Techniques, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3,no. 7, July (2014).

[32] T. Christopher, "A Study on Mining Lung Cancer Data for Increasing or Decreasing Disease Prediction Value by Using Ant Colony Optimization Techniques", Proceedings of the UGC Sponsored National Conference on Advanced Networking and Applications, (2015).

[33] Y.-C. Chen, W.-C. Ke, H.-W. Chiu Risk classification of cancer survival using ANN with gene expression data from multiple laboratories Compute Biol Med, Vol. 48, (2014), pp. 1–7.

[34] Yuvarani Mrs. P., Analysis of Lung Cancer Detection Algorithms - A Survey, Discovery the International journal, (2016), ISSN 2278 – 5469, EISSN 2278-5450.

[35] P. Thamilselvan, Dr. J. G. R. Sathiaseelan, An enhanced k nearest neighbor method to detecting and classifying MRI lung cancer images for large amount data, International Journal of Applied Engineering Research ISSN 0973-4562 Vol.11, No.6 (2016), pp 4223-4229

[36] K. Jayasurya, G. Fung, S. Yu, C. Dehing-Oberije, D. De Ruysscher, A. Hope, W. De Neve, Y. Lievens, P. Lambin, A. L. A. J. Dekker, Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy, T he International Journal of Medical Physics and Research, Vol. 37, no, 4, (2010).

[37] Ankit Agrawal, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi, Alok Choudhary, A Lung Cancer Outcome Calculator Using Ensemble Data Mining on SEER Data, ACM, (2011).

[38] Jinsa Kuruvilla, K. Gunavathi , Lung cancer classification using neural networks for CT images, ELSEVIER, (3013).

[39] Miroslav Marinov, Abu Saleh Mohammad Mosa, M.S, Illhoi Yoo, Ph.D, and Suzanne Austin Boren, Data-Mining Technologies for Diabetes: A Systematic Review, Journal of Diabetes Science and Technology, Vol.5, no.6, (2011).

[40] Rakesh Motka, Viral Parmar, Balbindra Kumar, A. R. Verma, " Diabetes Mellitus Forecast Using Different Data Mining Techniques", International conference on computer and Communication Technology, IEEE (2013).

[41] Ravi sankal, T.Jayakumari, ―prognosis of diabetes using data mining approach fuzzy c mean clustering and support vector machine‖-International journal of computer Trends and Technology (IJCIT) vol 11 number 2 may 2014.

[42] P.Radha, Dr.B.Srinivasan ―International Journal of Innovative science science Engineering & Technology Vol. 1 issue 6 august 2014.

[43] Sadri sa'di, Amanj maleki, Rami hashemi, Zahra panbechi, kamal chalabi, ―Comparison of Data Mining algorithms in the diagnosis of type- II diabetes‖ – International journal on computational science & Application(IJCSA) vol 5 no 5 october 2015.

[44] Durairaj M., Kalaiselvi G.,―Prediction of Diabetes Using Soft Computing Techniques- a Survey‖ International journal Of scientific & technology research, volume 4, ISSUE 03, PP.190-192, 2015.

[45] Dr.M.Renuka Devi, J.Maria Shyla, ―Analysis of various data mining Techniques to predict diabetes Mellitus‖,- International journal of Applied engineering Research‖, ISSn 0973-4562 vol 11 number 1 2016.

[46] Sudesh Rao, V. Arun Kumar, "Applying Data mining Technique to predict the diabetes of our future generations", ISRASE eXplore digital library, 2014.

[47] Mohtaram Mohammadi, Mitra Hosseini, Hamid Tabatabaee, "Using Bayesian Network for the prediction and Diagnosis of Diabetes", MAGNT Research Report, vol.2 (5), pp.892-902.

[48] P. Radha, Dr. B. Srinivasan, " Predicting Diabetes by consequence the various Data mining Classification Techniques", International Journal of Innovative Science, Engineering & Technology, vol. 1 Issue 6, August 2014, pp. 334-339.

[49] Madeeh Nayer Algedawy, Detecting Diabetes Mellitus using Machine Learning Ensemble, International Journal of Computer Systems (ISSN: 2394-1065), Volume 03– Issue 12, December, 2016.

[50] Nisha Yadav, Nisha Yadav, Virender Kumar Narang, Skin Diseases Detection Models using Image Processing: A Survey, International Journal of Computer Applications (0975 – 8887) Vol.137 No.12, March 2016.

[51] A.A.L.C. Amarathunga, E.P.W.C. Ellawala, G.N. Abeysekara, C. R. J. Amalraj, Expert System For Diagnosis Of Skin Diseases, INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 4, ISSUE 01, JANUARY 2015, ISSN 2277-8616 .

[52] Krupal S. Parikh, Trupti P. Shah , RahulKrishna Kota and Rita Vora, Diagnosing Common Skin Diseases using Soft Computing Techniques, International Journal of Bio-Science and Bio-Technology Vol.7, No.6 (2015), pp.275-286.

[53] L. Chang and C. H. Chen, "Applying decision tree and neural network to increase quality of dermatologic diagnosis", Expert Systems with Applications, Elsevier, vol. 36, (2009), pp. 4035-4041.

[54] Hatice cataloluk, Metin kesler," A Diagnostic Software Tool for Skin Diseases with Basic and Weighted K-NN" 978-1-4673-1448-0/12/$31.00 © IEEE 2012.

[55] L.Chang & C.H.Chen, "APPLYING DECISION TREE AND NEURAL NETWORK TO INCREASE QUALITY OF DERMATOLOGIC DIAGNOSIS", Expert Systems with Applications- Elsevier, Volume: 36, pp. 4035-4041, 2009.