USING DATA MINING TECHNIQUES IN HEART DISEASE DIAGNOSIS AND TREATMENT

Mai Shouman[1], Tim Turner[1], Rob Stocker[1]
[1]School of Engineering and Information Technology
University of New South Wales at the Australian Defence Force Academy
Northcott Drive, Canberra ACT 2600
m.shouman@student.adfa.edu.au, t.turner@adfa.edu.au, r.stocker@adfa.edu.au

Abstract— **The availability of huge amounts of medical data leads to the need for powerful data analysis tools to extract useful knowledge. Researchers have long been concerned with applying statistical and data mining tools to improve data analysis on large data sets. Disease diagnosis is one of the applications where data mining tools are proving successful results. Heart disease is the leading cause of death all over the world in the past ten years. Several researchers are using statistical and data mining tools to help health care professionals in the diagnosis of heart disease. Using single data mining technique in the diagnosis of heart disease has been comprehensively investigated showing acceptable levels of accuracy. Recently, researchers have been investigating the effect of hybridizing more than one technique showing enhanced results in the diagnosis of heart disease. However, using data mining techniques to identify a suitable treatment for heart disease patients has received less attention. This paper identifies gaps in the research on heart disease diagnosis and treatment and proposes a model to systematically close those gaps to discover if applying data mining techniques to heart disease treatment data can provide as reliable performance as that achieved in diagnosing heart disease.**

Keywords- E-health, Data Mining, Heart Disease Diagnosis and Treatment

I. INTRODUCTION

Data mining is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistical methods [1-5]. Data mining is rapidly growing successful in a wide range of applications such as analysis of organic compounds, financial forecasting, healthcare and weather forecasting [6]. Data mining in healthcare is an emerging field of high importance for providing prognosis and a deeper understanding of medical data [7]. Data mining applications in healthcare include analysis of health care centers for better health policy-making and prevention of hospital errors, early detection, prevention of diseases and preventable hospital deaths, more value for money and cost savings, and detection of fraudulent insurance claims [8]. Researchers are using data mining techniques in the diagnosis of several diseases such as diabetes [9], stroke [10], cancer [11], and heart disease [12].

Heart disease is the leading cause of death in the world over the past 10 years [13]. The European Public Health Alliance reported that heart attacks, strokes and other circulatory diseases account for 41% of all deaths [14]. The Economical and Social Commission of Asia and the Pacific reported that in one fifth of Asian countries, most lives are lost to non-communicable diseases such as cardiovascular diseases, cancers, diabetes and chronic respiratory diseases [15]. The Australian Bureau of Statistics reported that heart and circulatory system diseases are the first leading cause of death in Australia, causing 33.7% of all deaths [16]. Statistics of South Africa reported that heart and circulatory system diseases are the third leading cause of death in Africa [17].

Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amount of patients' data from which to extract useful knowledge, researchers have been using data mining techniques to help health care professionals in the diagnosis of heart disease [18-19]. Developing a tool to be embedded in the hospitals management system to help and give advice to the healthcare professionals in diagnosing and providing suitable treatment for heart disease patients is important. Several data mining techniques are used in the diagnosis of heart disease such as Naïve Bayes, Decision Tree, neural network, kernel density, automatically defined groups, bagging algorithm, and support vector machine showing different levels of accuracies [12, 20-24].

Although applying data mining in disease diagnosis and treatment is beneficial, less research has been done in identifying treatment plans for patients and especially for heart disease patients. Researchers have proven that hospitals do not provide the same quality of service even though they provide the same type of service [25]. Researchers are suggesting that applying data mining techniques in identifying effective treatments for patients can improve practitioner performance [26]. Researchers have been investigating applying different data mining techniques in the diagnosis of heart disease to identify which data mining technique can provide more reliable accuracy. There is no previous research that identifies which data mining technique can provide more reliable accuracy in identifying suitable treatment for heart disease patients.

This paper proposes a model for measuring if applying data mining techniques to heart disease treatment data can provide reliable performance as that achieved in diagnosing heart disease patients. The rest of the paper is divided as follows: section 2 provides an overview on using data mining techniques to help health care professionals in the diagnosis of heart disease, section 3 investigates future trends in using data mining techniques to help healthcare professionals in diagnosing and providing suitable treatments for heart disease

978-1-4673-0484-9/12/$31.00 ©2012 IEEE

patients, section 4 discusses the proposed research model which is followed by a summary section.

II. USING SINGLE AND HYBRID DATA MINING TECHNIQUES IN HEART DISEASE DIAGNOSIS

Statistical analyses have identified the risk factors associated with heart disease to be age, blood pressure, cholesterol, and smoking habit [27], total cholesterol [28], diabetes [29], hypertension and family history of heart disease [30], obesity, lack of physical activity, and high levels of smoking [31]. Knowledge of the risk factors associated with heart disease helps health care professionals to identify patients at high risk of having heart disease. Heart disease professionals store significant amounts of patients' data. It is important to analyze these datasets to extract useful knowledge. Data mining is an effective tool for analysing data to extract useful knowledge.

Different data mining techniques have been used to help health care professionals in the diagnosis of heart disease. Those most frequently used focus on classification: naïve bayes, decision tree, and neural network. Other data mining techniques are also used including kernel density, automatically defined groups, bagging algorithm, and support vector machine.

Table 1 shows a sample of different data mining techniques used in the diagnosis of heart disease over different heart disease datasets. The results of the different data mining research cannot be compared because they have used different datasets. However, over time a defacto benchmark data set has arisen in the literature: the Cleveland Heart Disease Dataset (CHDD http://archive.ics.uci.edu/ml/datasets/Heart+Disease). Results of trials on this dataset do allow comparison.

TABLE 1: A SAMPLE OF DATA MINING TECHNIQUES USED ON DIFFERENT HEART DISEASE DATASETS

Author	Year	Technique	Accuracy
Yan, et al.	2003	Multilayer Perceptron	63.6%
Andreeva, P.	2006	Naïve Bayes	78.563 %
		Decision Tree	75.738 %
		Neural network	82.773 %
		Kernel density	84.444 %
Palaniappan, et al.	2007	Naïve Bayes	95%
		Decision Trees	94.93%
		Neural Network	93.54%
De Beule, et al.	2007	Artificial neural network	82%
Tantimongcolwata, et al.	2008	Direct kernel self-organizing map	80.4%
		Multilayer Perceptron	74.5%
Hara, et al.	2008	Automatically Defined Groups	67.8%
		Immune Multi-agent Neural Network	82.3%
Sitar-Taut, et al.	2009	Naïve Bayes	62.03%
		Decision Trees	60.40%
Rajkumar, et al.	2010	Naive Bayes	52.33%
		KNN	45.67%
		Decision list	52%
Srinivas, et al.	2010	Naïve Bayes	84.14%
		One Dependency Augmented Naïve Bayes classifier	80.46%
Kangwanariyakul, et al.	2010	Back-propagation neural network	78.43%
		Bayesian neural network	78.43%
		probabilistic neural network	70.59%
		linear support vector machine	74.51%
		polynomial support vector machine	70.59%
		radial basis function kernel support vector machine	60.78%
Anbarasi, et al.	2010	Genetic with Decision tree	99.2%
		Genetic with Naïve Bayes	96.5%
		Genetic with Classification via clustering	88.3%

Table 2 illustrates a sample of data mining techniques used in the diagnosis of heart disease on the CHDD. Researchers have used data mining techniques on the heart disease benchmark dataset to extract trends and relationships between different variables such as blood pressure, cholesterol, and unstable angina [32-34].

TABLE 2: A SAMPLE OF DATA MINING TECHNIQUES USED ON THE CLEVELAND HEART DISEASE DATASET

Type	Author/ Year	Technique	Accuracy
Single	Cheung, 2001	Decision Tree	81.11%
		Naïve Bayes	81.48%
	Tu, et al., 2009	J4.8 Decision Tree	78.9%
		Bagging algorithm	81.41%
Hybrid	Polat et al., 2007	Fuzzy-AIRS–k-nearest neighbour	87%
	Das, et al., 2009	Neural network ensembles	89.01%

Recently, researchers started using hybrid data mining techniques in the diagnosis of heart disease. Polat et al., used fuzzy artificial immune recognition system and k-nearest neighbour in the detection of heart disease using the CHDD. The proposed model showed accuracy of 87% in the detection of heart disease patients [35] (Table 2). Das et al., used neural network ensembles in the diagnosis of heart disease using the CHDD showing accuracy of 89.01% [12] (Table 2). Kavitha et al. (2010) used a neural network and a genetic algorithm to detect the presence of heart disease on CHDD. In this model the neural network is trained using the genetic algorithm showing that the proposed hybridization is more stable [32]. This research did not calculate the model accuracy.

Comparison of single and hybrid data mining techniques in the diagnosis of heart disease on the CHDD shows different accuracies, with the hybrid techniques showing better accuracy than single techniques (Table 2). The best accuracy achieved using single data mining technique is 84.14% by naïve bayes [23]. However, the best accuracy achieved using hybrid data mining technique is 89.01% by neural network ensemble [12]. Hybridized data mining techniques are enhancing the accuracy of heart disease diagnosis.

Recently, we conducted research on the CHDD to demonstrate that using more sophisticated data mining techniques improves the accuracy of heart disease diagnosis [36]. Through a systematic investigation of several single data mining technique, different data discretization levels, the application of voting techniques, and reduced error pruning, we demonstrated increases in accuracy on all techniques assessed. Table 3 shows summary results from that work [36]:

TABLE 3: EXAMPLE RESULTS OF INCREASED ACCURACY ON BENCHMARK DATASET USING MORE SOPHISTICATED DATA MINING TECHNIQUES

		Sensitivity	Specificity	Accuracy
Shouman et al., 2011	Nine Voting Equal Frequency Discretization Gain Ratio Decision Tree	77.9%	85.2%	84.1%
Tu et al., 2009	J4.8 Decision Tree	72.01%	84.48%	78.9%
	Bagging Algorithm	74.93%	86.64%	81.41%

While not perfect (obviously), these results are very good. The ability to improve the accuracy through the application of more sophisticated techniques is also encouraging. Research is still needed for other data mining techniques such as kernel density and support vector machine on the benchmark dataset.

III. TRENDS IN USING DATA MINING TECHNIQUES IN HEART DISEASE

Although applying data mining is beneficial to healthcare [3, 23, 37], disease diagnosis [12, 21, 38], and treatment [39-41], few researches have investigated producing treatment plans for patients [40]. Accurate diagnosis and treatment given to patients have been major issues highlighted in medical services.

Recently, researchers started investigating using data mining techniques to handle the error and complexity of treatment processes for healthcare providers. Razali and Ali (2009) investigated generating treatment plans for acute upper respiratory infection disease patients using a decision tree. The model recommended treatment through giving drugs to patients showing accuracy of 94.73% [39]. Applying association rules and decision tree to treatment plans are showing acceptable performance. However, the comparison with other data mining techniques such as naïve bayes, neural network, and genetic algorithms still needs investigation. Saad Ali et al. (2010) presented the development of treatment plans to support treatment decision making for health care practitioners. The development of the treatment plan was generated on six common diseases using the decision trees technique showing an accuracy level that ranged from 77.97% to 91.67%. [40].

Although using data mining techniques in developing treatment plans for patients is showing acceptable performance, there has been little focus on treatment for heart disease patients. Kim et al. (2005) evaluated the current treatments for chronic heart failure using a decision tree and compared the results with those of large-scale clinical trials. They investigated which drugs can increase or decrease plasma level, fractional shortening, and left ventricular end-diastolic diameter in the cardiovascular disease [42]. However, they did not investigate using data mining techniques to identify the suitable treatment for heart disease patients.

As hybrid data mining techniques showed promising results in the diagnosis of heart disease, so researchers investigating the use of hybridized data mining techniques in identifying the suitable treatment for heart disease patients are likely to be fruitful. The selection of a suitable treatment for heart disease patients is inherently complex, particularly as overall treatment frequently involves multiple, often concurrent, elements. This complexity makes treatment recommendation from data mining very difficult. Data mining is suited to assist decision making when many variables must be assessed, such as multiple concurrent treatments, but usually to make a single selection. To facilitate data mining being successful, our approach to analyzing treatment options and formulating recommendations using data mining will focus on what is the next best treatment step within the current treatment plan. Recognizing that this loses some richness from the overall treatment plan, it does make the problem more like one of classification, for which data mining is well-suited.

In light of the success that data mining techniques, and particularly hybridized techniques, have had in classifying heart disease sufferers, it seems important that similar approaches are considered in the selection of appropriate treatment for heart disease sufferers. From this, we propose a systematic approach to assessing the effectiveness of hybridized data mining techniques to identify the suitable treatment for heart disease patients.

IV. PROPOSED RESEARCH MODEL

We propose that applying data mining techniques in identifying suitable treatments for heart disease patients is fruitful and needs further investigation. To evaluate if applying data mining techniques to heart disease treatment can provide as reliable performance as achieved in heart disease diagnosis, we propose the following approach (shown in Figure 1):

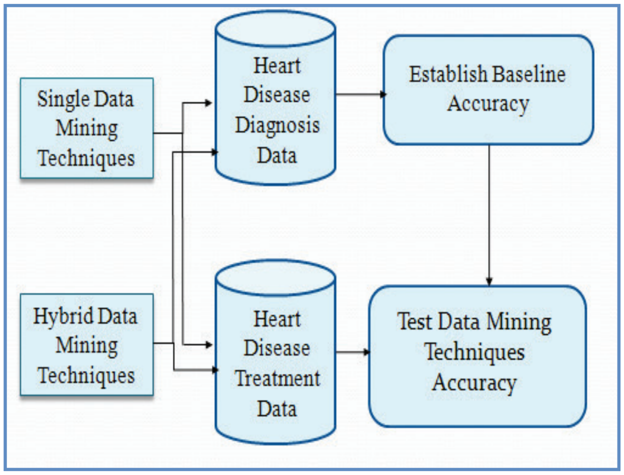
Single Data Mining Techniques
Hybrid Data Mining Techniques
Heart Disease Diagnosis Data
Heart Disease Treatment Data
Establish Baseline Accuracy
Test Data Mining Techniques Accuracy

Figure 1. Proposed Research Model

1. Apply single data mining techniques to heart disease diagnosis benchmark dataset to establish baseline accuracy for each single data mining technique in the diagnosis of heart disease patients.
2. Apply the same single data mining techniques used in heart disease diagnosis to heart disease treatment dataset to investigate if single data mining techniques can achieve equivalent (or better) results in identifying suitable treatments as that achieved in the diagnosis.
3. Apply hybrid data mining techniques to heart disease diagnosis benchmark dataset to establish baseline accuracy for each hybrid data mining technique in the diagnosis of heart disease patients.
4. Apply the same hybrid data mining techniques used in heart disease diagnosis to heart disease treatment dataset to investigate if hybrid data mining techniques can achieve equivalent (or better) results in identifying suitable treatments as that achieved in the diagnosis.

Applying the proposed model will help in answering some important questions including:

1. Can single and hybrid data mining techniques assist healthcare professionals in identifying suitable treatments for heart disease patients?
2. Which single data mining technique will give better accuracy in identifying suitable treatments for heart disease patients?
3. Will applying hybrid data mining techniques provide better accuracy than single data mining techniques in identifying suitable treatments?
4. Which hybrid data mining technique will give better accuracy in identifying suitable treatments for heart disease patients?
5. Are various data mining techniques of equivalent accuracy when compared across heart disease diagnosis and treatment selection?

V. SUMMARY

Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amounts of data, researchers are using data mining techniques in the diagnosis of heart disease. Although applying data mining techniques to help health care professionals in the diagnosis of heart disease is having some success, the use of data mining techniques to identify a suitable treatment for heart disease patients has received less attention. Also, applying hybrid data mining techniques has shown promising results in the diagnosis of heart disease, so applying hybrid data mining techniques in selecting the suitable treatment for heart disease patients needs further investigation. This paper identifies gaps in the research on heart disease diagnosis and treatment and proposes a model to systematically close those gaps to discover if applying data mining techniques to heart disease treatment data can provide as reliable performance as that achieved in diagnosing heart disease patients.

REFERENCES

1. Han, j. and M. Kamber, *Data Mining Concepts and Techniques*. 2006: Morgan Kaufmann Publishers.
2. Lee, I.-N., S.-C. Liao, and M. Embrechts, *Data mining techniques applied to medical information.* Med. inform, 2000.
3. Obenshain, M.K., *Application of Data Mining Techniques to Healthcare Data.* Infection Control and Hospital Epidemiology, 2004.
4. Sandhya, J., et al., *Classification of Neurodegenerative Disorders Based on Major Risk Factors Employing Machine Learning Techniques.* International Journal of Engineering and Technology, 2010. Vol.2, No.4.
5. Thuraisingham, B., *A Primer for Understanding and Applying Data Mining.* IT Professional IEEE, 2000.
6. Ashby, D. and A. Smith, *The Best Medicine?* Plus Magazine - Living Mathematics., 2005.
7. Liao, S.-C. and I.-N. Lee, *Appropriate medical data categorization for data mining classification techniques.* MED. INFORM., 2002. Vol. 27, no. 1, 59–67, .
8. Ruben, D.C.J., *Data Mining in Healthcare: Current Applications and Issues.* 2009.
9. Porter, T. and B. Green, *Identifying Diabetic Patients: A Data Mining Approach.* Americas Conference on Information Systems, 2009.
10. Panzarasa, S., et al., *Data mining techniques for analyzing stroke care processes.* Proceedings of the 13th World Congress on Medical Informatics, 2010.
11. Li L, T.H., Wu Z, Gong J, Gruidl M, Zou J, Tockman M, Clark RA, *Data mining techniques for cancer detection using serum proteomic profiling.* Artificial Intelligence in Medicine, Elsevier, 2004.
12. Das, R., I. Turkoglu, and A. Sengur, *Effective diagnosis of heart disease through neural networks ensembles.* Expert Systems with Applications, Elsevier, 2009. 36 (2009): p. 7675–7680.
13. World Health Organization. 2007 7-Febuary 2011]; Available from: http://www.who.int/mediacentre/factsheets/fs310.pdf.
14. European Public Health Alliance. 2010 7-February-2011]; Available from: http://www.epha.org/a/2352

15. ESCAP. 2010 7-February-2011]; Available from: http://www.unescap.org/stat/data/syb2009/9.Health-risks-causes-of-death.asp.

16. Australian Bureau of Statistics. 2010 7-February-2011]; Available from: http://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/E8510D1C8DC1AE1CCA2576F600139288/$File/33030_2008.pdf

17. Statistics South Africa. 2008 7-February-2011]; Available from: http://www.statssa.gov.za/publications/P03093/P030932006.pdf

18. Helma, C., E. Gottmann, and S. Kramer, *Knowledge discovery and data mining in toxicology.* Statistical Methods in Medical Research, 2000.

19. Podgorelec, V., et al., *Decision Trees: An Overview and Their Use in Medicine.* Journal of Medical Systems, 2002. Vol. 26.

20. Andreeva, P., *Data Modelling and Specific Rule Generation via Data Mining Techniques.* International Conference on Computer Systems and Technologies - CompSysTech, 2006.

21. Rajkumar, A. and G.S. Reena, *Diagnosis Of Heart Disease Using Datamining Algorithm.* Global Journal of Computer Science and Technology, 2010. Vol. 10 (Issue 10).

22. Sitar-Taut, V.A., et al., *Using machine learning algorithms in cardiovascular disease risk evaluation.* Journal of Applied Computer Science & Mathematics, 2009.

23. Srinivas, K., B.K. Rani, and A. Govrdhan, *Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks.* International Journal on Computer Science and Engineering (IJCSE), 2010. Vol. 02, No. 02: p. 250-255.

24. Yan, H., et al., *Development of a decision support system for heart disease diagnosis using multilayer perceptron.* Proceedings of the 2003 International Symposium on, 2003. vol.5: p. pp. V-709- V-712.

25. Youssef, F.N., D. Nel, and T. Bovaird, *Health care quality in hospitals.* Int. J. Health Care Qual. Assur, 1996. 9: 15-28.

26. Garg, A.X., et al., *Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: a Systematic Review.* . Journal of the American Medical Association, 2005. 293: p. 1223–1238.

27. Heller, R.F., et al., *How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Prevention Project.* BRITISH MEDICAL JOURNAL, 1984.

28. Wilson, P.W.F., et al., *Prediction of Coronary Heart Disease Using Risk Factor Categories.* American Heart Association Journal, 1998.

29. Simons, L.A., et al., *Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study.* Medical Journal of Australia, 2003. 178.

30. Salahuddin and F. Rabbi, *Statistical Analysis of Risk Factors for Cardiovascular disease in Malakand Division.* Pak. j. stat. oper. res., 2006. Vol.II: p. pp49-56.

31. Shahwan-Akl, L., *Cardiovascular Disease Risk Factors among Adult Australian-Lebanese in Melbourne.* International Journal of Research in Nursing, 2010. 6 (1).

32. Kavitha, K.S., K.V. Ramakrishnan, and M.K. Singh, *Modeling and design of evolutionary neural network for heart disease detection.* International Journal of Computer Science Issues (IJCSI), 2010. Vol. 7, Issue 5.

33. Parthiban, L. and R. Subramanian, *Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm.* International Journal of Biological and Life Sciences, 2007.

34. Patil, S.B. and Y.S. Kumaraswamy, *Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction.* International Journal of Computer Science and Network Security (IJCSNS), 2009. VOL.9 No.2.

35. Polat , K., S. Sahan, and S. Gunes, *Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing.* Expert Systems with Applications 2007. 32 p. 625–631.

36. Shouman, M., T. Turner, and R. Stocker, *Using decision tree for diagnosing heart disease patients.* 9th Australasian Data Mining Conference 2011. 121.

37. Lazakidou, A., *Web-Based Applications in Healthcare and Biomedicine*. 2010: Springer.

38. Tu, M.C., D. Shin, and D. Shin, *Effective Diagnosis of Heart Disease through Bagging Approach.* Biomedical Engineering and Informatics, IEEE, 2009.

39. Razali, A.M. and S. Ali, *Generating Treatment Plan in Medicine: A Data Mining Approach.* American Journal of Applied Sciences, 2009. 6 (2): 345-351.

40. Saad Ali, S.N., et al., *Developing treatment plan support in outpatient health care delivery with decision trees technique.* Springer-Verlag Berlin Heidelberg, 2010. Part II, LNCS 6441, pp. 475–482,.

41. Wright, A., E.S. Chen, and F.L. Maloney, *An automated technique for identifying associations between medications, laboratory results and problems.* Journal of Biomedical Informatics Elsevier, 2010. 43 (2010) 891–901.

42. Kim, J., et al., *A Novel Data Mining Approach to the Identification of Effective Drugs or Combinations for Targeted Endpoints—Application to Chronic Heart Failure as a New Form of Evidence-based Medicine.* Cardiovascular Drugs and Therapy, Springer 2005. 18 p. 483–489.